

COMPUTATIONAL MODELS OF ARGUMENT

Frontiers in Artificial Intelligence and Applications

The book series Frontiers in Artificial Intelligence and Applications (FAIA) covers all aspects of theoretical and applied Artificial Intelligence research in the form of monographs, doctoral dissertations, textbooks, handbooks and proceedings volumes.

The FAIA series contains several sub-series, including ‘Information Modelling and Knowledge Bases’ and ‘Knowledge-Based Intelligent Engineering Systems’. It also includes the biennial European Conference on Artificial Intelligence (ECAI) proceedings volumes, and other EurAI (European Association for Artificial Intelligence, formerly ECCAI) sponsored publications. An editorial panel of internationally well-known scholars is appointed to provide a high quality selection.

Series Editors:

J. Breuker, N. Guarino, J.N. Kok, J. Liu, R. López de Mántaras,
R. Mizoguchi, M. Musen, S.K. Pal and N. Zhong

Volume 287

Recently published in this series

- Vol. 286. H. Fujita and G.A. Papadopoulos (Eds.), New Trends in Software Methodologies, Tools and Techniques – Proceedings of the Fifteenth SoMeT_16
- Vol. 285. G.A. Kaminka, M. Fox, P. Bouquet, E. Hüllermeier, V. Dignum, F. Dignum and F. van Harmelen (Eds.), ECAI 2016 – 22nd European Conference on Artificial Intelligence, 29 August–2 September 2016, The Hague, The Netherlands – Including Prestigious Applications of Artificial Intelligence (PAIS 2016)
- Vol. 284. D. Pearce and H.S. Pinto (Eds.), STAIRS 2016 – Proceedings of the Eighth European Starting AI Researcher Symposium
- Vol. 283. R. Ferrario and W. Kuhn (Eds.), Formal Ontology in Information Systems – Proceedings of the 9th International Conference (FOIS 2016)
- Vol. 282. J. Mizera-Pietraszko, Y.-L. Chung and P. Pichappan (Eds.), Advances in Digital Technologies – Proceedings of the 7th International Conference on Applications of Digital Information and Web Technologies 2016
- Vol. 281. G. Chen, F. Liu and M. Shojafar (Eds.), Fuzzy System and Data Mining – Proceedings of FSDM 2015
- Vol. 280. T. Welzer, H. Jaakkola, B. Thalheim, Y. Kiyoki and N. Yoshida (Eds.), Information Modelling and Knowledge Bases XXVII
- Vol. 279. A. Rotolo (Ed.), Legal Knowledge and Information Systems – JURIX 2015: The Twenty-Eighth Annual Conference
- Vol. 278. S. Nowaczyk (Ed.), Thirteenth Scandinavian Conference on Artificial Intelligence – SCAI 2015

ISSN 0922-6389 (print)
ISSN 1879-8314 (online)

Computational Models of Argument

Proceedings of COMMA 2016

Edited by

Pietro Baroni

Department of Information Engineering, University of Brescia, Italy

Thomas F. Gordon

Fraunhofer FOKUS, Berlin, Germany

Tatjana Scheffler

UFS Cognitive Sciences, University of Potsdam, Germany

and

Manfred Stede

UFS Cognitive Sciences, University of Potsdam, Germany

IOS
Press

Amsterdam • Berlin • Washington, DC

© 2016 The authors and IOS Press.

This book is published online with Open Access and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).

ISBN 978-1-61499-685-9 (print)

ISBN 978-1-61499-686-6 (online)

Library of Congress Control Number: 2016948877

Publisher

IOS Press BV

Nieuwe Hemweg 6B

1013 BG Amsterdam

Netherlands

fax: +31 20 687 0019

e-mail: order@iospress.nl

Distributor in the USA and Canada

IOS Press, Inc.

4502 Rachael Manor Drive

Fairfax, VA 22032

USA

fax: +1 703 323 3668

e-mail: iosbooks@iospress.com

LEGAL NOTICE

The publisher is not responsible for the use which might be made of the following information.

PRINTED IN THE NETHERLANDS

Preface

The investigation of computational models of argument is a rich, interdisciplinary, and fascinating research field whose ultimate aim is to support the development of computer-based systems able to engage argumentation-related activities with human users or among themselves. This ambitious research goal involves the study of natural, artificial, and theoretical argumentation and, as such, requires openness to interactions with a variety of disciplines ranging from philosophy and cognitive science to formal logic and graph theory, to mention some.

The biennial International Conference on Computational Models of Argument (COMMA), reaching its sixth edition, provides since ten years a dedicated forum for presentation of the latest advancements in this multifaceted field, covering both basic research and innovative applications.

The first COMMA was supported by the EU 6th Framework Programme project ASPIC and was hosted by the University of Liverpool in 2006 with a vision for the future. After the event, a steering committee promoting the continuation of the conference was established and, since then, the steady growth of interest in computational argumentation research worldwide has gone hand in hand with the development of the conference itself and of related activities by its community.

Since the second edition, organized by IRIT in Toulouse in 2008, plenary invited talks by world-leading researchers and a software demonstration session became an integral part of the conference program.

The third edition, organized in 2010 by the University of Brescia in Desenzano del Garda, saw the addition of a best student paper award. The same year, the new journal *Argument and Computation*, closely related to the COMMA community, was started.

Since the fourth edition, organized by the Vienna University of Technology in 2012, an Innovative Application Track and a section for Demonstration Abstracts were included in the proceedings.

At the fifth edition, co-organized in 2014 by the Universities of Aberdeen and Dundee in Pitlochry, the main conference was preceded by the first Summer School on *Argumentation: Computational and Linguistic Perspectives*. The same year, the first International Competition on Computational Models of Argumentation, to be held in 2015, was launched.

This year COMMA is hosted by the University of Potsdam and the conference program is complemented by two satellite workshops, in addition to the second edition of the summer school. Moreover, reflecting the evolution of research publishing worldwide, COMMA 2016 proceedings will be Open Access.

The evolution of COMMA into an articulated event, however, is only subsidiary to the fulfillment of its mission, namely documenting and stimulating the advancement of knowledge and the development of applications in the field.

The past conference programs, along with the present one, give comfortable indications in this respect.

First of all, they have seen, since the very first edition, a balanced blend of theoretical and application-oriented works.

Further, in addition to “traditional” investigation topics in the field, like abstract argumentation frameworks, the conference has always included contributions concerning emerging trends and the development of new connections with other areas.

Among them, it is possible to mention the investigation of a variety of quantitative approaches to argumentation, in relationship with Bayesian networks, probability theory, or fuzzy logic. Also, we wish to mention the area of argument mining, the automatic detection and analysis of argumentation in linguistic data. While the term was barely known three years ago, this research has emerged as a fast-growing subfield of Computational Linguistics (CL), with a variety of specialized workshops having been formed over the past few years, and the topic has also been established in important CL conferences. So far, following the general trend in CL, the methods being applied to argument mining largely rely on machine learning over surface-oriented features of text; but there seems to be great potential in linking the text analysis also to the “deeper” phenomena – reasoning and inference – that render an argumentation plausible.

We conclude by remarking that the success of a conference depends on the contributions of many people.

We acknowledge steady support and encouragement by the COMMA Steering Committee.

We would like to thank the invited speakers, Jens Allwood, Anthony Hunter, and Marie-Francine Moens, for accepting our invitation and for witnessing, once again, the rich diversity of this area with their talks, covering respectively an insightful analysis of the normative and descriptive perspectives in argumentation studies, the promise and challenge of using computational persuasion for applications in behaviour change, and the formidable question of how can a machine acquire world and common sense knowledge for argument mining.

We are deeply grateful to the members of the Program Committee and to the additional reviewers for their invaluable efforts. Their reports and subsequent discussions led to the selection, out of 63 submissions, of 25 full papers and 17 short papers, to be included in the conference proceedings together with 10 demonstration abstracts. The submission and reviewing process has been managed through the EasyChair conference system, which we acknowledge for supporting COMMA since the first edition.

Last but not least, we thank all the authors for contributing to the success of the conference with their hard work and commitment.

Berlin/Brescia/Potsdam, July 2016

Pietro Baroni (Program chair)
 Thomas F. Gordon (Conference chair)
 Tatjana Scheffler (Local organization co-chair)
 Manfred Stede (Conference chair)

Programme Committee and Reviewers

Program Committee

Leila Amgoud	Gabriele Kern-Isberner
Ofer Arieli	Sébastien Konieczny
Kevin Ashley	Marie-Christine Lagasquie-Schiex
Katie Atkinson	Ho-Pun Lam
Ringo Baumann	John Lawrence
Trevor Bench-Capon	Beishui Liao
Philippe Besnard	Thomas Linsbichler
Floris Bex	Diane Litman
Stefano Bistarelli	Ana Gabriela Maguitman
Elizabeth Black	Jean-Guy Mailly
Alexander Bochman	Pierre Marquis
Guido Boella	Maria Vanina Martinez
Elise Bonzon	Nicolas Maudet
Gerhard Brewka	Sanjay Modgil
Katarzyna Budzyska	Pavlos Moraitis
Elena Cabrio	Timothy Norman
Martin Caminada	Nir Oren
Claudette Cayrol	Fabio Paglieri
Federico Cerutti	Simon Parsons
Carlos Chesñevar	Mikolaj Podlaszewski
Sylvie Coste-Marquis	Henri Prade
Jürgen Dix	Henry Prakken
Sylvie Doutre	Iyad Rahwan
Paul Dunne	Chris Reed
Wolfgang Dvořák	Tjitze Rienstra
Phan Minh Dung	Régis Riveret
Stefan Ellmauthaler	Odinaldo Rodrigues
Xiuyi Fan	Patrick Saint-Dizier
Dov Gabbay	Chiaki Sakama
Sarah Alice Gaggl	Francesco Santini
Alejandro García	Giovanni Sartor
Massimiliano Giacomin	Jodi Schneider
Lluís Godo	Claudia Schulz
Guido Governatori	Carles Sierra
Thomas F. Gordon	Guillermo R. Simari
Floriana Grasso	Mark Snaith
Nancy Green	Manfred Stede
Davide Grossi	Hannes Strass
Graeme Hirst	Katia Sycara
Anthony Hunter	Yuqing Tang
Souhila Kaci	Matthias Thimm
Antonis Kakas	Francesca Toni

Alice Toniolo
Paolo Torroni
Leon van der Torre
Bart Verheij
Srdjan Vesic
Serena Villata

Johannes Peter Wallner
Doug Walton
Emil Weydert
Stefan Woltran
Adam Wyner

Additional Reviewers

Jérôme Delobelle
Jean-Marie Lagniez
Marco Lippi
Nicolas Schwind
Tomas Trescak

Contents

Preface	v
<i>Pietro Baroni, Thomas F. Gordon, Tatjana Scheffler and Manfred Stede</i>	
Program Committee and Reviewers	vii
Invited Talks	
Argumentation, Activity and Culture	3
<i>Jens Allwood</i>	
Argumentation Mining: How Can a Machine Acquire World and Common Sense Knowledge?	4
<i>Marie-Francine Moens</i>	
Computational Persuasion with Applications in Behaviour Change	5
<i>Anthony Hunter</i>	
Innovative Applications	
Tweeties Squabbling: Positive and Negative Results in Applying Argument Mining on Social Media	21
<i>Tom Bosc, Elena Cabrio and Serena Villata</i>	
Dialog-Based Online Argumentation	33
<i>Tobias Krauthoff, Michael Baurmann, Gregor Betz and Martin Mauve</i>	
Understanding Group Polarization with Bipolar Argumentation Frameworks	41
<i>Carlo Proietti</i>	
Preferences in Argumentation for Statistical Model Selection	53
<i>Isabel Sassoan, Jeroen Keppens and Peter McBurney</i>	
An Argumentation Workflow for Reasoning in Ontology Based Data Access	61
<i>Bruno Yun and Madalina Croitoru</i>	
Regular Papers	
Argument Schemes for Reasoning About the Actions of Others	71
<i>Katie Atkinson and Trevor Bench-Capon</i>	
Verifiability of Argumentation Semantics	83
<i>Ringo Baumann, Thomas Linsbichler and Stefan Woltran</i>	
From Arguments to Constraints on a Bayesian Network	95
<i>Floris Bex and Silja Renooij</i>	

On Efficiently Enumerating Semi-Stable Extensions via Dynamic Programming on Tree Decompositions <i>Bernhard Bliem, Markus Hecher and Stefan Woltran</i>	107
An Ontology for Argumentation on the Social Web: Rhetorical Extensions to the AIF <i>Tom Blount, David E. Millard and Mark J. Weal</i>	119
Abstract Dialectical Argumentation Among Close Relatives <i>Alexander Bochman</i>	127
Argumentation Ranking Semantics Based on Propagation <i>Elise Bonzon, Jérôme Delobelle, Sébastien Konieczny and Nicolas Maudet</i>	139
Using Argument Features to Improve the Argumentation Process <i>Maximiliano C.D. Budán, Gerardo I. Simari and Guillermo R. Simari</i>	151
Minimal Cost Semantics in Argumentation Framework on Semiring Cost Assignment <i>Shuya Bundo and Kazunori Yamaguchi</i>	159
Spectral Techniques in Argumentation Framework Analysis <i>James Butterworth and Paul E. Dunne</i>	167
A Dialectical Approach for Argument-Based Judgment Aggregation <i>Martin Caminada and Richard Booth</i>	179
Towards a New Framework for Recursive Interactions in Abstract Bipolar Argumentation <i>Claudette Cayrol, Andrea Cohen and M-Christine Lagasquie-Schiex</i>	191
On the Effectiveness of Automated Configuration in Abstract Argumentation Reasoning <i>Federico Cerutti, Mauro Vallati and Massimiliano Giacomin</i>	199
Where Are We Now? State of the Art and Future Trends of Solvers for Hard Argumentation Problems <i>Federico Cerutti, Mauro Vallati and Massimiliano Giacomin</i>	207
Argumentation for Machine Learning: A Survey <i>Oana Cocarascu and Francesca Toni</i>	219
On the Acceptability Semantics of Argumentation Frameworks with Recursive Attack and Support <i>Andrea Cohen, Sebastian Gottifredi, Alejandro J. García and Guillermo R. Simari</i>	231
Explanation for Case-Based Reasoning via Abstract Argumentation <i>Kristijonas Čyras, Ken Satoh and Francesca Toni</i>	243
Quantifying the Difference Between Argumentation Semantics <i>Sylvie Doutre and Jean-Guy Mailly</i>	255
A Canonical Semantics for Structured Argumentation with Priorities <i>Phan Minh Dung</i>	263

Forbidden Sets in Argumentation Semantics <i>Paul E. Dunne</i>	275
I Heard You the First Time: Debate in Cacophonous Surroundings <i>Paul E. Dunne</i>	287
Mining Ethos in Political Debate <i>Rory Duthie, Katarzyna Budzynska and Chris Reed</i>	299
Argumentation as Information Input: A Position Paper <i>Dov Gabbay and Michael Gabbay</i>	311
Degrees of “in”, “out” and “undecided” in Argumentation Networks <i>Dov M. Gabbay and Odinaldo Rodrigues</i>	319
Formalizing Balancing Arguments <i>Thomas F. Gordon and Douglas Walton</i>	327
Assigning Likelihoods to Interlocutors’ Beliefs and Arguments <i>Seyed Ali Hosseini, Sanjay Modgil and Odinaldo Rodrigues</i>	339
A System for Dispute Mediation: The Mediation Dialogue Game <i>Mathilde Janier, Mark Snaith, Katarzyna Budzynska, John Lawrence and Chris Reed</i>	351
On ASPIC ⁺ and Defeasible Logic <i>Ho-Pun Lam, Guido Governatori and Régis Riveret</i>	359
Argument Analytics <i>John Lawrence, Rory Duthie, Katarzyna Budzynska and Chris Reed</i>	371
Argument Mining Using Argumentation Scheme Structures <i>John Lawrence and Chris Reed</i>	379
A Specialized Set Theoretic Semantics for Acceptability Dynamics of Arguments <i>Martin O. Moguillansky and Guillermo R. Simari</i>	391
Construction and Strength Calculation of Threats <i>Mariela Morveli-Espinoza, Ayslan T. Possebom and Cesar A. Tacla</i>	403
A Heuristic Strategy for Persuasion Dialogues <i>Josh Murphy, Elizabeth Black and Michael Luck</i>	411
Rethinking the Rationality Postulates for Argumentation-Based Inference <i>Henry Prakken</i>	419
Assessing Weight of Opinion by Aggregating Coalitions of Arguments <i>Pavithra Rajendran, Danushka Bollegala and Simon Parsons</i>	431
Perfection in Abstract Argumentation <i>Christof Spanring</i>	439
Gödel Fuzzy Argumentation Frameworks <i>Jiachao Wu, Hengfei Li, Nir Oren and Timothy J. Norman</i>	447

Demonstrations

DALEK: A Tool for Dialectical Explanations in Inconsistent Knowledge Bases	461
<i>Abdallah Arioua, Madalina Croitoru and Patrice Buche</i>	
ConArg: A Tool for Classical and Weighted Argumentation	463
<i>Stefano Bistarelli, Fabio Rossi and Francesco Santini</i>	
Efficient and Off-The-Shelf Solver: jArgSemSAT	465
<i>Federico Cerutti, Mauro Vallati and Massimiliano Giacomini</i>	
Generating Structured Argumentation Frameworks: AFBenchGen2	467
<i>Federico Cerutti, Massimiliano Giacomini and Mauro Vallati</i>	
A System for Supporting the Detection of Deceptive Reviews Using Argument Mining	469
<i>Oana Cocarascu and Francesca Toni</i>	
DIAMOND 3.0 – A Native C++ Implementation of DIAMOND	471
<i>Stefan Ellmauthaler and Hannes Strass</i>	
GrappaVis – A System for Advanced Graph-Based Argumentation	473
<i>Georg Heissenberger and Stefan Woltran</i>	
The ARGTEACH Web-Platform	475
<i>Claudia Schulz and Dragos Dumitrache</i>	
Gorgias-B: Argumentation in Practice	477
<i>Nikolaos I. Spanoudakis, Antonis C. Kakas and Pavlos Moraitis</i>	
The RationalGRL Toolset for Goal Models and Argument Diagrams	479
<i>Marc van Zee, Diana Marosin, Floris Bex and Sepideh Ghanavati</i>	
Subject Index	481
Author Index	483

Invited Talks

This page intentionally left blank

Argumentation, Activity and Culture

Jens ALLWOOD ^{a,1}

^a *SCCIIIL Interdisciplinary Center, University of Gotheburg, Sweden*

Argument mining is difficult for many reasons, but one of the main reasons is that real arguments are not purified and easily recognizable according to the normative criteria of logos and dialectics. Instead, most actually occurring arguments are guided by many concerns, some of which were identified in traditional rhetoric (where logos is only one concern). More descriptive approaches, like Argumentation schemes (Walton), Conversational Analysis (CA), Activity based Communication Analysis (ACA) Critical Discourse Analysis (CDA) and Intercultural rhetoric have identified others.

However, the most common way of studying argumentation through the ages has been normative, i.e. how we should or should not argue from some point of view. This can be contrasted with a study of how we actually argue – a descriptive study.

In my talk, I, thus, start by discussing how we should study argumentation. I contrast a normative perspective with a descriptive perspective and first briefly consider some of the normative ideas about how we should argue – “positive normativity” and then consider some of the ideas concerning how we should not argue – “negative normativity”.

I then turn to consider some of the classical, mainly Aristotelian, ideas about rhetoric (logos, ethos and pathos) and his pointing out of the importance of “kairos” – how the means of persuasion are influenced by situational and other background factors.

Following this, I discuss some contributions from the more descriptive approaches mentioned above which can be used in studies of argumentation. In the third section of the talk, building on a combination of ideas in the preceding discussion, I present some general steps that could be taken in identifying and analyzing argumentation and in a fourth section this is exemplified by discussing some examples of a rhetorical analysis of information and argumentation concerning obesity, given on official web sites in Malaysia and Sweden.

¹Corresponding Author: Jens Allwood, SCCIIIL Interdisciplinary Center, University of Gotheburg, Sweden;
E-mail: jens.allwood@gu.se

Argumentation Mining: How Can a Machine Acquire World and Common Sense Knowledge?

Marie-Francine MOENS ^{a,1}

^a*Department of Computer Science, KU Leuven, Belgium*

Keywords. Natural language understanding, Knowledge acquisition, Machine learning

Argumentation mining regards an advanced form of human language understanding by the machine. This is a challenging task for a machine. When sufficient explicit discourse markers are present in the language utterances, the argumentation can be interpreted by the machine with an acceptable degree of accuracy. However, in many real settings, the task is much more difficult due to the lack or ambiguity of the discourse markers, and the fact that a substantial amount of knowledge needed for the correct recognition of the argumentation, its components and their relationships is not explicitly present in the text, but makes up the background knowledge that humans possess when interpreting language. The lecture focuses on how the machine can automatically acquire such knowledge.

In this lecture we consider argumentation mining from written text. First, we give an overview of the latest methods for human language understanding that map language to a formal knowledge representation that facilitates other tasks (for instance, a representation that is used to visualize the argumentation or that is easily shared in a decision or argumentation support system). Most current systems are trained on texts that are manually annotated. Then we go deeper into the new field of representation learning that nowadays is very much studied in computational linguistics. This field investigates methods for representing language as statistical concepts or as vectors, allowing straightforward methods of compositionality. The methods often use deep learning and its underlying neural network technologies to learn concepts from large text collections in an unsupervised way (i.e., without the need for manual annotations). We show how these methods can help the argumentation mining process, but also demonstrate that these methods are still insufficient to automatically acquire the necessary background knowledge and more specifically world and common sense knowledge. We propose a number of ways to improve the learning from textual, visual or database data, and discuss how we can integrate the learned knowledge in the argumentation mining process.

¹Corresponding Author: Marie-Francine Moens, Department of Computer Science, KU Leuven, Celestijnenlaan 200A, B-3001 Heverlee, Belgium ; E-mail: sien.moens@cs.kuleuven.be

Computational Persuasion with Applications in Behaviour Change

Anthony HUNTER

*Department of Computer Science,
University College London,
Gower Street, London WC1E 6BT, UK
anthony.hunter@ucl.ac.uk*

Abstract. Persuasion is an activity that involves one party trying to induce another party to believe something or to do something. It is an important and multifaceted human facility. Obviously, sales and marketing is heavily dependent on persuasion. But many other activities involve persuasion such as a doctor persuading a patient to drink less alcohol, a road safety expert persuading drivers to not text while driving, or an online safety expert persuading users of social media sites to not reveal too much personal information online. As computing becomes involved in every sphere of life, so too is persuasion a target for applying computer-based solutions. An automated persuasion system (APS) is a system that can engage in a dialogue with a user (the persuadee) in order to persuade the persuadee to do (or not do) some action or to believe (or not believe) something. To do this, an APS aims to use convincing arguments in order to persuade the persuadee. Computational persuasion is the study of formal models of dialogues involving arguments and counterarguments, of user models, and strategies, for APSs. A promising application area for computational persuasion is in behaviour change. Within healthcare organizations, government agencies, and non-governmental agencies, there is much interest in changing behaviour of particular groups of people away from actions that are harmful to themselves and/or to others around them.

Keywords. Computational persuasion; Persuasion dialogues; Persuasive arguments; Dialogical argumentation; Computational models of argument; Probabilistic argumentation; Argumentation strategies.

1. Introduction

Persuasion is an activity that involves one party trying to get another party to do (or not do) some action or to believe (or not believe) something. It is an important and multifaceted human facility. Consider, for example, a doctor persuading a patient to drink less, a road safety expert persuading drivers to not text while driving, or an online safety expert persuading users of social media sites to not reveal too much personal information.

In this paper, I discuss some aspects of the notion of persuasion, and explain how this leads to the idea of computational persuasion. Computational models of

argument are central to the development of computational persuasion. I briefly review some key aspects of computational models of argument, and highlight some topics that need further development. I then briefly cover behaviour change as a topic that we can apply methods from computational persuasion, and evaluate the progress in the field.

2. What is persuasion?

The aim of persuasion is for the persuader to change the mind of the persuadee. Some kinds of interaction surrounding persuasion include: Persuader collecting information, preferences, etc from the persuadee; Persuader providing information, offers, etc to the persuadee; Persuader winning favour (e.g. by flattering the persuadee, by making small talk, by being humorous, etc); But importantly, arguments are the essential structures for presenting the claims (and counter claims) in persuasion. An argument-centric focus on persuasion leads to a number of inter-related aspects (see list below) that need to be taken into account, any of which can be important in bringing about successful persuasion.

Persuader The nature of the persuader can be important. From a rational perspective, seemingly good features of a persuader are that s/he has relevant authority, expertise, or knowledge, and seemingly poor features of a persuader are that s/he is attractive, witty, or a celebrity. However, in practice, different persuadees respond to different features. For instance, a teenager is unlikely to be convinced by a government safety expert to wear a helmet when on a bike, but may be influenced by a celebrity to do so.

Language The choice of language in argumentation can be important. This goes from choice of words (e.g. use of *freedom fighter* versus *terrorist*), to choice of metaphor, or use of irony[17].

Psychology The use of psychological techniques can be important [16] such as: Reciprocation (e.g. doing a small favour for someone is more likely to result in a big favour being obtained in return); Consistency (e.g. getting expressed support for a cause, prior to asking for material support is more likely to be successful); And social proof (e.g. treating dog phobia in children by showing videos of children playing happily with children).

Personality Determining the personality of the persuadee can be important. Consider for example persuading someone to vote in the national election: If the person “follows the crowd”, then tell them that the majority of the population voted in the last election, whereas if the person “follows rules rigorously”, then tell them that it is their duty to vote. Mistaking the personality trait can have a negative effect on the chances of successful persuasion.

Rationality Presenting rational arguments can be important. If a persuader wants to convince the persuadee of an argument (a persuasion argument), then this includes acceptability of the persuasion argument (against counterarguments), believing the premises of the persuasion argument, fit of persuasion argument with agenda, goals, preferences, etc, quality of constellation of arguments considered (balance, depth, breadth, understandability, etc).

Emotion Presenting emotional arguments can be important. For example, *you have a good income, and so you should feel guilty if you do not donate money to this emergency appeal by Médecins Sans Frontières*. As another example, *your parents will be proud of you if you complete your thesis and get your PhD award*. Note, emotional arguments contrast with evidential/logical arguments (e.g. *You will have a much higher chance of getting a highly paid job if you complete your thesis and get your PhD award*).

The above dimensions that can affect the success of argumentation can be considered together in the following criterion for successful persuasion.

Selectivity Persuasion does not involve exhaustive presentation of all possible arguments [8]. Rather it requires careful selection of arguments that are most likely to be efficacious in changing the mind of the persuadee. Deciding on which arguments to select depends on diverse features of the arguments and the persuadee such as the nature of the persuader, the language of the arguments, use of psychological techniques, personality of the persuadee, use of rational and/or emotional argumentation, etc.

Being selective does not mean that argumentation needs to be constrained in any way other than being the most efficacious for persuasion. In particular, I would like to make the following claim.

Persuasion is not normative There are no underlying rules or principles to the use of argumentation in persuasion. This means for instance that arguments can be inconsistent, irrational, untrue, etc. if they persuade. Though inconsistent, irrational, untrue arguments may be counter-productive with some audiences, as well as being potentially problematic from moral, ethical, and regulatory perspectives.

A corollary of the above claim is that how convincing an argument is does not equal how correct it is. For example, arguments like *homeopathy focuses on processes of health and illness rather than states, and therefore it is better than regular medicine* and *the sheer weight of anecdotal evidence gives rise to the common-sense notion that there must be some basis for homeopathic therapies by virtue of the fact that they have lasted this long* can be convincing for some audiences.

3. What is computational persuasion?

An **automated persuasion system** (APS), i.e. a persuader, is a system that can engage in a dialogue with a user, i.e. a persuadee, in order to persuade that persuadee to do (or not do) some action or to believe (or not believe) something. To do this, an APS aims to use convincing arguments in order to persuade the persuadee. The dialogue may involve moves including queries, claims, and importantly, arguments and counterarguments, that are presented according to some protocol. Whether an argument is convincing depends on the context, and on the characteristics of the persuadee. An APS maintains a model of the persuadee,

and this is harnessed by the strategy of the APS in order to choose good moves to make in the dialogue.

Computational persuasion is the study of formal models of dialogues involving arguments and counterarguments, of persuadee models, and strategies, for APSs. Therefore, developments in computational persuasion build on computational models of argument. Note, the aim of computational persuasion is not to produce models of human persuasion (c.f. [11]), rather it is to produce models of persuasion that can be used by computers to persuade humans, and that they can be shown to have a reasonable success rate in some persuasion goal (i.e. that a reasonable proportion of the users are persuaded by the arguments and therefore do the action or accept the belief).

3.1. What do computational models of argument offer?

Computational persuasion is based on computational models of argument. These models are being developed to reflect aspects of how humans use conflicting information by constructing and analyzing arguments. A number of models have been developed, and some basic principles established. We can group much of this work in four levels as follows (with only examples of relevant citations).

Dialectical level Dialectics is concerned with determining which arguments win in some sense. In abstract argumentation, originally proposed in the seminal work by Dung [23], arguments and counterarguments can be represented by a graph. Each node denotes an argument, and each arc denotes one argument attacking another argument. Dung defined some principled ways to identify extensions of an argument graph. Each extension is a subset of arguments that together act as a coalition against attacks by other arguments. An argument in an extension is, in a sense, acceptable. Methods for argument dynamics ensure that specific arguments hold in the extensions of the argument graph such as epistemic enforcement in abstract argumentation [4,3,18], revision of argument graphs [19,20], and belief revision in argumentation (e.g. [14,27,10,22]).

Logical level At the dialectic level, arguments are atomic. They are assumed to exist, but there is no mechanism for constructing them. Furthermore, they cannot be divided or combined. To address this, the logical level provides a way to construct arguments from knowledge. At the logical level, an argument is normally defined as a pair $\langle \Phi, \alpha \rangle$ where Φ is a minimal consistent subset of the knowledgebase (a set of formulae) that entails α (a formula). Here, Φ is called the support, and α is the claim, of the argument. Hence, starting with a set of formulae, arguments and counterarguments can be generated, where a counterargument (an argument that attacks another argument) either rebuts (i.e. negates the claim of the argument) or undercuts (i.e. negates the support of the argument). A range of options for structured argumentation at the logic level have been investigated (see [9,61,64,28] for tutorial reviews of some of the key proposals).

Dialogue level Dialogical argumentation involves agents exchanging arguments in activities such as discussion, debate, persuasion, and negotiation. Starting with [31,43], dialogue games are now a common approach to characterizing argumentation-based agent dialogues (e.g. [1,12,21,24,45,46,50,51,65]). Dialogue games are normally made up of a set of communicative acts called moves, and a

protocol specifying which moves can be made at each step of the dialogue. Dialogical argumentation can be viewed as incorporating logic-based argumentation, but in addition, dialogical argumentation involves representing and managing the locutions exchanged between the agents involved in the argumentation. The emphasis of the dialogical view is on the interactions between the agents, and on the process of building up, and analyzing, the set of arguments until the agents reach a conclusion. See [52] for a review of formal models of persuasion dialogues and [62,13] for reviews and analyses of strategies in dialogical argumentation.

Rhetorical level Normally argumentation is undertaken in some wider context of goals for the agents involved, and so individual arguments are presented with some wider aim. For instance, if an agent is trying to persuade another agent to do something, then it is likely that some rhetorical device is harnessed and this will affect the nature of the arguments used (e.g. a politician may refer to *investing in the future of the nation's children* as a way of persuading colleagues to vote for an increase in taxation). Aspects of the rhetorical level include believability of arguments from the perspective of the audience [32], impact of arguments from the perspective of the audience [33], use of threats and rewards [2], appropriateness of advocates [34], and values of the audience [5,6,48].

So computational models of argument offer a range of formal systems for generating and comparing arguments, and for undertaking this in a dialogue.

3.2. Shortcomings in the state of the art

However there are shortcomings in the state of the art of computational models of argument for application in persuasion. The current state of the literature does not adequately offer the following and hence there are some exciting research challenges to be addressed if we are to deliver computational persuasion.

Domain knowledge A formalization of domain knowledge appropriate for constructing arguments concerning behaviour change (e.g. a formalism for representing persuadee preferences, persuadee goals, persuadee preferences, system persuasion goals, and system knowledge concerning actions that can address persuadee goals, etc) though the multiagent communities offer proposals that might be adapted for our needs.

Persuasion protocols Protocols that take account of humans unable to make rich input (since we are not supporting free text input from the persuadee).

Persuadee models Persuadee models that allow the persuasion system to construct a model of the persuadee's beliefs and preferences, to qualify the probabilistic uncertainty of that model, and to update that model and the associated uncertainty as the dialogue progresses, though some promising proposals could contribute to our solution (e.g. [29,36,58,38]).

Persuasion strategies Strategies for persuasion that harness the persuadee model to find optimal moves to make at each stage (trading the increase in probability of successfully persuading the persuadee against the raised risk that the persuadee disengages from the dialogue as it progresses).

In order to focus research on addressing these shortcomings, we can consider how computational persuasion can be developed and evaluated in the context of behaviour change applications.

Field	Examples of behaviour change topic
Healthy life-styles	eating fewer calories, eating more fruit and veg, doing more exercise, drinking less alcohol
Addiction management	gambling, smoking, drugs
Treatment compliance	self-management of diabetes, taking vaccines, completing course of antibiotics
Personal finance	borrowing less, saving more
Education	starting or continuing with a course, studying properly
Energy efficiency	reducing electricity consumption, installing home insulation
Citizenship	voting, recycling, contributing to charities, wasting less food
Safe driving	not exceeding speed limits, not texting while driving
Anti-social behaviour	aggression, vandalism, racism, sexism, trolling

Table 1. Some examples where people could change their behaviour and for which there would be a substantial quantifiable benefit to themselves, and/or to society.

4. What is behaviour change?

There is a wide variety of problems that are dangerous or unhealthy or unhelpful for an individual, or for those around them, and that are expensive to government and/or to society (see Table 1 for examples). For each type of problem, we can conceivably tackle a small proportion of cases with substantial benefit to individuals, government and society using techniques for behaviour change.

Many organizations are involved in behaviour change, and many approaches are used to persuade people to change their behaviour including counselling, information resources, and advertising. Many diverse factors can influence how such approaches can be used effectively in practice such as the following.

- Perceived social norms (e.g. everyone drives above the speed limit).
- Social pressure (e.g. my friends laugh at me if I drive slowly).
- Emotional issues (e.g. speeding is cool).
- Agenda (e.g. I am always late for everything, and so I have to speed).
- Perception of an issue (e.g. I am a good driver even if I speed).
- Opportunities to change behaviour (e.g. access to a race track on which to drive fast instead of driving fast on ordinary roads).
- Attitude to persuader (e.g. I listen to Lewis Hamilton not a civil servant).
- Attitude to information (e.g. I switch off if I am given statistics).

As computing becomes involved in every sphere of life, so too is persuasion a target for applying computer-based solutions. There are **persuasion technologies** that have come out of developments in human-computer interaction research (see for example the influential work by Fogg [26]) with a particular emphasis on addressing the need for systems to help people make positive changes to their behaviour, particularly in healthcare and healthy life-styles.

Many of these persuasion technologies for behaviour change are based on some combination of questionnaires for finding out information from users, provision of information for directing the users to better behaviour, computer games to enable users to explore different scenarios concerning their behaviour, provision

of diaries for getting users to record ongoing behaviour, and messages to remind the persuadee to continue with the better behaviour.

Interestingly, argumentation is not central to the current manifestations of persuasion technologies. The arguments for good behaviour seem either to be assumed before the persuadee accesses the persuasion technology (e.g. when using diaries, or receiving email reminders), or arguments are provided implicitly in the persuasion technology (e.g. through provision of information, or through game playing). So explicit consideration of arguments and counterarguments are not supported with existing persuasion technologies. This creates interesting opportunities for computational persuasion to develop APSs for behaviour change where arguments are central.

5. How can computational persuasion be applied?

Computational models of argument drawing on ideas of abstract argumentation, logical argumentation, dialogical argumentation, together with techniques for argument dynamics and for rhetorics, offer an excellent starting point for developing computational persuasion for applications in behaviour change.

I assume that an APS for behaviour change is a software application running on a desktop or mobile device. Some difficult challenges to automate persuasion via an app are the following.

1. Need asymmetric dialogues without natural language interface.
2. Need short dialogues to keep engagement.
3. Need well-chosen arguments to maximize impact.
4. Need to model the user in order to be able to optimize the dialogue.
5. Need to learn from previous interactions with the agent or similar agents.
6. Need to model the domain to generate arguments/counterarguments.

The dialogue may involve steps where the system finds out more about the persuadee's beliefs, intentions and desires, and where the system offers arguments with the aim of changing the persuadee's beliefs, intentions and desires. The system also needs to handle objections or doubts (represented by counterarguments) with the aim of providing a dialectically winning position. To illustrate how a dialogue can lead to the presentation of an appropriate context-sensitive argument consider the example in Table 2. In this, only the APS presents arguments, and when it is the user's turn s/he can only answer questions (e.g. yes/no questions) or select arguments from a menu. In Figure 1, a dialogue step is illustrated where a user can state the degree of agreement or disagreement in an argument.

Arguments can be automatically generated from a knowledgebase. For this, we can build a knowledgebase for each domain, though there are many commonalities in the knowledge required for each behaviour change application.

- Persuadee beliefs (e.g. cakes give a sugar rush).
- Persuadee preferences (e.g. burgers are preferred to apples).
- Behavioural states (e.g. persuadee's weight, exercise regime, etc.).
- Behavioural actions (e.g. eat a piece of fruit, eat a piece of cake, walk 1km).
- Behavioural goals (e.g. lose 10Kg by Christmas, reduce sugar intake).

Step	Who	Move
1	APS	To improve your health, you could join an exercise class
2	User	Exercise classes are boring
3	APS	For exciting exercise, you could do an indoor climbing course
4	User	It is too expensive
5	APS	Do you work?
6	User	No
7	APS	If you are registered unemployed, then the local sports centre offers a free indoor climbing course
8	APS	Would you try this?
9	User	Yes

Table 2. Simple example of an asymmetric dialogue between a user and an APS. As no natural language processing is assumed, the arguments posted by the user are actually selected by the user from a menu provided by the APS.

Since you do little exercise, you should do a regular exercise class

When I do exercise, I get very hungry and I put on weight

Strongly agree

Agree

Neither agree nor disagree

Disagree

Strongly disagree

Figure 1. Interface for an asymmetric dialogue move for asking the user's belief in an argument. The top argument is by the APS, and the second argument is a counterargument presented by the APS. The user uses the menu to give his/her belief in the counterargument.

To represent and reason with the domain knowledge, we can harness a form of BDI calculus in predicate logic for relating beliefs, behavioural goals, and behavioural states, to possible actions. We can then use the calculus with logical argumentation to generate arguments for persuasion. A small example of an argument graph that we might want to generate by this process is given in Figure 2 including the persuasion goal *giving up smoking will be good for your health*.

To support the selection of arguments, we require persuadee models. For this,

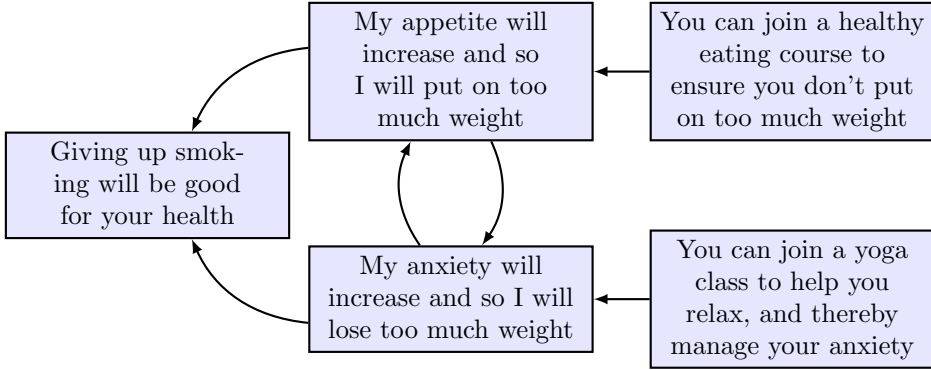


Figure 2. Example of an argument graph for persuasion

we can establish the probabilistic uncertainty associated with the APS model of the persuadee's beliefs, behavioural state, behavioural goals, preferences, and tendencies etc by asking the persuadee appropriate questions, by considering previous usage of the APS by the persuadee, and by the general class of the persuadee (i.e. by assignment to a built-in model learned from a class of similar users). Key possible dimensions for modelling uncertainty are given in Table 3.

Two main approaches to probabilistic argumentation are the constellations and the epistemic approaches [37].

- In the constellations approach, the uncertainty is in the topology of the graph (see for example [42,35]). As an example, this approach is useful when one agent is not sure what arguments and attacks another agent is aware of, and so this can be captured by a probability distribution over the space of possible argument graphs.
- In the epistemic approach, the topology of the argument graph is fixed, but there is uncertainty about whether an argument is believed [63,37,41]. A core idea of the epistemic approach is that the more likely it is to believe in an argument, the less likely it is to believe in an argument attacking it. The epistemic approach can give a finer grained version of Dung's approach, and it can be used to give a valuable alternative to Dung's approach. For example, for a graph containing arguments A and B where B attacks A , it might be the case that a user believes A and not B , and if so the epistemic extension (the set of believed arguments) would be $\{A\}$ which is in contrast the Dung's approach where the only extension is $\{B\}$.

There are approaches to bringing probability theory into systems for dialogical argumentation. A probabilistic model of the opponent has been used in a dialogue strategy allowing the selection of moves for an agent based on what it believes the other agent is aware of [57]. In another approach to probabilistic opponent modelling, the history of previous dialogues is used to predict the arguments that an opponent might put forward [29]. Though further avenues need to be explored.

The constellations approach can model the uncertainty about the structure of the graph in the persuadee mind. We can update the model with each argu-

Type of uncertainty	Modelling technique
Beliefs of persuadee	Epistemic approach
Arguments/attacks known by persuadee	Constellations approach
Moves that persuadee makes	PFSMs/POMDPs
Risk of disengagement	Markov models

Table 3. Possible dimensions of uncertainty in models of persuadee

ment/attack presented. Also, we can use expected utility to identify best choice of argument/attack to present [40].

The epistemic approach is useful for asymmetric dialogues where the user is not allowed to posit arguments or counterarguments [39]. So the only way the user can treat arguments that s/he does not accept is by disbelieving them. In contrast, in symmetric dialogues, the user could be allowed to posit counterarguments to an argument that s/he does not accept. The distribution can be updated in response to moves made (posits, answers to queries, etc) using different assumptions about the persuadee (credulous, skeptical, rational, etc). The aim is to choose moves that will increase belief in positive persuasion goals or decrease belief in negative persuasion goals.

For modelling the possible dialogues that might be generated by a pair of agents, a probabilistic finite state machine can represent the possible moves that each agent can make in each state of the dialogue assuming a set of arguments that each agent is aware of [38]. Each state is composed of the public state of the dialogue (e.g. what has been said) and the private state of each participant (e.g. the arguments they believe). We can find optimal sequences of moves by handling uncertainty concerning the persuadee using partially observable markov decision processes (POMDPs) when there is uncertainty about the private state of the persuader [30].

A strategy for an APS needs to find the best choice of move at each stage where best is determined in terms of some combination of the need to increase the likelihood that the persuadee is persuaded by the goal of the persuasion, and the need to decrease the likelihood that the persuadee disengages from the dialogue. For instance, at a certain point in the dialogue, the APS might have a choice of two arguments *A* and *B* to present. Suppose *A* involves further moves to be made (e.g. supporting arguments) whereas *B* is a single posit. So choosing *A* requires a longer dialogue (and higher probability of disengagement) than *B*. However, if the persuadee keeps to the end of each dialogue, then it is more likely that the persuadee believes *A* than *B*. An APS should present arguments and counterarguments that are informative, relevant, and believable, to the persuadee. If the APS presents uninformative, irrelevant, or unbelievable arguments (from the perspective of the persuadee), the probability of successful persuasion is reduced, and it may alienate the persuadee. A choice of strategy depends on the protocol, and on the kind of dynamic persuadee model. Various parameters can be considered in the strategy such as the preferences of the persuadee, the agenda of the persuadee, etc.

Probabilistic models of the opponent have been used in some strategies allowing the selection of moves for an agent based on what it believes the other

agent believes [36]. Utility theory has also been considered in argumentation (for example [54,59,44,49]) though none of these represents the uncertainty of moves made by each agent in argumentation. Probability theory and utility theory (using decision theory) has been used in [40] to identify outcomes with maximum expected utility where outcomes are specified as particular arguments being included or excluded from extensions. Strategies in argumentation have also been analyzed using game theory [53,55,25], though these are more concerned with issues of manipulation, rather than persuasion.

Given that we need to consider multiple dimensions in identifying a more convincing argument (e.g. whether an argument is believed, whether an argument is undefeated, whether it is relevant, whether it relates to the goals of the persuadee, etc), there is a need to generalize the existing proposals for strategies for argumentation.

6. Discussion

Computational persuasion, being based on computational models of argument, is a promising approach to technology for behaviour change applications. Developing an APS involves research challenges including: undertaking the dialogue without using natural language processing; having an appropriate model of the domain in order to identify arguments; having an appropriate dynamic model of the persuadee; and having a strategy that increases the probability of persuading the persuadee. Furthermore, with even a modest set of arguments, the set of possible dialogues can be enormous, and so the protocols, persuadee models, and strategies need to be computationally viable.

In the short-term, we may envisage that the dialogues between an APS and a user involve limited kinds of interaction. For example, the APS manages the dialogue by asking queries of the persuadee, where the allowed answers are given by a menu or are of restricted types (e.g. age), and by positing arguments, and the persuadee may present arguments that are selected from a menu presented by the APS. Obviously richer natural language interaction would be desirable, but it is not feasible in the short-term. Even with such restricted asymmetric dialogues, it may be possible that effective persuasion can be undertaken, and furthermore, we need to investigate this conjecture empirically with participants.

There are some investigations of computational models of argument with participants. In a study by Rahwan *et al* [56], participants were given argument graphs and asked about their confidence in specific arguments being acceptable or unacceptable. Interestingly, for an unattacked argument A that is then attacked by a new argument B , the confidence in A being acceptable does not necessarily fall to zero (as would be predicted by the usual dialectical semantics for abstract argumentation). Then if a further new argument C is added that attacks B , the confidence in A being acceptable does not necessarily rise to 1 (as would be predicted by the usual dialectical semantics for abstract argumentation). In another study, Cerutti *et al* [15], investigated how well an approach to structured argumentation by Prakken and Sartor models how a group of participants reason with three different argumentation scenarios. Their results showed that a corre-

spondence between the acceptability of arguments by participants and the justification status predicated by the structured argumentation in the majority of the cases. But in some cases, the implicit knowledge about domains could substantially affect this. In a study of argumentation dialogues, Rosenfeld and Kraus [60] undertook studies with participants in order to develop a machine learning-based approach to predict the next move a participant would make in a dialogue. Emotion in argumentation has also been the subject of a study with participants in a debate where the emotional state was estimated from EEG data and automated facial expression analysis. In this study, Benlamine *et al* [7] showed for instance that the number and the strength of arguments, attacks and supports exchanged between a participant could be correlated with particular emotions of the participant. There are also relevant studies investigating the efficacy of using arguments as a way of persuading people when compared with other counselling methods indicating that argumentation may have disadvantages if used inappropriately [47]. Whilst these studies only consider some aspects of computational models of argument, they point to the need for further studies with participants if we are to develop a well-understood and well-grounded framework for computational persuasion.

Acknowledgements

This research is part-funded by EPSRC grant EP/N008294/1 *Framework for Computational Persuasion*¹

References

- [1] L. Amgoud, N. Maudet, and S. Parsons. Arguments, dialogue and negotiation. In *Proceedings of ECAI'00*, pages 338–342. IOS Press, 2000.
- [2] L. Amgoud and H. Prade. Formal handling of threats and rewards in a negotiation dialogue. In *Proceedings of AAMAS'05*, pages 529–536, 2005.
- [3] R. Baumann. What does it take to enforce an argument? minimal change in abstract argumentation. In *Proc. of ECAI'12*, pages 127–132, 2012.
- [4] R. Baumann and G. Brewka. Expanding argumentation frameworks: Enforcing and monotonicity results. In *Proc. of COMMA'10*, pages 75–86, 2010.
- [5] T. Bench-Capon. Persuasion in practical argument using value based argumentation-frameworks. *Journal of Logic and Computation*, 13(3):429–448, 2003.
- [6] T. Bench-Capon, S. Doutre, and P. Dunne. Audiences in argumentation frameworks. *Artificial Intelligence*, 171(1):42–71, 2007.
- [7] S. Benlamine, M. Chaouachi, S. Villata, E. Cabrio, C. Frasson, and F. Gandon. Emotions in argumentation: an empirical evaluation. In *Proc. of IJCAI'15*, pages 156–163, 2015.
- [8] Ph. Besnard and A. Hunter. *Elements of Argumentation*. MIT Press, 2008.
- [9] Ph. Besnard and A. Hunter. Constructing argument graphs with deductive arguments: a tutorial. *Argument and Computation*, 5(1):5–30, 2014.
- [10] P. Bisquert, C. Cayrol, F. Dupin de Saint-Cyr, and M.-C. Lagasquie-Schiex. Enforcement in argumentation is a kind of update. In *Proc. of SUM'13*, volume 8078 of *LNCS*, pages 30–42. Springer, 2013.
- [11] P. Bisquert, M. Croitoru, and F. Dupin de Saint-Cyr. Four ways to evaluate arguments according to agent engagement. In *Proc. of Brain Informatics and Health*, volume 9250 of *LNCS*. Springer, 2015.

¹For more information on the project, see www.computationalpersuasion.com.

- [12] E. Black and A. Hunter. An inquiry dialogue system. *Autonomous Agents and Multi-Agent Systems*, 19(2):173–209, 2009.
- [13] E. Black and A. Hunter. Reasons and options for updating an opponent model in persuasion dialogues. In *Proc. of TAFA'15*, volume 9524 of *LNCS*, pages 21–39. Springer, 2015.
- [14] C. Cayrol, F. Dupin de Saint-Cyr, and M.-C. Lagasquie-Schiex. Change in abstract argumentation frameworks: Adding an argument. *Journal of Artificial Intelligence Research*, 38:49–84, 2010.
- [15] F. Cerutti, N. Tintarev, and N. Oren. Formal arguments, preferences, and natural language interfaces to humans: an empirical evaluation. In *Proc. of ECAI'14*, pages 207–212, 2014.
- [16] R. Cialdini. *Influence: The Psychology of Persuasion*. HarperCollins, 1984.
- [17] R. Cockcroft and S. Cockcroft. *Persuading People*. Macmillan, 1992.
- [18] S. Coste-Marquis, S. Konieczny, and J-G. Maily. Extension enforcement in abstract argumentation as an optimization problem. In *Proc. of IJCAI'15*, pages 2876–2882, 2014.
- [19] S. Coste-Marquis, S. Konieczny, and J-G. Maily. On the revision of argumentation systems: Minimal change of argument statuses. In *Proc. of KR'14*, pages 72–81, 2014.
- [20] S. Coste-Marquis, S. Konieczny, and J-G. Maily. A translation-based approach for revision of argumentation frameworks. In *Proc. of JELIA'14*, pages 77–85, 2014.
- [21] F. Dignum, B. Dunin-Keplicz, and R. Verbrugge. Dialogue in team formation. In *Issues in Agent Communication*, pages 264–280. Springer, 2000.
- [22] M. Diller, A. Haret, T. Linsbichler, S. Rümmele, and S. Woltran. An extension-based approach to belief revision in abstract argumentation. In *Proc. of IJCAI'15*, pages 2926–2932, 2015.
- [23] P. Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming, and n-person games. *Artificial Intelligence*, 77:321–357, 1995.
- [24] X. Fan and F. Toni. Assumption-based argumentation dialogues. In *Proc. of IJCAI'11*, pages 198–203, 2011.
- [25] X. Fan and F. Toni. Mechanism design for argumentation-based persuasion. In *Proc. of COMMA'12*, pages 322–333, 2012.
- [26] B. Fogg. Persuasive computers. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*, pages 225–232. CHI, 1998.
- [27] D. Gabbay and O. Rodrigues. A numerical approach to the merging of argumentation networks. In *Proc. of CLIMA'12*, volume 7486 of *LNCS*, pages 195–212. Springer, 2012.
- [28] A. Garcia and G. Simari. Defeasible logic programming: Delp-servers, contextual queries, and explanations for answers. *Argument and Computation*, 5(1):63–88, 2014.
- [29] C. Hadjinikolis, Y. Siantos, S. Modgil, E. Black, and P. McBurney. Opponent modelling in persuasion dialogues. In *Proc. of IJCAI'13*, page 164170, 2013.
- [30] E. Hadoux, A. Beynier, N. Maudet, P. Weng, and A. Hunter. Optimization of probabilistic argumentation with markov decision models. In *Proc. of IJCAI'15*, 2015.
- [31] C. Hamblin. Mathematical models of dialogue. *Theoria*, 37:567–583, 1971.
- [32] A. Hunter. Making argumentation more believable. In *Proc. of AAAI'04*, pages 269–274, 2004.
- [33] A. Hunter. Towards higher impact argumentation. In *Proc. of AAAI'04*, pages 275–280, 2004.
- [34] A. Hunter. Reasoning about the appropriateness of proponents for arguments. In *Proc. of AAAI'08*, pages 89–94, 2008.
- [35] A. Hunter. Some foundations for probabilistic abstract argumentation. In *Proc. of COMMA'12*, pages 117–128. IOS Press, 2012.
- [36] A. Hunter. Modelling uncertainty in persuasion,. In *Proc. of SUM'13*, volume 8078 of *LNCS*, pages 57–70. Springer, 2013.
- [37] A. Hunter. A probabilistic approach to modelling uncertain logical arguments. *International Journal of Approximate Reasoning*, 54(1):47–81, 2013.
- [38] A. Hunter. Probabilistic strategies in dialogical argumentation. In *Proceedings of SUM'14*, volume 8720 of *LNCS*, pages 190–202. Springer, 2014.
- [39] A. Hunter. Modelling the persuadee in asymmetric argumentation dialogues for persua-

- sion. In *Proc. of IJCAI'15*, 2015.
- [40] A. Hunter and M. Thimm. Probabilistic argument graphs for argumentation lotteries. In *Computational Models of Argument (COMMA'14)*, 2014.
 - [41] A. Hunter and M. Thimm. Probabilistic argumentation with incomplete information. In *Proc. of ECAI'14*, pages 1033–1034, August 2014.
 - [42] H. Li, N. Oren, and T. J. Norman. Probabilistic argumentation frameworks. In *Proc. of TAFE'11*, 2011.
 - [43] J. Mackenzie. Question begging in non-cumulative systems. *Journal of Philosophical Logic*, 8:117–133, 1979.
 - [44] P. Matt and F. Toni. A game-theoretic measure of argument strength for abstract argumentation. In *Proceedings of JELIA'08*, volume 5293 of *LNCS*, pages 285–297, 2008.
 - [45] P. McBurney and S. Parsons. Games that agents play: A formal framework for dialogues between autonomous agents. *Journal of Logic, Language and Information*, 11:315–334, 2002.
 - [46] P. McBurney, R. van Eijk, S. Parsons, and L. Amgoud. A dialogue-game protocol for agent purchase negotiations. *Journal of Autonomous Agents and Multi-Agent Systems*, 7:235–273, 2003.
 - [47] H. Nguyen and J. Masthoff. Designing persuasive dialogue systems: Using argumentation with care. In *Proceedings of Persuasive technology'08*, pages 201–212, 2008.
 - [48] N. Oren, K. Atkinson, and H. Li. Group persuasion through uncertain audience modelling. In *Proc. of COMMA'12*, pages 350–357, 2012.
 - [49] N. Oren and T. Norman. Arguing using opponent models. In *Proc. of ArgMAS'09*, volume 6057 of *LNCS*, pages 160–174, 2009.
 - [50] S. Parsons, M. Wooldridge, and L. Amgoud. Properties and complexity of some formal inter-agent dialogues. *Journal of Logic and Computation*, 13(3):347–376, 2003.
 - [51] H. Prakken. Coherence and flexibility in dialogue games for argumentation. *Journal of Logic and Computation*, 15(6):1009–1040, 2005.
 - [52] H. Prakken. Formal systems for persuasion dialogue. *Knowledge Engineering Review*, 21(2):163–188, 2006.
 - [53] I. Rahwan and K. Larson. Mechanism design for abstract argumentation. In *Proc. of AAMAS'08*, pages 1031–1038, 2008.
 - [54] I. Rahwan and K. Larson. Pareto optimality in abstract argumentation. In *Proc. of AAAI'08*, 2008.
 - [55] I. Rahwan, K. Larson, and F. Tohmé. A characterisation of strategy-proofness for grounded argumentation semantics. In *Proc. of IJCAI'09*, pages 251–256, 2009.
 - [56] I. Rahwan, M. Madakkatell, J. Bonnefon, R. Awan, and S. Abdallah. Behavioural experiments for assessing the abstract argumentation semantics of reinstatement. *Cognitive Science*, 34(8):1483–1502, 2010.
 - [57] T. Rienstra. Towards a probabilistic dung-style argumentation system. In *Proc. of Agreement Technologies (AT'12)*, 2012.
 - [58] T. Rienstra, M. Thimm, and N. Oren. Opponent models with uncertainty for strategic argumentation. In *Proc. of IJCAI'13*. IJCAI/AAAI, 2013.
 - [59] R. Riveret, H. Prakken, A. Rotolo, and G. Sartor. Heuristics in argumentation: A game theory investigation. In *Proc. of COMMA'08*, pages 324–335, 2008.
 - [60] A. Rosenfeld and S. Kraus. Providing arguments in discussions based on the prediction of human argumentative behavior. In *Proc. of AAAI'15*, pages 1320–1327, 2015.
 - [61] S. Modgil and H. Prakken. The ASPIC+ framework for structured argumentation: a tutorial. *Argument and Computation*, 5(1):31–62, 2014.
 - [62] M. Thimm. Strategic argumentation in multi-agent systems. *Kunstliche Intelligenz*, 28:159–168, 2014.
 - [63] M. Thimm. A probabilistic semantics for abstract argumentation. In *Proc. of ECAI'12*, August 2012.
 - [64] F. Toni. A tutorial on assumption-based argumentation. *Argument and Computation*, 5(1):89–117, 2014.
 - [65] D. Walton and E. Krabbe. *Commitment in Dialogue: Basic Concepts of Interpersonal Reasoning*. SUNY Press, 1995.

Innovative Applications

This page intentionally left blank

Tweeties Squabbling: Positive and Negative Results in Applying Argument Mining on Social Media

Tom BOSCH^a, Elena CABRIO^a, Serena VILLATA^b

^a Université Côte d'Azur, Inria, CNRS, I3S, France.

e-mail: tom.bosc@inria.fr; elena.cabrio@unice.fr

^b Université Côte d'Azur, CNRS, Inria, I3S, France.

e-mail: villata@i3s.unice.fr

Abstract. The problem of understanding the stream of messages exchanged on social media such as Facebook and Twitter is becoming a major challenge for automated systems. The tremendous amount of data exchanged on these platforms as well as the specific form of language adopted by social media users constitute a new challenging context for existing argument mining techniques. In this paper, we describe an ongoing work towards the creation of a complete argument mining pipeline over Twitter messages: (i) we identify which tweets can be considered as arguments and which cannot, (ii) over the set of tweet-arguments, we group them by topic, and (iii) we predict whether such tweets support or attack each other. The final goal is to compute the set of tweets which are widely recognized as accepted, and the different (possibly conflicting) viewpoints that emerge on a topic, given a stream of messages.

Keywords. Argument mining, Social media, Supervised classification approaches

1. Introduction

Argumentation has come to be increasingly central as a main study within Artificial Intelligence, due to its ability to conjugate representational needs with user-related cognitive models and computational models for automated reasoning. An important source of data for many of the disciplines interested in such studies is the Web, and social media in particular. Newspapers, microblogs, online debate platforms and social networks provide an heterogeneous flow of information where natural language arguments can be identified and analyzed. The availability of such data, together with the advances in Natural Language Processing and Machine Learning, supported the rise of a new research area called *argument mining*, whose main goal is the automated extraction of natural language arguments and their relations from generic textual corpora, with the final purpose of providing machine-processable data for computational models of argument.

Despite the increasing amount of argument mining approaches [21], none of them has tackled the challenge of extracting arguments and their relations on social media like Twitter or Facebook. Such a kind of natural language arguments raises further issues in

addition to the standard problems faced by argument mining approaches typically dealing with newspapers, novels or legal texts: messages from Twitter are squeezed, noisy and often unstructured. More specifically, the following issues have to be considered: *i)* the 140-characters limit forces users to express their ideas very succinctly; *ii)* the quality of the language in Twitter is deteriorated, including a lot of variants in spelling, mistakes and abbreviations, and *iii)* Twitter's API filters tweets on hashtags but cannot retrieve all the replies to these tweets if they do not contain the same hashtags.

In this paper, we provide a preliminary answer to the following research question: *how to extract the arguments and predict the relations among them on Twitter data?* and we highlight the open challenges still to be addressed. We consider both the two main stages in the typical argument mining pipeline, from the unstructured natural language documents towards structured data: we first detect arguments within the natural language texts from Twitter, the retrieved arguments will thus represent the nodes in the final argument graph returned by the system, and second, we predict what are the relations, i.e., *attack* or *support*, holding between the arguments identified in the first stage.

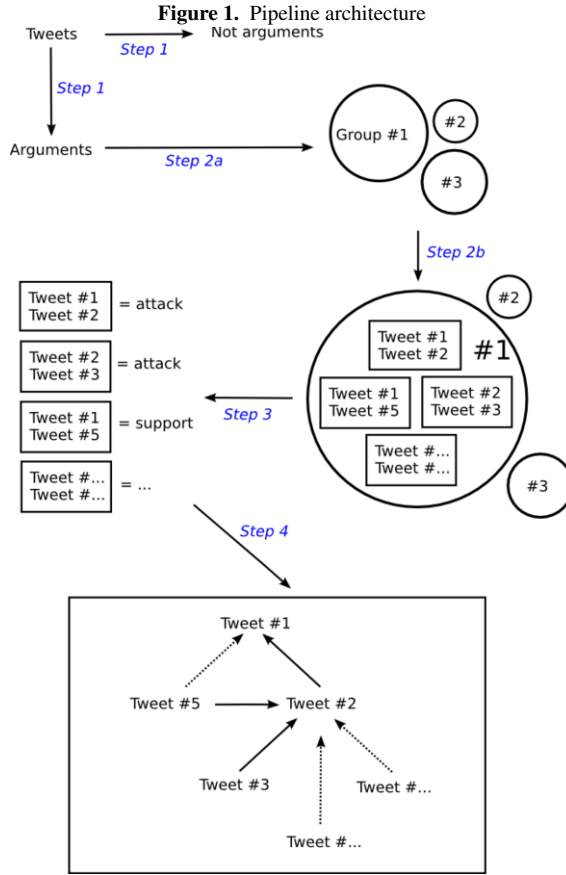
The main advantage of our approach is that it provides a whole argument mining pipeline to analyze flows of tweets, allowing for the application of reasoning techniques over the output structured data, like the identification of the set of widely accepted arguments or trends analysis. However, being it an ongoing work, we highlight in this paper both positive and negative results in applying argument mining on Twitter data, analyzing solutions and potential alternatives to be explored.

The paper is organized as follows. Section 2 presents our argument mining framework and its evaluation, and Section 3 compares the proposed approach with the related literature. Section 3 describes relevant works in the literature, while Conclusions end the paper, drawing final remarks and describing future work.

2. Argument Mining on Twitter

The argument mining pipeline we propose, visualized in Figure 1, is composed of four main steps, that consist in: *i)* separating tweet-arguments from non-argument tweets; *ii)* grouping tweet-arguments discussing about the same issue, and create pairs of arguments; *iii)* predicting the relations of *attack* and *support* among the tweets in the pairs; and *iv)* building argumentation graphs.

First of all, we need to clarify what we mean by *argument* in this paper: an argument gives a reason to support a claim that is questionable, or open to doubt. In the computational models of argument field, an argument is made of three components: the *premises* representing the reason, a *conclusion* which is the supported claim, and a *relation* showing how the premises lead to this conclusion. Facing the issue of dealing with Twitter data, i.e., dealing with textual arguments of length inferior or equal to 140 characters, we (almost) never find such a kind of complete structure of the arguments. We have thus labeled as arguments all those text snippets providing a portion of a standard argument structure, e.g., opinions under the form of claim, data like in the Toulmin model [30], or persuasive conclusions. Future work includes the “composition” of such elements to build a single well-structured argument. Second, it is worth noticing that the support and the attack relations are not symmetric: we considered the temporal dimension to decide the direction of these relations, i.e., a tweet that is proposed at time $t + 1$ attacks (resp.



supports) a tweet which has been provided at time t . In the following, each step of the pipeline is described in detail, together with the experimented approach, and the obtained results of this ongoing work.

Dataset. Up to our knowledge, DART [7] is the only existing dataset of arguments and their relations on Twitter, therefore it has been chosen to test our pipeline. It is composed of:

- (a) 4000 tweets annotated as argument/not argument: 1000 tweets for each of the following 4 topics: the letter to Iran written by 47 senators on 10/03/2015; the referendum in Greece for or against Greece leaving European Union on 10/07/2015; the release of Apple Watch on 10/03/2015; the airing of episode 4 (season 5) of the serie Game of Thrones on 4/05/2015. A tweet is annotated as argument if it contains an opinion or factual information, or if it is a claim expressed as question (rhetorical questions, attempts to persuade, containing sarcasms/irony). The argument annotation task is carried out on a single tweet and not on subparts of it. A text containing an opinion is considered as an argument. For example, in the following tweet the opinion of the author is clearly expressed in the second sentence (i.e., *I won't be running out to get one*):

RT @mariofraioli: What will #AppleWatch mean for runners? I can't speak for everyone, but I won't be running out to get one. Will you? <http://t.co/xBpj0HWK>

We consider as arguments also claims expressed as questions (either rhetorical questions, attempts to persuade, containing sarcasm or irony), as in the following example:

RT @GrnEyedMandy: What next Republicans? You going to send North Korea a love letter too? #47Traitors

or:

Perhaps Apple can start an organ harvesting program. Because I only need one kidney, right? #iPadPro #AppleTV #AppleWatch

Tweets containing factual information are annotated as arguments, given that they can be considered as premises or conclusions. For example:

RT @HeathWallace: You can already buy a fake #AppleWatch in China <http://t.co/WpHEDqYuUC> via @cnnnews @mr_gadget <http://t.co/WhcMKuM>

Defining the amount of world knowledge needed to determine whether a text is a fact or an opinion when it contains unknown acronyms and abbreviations can be pretty tough. Consider the following tweet:

RT @SaysSheToday: The Dixie Chicks were attacked just for using IA right to say they were ashamed of GWB. They didn't commit treason like the #47Senators

where the mentioned entities *The Dixie Chicks*, *GWB*, and *IA right* are strictly linked to the US politics, and hardly interpreted by people out of the US politics matters. In this case, annotators are asked to suppose that the mentioned entities exist, and focus on the phrasing of the tweets.

However, if tweets contain pronouns only (preventing the understanding of the text), we consider such tweets as not “self-contained”, and thus non arguments. It can be the case of replies, as in the following example, in which the pronoun *he* is not referenced anywhere in the tweet.

@FakeGhostPirate @GameOfThrones He is the one true King after all ;)

For tweets containing an advertisement to push into visiting a web page, if an opinion or factual information is also present, then the tweet is considered as an argument, otherwise it is not. Consider the following example:

RT @NewAppleDevice: Apple's smartwatch can be a games platform and here's why <http://t.co/uIMGDyw08I>

It contains factual information that can be understood even without visiting the link. On the contrary, the following tweet is not an argument, given that it does not convey an independent message while excluding the link:

For all #business students discussing #AppleWatch this morning. Give it a test drive thanks to @UsVsTh3m: <http://t.co/x2bGc9j1GI>.

- (b) 2181 tweet-arguments on the Apple Watch release classified in 7 categories (i.e. *features (F)*, *price (P)*, *look (L)*, *buying announcement (B)*, *advertisement (A)*, *forecast on the product success (S)*, *news (N)*, *others (O)*) (see Table 3). Moreover, the tweets contained in the category *features* have been grouped in the following more fine-grained categories: *health*, *innovation*, *battery*.
- (c) 1891 pairs of tweet-arguments of the categories: *price*, *health*, *look*, *predictions* annotated with the following relations: *support* (446), *attack* (122), *unknown* (1323). After a first annotation round to test the guidelines provided in [9], we realized that a few additional instructions should be added with the aim to consider the specificity of the Twitter scenario. The instructions we introduced are as follows: If both Tweet-A and Tweet-B in a pair are factual tweets, and they are related to the same issue, the pair must be annotated as *support*, as in:

Tweet-A: .@AirStripmHealth + #AppleWatch provides HIPPA compliant capabilities for physicians, mothers, babies, and more #AppleEvent

Tweet-B: accessible heart rate monitors and opinions on that #iWatch #apple #accessibility #ios <https://t.co/ySYM8dkOPf> via @audioBoom

If both Tweet-A and Tweet-B in a pair are opinion tweets, and they are related to the same issue, the pair must be annotated as *support*, as in:

Tweet-A: Think of how much other stuff you can buy with the money you spend on an #AppleWatch

Tweet-B: #AppleWatch Tempting, but not convinced. #appletv Yes.
#iPhone6sPlus No plan to upgrade #iPadPro little high price, wait & watch

If Tweet-B is a factual tweet, and Tweet-A is an opinion on the same issue, the pair must be annotated as *support*, as in:

Tweet-A: Wow. Your vitals on your iwatch. That's bonkers. #AppleEvent

Tweet-B: accessible heart rate monitors and opinions on that #iWatch #apple #accessibility #ios <https://t.co/ySYM8dkOPf> via @audioBoom

If Tweet-A is a factual tweet, and Tweet-B expresses someone's wishes to buy the product or an opinion about it, the pair must be annotated as *unknown*, as in:

Tweet-A: Mom can listen to baby's heart rate with #AppleWatch #airstrip

Tweet-B: Wow!!! Look at what the #Ap, pleWatch can do for #doctors that's amazing! Seeing their vitals? I just got chills! In a good way #AppleEvent

Concerning the annotation of the arguments/non arguments, in the reconciliation phase among the three students annotators, the label that was annotated by at least 2 annotators out of 3 was chosen (majority voting mechanism). If all the annotators disagree or if more than one annotator labels the tweet as unknown, then such tweet is discarded. The inter-annotator agreement has been calculated between the expert annotators and the reconciled student annotations on 250 tweets of the first batch, resulting in $\alpha_{47\text{traitors}} = 0.81$ (Krippendorff's α handles missing values, the label "unknown" in our case). Concerning the pair annotation with the support/attack/unknown relations, the inter-annotator agreement has been calculated on 99 pairs (33 pairs randomly extracted from each of the three first topics), resulting in Krippendorff $\alpha = 0.67$.

Dataset	# tweet-arg.	not-arg.	unknown	total
Training set	2079	829	92	3000
Test set	623	352	25	1000

Table 1. Statistics of dataset (a)

Approach	Average F1
baseline	0.64
baseline + tokens	0.66
baseline + tokens + bigrams tokens	0.67

Table 2. Validation of the model and feature use

2.1. Step 1: Argument identification.

The first task in our pipeline is the binary classification of tweets as argument/non argument. To train a generic, domain-independent argument detector, we separate the training, validation and test data according to the topics of dataset (a) to avoid overfitting. We train and validate on the first three topics, and we test on the Apple Watch dataset (Table 1 provides some statistics on the data). We ignore tweets classified as unknown. We use 3-fold cross-validation (we alternately train the model on the tweets of the first two topics and leave the third topic out as a validation set) with randomized hyperparameter search [3].¹ Because the classes are unbalanced and the balance is not necessarily the same across all datasets, the training phase weights the errors inversely proportional to class frequencies.

As baseline, we use raw character counts as features (causing smileys, capital letters, punctuation marks to influence the model). Then, tweets have been tokenized with Twokenize² and annotated with their PoS applying Stanford POS tagger. POS tags are then used as features, as well as bigrams of tags. As a baseline model, we train a logistic regression model³ on these features only.

We also augment features with normalized tokens and bigrams of tokens, and this effectively improves over the baseline (see Table 2). The best model (Logistic regression, L2-penalized with $\lambda = 100$) is obtained by using all the features and re-training on the 3 folds. It yields an F1-score of 0.78 over the test set, that can be considered as satisfactory. The difference between the average F1-score over the validation set (see Table 2) and the F1 over the test set is due to the addition of the tweets of the validation set (around 1000 additional tweets) for training the final model.

¹A randomized hyperparameter search samples parameter settings a fixed number of times and has been found to be more effective in high-dimensional spaces than exhaustive search.

²<http://www.cs.cmu.edu/~ark/TweetNLP/>

³Like all regression analyses, the logistic regression is a predictive analysis. It is used to describe data and to measure the relationship between one dependent variable and one or more independent variables by estimating probabilities using a logistic function, i.e., the cumulative logistic distribution.

	O	A	B	F	L	N	P	S
#	720	175	370	619	205	65	189	112

Table 3. Statistics on dataset (b), # tweets

	F	L	P	S
average F1-score (train set)	0.36	0.57	0.60	0.15
F1-score (test set)	0.56	0.58	0.60	0.00

Table 4. Classification results (step 2)

2.2. Step 2: Pairs creation.

Once we are able to identify tweet-argument, we create pairs of them to predict the relations among them. Given a stream of tweets, it would be impossible to apply a naive approach comparing all the pairs of tweets, since this would lead to the creation of numerous unrelated pairs.

To deal with this issue, we firstly tested the solution of clustering the tweets into *sub-topics*, and then create pairs from these sub-topics. The major problem that we faced is the difficulty of automatically finding meaningful sub-topics. We tested both Latent Dirichlet Allocation⁴ [6] and more powerful models such as Correlated Topic Models⁵ [5], but the interpretability of the clusters did not improve [11].

Instead, since we have classified goldstandard data for Apple Watch (dataset (b), see Table 3), we decided to focus on this topic only, and turn the clustering problem into a classification problem. Another possibility would have been to tune the hyperparameters before applying the clustering algorithms to retrieve the annotated categories, but given the small size of the goldstandard, we could not explore that direction further.

In particular, we focus on categories F (features), L (look), P (price) and S (predictions about the success of the product) because they contain the most interesting tweets. We use the same features and same hyperparameters selection scheme as in step 1. The training set contains 2031 tweets, and the test set contains 150 tweets. The 3 folds are randomly created across all the training set, and we take the average of all the macro F1-scores on all the folds to select the best model. We use regularized logistic regression and the results obtained by the best model (L1-penalized with $\lambda = 100$) are reported in Table 4 for each category, averaged over all the folds. As can be observed, some categories are harder to predict than others, but the performance on the easy classes (F, L, P here) are quite satisfactory. A paraphrase detection tool could be added at this step to deduplicate similar tweets and give more weights to the arguments that are often used in subsequent steps.

⁴Latent Dirichlet allocation is a generative probabilistic model of a corpus. The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words.

⁵Correlated Topic Models use a more flexible distribution for the topic proportions that allows for covariance structure among the components. This gives a more realistic model of latent topic structure where the presence of one latent topic may be correlated with the presence of another.

2.3. Step 3: Relation detection.

Given the pairs of tweet-arguments returned by step 2, the next step consists in predicting the relation holding between the tweets in a pair. Dataset (c) contains ~ 600 tweets each for *look*, *price* and *health* categories of the Apple Watch: we put pairs concerning the product price in the test set, whereas all the other tweets are in the training set. An additional validation set contains 100 tweets on the user predictions on the product success.

Given the closeness of the task with textual entailment [9], we decide to explore first a prediction of the support and attack relations using the Excitement Open Platform (EOP)⁶ for recognizing textual entailment. The intuition is to consider the support relation as an entailment, and the attack relation as a contradiction, following the approach in Cabrio and Villata [8].

In addition, following the same guidelines proposed by [9], pairs are also annotated according to the Recognizing Textual Entailment (RTE) framework, i.e., pairs linked by a support relation as *entailment/non-entailment*, and pairs linked by an attack relation as *contradiction/non-contradiction*.

However, given the specificity of Twitter data and the fact that predicting support and attack relations is not the same as recognizing entailment, results were far from being satisfying (see Table 5), also due to the huge number of unrelated pairs (tagged as unknown in Dataset (c)). Then we decided to implement a neural sequence classifier inspired by [26]. We encode the tokens as precomputed GloVe embeddings⁷ [24] of size 200. When a token does not have an embedding, we generate a random embedding according to a multivariate normal distribution with empirical mean and variance of existing embeddings.

Such a neural classifier is an encoder-decoder architecture with two distinct Long Short-Term Memory networks⁸ (LSTM) [16], where we pass the last hidden-state of the first LSTM to initialize the second. The probabilities over the 3 categories are given by a softmax function, i.e., a function which takes as input a C -dimensional vector z and outputs a C -dimensional vector y of real values between 0 and 1, at the output layer of the second LSTM at the last pass. Our objective is cross entropy, and we oversample the attack and support categories so that the probability of drawing a tweet from a category is uniform on the three categories. We use Stochastic Gradient Descent with Adam⁹ [17] to optimize. We periodically test our model against the validation set, and stop the training when the validation error stops improving. We select the best performing model on the validation set. However, also in this case, results are not satisfying (see Table 5).

We realize that such classification step on Twitter is pretty hard, even for human. As an example, consider the following pair:

⁶<http://hltfbk.github.io/Excitement-Open-Platform/>

⁷GloVe is an unsupervised learning algorithm for obtaining vector representations for words. Training is performed on aggregated global word-word co-occurrence statistics from a corpus, and the resulting representations showcase interesting linear substructures of the word vector space.

⁸Long Short-Term Memory networks are a special kind of Recurrent Neural Networks, capable of learning long-term dependencies.

⁹Adam is an algorithm for first-order gradient-based optimization of stochastic objective functions, based on adaptive estimates of lower-order moments.

Model	EOP (MaxEnt)	Neural model
F1-score Support	0.17	0.20
F1-score Attack	0.0	0.16

Table 5. Comparing the two models

T1: *Can't believe the designers of #AppleWatch didn't present a better shaped watch. It's still too clunky looking & could've been more sleek.*

T2: *@APPLEOFFICIAL amazing product updates. Apple TV looks great. BUT! Please make a bigger iWatch! Not buying it until it's way bigger.*

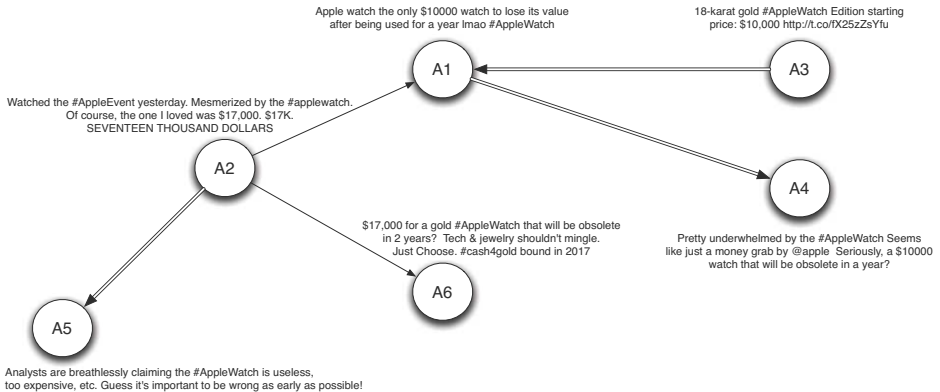
On the one hand, the tweets agree in that the watch is not properly sized. On the other hand, they disagree since one user finds it too big and the other one too small, which are opposite viewpoints.

The neural model is more promising because it can be easily used in a semi-supervised settings, but the lack of a large-sized corpus is a huge hurdle for training such a model (however, there is a huge amount of data in the DART dataset that has not been labeled yet, for which an annotation effort should be considered).

2.4. Step 4: Graph building

We can now build an *argument graph* whose nodes are the arguments and whose edges are the predicted relations (supports/ attacks). An example of such a graph is visualized in Figure 2, where an extract of the tweets for the iWatch topic is presented. It is easy to note that such a kind of visualization allows for a deeper understanding of the ongoing Twitter discussion, and would provide a valuable support for social media content analysis.

Figure 2. Example of argumentation graph (where single edges represent attack and double ones represent support) resulting from the identified arguments and predicted relations for the iWatch topic.



The last step of the pipeline consists in applying argumentation semantics to identify the set(s) of accepted arguments. Several systems can be adopted to perform such a computation in a scalable way, as those participating to the ICCMA challenge [29]. In our framework, we used the ASPARTIX-D system¹⁰, after the flattening of the bipolar

¹⁰https://ddl.inf.tu-dresden.de/web/Sarah_Alice_Gaggl/ASPARTIX-D

argumentation framework to an abstract Dung-like argumentation framework, as done in [9]. This step returns the set of acceptable arguments such that the different (coherent) viewpoints expressed through the tweets are highlighted, as well as the identifiable attack points in the stream.

Some considerations can be drawn about the resulting graphs. First of all, graphs are, differently from [10] for instance, rather sparse, meaning that they do not present a star structure. They are more like a set of subgraphs connected with each other, where each subgraph concerns a different sub-issue of the general topic, i.e., the price of the Hermes iWatch band inside the *Price* issue of the iWatch topic. This is a specificity of Twitter discussions being them a continuous stream of messages. Second, as for the case of the debates extracted in [10], no cycle is present.

3. Related Work

The first stage of the argument mining pipeline is to detect arguments within the input texts. Many approaches have recently tackled such challenge adopting different methodologies, e.g., SVM [22,23,28,12,20], Naïve Bayes classifiers [4], Logistic Regression [18].

The second stage consists in predicting what are the relations holding between the arguments identified in the first stage. This is an extremely complex task, as it involves high-level knowledge representation and reasoning issues, and, for this reason, existing approaches assume simplifying hypotheses, like the fact that evidence is always associated with a claim [2].

However, all these approaches do not tackle the challenge of applying argument mining to Twitter data. Argumentation is applied to Twitter by [13] who extract a particular version of arguments they called “opinions” based on incrementally generated queries. Their goal is to detect conflicting elements in an opinion tree to avoid potentially inconsistent information. Both the goal and the adopted methodology is different from the one we present in this paper.

Finally, to tackle these challenging tasks, high-quality annotated corpora are needed, see [25,22,18,2,27,10,14], to be used as a training set for any kind of aforementioned prediction. None of these corpora deals with Twitter data. An exhaustive state of the art about argument mining techniques and applications is in [21].

4. Conclusions

In this paper, we present an ongoing work to apply the argument mining pipeline on Twitter data. This challenging task can be divided into the following three sub-tasks: *i*) the identification of tweet-arguments from non argumentative tweets in the stream of tweets, *ii*) the composition of tweet-arguments into meaningful pairs where pairs of completely unrelated tweet-arguments are discarded, and *iii*) the prediction of the relation, i.e., support or attack, between the tweet-arguments in a pair. While we achieved satisfiable results concerning sub-tasks (*i*) and (*ii*), negative results are shown even by applying different strategies to sub-task (*iii*). Even if we know that negative results do not convey to solutions, we believe that they represent an unavoidable step in an emerging research

topic as argument mining is, and they provide a useful guide to the further exploration of the faced challenge. This is why we report them in this paper.

Investigating potential solutions to this open issue is our main future work direction. To address this argument structure prediction task, we are exploring the application of relation classification in discourse analysis techniques [19], and semantic textual similarity estimation techniques [1]. Another open challenge in dealing with Twitter is about big data: Twitter provides a very large data collection that raises the issue of the scalability of the applied argument mining techniques. Making our framework robust and scalable enough to process the Twitter streams of data is another future research line.

Acknowledgement

The work carried out in this paper is partially funded by the start-up Vigiglobe (Sophia Antipolis, France).

References

- [1] Palakorn Achananuparp, Xiaohua Hu, and Xiaojiong Shen. The evaluation of sentence similarity measures. In Il-Yeol Song, Johann Eder, and Tho Manh Nguyen, editors, *Data Warehousing and Knowledge Discovery, 10th International Conference, DaWaK 2008, Turin, Italy, September 2-5, 2008, Proceedings*, volume 5182 of *Lecture Notes in Computer Science*, pages 305–316. Springer, 2008.
- [2] Ehud Aharoni, Anatoly Polnarov, Tamar Lavee, Daniel Hershcovich, Ran Levy, Rutu Rinott, Dan Gutfreund, and Noam Slonim. A benchmark dataset for automatic detection of claims and evidence in the context of controversial topics. In *Proceedings of the First Workshop on Argumentation Mining*, page 6468. Association for Computational Linguistics, 2014.
- [3] James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *J. Mach. Learn. Res.*, 13(1):281–305, February 2012.
- [4] Or Biran and Owen Rambow. Identifying justifications in written dialogs by classifying text as argumentative. *Int. J. Semantic Computing*, 5(4):363–381, 2011.
- [5] David M. Blei and John D. Lafferty. Correlated topic models. In *In Proceedings of the 23rd International Conference on Machine Learning*, pages 113–120. MIT Press, 2006.
- [6] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003.
- [7] Tom Bosc, Elena Cabrio, and Serena Villata. Dart: a dataset of arguments and their relations on twitter (accepted for publication). In *Proceeding of LREC 2016*, 2016.
- [8] Elena Cabrio and Serena Villata. Combining textual entailment and argumentation theory for supporting online debates interactions. In *The 50th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, July 8-14, 2012, Jeju Island, Korea - Volume 2: Short Papers*, pages 208–212, 2012.
- [9] Elena Cabrio and Serena Villata. A natural language bipolar argumentation approach to support users in online debate interactions. *Argument & Computation*, 4(3):209–230, 2013.
- [10] Elena Cabrio and Serena Villata. Node: A benchmark of natural language arguments. In Simon Parsons, Nir Oren, Chris Reed, and Federico Cerutti, editors, *Computational Models of Argument - Proceedings of COMMA 2014, Atholl Palace Hotel, Scottish Highlands, UK, September 9-12, 2014*, volume 266 of *Frontiers in Artificial Intelligence and Applications*, pages 449–450. IOS Press, 2014.
- [11] Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L Boyd-Graber, and David M Blei. Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems*, pages 288–296, 2009.
- [12] Judith Eckle-Kohler, Roland Kluge, and Iryna Gurevych. On the role of discourse markers for discriminating claims and premises in argumentative discourse. In Lluís Màrquez, Chris Callison-Burch, Jian Su, Daniele Pighin, and Yuval Marton, editors, *Proceedings of the 2015 Conference on Empirical Meth-*

- ods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 2236–2242. The Association for Computational Linguistics, 2015.
- [13] Kathrin Grosse, María Paula González, Carlos Iván Chesñevar, and Ana Gabriela Maguitman. Integrating argumentation and sentiment analysis for mining opinions from twitter. *AI Commun.*, 28(3):387–401, 2015.
 - [14] Ivan Habernal, Judith Eckle-Kohler, and Iryna Gurevych. Argumentation mining on the web from information seeking perspective. In Elena Cabrio, Serena Villata, and Adam Wyner, editors, *Proceedings of the Workshop on Frontiers and Connections between Argumentation Theory and Natural Language Processing, Forlì-Cesena, Italy, July 21-25, 2014.*, volume 1341 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2014.
 - [15] Jan Hajic and Junichi Tsujii, editors. *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, August 23-29, 2014, Dublin, Ireland*. ACL, 2014.
 - [16] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.
 - [17] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
 - [18] Ran Levy, Yonatan Bilu, Daniel Hershcovich, Ehud Aharoni, and Noam Slonim. Context dependent claim detection. In Hajic and Tsujii [15], pages 1489–1500.
 - [19] Ziheng Lin, Min-Yen Kan, and Hwee Tou Ng. Recognizing implicit discourse relations in the penn discourse treebank. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, EMNLP 2009, 6-7 August 2009, Singapore, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 343–351. ACL, 2009.
 - [20] Marco Lippi and Paolo Torroni. Context-independent claim detection for argument mining. In Qiang Yang and Michael Wooldridge, editors, *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015*, pages 185–191. AAAI Press, 2015.
 - [21] Marco Lippi and Paolo Torroni. Argumentation mining: State of the art and emerging trends. *ACM Transactions on Internet Technology*, 2016.
 - [22] Raquel Mochales Palau and Marie-Francine Moens. Argumentation mining. *Artif. Intell. Law*, 19(1):1–22, 2011.
 - [23] J. Park and C. Cardie. Identifying appropriate support for propositions in online user comments. In *Proceedings of the First Workshop on Argumentation Mining*. Association for Computational Linguistics, 2014.
 - [24] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics.
 - [25] Chris Reed and Glenn Rowe. Araucaria: Software for argument analysis, diagramming and representation. *International Journal on Artificial Intelligence Tools*, 13(4):983, 2004.
 - [26] Tim Rocktäschel, Edward Grefenstette, Karl Moritz Hermann, Tomáš Kočiský, and Phil Blunsom. Reasoning about entailment with neural attention. In *International Conference on Learning Representations*, 2016.
 - [27] Christian Stab and Iryna Gurevych. Annotating argument components and relations in persuasive essays. In Hajic and Tsujii [15], pages 1501–1510.
 - [28] Christian Stab and Iryna Gurevych. Identifying argumentative discourse structures in persuasive essays. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 46–56. ACL, 2014.
 - [29] Matthias Thimm and Serena Villata. System descriptions of the first international competition on computational models of argumentation (ICCMa’15). *CoRR*, abs/1510.05373, 2015.
 - [30] S.E. Toulmin. *The Uses of Argument*. Cambridge University Press, 2003.

Dialog-Based Online Argumentation

Tobias KRAUTHOFF^{a,1}, Michael BAURMANN^b, Gregor BETZ^c and
Martin MAUVE^a

^a*Department of Computer Science, University of Düsseldorf, Germany*

^b*Department of Sociology, University of Düsseldorf, Germany*

^c*Department of Philosophy, Karlsruhe Institute of Technology, Germany*

Abstract. In this position paper we propose a novel approach to online argumentation. It avoids the pitfalls of unstructured systems such as asynchronous threaded discussions and it is usable by any participant without training while still supporting the full complexity of real-world argumentation. The key idea is to let users exchange arguments with each other in the form of a time-shifted dialog where arguments are presented and acted upon one-at-a-time. We highlight the key research challenges that need to be addressed in order to realize such a system and provide first solutions for those challenges.

Keywords. online argumentation, dialog-based approach, computer science, collaborative argumentation, collaborative work, dialog games

1. Introduction

Argumentation, the rational exchange of positions, reasons and justifications, is a vital tool whenever a group of two or more persons needs to decide on a course of action, to determine what to accept as truth, to agree on a set of shared values or to simply reach a common understanding of what the positions of the members of the group are. The Internet has provided the basic infrastructure to enable argumentation for all kinds of groups, no matter how large these groups are, where the members of the group are located or at what time they choose to participate.

Unfortunately, this basic infrastructure has not yet led to the hoped for spreading of rational exchange of arguments. In fact, quite the opposite seems to be true. A quick glance at the discussion section of online-news-media as well as blogs and social media sites shows that the expression of opinions, disputes and controversies in the Internet are often anything but rational. They lack structure and clarity, suffer from frequent repetition of similar arguments, conflate diverse aspects of a subject or are biased, irrelevant, emotionally heated and ill-informed. Furthermore they encourage the balkanization of the participants and they do not scale to large numbers of users. It has been argued [1,2] that this may be due to the predominant use of forum-based systems which rely on the input of free text.

As a consequence there have been several attempts to provide better support for online argumentation. However, so far, none of them has had really significant practical

¹Corresponding Author: Tobias Krauthoff, Department of Computer Science, Universitätsstraße 1, 40225 Düsseldorf, Germany; E-mail: krauthoff@cs.uni-duesseldorf.de

impact. One important reason for this may be due to the fact that forum-based systems offer something that other systems do not: they allow for a highly complex exchange of arguments and counter-arguments with an intuitive statement-reply scheme. Other approaches to online argumentation either do not capture the full complexity of argumentation (e.g., pro/con lists) or they require that the user is trained in operating a rather complex technical tool (e.g., the cooperative creation of an argument map).

In this paper we describe a novel approach to support online argumentation, that does not require any prior knowledge or training from the user and avoids the shortcomings of forum-based systems while still allowing complex argumentation. The main idea is to guide participants through the arguments provided by other users so that they perform a time shifted dialog with those that have participated before them. The system is driven by a formal data structure capturing the full complexity of argumentation. The user interaction, however, has the structure of a regular dialog as it is performed in everyday life. It is the task of system – and not of the participants – to translate between those two views. We call this approach dialog-based online argumentation.

A full realization of dialog-based online argumentation requires the solution of several hard problems. We cannot hope to solve all of them in one pass. Therefore the contributions of the paper are limited to: (1) the presentation of how we envision that dialog-based online argumentation should work, (2) the identification of the main research challenges that need to be addressed to realize this approach, (3) a description of first ideas on how they can be addressed.

This paper is structured as follows. In Section 2 we give a brief overview of related work in the area of online argumentation. The general idea of dialog-based online argumentation is presented in Section 3. Section 4 defines the terms that we use in the remainder of the paper. The key challenges that need to be addressed in order to realize dialog-based online argumentation are described in Sections 5, 6 and 7. We conclude the paper with a summary and an outlook to future work in Section 8.

2. Related Work

Forum-based approaches, also called asynchronous threaded discussions, allow participants to exchange arguments by means of a sequence of text contributions. In the past those approaches have encountered much criticism: in particular they are believed to lead to a high degree of redundancy and balkanization while scaling poorly with the number of involved participants [1]. However, in practice they are, by far, the most commonly used approach to support online argumentation.

Online systems for argument mapping enable participants to structure their arguments and the relation between them in an argument map. Examples are Carneades [3,4], Deliberatorium [5] and ArguNet [6]. While those systems do avoid the shortcomings of forum-based approaches, they require the users to become familiar with their notation and the semantics of formal argumentation. Therefore, in practice, they are used by experts or students who want to learn about the logic of argumentation rather than by average participants that want to take part in an online argumentation.

It has been suggested, e.g. ConsiderIt [7], to use online pro and contra lists to aid collective decision making processes. These lists work very well for evaluating a given proposal. However, they are not suitable to deal with more general positions and alternatives since they do not support the exchange of arguments and counter arguments.

The idea of engaging in a formalized dialog to exchange arguments is used by so called dialog games. In these games the participants follow a set of rules to react to the statements of each other, e.g. [8]. They are commonly used as a teaching method and were originally developed without any computers in mind. However, with the ability to let computers enforce the rules, they gained significant attention. A good overview of the current state of development of digital dialog games is given in [9]. In contrast to our work, dialog games look at the real-time interaction between users in order to learn something about a subject at hand. They do not seek to provide better instruments for online argumentation.

In addition to the main classes of ideas presented above there are three individual systems that are related to our work. The first one is the *Structured Consultation Tool* (SCT) [10]. Its primary goal is to allow a government agency to elaborate and present a justification for a given action. Members of the public can then evaluate that reasoning in a step-by-step fashion. While the SCT explicitly seeks feedback on the arguments provided by the government agency it does so in a questionnaire kind of way. This is valid for gathering feedback on government proposals, but it is unsuitable for an online argumentation, where the dynamic exchange of arguments is the main focus.

The Carneades Opinion Formation and Polling Tool [11] is part of the Carneades argumentation mapping system. It allows participants to provide structured, questionnaire-style feedback on a given argumentation consisting of multiple arguments and positions put forward by - potentially - many agents. This tool can be regarded as a generalization of the SCT. As with the SCT the questionnaire-style feedback is well suited for an evaluation of government activities by citizens but it does not fit the idea of an online argumentation amongst peers.

The third system that is related to our work is Arvina [12] and its predecessor MAGtALO [13]. Both systems allow a user to conduct a dialog between robots and humans. As a basis they use an existing argumentation specified in a formal language [14] where the positions and arguments of some real-world persons are marked. A robot can use this information to argue with human participants. The participants can query the robots and each other. In contrast to the system we envision Arvina and MAGtALO are driven by the questions of the users. Thus there is no need for the users to react to replies from the system by providing their own arguments.

3. Large Scale Online Discussions as a Dialog between a System and many Users

The primary goal of dialog-based online argumentation is to enable any user without prior knowledge or training, to participate efficiently in a large-scale online argumentation. At the same time dialog-based online argumentation avoids or at the very least reduces the problems that plague unstructured online argumentation such as a high level of redundancy, balkanization, and logical fallacies.

The foundation of dialog-based online argumentation is a novel way to navigate an existing set of arguments pertaining to a given subject. Instead of presenting many arguments at once – in maps or lists of arguments – the user is shown only a single argument at a time. It is then possible to select a response from a list of alternatives. Based on this response and, possibly, the data gathered from the responses of other participants, the system selects the next argument that is shown to the user. In this way the user and

the system perform a dialog where the system selects arguments that are likely to be of interest to the user and the user provides feedback on those arguments.

Both, the user and the system, profit from the dialog. The user is efficiently guided towards those arguments that are particularly relevant for her. If done right, this should also eliminate redundancy and balkanization and reduce the occurrence of logical fallacies. The system, on the other hand, will increase its knowledge base with every response from a participant. This can then be used to improve the selection of arguments for the next user and to provide a summary of the online argumentation at hand.

There are at least two obvious research questions when considering the foundation of dialog-based online argumentation: How should the next argument be determined that is presented to the user? And: How should the list of responses look like that the participants can choose from? We will touch upon both questions later.

Dialog-based online argumentation, as described so far, requires a fixed set of arguments that is pre-constructed by experts. In many application scenarios this will be entirely sufficient. It will allow users to form their own opinion regarding the presented arguments and ultimately make a decision on which position to support. The system, on the other hand, will be able to learn about the popularity and perceived interdependency of arguments and positions.

However, for a genuine online argumentation the system has to allow participants to add their own arguments. This raises the question how user input can be integrated in a way that enables the navigation of arguments to operate on user-provided arguments. After all, the users are not schooled in argumentation (software) and will not articulate their views in a formally standardized language. This is the third main challenge for realizing dialog-based online argumentation.

The following sections will give an overview of the terms as well as the three challenges and potential solutions. We are currently in the process of developing a first prototype of dialog-based online argumentation which can be accessed at <https://dbas.cs.uni-duesseldorf.de/>.

4. Terms and Data Structure

In the following we define the terms that will be used to describe the main aspects of dialog-based online argumentation. Their definition also describes the underlying data structure of our implementation.

Every online argumentation is identified by a **topic**. An example of a topic could be: “Our town needs to cut spending. Please discuss ideas how this should be done”. **Statements** are the most basic primitives used in an online discussion. Examples for statements are: “We should shut down university park” or “Shutting down university park will save \$ 100,000 a year”. Individual participants might consider a given statement to be true or false. A **position** is a prescriptive statement, i.e., a statement which recommends or demands that a certain action be taken. “We should shut down university park” is an example for a position.

While experimenting with an early prototype we realized, that we need a somewhat unusual definition of the term “argument”. First of all, there is argumentation for or against statements. This leads to the well-known premise-conclusion-structure of an argument, where both premises and conclusions are statements or negations of statements.

For example: “Shutting down university park will save \$ 100.000 a year, therefore we should shut down university park” would be an argument, where “Shutting down university park will save \$ 100.000 a year” is the premise and “we should shut down university park” is the conclusion. With this structure it is straight forward to support attacks and rebuttals. An attack is an argument with a conclusion that is the negation of a premise of another argument, while a rebuttal is an argument with a conclusion that is the negation of the conclusion of another argument.

Furthermore, there are arguments that target the validity of other arguments by undercutting attacks. An undercutting attack is an argument that does not reason about statements in another argument but question that a certain statement really supports a conclusion. An example would be: “Yes, drug dealers are using the park to sell drugs but this is not a good reason for shutting down university park since we should not give in to criminals”. In this example the premise is “We should not give in to criminals” while the conclusion is the negative form of the argument “We should shut down university park because criminals use university park to sell drugs”.

As a consequence we use the following definition: an **argument** consists of one or more statements (or their negations) that form the **premise(s)** and one statement or another argument (or the negation of any of those two) that form the **conclusion**.

Together, arguments and statements form a (partially connected) *web of reasons* (WoR).

5. Challenge: Providing Feedback

The most basic building block of dialog-based online argumentation is gathering feedback from a user regarding a given argument. This is done by asking a question derived from the statements pertaining to the argument in the WoR. For example, if the premise of the argument is “Criminals use University Park to sell drugs” and the conclusion is “We should shut down University Park” the question generated by the system could be “What do you think about the following argument: ‘We should shut down University Park’ because ‘criminals use University Park to sell drugs’?”

The system then offers a set of answers from which the user can choose. This set has to be constructed in a way that enables an untrained user to provide precise feedback on the argument. A simple choice between: “I agree with this argument” and “I do not agree with this argument” could certainly be made by an untrained participant. However, both statements are not precise and have little significance. For example “I do not agree with this argument” might refer to several distinct scenarios: the user might disagree with the premise, the user might think that the conclusion is not supported by the premise or the user might consider this to be a valid argument but that it is weaker than other arguments supporting the negation of its conclusion.

In order to get precise and meaningful feedback from the user, the system has to differentiate between the scenarios by means of a set of answers that the user can choose from. Experiments with a prototype system that allowed users to react to arguments of a pre-constructed online-argumentation led us to two observations: (1) We need to add alternatives that are not commonly mentioned in argumentation theory, such as “I don’t care about this argument.” (2) We have to take into account that giving feedback on an argument is a two step process. The first step is mandatory and requires just a single

click from the user to determine his initial reaction to the argument. For example, the user could choose: “Yes, criminals use University Park to sell drugs, but I do not think that this is a valid reason to close down University Park.” The second step is selecting or entering a statement that supports the choice taken in the first step. For the given example this might be “Because we should not give in to criminals.” The second step is only available if the selection in the first step allows for a follow up statement and the user can choose to skip it. Separating the two steps facilitates very fast feedback and a clean user interface design.

On the basis of this general approach, the options can be examined that a participant should have in the first step of providing feedback. We propose the following: (1) Reject the premise. (2) Accept the premise and, as a consequence, the conclusion. (3) Accept the premise but disagree that this leads to accepting the conclusion. (4) Accept the premise but state that there is a stronger argument that leads to rejecting the conclusion. (5) Do not care about the argument.

Once the user has selected an answer the system can use this to update the internal information of the WoR and to select the next argument that is presented to the user.

6. Challenge: Navigating the Web of Reasons

The second major challenge for dialog-based online argumentation is how the system should select the arguments that are presented to the participant. We believe that addressing this challenge will have to be a competitive process between different approaches. Any navigation, however, will consist of two parts: (1) bootstrapping the dialog by identifying the first argument that should be presented to a given user and (2) selecting the next argument based on the prior actions of the user.

6.1. Bootstrapping

The first thing that the system needs to do when a new users wants to participate in the online discussion is to choose an initial argument to present to the user. This is challenging since the system has no information on the user, yet.

One fairly straightforward solution is to simply ask the participant for an initial position he is interested in. This is the starting point in the WoR. The user is then invited to indicate his attitude towards this position: he can support or attack the position or investigate existing arguments regarding this position.

If the supporting or the attacking option is chosen, the user is asked to select or provide a statement explaining his choice. This statement is used as the premise and the position (or its negation) is the conclusion. Thereby the first argument is complete and bootstrapping is finished. If the user chooses to investigating existing arguments, the system instead selects an initial argument from the WoR where the position (or its negation) under consideration is the conclusion. We have implemented this approach in our prototype and it works surprisingly well.

6.2. Selecting the Next Argument

The selection of the next argument that is presented to the participant can be based on many sources of information. In particular it could rely on the history of actions that

this specific participant has performed and the knowledge gained by the actions of other users. Different kinds of selection criteria could operate on this basis. Furthermore, the selection of arguments might be influenced by the need of the system to learn more about specific arguments or the desire to keep the participant interested in continuing the dialog.

At the moment we use a very simple approach which, nonetheless, illustrates the potential of our idea. We look at the participation history of a user to identify the most recent argument that she provided or supported. Then we search the WoR for an argument of prior users which challenges (attack, rebut or undercut) the argument of the current user. This argument is shown to the current user who then has the opportunity to react to it and thereby provides the next argument. This process continues until the WoR contains no counter argument to the argument of the current user. The overall intention is to simulate a real discussion where participants challenge the arguments of other participants.

7. Challenge: Accepting New Arguments

The key to incorporating new arguments in dialog-based online argumentation is to nudge the users to provide arguments themselves and to connect them with existing arguments in an appropriate way. Currently, we use four mechanisms for this purpose. First, users can enter their own statements only within the dialog. This ensures that whatever statement the user enters, it is automatically connected to the WoR in an appropriate fashion. Second, we apply sentence openers to frame the statements of the users. In this way the user is guided towards making structured and well-formed statements. Third, we automatically match the text entered by users with existing statements in the WoR by means of the Levenstein distance [15] and display the users the top results while they enter their statement. Users can then select one of these results instead of completing their own statement. Finally, whenever users employ the keyword “and” in a premise, the system asks the user if this is in fact a single statement or a sequence of statements. The reply to this question helps the system to identify arguments that have multiple premises.

8. Conclusion and Future Work

In this paper we have presented the idea of dialog-based online argumentation as a time-shifted dialog between the individual users participating in an online argumentation. We have identified the three main challenges that need to be solved in order to realize this idea: providing feedback on existing arguments, selecting the next argument that should be presented to the user and incorporating user input. For each challenge we have provided an initial solution and we have developed a first prototype implementing them.

While the work presented in this paper is sufficient to provide a first glimpse at dialog-based online argumentation, there is a multitude of further research questions that have not yet been addressed. We believe that in particular the selection strategies for the next argument provides a lot of research opportunities. New solutions and inspirations in this area might be derived, e.g., from argumentation theory, the studies on bounded rationality and fallacies of group deliberation, “wisdom of the crowd approaches” or the research area of recommender systems. We also expect that novel ways to embed dialog-

based online argumentation in regular web-content such as blogs or online newspapers will be part of the future work in this area.

Finally, and possibly most importantly, it will be pivotal to perform an empirical evaluation of dialog-based online argumentation in real-life settings.

Acknowledgements

This work was done in the context of the graduate school on online participation, funded by the ministry of innovation, science and research in North Rhine Westphalia, Germany. We would like to thank the reviewers for their very helpful comments. Since this work was accepted as a short paper, we were not able to extend it in the way they suggested. We will prepare a full version of this paper that includes both those changes and a full description of our prototype implementation.

References

- [1] M. Klein, "Using Metrics to Enable Large-Scale Deliberation," *A workshop of the ACM Group 2010 Conference*, 2010.
- [2] P. Spada, M. Klein, R. Calabretta, L. Iandoli, and I. Quinto, "A First Step toward Scaling-up Deliberation: Optimizing Large Group E-Deliberation using Argument Maps," 2014.
- [3] T. F. Gordon and D. Walton, "The Carneades Argumentation Framework – Using Presumptions and Exceptions to Model Critical Questions," *6th Comput. Model. Nat. argument Work. (CMNA), Eur. Conf. Artif. Intell.*, vol. 6, pp. 5–13, 2006. [Online]. Available: <http://cgi.csc.liv.ac.uk/~floriana/CMNA6/CMNA06Gordon.pdf>
- [4] T. F. Gordon, "Carneades - tools for argument (re)construction, evaluation, mapping and interchange," <http://carneades.github.io/>, 2015, [Online, Last access 2016-03-13].
- [5] L. Klein, Mark; Iandoli, "Supporting Collaborative Deliberation Using a Large-Scale Argumentation System: The MIT Collaboratorium," 2008.
- [6] D. C. Schneider, C. Voigt, and G. Betz, "Argunet – A software tool for collaborative argumentation analysis and research," 2006. [Online]. Available: <http://cmna.csc.liv.ac.uk/CMNA7/papers/Schneider.pdf>
- [7] T. Kriplean, J. Morgan, D. Freelon, A. Borning, and L. Bennett, "Supporting Reflective Public Thought with ConsiderIt," in *Proc. ACM 2012 Conf. Comput. Support. Coop. Work - CSCW '12*. ACM Press, 2012, pp. 265–274. [Online]. Available: <http://dl.acm.org/citation.cfm?doid=2145204.2145249>
- [8] S. Wells, "Supporting Argumentation Schemes in Argumentative Dialogue Games," *Studies in Logic, Grammar and Rhetoric*, vol. 36, no. 1, pp. 171–191, 2014.
- [9] T. Yuan, D. Moore, C. Reed, A. Ravenscroft, and N. Maudet, "Informal logic dialogue games in human-computer dialogue," *The Knowledge Engineering Review*, vol. 26, no. 02, pp. 159–174, 2011.
- [10] T. Bench-Capon, K. Atkinson, and A. Wyner, "Using Argumentation to Structure E-Participation in Policy Making," vol. 8980, pp. 1–29, 2015. [Online]. Available: <http://link.springer.com/10.1007/978-3-662-46485-4>
- [11] T. F. Gordon, "Structured Consultation with Argument Graphs," 2013. [Online]. Available: http://publica.fraunhofer.de/eprints/urn_nbn_de_0011-n-2856073.pdf
- [12] F. Bex, J. Lawrence, M. Snaith, and C. Reed, "Implementing the Argument Web." *Commun. ACM*, vol. 56, no. 10, pp. 66–73, 2013.
- [13] S. Wells and C. Reed, "MAGtALO: an Agent-Based system for persuasive online interaction," in *AISB 2008 Convention Communication, Interaction and Social Intelligence*, vol. 1. Citeseer, 2008, p. 29.
- [14] F. Bex, J. Lawrence, and C. Reed, "Generalising argument dialogue with the Dialogue Game Execution Platform," in *COMMA*, 2014, pp. 141–152.
- [15] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," in *Soviet physics doklady*, vol. 10, no. 8, 1966, pp. 707–710.

Understanding Group Polarization with Bipolar Argumentation Frameworks

Carlo PROIETTI

Lund University

Abstract. Group polarization occurs when an initial attitude or belief of individuals becomes more radical after group discussion. Polarization often leads subgroups towards opposite directions. Since the 1960s this effect has been observed and repeatedly confirmed in lab experiments by social psychologists. Persuasive Arguments Theory (PAT) emerged as the most convincing explanation for this phenomenon. This paper is a first attempt to frame the PAT explanation more formally by means of Bipolar Argumentation Frameworks (BAFs). In particular, I show that polarization may emerge in a BAF by simple and rational belief updates by participants.

Keywords. Group Polarization, Persuasive Arguments Theory, Bipolar Argumentation Frameworks, Value-Based Argumentation Frameworks.

Introduction

Group-induced attitude polarization, also known as *risky shift* ([27]) occurs “when an initial tendency of individual group members toward a given direction is enhanced following group discussion. For example, a group of moderately profeminist women will be more strongly profeminist following group discussion” ([15]). This phenomenon occurs very often in real-life scenarios such as political debate ([28]) or discussion on virtual forums ([29]). Polarization often leads subgroups towards opposite directions, a phenomenon called *bipolarization*. Therefore, it speaks against the assumption that debate among informed individuals should lead to consensus and be truth-conducive. The fundamental question to ask is whether polarization is intrinsically irrational or not. A second question is whether it may happen in situations of perfect communication within a group. Both questions are very complex to disentangle insofar as rationality is a vaguely defined concept. However, formal approaches, as the one I adopt here, provide enough tools to capture the notion of rational update of information and therefore allow asking the question as to whether polarization may happen in situations of rational update by individuals.

Group polarization needs not to be confused with a similar phenomenon, called *belief polarization*.¹ The essential difference lies in the fact that debate and argumentation are essential ingredients of the former but not of the latter.

Large field experiments, mostly conducted in the 1970s, isolated two main concurrent explanations for this phenomenon. The first one builds upon *Social Comparison Theory* and the second upon *Persuasive Arguments Theory* (PAT). According to Social Comparison explanations, such as [26], polarization may arise in a group because individuals are motivated to perceive and present themselves in a favorable light in their social environment. To this end, people tend to take a position which is similar to everyone else but a bit more extreme. The PAT explanation ([30]) assumes instead that individuals become more convinced of their view when they hear novel and persuasive arguments in favor of their position, and therefore “Group discussion will cause an individual to shift in a given direction to the extent that the discussion exposes that individual to persuasive arguments favoring that direction” ([15]).

Both Social Comparison Theory and PAT have inspired multi-agent simulation models of opinion formation meant to explain bipolarization effects. Models inspired by Social Comparison explanations typically assume that agents are positively influenced by their ingroup members and negatively influenced by outgroup members ([12]). Alternatively, some models presuppose that the agents’ opinions come closer to opinions of similar degree and instead shift away from opinions of a too different degree ([16]). Models inspired by PAT do not assume negative influence of any kind, but presuppose homophily, i.e. stronger interaction with like-minded individuals ([24]), or biased assimilation of arguments ([22]). Both kind of models can explain bipolarization effects. However, models based on social comparison fall back on a much stronger assumption. Furthermore, empirical research showing the presence of negative influence in social interaction is not immune from criticisms ([19]).

Other than being the most recognized by psychologists nowadays, the PAT explanation is also of main interest for answering our questions. Indeed it posits that polarization may arise by a rational process due to individuals refining their argumentative skills. However, the exact mechanisms of how this process may unfold are still unclear. To understand polarization we need to decompose it into its basic ingredients, i.e. (a) a plurality of agents, (b) a debated issue, (c) possibly different prior opinions held by the agents about the debated issue, (d) *pro* and *contra* arguments – possibly related with each other by relations of refutation, support, counterattack etc. – and (e) update, by the agents, of their argumentative basis.

All such ingredients can be formally framed by the help of Argumentation Frameworks ([10]), more specifically via *Bipolar Argumentation Frameworks* (BAFs) introduced by [3]. A BAF consists of a graph where nodes are arguments and directed links represent either *supports* or *attacks* among them. A specific BAF is originally meant to represent a completed process of argumentation, i.e. the situation where “everything is on the table”. Here we give BAFs a dynamic

¹Belief polarization ([23]) happens when two parties are lead to more extreme disagreement after considering the same evidence. Formal approaches based on *Bayesian networks* have already shown that this phenomenon needs not to be irrational ([17]).

turn in order to understand the steps of an argumentative debate among n agents. Indeed, given a BAF A , the information available to the participants to a debate can be represented as a subgraph of A . The result of a debate/exchange of arguments between two agents j and k can be framed as an operation on their respective subgraphs. It is very easy to show, even in this purely qualitative framework, that polarization may easily emerge throughout a debate.

I proceed as follows. Section 1 reviews the structure of some lab experiments meant to show the emergence of group polarization and to test the PAT explanation. Section 2 introduces BAFs and shows how to frame a debate and argumentative update in a group of n agents. It is shown how polarization towards opposite directions can arise due to incomplete communication in a group. Section 3 shows that polarization can also emerge in situations of full communication due to individual biases. Section 4 concludes by presenting some further research questions that can be answered by appeal to Argumentation Frameworks.

1. Group polarization in the lab

Many experimental studies have been conducted to show that persuasive and novel arguments can induce polarization ([30]). Such experiments have a more or less standard structure. Test subjects are presented with a binary choice between two options A and B, where A is a low-risk low-gain option and B is high-risk high-gain. Test subjects should provide their initial odds for switching from A to B.² Subjects are also asked to write down arguments *pro* and *contra* the decision of switching from A to B. Arguments are then circulated among the participants who should rank them on the basis of their *persuasiveness* and *novelty*. Participants are then asked again to give their odds for switching from A to B. The difference between the (average value of the) prior odds and the (average value of the) posterior odds gives the measure of polarization towards A or B. The same test is repeated over different pairs A and B: some pairs typically show polarization toward A while others toward B.

Experimental results established some important correlations:

- (a) Prior to group discussion there exists a *culturally given pool of arguments* that determines the initial propensity of individuals towards A or towards B.
- (b) The number and persuasiveness of the arguments *pro* (*contra*) are strongly correlated with the initial choice of odds in one direction or the other.
- (c) Sharing of arguments is a necessary condition for polarization.
- (d) Persuasiveness and novelty of the shared arguments *pro* or *contra* are strongly correlated with polarization in one direction or the other.
- (e) Actual face to face debate among subjects does not increase polarization

Points (a) to (d) provide evidence in favor of PAT, while point (e) speaks against the social comparison explanation.

²Typically, test subjects should rate in a 1 to 10 scale how inclined they are to switch from A to B.

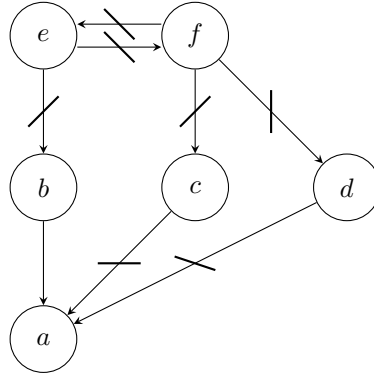


Figure 1. An example of \mathcal{BAF} . Labelled nodes represent arguments. Relations of support between arguments are indicated with a plain edge, while relations of attack are indicated with a barred one.

Pro and contra arguments play an essential role in this picture³ and the experiments thus far presented are quite convincing. However, to fully understand the impact and the role of arguments in polarization we need a more fine-grained picture. As a first important point, it is simplistic to categorize argumentative moves in a debate simply as pro or contra. Arguments in a debate usually form a complex network, e.g. some argument x undermines y which in turn supports z (and therefore x also undermines z). To better estimate the impact of an argument in a debate we should then assess its impact on the overall network, and this is something that Argumentation Frameworks allow us to do. Secondly, we need to represent the network dynamics as generated by debate. We shall deal with both these issues in the next section.

2. Bipolar Argumentation Frameworks

Bipolar Argumentation Frameworks [3] are defined as follows.

Definition 1 (BAF) A Bipolar Argumentation Framework \mathcal{BAF} is a triple $(\mathcal{A}, \mathcal{R}^a, \mathcal{R}^s)$ where \mathcal{A} is a finite and non-empty set of arguments and $\mathcal{R}^a, \mathcal{R}^s \subseteq \mathcal{A} \times \mathcal{A}$

Here \mathcal{R}^a and \mathcal{R}^s are binary relations over \mathcal{A} , called the *attack* and the *support* relation. $a\mathcal{R}^ab$ means argument a attacks argument b , while $a\mathcal{R}^sb$ means a supports b . An example of a BAF is provided by Figure 1. Relations of support between arguments are represented by a plain directed edge, while relations of attack by a barred one. Here, for example, argument a receives support from b which, in its turn is attacked by e . In an intuitive sense, a is therefore indirectly attacked by e , which undermines one of its supports. Therefore, with respect to Dung's original

³It is important to stress that in such context, as in everyday discussions, 'pro' and 'contra' are quite independent notions. No specific constraint is given such as ,e.g., an argument pro A is an argument contra not-A. Therefore, in a formal context, we need to represent pro and contra as two independent binary relations among arguments.

framework we have more complex types of attack than the simple \mathcal{R}^a . They fall under two general categorizations, provided by the following definition (see [4]).

- Definition 2 (Complex attacks)** (i) *There is a supported attack from a to b if there is a sequence $a_1 \mathcal{R}_1 \dots \mathcal{R}_{n-1} a_n$, $n \geq 3$, with $a_1 = a$, $a_n = b$, $\forall i = 1, \dots, n-2$ $\mathcal{R}_i = \mathcal{R}^s$ and $\mathcal{R}_{n-1} = \mathcal{R}^a$.*
- (ii) *There is a secondary attack from a to b if there is a sequence $a_1 \mathcal{R}_1 \dots \mathcal{R}_{n-1} a_n$, $n \geq 3$, with $a_1 = a$, $a_n = b$, $\forall i = 2, \dots, n-1$ $\mathcal{R}_i = \mathcal{R}^s$ and $\mathcal{R}_1 = \mathcal{R}^a$.*

In other words, a supported attack consists of an attack preceded by a chain of supports, while a secondary attack is a simple attack followed by a chain of supports. We shall use the term ‘attack’ to indicate both simple and complex attacks.

Given a particular $\mathcal{BAF} = (\mathcal{A}, \mathcal{R}^a, \mathcal{R}^s)$, its generating set \mathcal{A} is meant to represent an argumentative pool (the “culturally given pool of arguments” from Section 2). A debated *issue* can therefore be regarded as a specific subset of \mathcal{A} ; in our examples we shall use the singleton set $\{a\}$ as our debated issue.

In this framework, the acceptability of an argument depends on its membership of some sets, usually called *solutions* (or *extensions*). Solutions should have some specific properties. The basic ones among them are *conflict-freeness* and *collective defense* of their own arguments. Intuitively, conflict-freeness means that a set of arguments is coherent, in the sense that no argument attacks another in the same set.⁴

Definition 3 (Conflict-freeness) *A set S is conflict-free if there is no $a, b \in S$ s.t. a attacks b .*

The largest conflict-free sets in \mathcal{BAF} of Figure 1 are $\{a, b, f\}$ and $\{c, d, e\}$. A solution should also be able to defend its arguments against external attacks. Such feature is provided by the definition of collective defense.

Definition 4 (Collective defense) *A set S defends collectively an argument a if for all b such that b attacks a there is a $c \in S$ s.t. c attacks b .*

These two notions are the basis of most of the solution concepts in the standard Dung’s framework (admissibility, preferredness, stability and groundedness). Related solution concepts for BAF have been worked out by [3] and [4]. For our present purposes we need only to introduce the basic notion of d-admissibility (see [3]).⁵

Definition 5 (d-admissibility) *Let $S \subseteq \mathcal{A}$. S is d-admissible iff S is conflict-free and defends all its elements*

We can see from our example of Figure 1 that two maximal different solutions are admissible: the sets $\{a, b, f\}$ and $\{c, d, e\}$. Argument a belongs to the first but

⁴A stronger notion of coherence is also provided in [3] under the name of ‘safety’. However, we only need to introduce conflict-freeness for our present purposes.

⁵The letter ‘d’ stands for Dung. Indeed, two other notions of admissibility, c-admissibility and s-admissibility, are introduced in [3].

not to the second. Indeed the two sets represent quite opposite positions. If we see our example as the final stage of a debate, participants are in a difficult stand: they have to decide which solution to accept, and such solutions are opposite. However, there are many preliminary steps in a debate where polarization may emerge and participants can be pushed in one direction or the other. Our task for the next Section is precisely to clarify this process.

2.1. The dynamics of a debate

If we regard our \mathcal{BAF} of Figure 1 as the final stage of a debate, then the cognitive state of someone entering the debate should be seen as a partial representation of such BAF: an individual may not be aware of some arguments on the table. She may also not be aware that some argument attacks or supports another. She may even have different opinions and think that some argument attacks another while this is not the case. If we rule out the latter option – which is reasonable to do in our context – then the state of an individual entering a debate is best represented as a *subgraph* of the larger \mathcal{BAF} .⁶ By consequence, the initial setup of a debate among n agents can be encoded as a multiagent scenario where agents' states are represented by a subgraph of a given BAF. This gives rise to the following definition.

Definition 6 (Multiagent scenario) *Given \mathcal{BAF} , a multiagent scenario is a vector $(\mathcal{BAF}_1, \dots, \mathcal{BAF}_n)$ of BAFs where each \mathcal{BAF}_i (for $1 \leq i \leq n$) is a subgraph of \mathcal{BAF}*

Once a multiagent scenario is set we need to model the successive steps of information exchange in a debate. There are many ways agents could merge new information when such information disagrees with the information they have (see [6]). All of the known merging procedures have some problematic aspect and none of them satisfies all the intuitive properties of an aggregation process (see [7] and [8]). However, in our scenario there is no disagreement possible on whether an arguments attacks or supports another argument. When the situation is such, an argumentative update after an exchange among n agents is modelled simply as the *union* of the participants' respective graphs.

Definition 7 (Argumentative update) *Given a vector $(\mathcal{BAF}_1, \dots, \mathcal{BAF}_n)$ of BAFs we define, for each i , the update after information exchange as $\mathcal{BAF}_i^* = (\bigcup_{j=1}^n \mathcal{A}, \bigcup_{j=1}^n \mathcal{R}_i^a, \bigcup_{j=1}^n \mathcal{R}_i^s)$*

It is very easy to see, even in this purely qualitative framework, that polarization may easily emerge through debate. Consider a simple example of an exchange on a specific issue a with two agents 1 and 2 where both have arguments against a . Suppose that their respective initial states are represented by $\mathcal{BAF}_1 = (\{a, c\}, \{(c, a)\}, \emptyset)$ (Figure 2a) and $\mathcal{BAF}_2 = (\{a, d\}, \{(d, a)\}, \emptyset)$ (Figure

⁶Analogous approaches have been extensively developed by [25], [2] and [9] to encode multi-agent debate dynamics with argumentations systems. Here too the knowledge base of an agent is encoded by a BAF. The agent's knowledge base is a subset of a larger *universe* ([9]) or *universal argumentation framework* ([25] and [2]) whose role is analogous to our argumentative pool.

2b). $\mathcal{BAF}_1 \cup \mathcal{BAF}_2$ is clearly $(\{a, c, d\}, \{(c, a), (d, a)\}, \emptyset)$. Both 1 and 2 have a new arguments against a . In other words, both get more radical and, therefore, the group “shifts” in the direction against a . This dynamic is typically called an *echo chamber*: people become more radical than their original position because they share information with other people who have similar views. Needless, to say, an echo chamber may lead the group towards the opposite direction as well. This happens when people with arguments in favor of a discuss together.

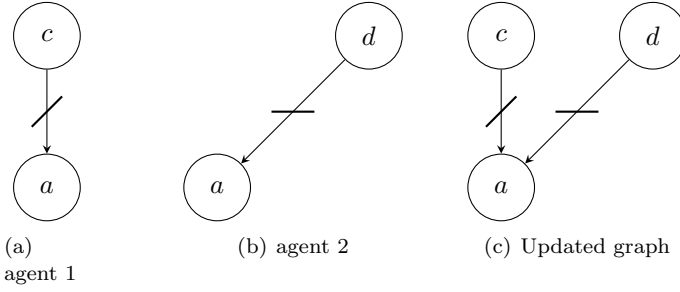


Figure 2. Argumentative update for agents 1 and 2

A typically suggested policy to prevent echo chambers is to diversify opinions by favoring the interaction of people with different priors.⁷ Back to our example, it is easy to see the effect of such mixing if we add a third agent with an argument in support of a to our debate at its initial state. Suppose indeed that agent 3’s initial state is $\mathcal{BAF}_3 = (\{a, b\}, \emptyset, \{(b, a)\})$. Then the argumentative update will be as in Figure 3. Here the echo chamber effect is prevented. Indeed, both arguments *pro*

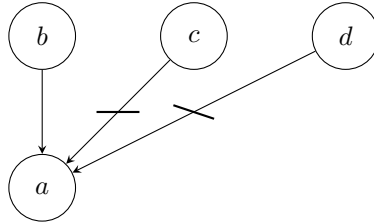


Figure 3. Argumentative update in a mixed group

and *contra* a are available to everybody, the debate is closer to a “tie” and there is no straightforward solution to choose at this stage. However this is not the end of the story. There is much psychological evidence to the fact that polarization can happen at this point too. Indeed people polarize towards opposite directions also in situations of high connectivity, e.g. online political debates ([28]). To see how this is possible we shall incorporate in our framework two explanatory clues provided by social psychology and legal reasoning.

⁷For example, larger representation of minorities in panels and decision committees goes in this direction.

3. Psychological processes and values

For purposes of decision making, agents 1, 2 and 3 in our example often need to break the tie and decide which arguments to save as more relevant when contrasting information is available. There are at least two ways this is done in real life scenarios as we shall see in this section.

3.1. Cognitive dissonance

The presence of inconsistent information usually makes individuals uncomfortable and motivates them to reduce so-called *cognitive dissonance* ([11]). This can happen in different manners. People may avoid information which would likely increase the dissonance. They may also discard evidence against their prior beliefs. Or else, they may devote more scrutiny to hypotheses and explanations that speak against their prior beliefs [13].

The third possibility seems to explain belief polarization without necessarily assuming that individuals are irrational ([18]). We can easily explain how this works in our framework by reference to our example. Agents 1, 2 and 3 are all in the same state after their first argumentative update (Figure 3). However, their initial state was quite different: agents 1 and 2 had evidence against a , while agent 3 had evidence in favor of a . In addition to that, more arguments are potentially available in their pool, such as e and f in Figure 1. Agent 1 reached her present state by receiving arguments b and d as new information. In an intuitive sense b speaks against her prior beliefs, while d does not. What should then happen when agent 1 scrutinizes b more closely? Intuitively, she should be more likely to find out arguments that undermine b if any. But argument e attacks b in our pool \mathcal{A} of arguments. It may therefore be likely that agent 1 ends up as in Figure 4(b). Here admissible sets are $\{c, d, e\}$ and its subsets and all of them (directly or indirectly) attack a .

On the other hand agent 3 reached her present state by incorporating arguments c and d , which both go against her prior belief. Therefore, she is likely to find out arguments that undermine c and d , if any. Such an argument is f . It is therefore likely that agent 3 ends up as in Figure 5(a), where admissible sets are $\{a, b, f\}$, $\{b, f\}$, $\{b\}$ and $\{f\}$. Such sets contain only arguments supporting a . Agent 1 and 3 will therefore disagree and polarization is back again.

Such a way of updating takes into account not only the agent's present state but also the previous ones. This, of course, is not the full story of how agents may scrutinize new evidence that contrasts with their prior beliefs, but it tells us that polarization may be very resilient and difficult to contrast even when people with different priors interact in an open and large debate.

3.2. Values

In cases like the one represented in Figure 1 a dispute cannot be settled. Indeed, there are two maximal disjoint admissible solutions for the graph: $\{c, d, e\}$ and $\{a, b, f\}$. As stressed by [1], this is often the case in contexts of practical reasoning, law or ethical debate, which are also contexts where polarization often arises. In many cases the dispute is solved by appeal to the arguments' intrinsic value.

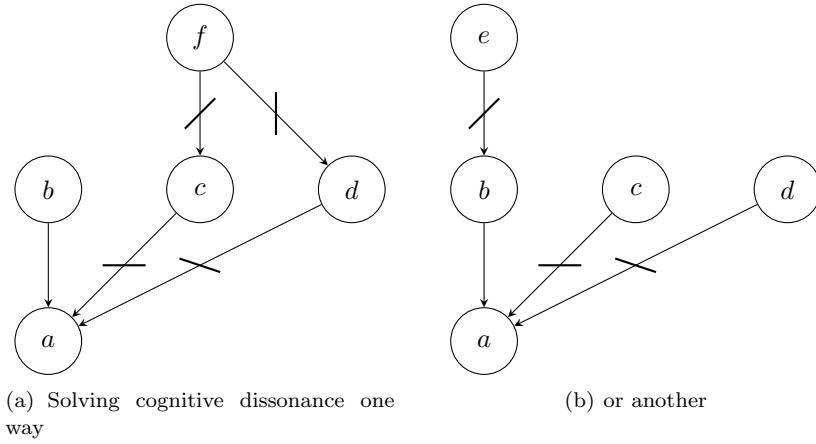


Figure 4. How subjects may solve cognitive dissonance

Often no conclusive demonstration of the rightness of one side is possible: both sides will plead their case, presenting arguments for their view as to what is correct. Their arguments may all be sound. But their arguments will not have equal value for the judge charged with deciding the case: the case will be decided by the judge preferring one argument over the other. And when the judge decides the case, the verdict must be supplemented by an argument, intended to convince the parties to the case, fellow judges and the public at large, that the favoured argument is the one that should be favoured. ([1], p.429-430)

Arguments are very often attached with *values* in public debate too. As an example, an argument against gun-control may be often associated with *individual freedom*, while arguments for gun-control have a special inclination towards *non-violence*.⁸ Indeed, different groups of people may hold different value rankings. This is an explanatory clue for polarization in many contexts. To make this point we need to define Value-based Bipolar Argumentation Frameworks (VBAF). This is done by expanding Bench-Capon's definition in [1].

Definition 8 (VBAF) A Value-based Bipolar Argumentation Framework \mathcal{VBAF} is a tuple $(\mathcal{A}, \mathcal{R}^a, \mathcal{R}^s, V, val, P)$ where \mathcal{A} , \mathcal{R}^a and \mathcal{R}^s are as before, V is a set of values, val is an assignment $\mathcal{A} \rightarrow V$ and P is a set of “possible audiences” where $p \in P$ is a ranking on V

Given a set V of values (e.g. freedom, non-violence etc.), arguments are associated to them by means of the function val . A possible audience p represents the specific ranking an individual or a group assigns to such values. Relative to a specific audience an argument a can properly attack or support b only when the value of a is greater or equal to the value of b . More formally, the following definition applies.

⁸This doesn't mean that arguments for different sides are always associated with different values. Quite often, indeed, to make a “good” move in a debate is to attack the opposite side with an argument who has value for the other side.

Definition 9 (Strong attack and strong support) For all $a, b \in \mathcal{A}$ and $p \in P$

- (i) a strongly attacks b for audience p iff $a\mathcal{R}^a b$ and not $val(a) <_p val(b)$
- (ii) a strongly supports b for audience p iff $a\mathcal{R}^s b$ and not $val(a) <_p val(b)$

Going back to our main example in Figure 1, we can easily show how this may generate polarization of two different audiences. Suppose we have only two values, which we label by two different colors, e.g. red and blue. Our V is then $\{red, blue\}$. We also suppose that $val = \{(a, red), (b, red), (c, blue), (d, blue), (e, blue), (f, red)\}$. Finally, we assume that agents belong to two audiences p_1 and p_2 where $blue <_{p_1} red$ and $red <_{p_2} blue$.

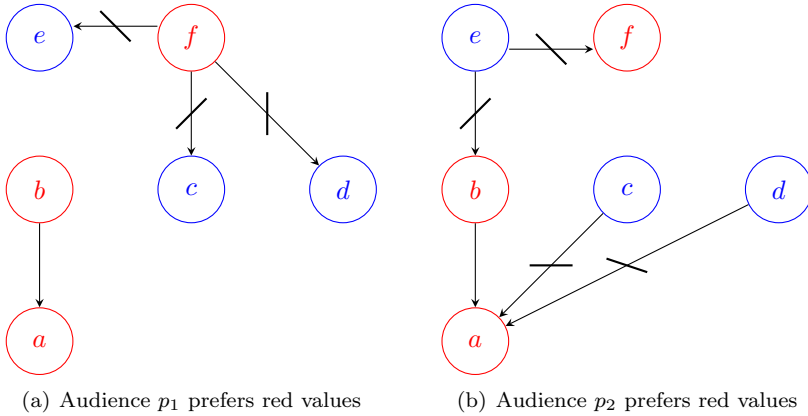


Figure 5. Two values

When the situation at the final stage of the debate is as in Figure 1 the two audiences may adopt two different solutions based on their value rankings. Audience p_1 will come out to the $\mathcal{VB}\mathcal{A}\mathcal{F}$ represented in Figure 5(a) while audience p_2 will converge to the one represented in Figure 5(b). For p_1 a belongs to an admissible solution while this is clearly not the case for p_2 .

All in all, there are many possible explanatory clues for group polarization. BAFs and their dynamics are an adequate tool for capturing most of them.

4. Conclusions and future work

Group polarization is a very complex phenomenon and this paper constitutes only an initial stage of a formal research on this problem. Our main aim was to show that bipolar argumentation frameworks are an adequate tool for framing the steps of a polarization process. We have shown that in some simple scenarios polarization may be captured at a very intuitive level by a simple process of argumentative update. However much work in many directions is left to do in future research. First, we have left out all the quantitative aspects which are a fundamental ingredient of group polarization. Indeed, polarization of attitudes means that argumentative updates induce an increase of the likelihood that individuals

will settle an issue in one way or another. A measure of such likelihood is therefore needed. *Probabilistic Argumentation Frameworks* [21] and *Graded Semantics* [14] are a useful tool for providing such measures and to investigate how likelihood is influenced by argumentative dynamics. Further insights for implementation can be provided by *Social Argumentation Frameworks* [20] and [5]. Such structures are an extension of Argumentation Frameworks meant to model and assess on-line debates, where *pro* and *contra* votes are associated to arguments. As a most interesting aspect, [20] provides a fine-grained semantics to compute one arguments strength as a function of the structure of the graph and the social opinion expressed through the votes.

Argumentative dynamics are a second main field of inquiry to understand polarization. In our examples, we adopted union of graphs as a straightforward policy of argumentative update. However, as stressed in Section 2, this only works under specific conditions. It won't work in more complex situations where participants receive information which is inconsistent with their prior belief state. To handle such situations more complex operations of graph merging are needed, which are provided by [6],[7] and [8].

References

- [1] T.J. Bench-Capon. Persuasion in Practical Argument Using Value-based Argumentation Frameworks. *Journal of Logic and Computation*, 13(3): 430–448, 2003.
- [2] M. Caminada and C. Sakama. On the Issue of Argumentation and Informedness. *2nd International Workshop on Argument for Agreement and Assurance (AAA 2015)*, 2015.
- [3] C. Cayrol and M.C. Lagasquie-Schiex. On the Acceptability of Arguments in Bipolar Argumentation Frameworks. *Lecture Notes in Computer Science*, 3571: 378–389, 2005.
- [4] C. Cayrol and M.C. Lagasquie-Schiex. Bipolarity in Argumentation Graphs: Towards a better Understanding. *International Journal of Approximate Reasoning*, 54(7): 876–899, 2013.
- [5] M. Correia, J. Cruz and J. Leite. On the Efficient Implementation of Social Abstract Argumentation. *ECAI 2014*: 225–230, 2014.
- [6] S. Coste-Marquis, C. Devred, S. Konieczny, M.C. Lagasquie-Schiex and P. Marquis. On the merging of Dung's argumentation systems. *Artificial Intelligence*, 171: 730–753, 2007.
- [7] J. Delobelle, S. Konieczny and S. Vesic. On the Aggregation of Argumentation Frameworks. *IJCAI 2015*: 2911–2917, 2015.
- [8] J. Delobelle, A. Haret, S. Konieczny, J. Mailly, J. Rossit. and S. Woltran. Merging of Abstract Argumentation Frameworks. *KR 2016*: 33–42, 2016.
- [9] F. Dupin de Saint-Cyr, P. Bisquert, C. Cayrol and M.C. Lagasquie-Schiex. Argumentation update in YALLA (Yet Another Logic Language for Argumentation). *International Journal of Approximate Reasoning*, 75: 57–92, 2016.
- [10] P.M. Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence* 77 (2): 321–357, 1995.
- [11] L. Festinger. *A Theory of Cognitive Dissonance*. Stanford, CA: Stanford University Press, 1957.
- [12] A. Flache and M.W. Macy. Small world and cultural polarization. *Journal of Mathematical Sociology* 35: 146–176, 2011.
- [13] T. Gilovich. *How we know what isn't so*. The Free Press, New York, 1991.
- [14] D. Grossi and S. Modgil. On the Graded Acceptability of Arguments. *Proceedings of the IJCAI 2015*: 868–874, 2015.
- [15] D.J. Isenberg. Group Polarization: A critical review and a Meta-Analysis. *Journal of Personality and Social Psychology* 50 (6): 1141–1151, 1986.

- [16] W. Jager and F. Amblard Uniformity, bipolarization and pluriformity captured as generic stylized behavior with an agent-based simulation model of attitude change. *Computational & Mathematical Organization Theory* 10: 295–303, 2004.
- [17] A. Jern, K.K. Chang and C. Kemp Belief Polarization is not always irrational. *Psychological Review* 121(2): 206–224, 2014.
- [18] T. Kelly Disagreement, Dogmatism, and Belief Polarization. *Journal of Philosophy* 105(10): 611–633, 2008.
- [19] Z. Krizan and R.S. Baron Group polarization and choice-dilemmas: How important is self-categorization? *European Journal of Social Psychology* 37: 191–201, 2007.
- [20] J. Leite and J. Martins Social Abstract Argumentation. *IJCAI 2011*: 2287–2292, 2011.
- [21] H. Li, N. Oren and T.J. Norman Probabilistic Argumentation Frameworks *Lecture Notes in Computer Science* 7132: 1–16, 2011.
- [22] Q. Liu, J. Zhao and X. Wang Multi-agent model of group polarisation with biased assimilation of arguments *Control Theory & Applications, IET* 9.3: 485–492, 2014.
- [23] C. Lord, L. Ross and M. Lepper Biased Assimilation and Attitude Polarization: The Effects of Prior Theories on Subsequently Considered Evidence. *Journal of Personality and Social Psychology* 37 (11): 2098–2109, 1979.
- [24] M. Mäs and A. Flache Differentiation without Distancing. Explaining Bi-Polarization of Opinions without Negative Influence. *PLoS ONE* 8 (11): e74516, 2013.
- [25] C. Sakama. Dishonest Arguments in Debate Games. *COMMA 2012*, 75: 177–184, 2012.
- [26] G.S. Sanders and R.S. Baron Is social comparison irrelevant for producing choice shifts? *Journal of Experimental Social Psychology* 13: 303–314, 1977.
- [27] J.A. Stoner A comparison of individual and group decision involving risk MA thesis, Massachusetts Institute of Technology, 1961.
- [28] C. Sunstein. *Why societies need Dissent*. Cambridge, Harvard University Press, 2003.
- [29] S. Yardi, D. Boyd, Dynamic Debates: An analysis of group polarization over time on Twitter *Bulletin of Science, Technology and Society* 30 (5), pp. 316–27, 2010.
- [30] A. Vinokur and E. Burnstein Effects of partially shared persuasive arguments on group-induced shifts, *Journal of Personality and Social Psychology* 29 (3): 305–15, 1974.

Preferences in Argumentation for Statistical Model Selection

Isabel SASSOON^{a,1}, Jeroen KEPPENS^a and Peter MCBURNEY^a

^a*Department of Informatics, King's College London*

Abstract. The increase in routine clinical data collection coupled with an expectation to exploit this in support of evidence based decision making creates a need for an intelligent model selection system to support clinicians when analysing data because clinicians often lack the statistical expertise to do this independently. In a previous position paper, an argumentation based approach to devise a decision support system for such an application was introduced. This approach ignored the relative strength of arguments for and against alternative models. This paper demonstrates how an extended argumentation framework can be employed to capture and reason with statistical and research domain knowledge that affects the relative strength of arguments. The approach is validated by means of a real-world case study.

Keywords. argumentation schemes, preferences, automated analysis, decision support, model selection

1. Introduction

Answering a research question through statistical data analysis normally involves applying a particular statistical model or technique to the data. Software packages make the application of statistical methods easy but it is hard to determine which one to employ. The suitability of statistical approaches depends on the research question at hand, the assumptions underpinning the approaches and the extent to which they are satisfied by the data. Assessing the latter requires both statistical domain knowledge and an understanding of the data and how it has been collected.

This paper is part of a broader project that aims to address this problem by means of an intelligent decision support system to aid with model selection. For the purposes of this paper, it is assumed that a clinician aims to analyse a research question by means of an existing data set. Sometimes, clinicians interact with statistical concepts at the design stage of a study, before data has been collected. Extending the approach to the latter scenario is left for future work. The research questions of interest extend beyond system identification by finding a "best-fit" model for the data, and include hypothesis testing and other methods where the conclusions derived from statistical analysis are only valid in so far as an appropriate model has been applied.

Previously, we have proposed an approach to employ computational models of argumentation to identify the reasons to accept or reject the use of a statistical model [10].

¹Corresponding Author: Isabel Sassoon, Department of Informatics, King's College London, Strand, WC2R 2LS London, UK; E-mail: isabel.sassoon@kcl.ac.uk.

Using Dung’s argumentation framework [6], the resulting models enable identification of sets of accepted arguments and associated models. This approach ignores the relative strength of arguments. Statistical and application domain knowledge can inform appraisals of argument strength and can be modelled by means of preferences over arguments. This paper aims to address where such preferences emanate from, how they should be represented and how we might reason with them using existing argumentation approaches. The approach is validated by means of a case study in the medical domain, the initial results of which were published by Schilling et. al. [11].

2. Background

Within the clinical domain, clinicians are able to query and access many databases to explore and test research questions or perform hypothesis testing. A systematic review highlighted that while reporting of survival analysis results in journal publications had increased and the quality of the reporting of statistical analysis was improving slowly, only a low proportion of articles mention validation of model assumptions prior to use [1]. In our previous work we addressed the issue of the models to consider on the grounds of achieving the desired analytical objective and the underlying critical assumption testing. However as preferences are an important element in decision making, especially collective decision making, there is a need to leverage them as part of the model selection process. Within our model selection process preferences will be used to support the selection of the most appropriate model when more than one is possible, and given the clinician’s research question and data.

In our position paper, we proposed an architecture for an argumentation based system to support the model selection process through the use of a knowledge base and an argumentation scheme [10]. A core component of this system is a statistical knowledge base (SKB) that defines the relations between research question type (R), research objectives (O), models (M) and assumptions (A). The SKB holds facts linking R, O, M, A in a way that will support the queries from the argumentation schemes. The SKB specifies multiple research question types. Each is linked to the objectives O that can fulfil that research question R . Models M are defined and linked to the respective objectives they are suitable for. For each model the critical assumptions that must be satisfied for the model to be applicable are identified. The relations and contents of the SKB are derived from statistical theory and best practice.

The elements of the SKB are denoted as follows:

- The set of *models*: $M = \{m_1, \dots, m_K\}$
- The set of *assumptions*: $A = \{a_1, \dots, a_P\}$
- The set of *objectives*: $O = \{o_1, \dots, o_Q\}$

The following relationships are defined in the SKB:

- $F : M \times O$ where $(m_k, o_q) \in F$ iff m_k fulfils objective o_q
- $C : M \times A$ where $(a_p, m_k) \in C$ iff a_p is a critical assumption for m_k
- $O : O \times O$ where $(o_r, o_q) \in O$ iff o_r is an alternative objective to o_q

A key benefit of the architecture proposed in [10] is that it differentiates knowledge into domain and problem specific information to be provided by the clinician, the prob-

lem independent domain specific statistical knowledge base and problem and domain independent argumentation schemes. This facilitates maintainability of the approach. However, our approach ignored subtle differences between the applicability of plausible models to a problem, such as the extent to which non-critical assumptions are not satisfied and contextual information that affects a model's suitability to meet the research objective. This work aims to capture such subtleties by modelling them by means of preferences over arguments.

A number of distinct approaches to represent and reason with preferences over arguments have been devised. Key approaches include Preference Argumentation Frameworks (PAF) [2, 4], Value-Based Argumentation Frameworks (VAF) [5] and Extended Argumentation Frameworks (EAFs) [9].

In VAFs, arguments are said to promote values and preferences over arguments derived from a preference ordering over values. Because the intelligent decision support system proposed herein aims to enable clinicians to answer research questions *objectively* supported by data, the choice of statistical model for performing an analysis rarely involves a conflict of values². Thus, while VAFs enable a broad range of scenarios to be analysed, they are not a good fit for the problem at hand. Therefore, the remainder of this section focusses on PAFs, specifically in its incarnation of Argumentation Frameworks based on Contextual Preferences (CPAFs) [3], and EAFs [9].

3. Method

The objective of this paper is to define a preference ordering $Pref : M \times M$ over a set of models $M = \{m_1, \dots, m_n\}$. However, such an ordering or orderings are not necessarily defined over the models directly. This section examines where the preferences for statistical model selection stem from, how they should be represented and which argumentation framework is suitable to infer decision support information based on those preference.

One source for preference orders is the statistical theory underpinning each model and dictating which models perform better when certain conditions are present in the data or the research question. For example, certain types of model are more resilient to particular features in the data, e.g. censoring or the proportion of case data lost to follow up, whereas others tend to become unreliable in such circumstances. Here, the presence of a particular feature causes a preference ordering over statistical models to arise. This relationship between a feature and an associated preference ordering is a matter of statistical knowledge. The presence of the feature may be determined by applying a test on the data or needs to be elicited from domain knowledge. In what follows, such preferences are called feature-based preferences.

A second source of preference orders is derived from model intent. There are different reasons for building a model when answering a research question. McBurney [8] explores the different purposes or reasons why a model can be used. In the context of statistical analysis the two most common intents for building a model on data are the need to predict or the need to explain (understand) the data. This is also covered in detail in [12]. In her article, Shmueli tackles the distinction between explanatory modelling and predictive modelling in detail and the implications these have on the choice of model

²It is understood this would be different in a scenario where statistical analysis aims to serve a *political* agenda.

CD1	Model	P_1
absent	m_1 KM	unaffected
	m_2 PH	unaffected
	$m_4 \chi^2$	unaffected
light	m_1 KM	unaffected
	m_2 PH	unaffected
	$m_4 \chi^2$	affected
heavy	m_1 KM	affected
	m_2 PH	unaffected
	$m_4 \chi^2$	affected

Table 1. P_1 for model resilience to censoring

CD2	Model	P_2
predict	m_1 KM	avoid
	m_2 PH	suitable
	$m_4 \chi^2$	avoid
explain	m_1 KM	suitable
	m_2 PH	suitable
	$m_4 \chi^2$	neutral

Table 2. P_2 for model intent

to use. The definition of a good model will differ depending on whether we are looking for explanatory or predictive power, and this will reflect itself in an order of preference between models that can achieve a specific analytic objective. This preference order between models will change depending on the intent (purpose) of the analysis. In what follows, such preferences are called intent-based preferences.

Finally, there may be preference orders that are derived from the clinicians themselves. This could be due to the fact they are more familiar with a model, or that the literature they reference most makes use of a particular model. These preference orders can arise when more than one clinician is involved in an analysis and are an important factor within the decision making process. In what follows, such preferences are called domain-based preferences.

To incorporate preferences into the approach, the statistical knowledge base (SKB) introduced in [10] is extended with

- A set of context domains $CD = \{CD_1, \dots, CD_H\}$. Each CD_h is a set of mutually exclusive contexts.
- A set of totally ordered sets of performance measures $P = \{P_1, \dots, P_H\}$. Each P_h contains a set of measures $p_{h1} \prec \dots \prec p_{hj_h}$ by means of which a model's performance is assessed in a specific context.
- A set of performance function $PF = \{PF_1, \dots, PF_H\}$, such that each $PF_i : CD_i \times M \mapsto P_i$.

For example, the feature-based preference "resilience to censoring" can be modelled by a context domain $CD_1 = \{\text{absent}, \text{light}, \text{heavy}\}$ where the elements in the set correspond to features indicating distinct degrees to which censoring is present in the data. These can be defined more precisely in terms of proportion of records in the data affected but we avoid doing so to keep the example simple. The corresponding performance measure might be defined as $P_1 = \{\text{unaffected}, \text{affected}\}$. Table 1 presents an example of a performance function.

An example of an intent-based preference is $CD_2 = \{\text{predict}, \text{explain}\}$ where the performance measures would be defined as $P_2 = \{\text{suitable}, \text{neutral}, \text{avoid}\}$, as in Table 2. This can also be defined for domain-based preferences CD_3 where $P_3 = \{\text{preferred}, \text{neutral}\}$.

To construct an argumentation model based on the extended statistical knowledge base, first the set of context domains CD for the problem at hand must be established. CD contains contexts taken from the context domains in $\{CD_1, \dots, CD_H\}$. Formally,

$CD \subseteq CD_1 \cup \dots \cup CD_H$. Whether a context is relevant to a problem is derived by applying a test on the data, elicited from the domain expert/clinician or elicited from the research question. Where identification of the context is not straightforward, the contexts in CD provide hooks (conclusions) for further arguments about the appropriate statistical model.

Let $\langle Args, R \rangle$ be an argumentation framework produced using the method described previous in [10]. Such a model can now be extended to an EAF [9] $\langle Args, R', D \rangle$ by defining:

- $R' = R \cup \{(c_{ij}, c_{ik}) | c_{ij}, c_{ik} \in CD \cap CD_i, c_{ij} \neq c_{ik}\}$. Intuitively, R is extended with a symmetric attack relationship between each distinct pair of contexts in CD from the same context domain CD_i .
- $D = \{(c_{ij}, (m_1, m_2)) | c_{ij} \in CD, PF_i(c_{ij}, m_1) \prec PF_i(c_{ij}, m_2)\}$. Intuitively, an attack relationship $c_{ij} \twoheadrightarrow (m_1 \rightarrow m_2)$ is added for each attack of a model m_2 by a model m_1 where a context c_{ij} justifies a preference of m_2 over m_1 .

The model can be enhanced further to take into account an importance order I of the context domains, if this is available. Let $\langle Args, R, D \rangle$ be the EAF, this can be extended to include I the importance of the context domains order by defining I as a complete or partial order on $CD \times CD$.

4. Case Study

The example used in this case study is derived from the ongoing collaboration with the Head and Neck Department at Guy's Hospital, King's College London (UK). The first published output of this work is in [11], and relies on a rich data set collected as part of the Sentinel European Node Trial (SENT). This data was collected as an observational study across 14 european centres and recruited a total of 415 patients who met the entrance criteria at diagnosis. The study commenced in 2005 and involves over 40 clinicians across the participating hospitals. The centres are periodically updating the current status of the patients in the trial. The main motivation for this trial was to assess whether sentinel node biopsy is a reliable and safe diagnostic technique in patients with early stage oral squamous cell carcinoma. The first output from this data answers the primary objective on patients with the potential for at least 3 years of follow up.

The data collected offers a cohort of data that can be exploited in support of answering many more clinician research questions or secondary objectives. There are a number of such analyses in progress initiated by different clinicians involved in SENT. An example of such a secondary analysis will be used as the case study in this paper. The research question is to identify whether there is a difference in survival between patients (within the SENT trial) who had so-called adjuvant therapy (such as Radiotherapy or Chemotherapy) to those that did not have any additional treatment.

By means of the approach presented in previous work [10], an argumentation framework $\langle Args, R \rangle$ is produced where $Args = \{m_1, m_2, m_4\}$, where each m_i is an argument supporting the use of a particular model and $R = \{(m_1, m_2), (m_1, m_4), (m_2, m_4), (m_2, m_1), (m_4, m_1), (m_4, m_2)\}$, which is the set of pairwise attacks between alternative models. Note that this is a substantial simplification of the argumentation model presented in [10]. The underlying assumptions have been omitted from the model as they are not necessary to understand how preferences are added.

To incorporate preferences over the models m_1 , m_2 and m_4 , four context domains need to be considered. The first (CD_1) corresponds to censoring. A query on the data has determined the presence of heavy censoring. Censoring can affect the reliability of the estimates obtained from some models, in this case both m_1 and m_4 are affected by heavy censoring. Using the context domain and performance function from Table 2, the following preference arguments c_{ij} arise: $c_{11} \rightarrow (m_1 \rightarrow m_2)$, $c_{12} \rightarrow (m_4 \rightarrow m_2)$.

The preference argument c_{11} is derived from the CD_1 and it attacks the attack of (m_1, m_2) .

The second context domain (CD_2) corresponds to intent. In this case, the intent of the study is to explore or explain the data, therefore the context domain for model intent is relevant and preferences arising from the intent of explaining will be used. Using the context domain and performance function from Table 2, the following preference arguments arise: $c_{21} \rightarrow (m_4 \rightarrow m_1)$, $c_{22} \rightarrow (m_4 \rightarrow m_2)$.

The remaining context domains (CD_3 and CD_4) stem from clinician preferences. These are preferences expressed by different clinicians and result in a set of preference arguments that attack the attack of all arguments in support of all models except the one expressed by the clinician. The following preference arguments arise: $c_{31} \rightarrow (m_1 \rightarrow m_2)$, $c_{32} \rightarrow (m_4 \rightarrow m_2)$, $c_{41} \rightarrow (m_1 \rightarrow m_4)$, $c_{42} \rightarrow (m_2 \rightarrow m_4)$.

Finally, an importance ordering I , specifying that $CD_1 \succ CD_2 \succ CD_3 \succ CD_4$ is added to the argument framework. Depending on which context domains are pertinent to a specific analysis there may be an order on the context domain. The context domains that relate to statistical theory are more important in model selection than clinician preferences.

To recommend the most suitable model to apply for this analysis we would require a complete extension of this framework, which contains arguments in support of one model only. Without considering any preference arguments this argumentation framework contains only arguments that symmetrically attack each other. The introduction of preferences will enable the strengths of the arguments to be taken into consideration.

Applying CPAF to this argumentation framework using the above model yields a recommendation for the use of model m_2 , irrespective of the approach used. The application of EAF to this situation does not yield any stable extensions, except the empty set. This is due to the relative importance of the preferences emanating from the different context domains not being exploited.

The preferences can be resolved in order to determine the recommended model by initially only considering the preference arguments from the most important context domain (CD_1). The preference arguments in the EAF attack the existing attacks between arguments in support of the models and their effect on the argumentation framework can be seen in Figure 1. In this case, m_2 is the only argument that is not strictly defeated and as such this would be the recommended model to be used, this would represent the stable extension to the argumentation framework. In this EAF, the justification to its choice over m_1 and m_4 is given by the context domain used in order to resolve this. In this case the recommendation of m_2 over the other models is explained by it being preferred under conditions of censoring.

If we assume that the order over the context domains is not known, then the extensions for the EAF can be computed for each CD_i in turn. The resulting extensions would be: $S_1 = \{m_2\}$, $S_2 = \{m_2\}$, $S_3 = \{m_2\}$ and $S_4 = \{m_4\}$ where S_i corresponds to the stable extension for CD_i . In other words model m_4 would only be selected in a situation where the preferences of clinician 2 are prioritised over all other contexts.

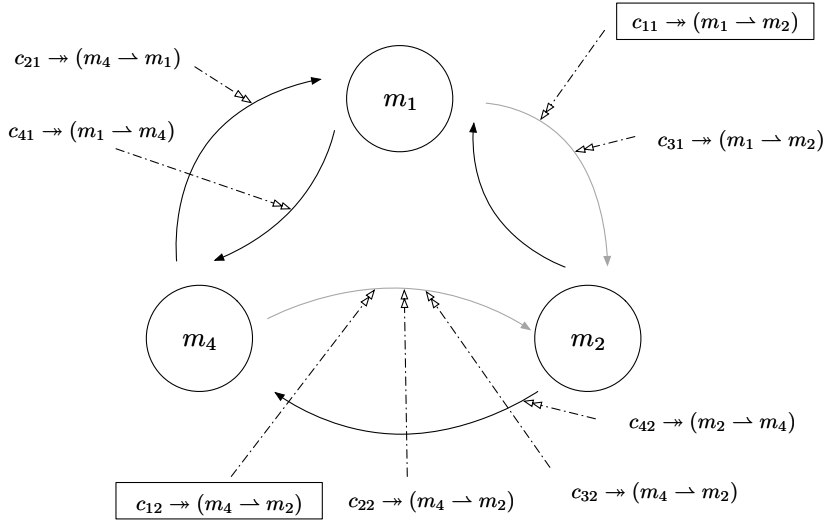


Figure 1. The preference arguments, considering only the preference arguments from context domain CD_1

5. Conclusion

This paper has presented an approach to develop a decision support tool to aid domain experts who collect data as part of their professional practice with choosing statistical techniques for analysis. This work has built on earlier work presented in [10]. Our proposed methods support the statistical model selection process by enabling contrasting preference orderings to be accounted for and reasoned with in order to recommend the most suitable model. This is achieved through EAFs and an extended statistical knowledge base. This approach can also take into account the relative importance of the different preference context domains, if this is applicable to the situation. Our proposed methodology for the inclusion of preferences enables the different types of preferences and their potential conflicts to be leveraged within the statistical model selection process, without statistical, informatics or administrative support.

The use of clinical preferences and argumentation to support decision making by clinicians has also been explored by Hunter et al [7]. In this paper, the aim is to offer the clinician the facility to aggregate evidence whilst taking into consideration the clinician's own assessment of the strength or weaknesses of each item of evidence. A clinician's preference may stem from the source of the evidence and is applied to the evidence used to evaluate the arguments, not on the arguments themselves. This method was evaluated by means of an actual trial with clinicians. The difference between our situation and the scenario considered in this paper is that in our case the preferences are not completely dependent on the clinician's view.

A prototype of the proposed system is being developed. This will offer the opportunity for the evaluation of the system using a range of case studies. Future work will focus on developing an ontology in support of a more flexible input method for the clinician's research question. This would enable clinicians to formulate their research questions us-

ing the terminology they may be more familiar with, as the ontology would relate it to the key concepts required by the proposed system to proceed with model selection. We also plan to address situations where the assumptions about the data available are removed. In such situations the data required to answer a research question may need to be extracted from multiple disparate sources, which may vary in provenance and quality. This would require methods able to handle multiple data sources, data matching, data quality and their impact on the proposed method for statistical model selection.

6. Acknowledgements

The authors would like to thank Mark McGurk for his generous support on this project.

References

- [1] Abaira, V., Muriel, A., Emparanza, J. I., Pijoan, J. I., Royuela, A., Plana, M. N., Cano, A., Urreta, I., and Zamora, J. (2013). Reporting quality of survival analyses in medical journals still needs improvement. a minimal requirements proposal. *Journal of Clinical Epidemiology*, 66(12):1340–1346.
- [2] Amgoud, L. and Cayrol, C. (2002). A reasoning model based on the production of acceptable arguments. *Annals of Mathematics and Artificial Intelligence*, 34(1-3):197–215.
- [3] Amgoud, L., Parsons, S., and Perrussel, L. (2000). An argumentation framework based on contextual preferences. In *Proc. of FAPR'00, London*, pages 59–67.
- [4] Amgoud, L. and Vesic, S. (2009). Repairing preference-based argumentation frameworks. In *IJCAI*, pages 665–670.
- [5] Bench-Capon, T. J. (2003). Persuasion in practical argument using value-based argumentation frameworks. *Journal of Logic and Computation*, 13(3):429–448.
- [6] P. M. Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77(2):321–358.
- [7] Hunter, A. and Williams, M. (2010). Using clinical preferences in argumentation about evidence from clinical trials. In *Proceedings of the 1st ACM International Health Informatics Symposium, IHI '10*, pages 118–127, New York, NY, USA. ACM.
- [8] McBurney, P. (2012). What are models for? In *Proceedings of the 9th European Conference on Multi-Agent Systems, EUMAS'11*, pages 175–188, Berlin, Heidelberg. Springer-Verlag.
- [9] Modgil, S. (2009). Reasoning about preferences in argumentation frameworks. *Artificial Intelligence*, 173(9–10):901 – 934.
- [10] Sassoon, I., Keppens, J., and McBurney, P. (2014). Towards argumentation for statistical model selection. In *5th International Conference on Computational Models of Argument*, pages 67–74. IOS Press.
- [11] Schilling, C., Stoeckli, S. J., Sloan, P., McGurk, M., Sassoon, I., and others (2015). Sentinel european node trial (sent): 3-year results of sentinel node biopsy in oral cancer. *European Journal of Cancer*, 51(18):2777 – 2784.
- [12] Shmueli, G. (2010). To explain or to predict? *Statistical Science*, pages 289–310.

An Argumentation Workflow for Reasoning in Ontology Based Data Access

Bruno YUN^a, Madalina CROITORU^{a,1}

^a INRIA Graphik/LIRMM University Montpellier, France

Abstract. In this paper we demonstrate how to benefit from structured argumentation frameworks and their implementations to provide for reasoning capabilities of Ontology Based Data Access systems under inconsistency tolerant semantics. More precisely, given an inconsistent $Datalog^{\pm}$ knowledge base we instantiate it using the $ASPIC^+$ framework and show that the reasoning provided by $ASPIC^+$ is equivalent to the main inconsistent tolerant semantics in the literature. We provide a workflow that shows the practical interoperability of the logic based frameworks handling $Datalog^{\pm}$ and $ASPIC^+$.

Keywords. Applications and Structured Argumentation and $Datalog^{\pm}$ -

Introduction and Motivation

Ontology Based Data Access (OBDA) is a popular setting used by many Semantic Web applications that encodes the *access to data sources using an ontology (vocabulary)* [17, 18, 9]. The use of the ontology will help obtain a unified view over heterogeneous data sources. Moreover, the ontology will enable the exploitation of implicit knowledge not explicitly stored in the data sources alone. One of the main difficulties in OBDA consists in dealing with potentially inconsistent union of facts (data sources). Reasoning with inconsistency needs additional mechanisms because classical logic will infer everything out of *falsum*. It is classically assumed (and a hypothesis that we will also follow in this paper) that the inconsistency in OBDA occurs at the fact level and not due to the ontology [17, 18, 9]. The *facts are error prone* due to their unrestrained provenance while *ontologies are considered agreed upon* as shared conceptualisations.

We consider here two main methods of handling inconsistency. On one hand (and inspired from database research) we consider *repair based techniques*. A repair is a maximally consistent set of facts. Reasoning with inconsistency using repairs relies on reasoning with repairs and combining the results using various methods (called *inconsistency tolerant semantics*). [5, 7, 13] Despite them being the mainstream techniques for OBDA reasoning, the main drawback of inconsistent tolerant semantics is the *lack of implementations* excepting a few dedicated approaches to particular semantics [14, 6].

A second method consists of using argumentation techniques. A Dung argumentation system [12] is a pair $\mathcal{AS} = (\mathcal{A}, \mathcal{C})$, where \mathcal{A} is a set of arguments and $\mathcal{C} \subseteq \mathcal{A} \times \mathcal{A}$ is a binary attack relation on them.

¹Corresponding Author: E-mail: croitoru@lirmm.fr

In this paper we demonstrate how to benefit from structured argumentation frameworks and their implementations to provide for reasoning capabilities of OBDA systems under inconsistency tolerant semantics. More precisely, given an inconsistent $Datalog^\pm$ [8] knowledge base we instantiate it using the $ASPIC^+$ framework [15] and show that the reasoning provided by $ASPIC^+$ is equivalent to the main inconsistent tolerant semantics in the literature. The *significance of this work* is a proposed workflow that will enable $Datalog^\pm$ frameworks to handle inconsistencies in knowledge bases by means of the $ASPIC^+$ framework. We use two frameworks:

- **Graal**, a Java toolkit dedicated to querying knowledge bases within the framework of $Datalog^\pm$ and maintained by GraphIK team. Graal takes as input a Dlgp file and a query and answer the query using various means (saturation, query rewriting). This toolkit can be found at <https://graphik-team.github.io/graal/>.
- **ACL's ASPIC** project that takes as input a query and $ASPIC^+$ knowledge base, i.e. rules (strict and defeasible), ordinary premises, axioms and preferences. The output is the answer to the query. This inference engine can be found at <http://aspic.cossac.org/components.html>.

We use Graal's representation of a knowledge base and construct the necessary input for the $ASPIC^+$ argumentation inference engine. The difficulty of this work resides in the definition of the mapping (the contrariness function and the way facts and rules are handled) that ensures the semantic equivalence proved in the next sections.

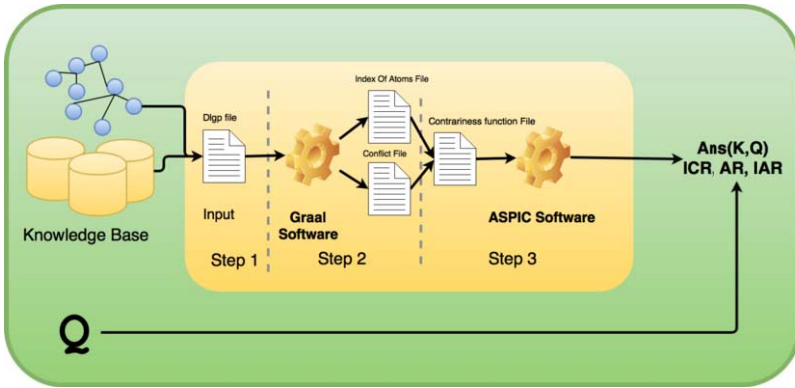


Figure 1. Interoperability Workflow of ACL and Graal.

In Figure 1 the interoperability workflow of the Graal software and the $ASPIC^+$ implementation are shown. Let us detail here how the workflow functions:

- **Step 1.** The input of the software is a $Datalog^\pm$ knowledge base obtained from the OBDA setting (that considers several data sources unified under the same ontology). The *dlg*p file that encodes this knowledge base (a textual format for the existential rule / Datalog framework) is parsed by the Graal framework. Each line in a *dlg*p file corresponds either to a fact, existential rule, negative constraint or conjunctive query. Please note that a complete grammar of the *dlg*p format is available here: https://graphik-team.github.io/graal/papers/datalog+_v2.0_en.pdf.

- **Step 2 and 3.** The intermediary files are meant to serve as input for the contrariness function computation in Step 3. The contrariness function encodes the conflicts between atoms. This will be formally defined in the next section. Please note that one of the main difficulties of this work is properly defining the contrariness function such that the produced results are sound and complete with respect to inconsistent tolerant semantics.
- **Querying.** The output of this inference engine is the answer to the query w.r.t the inconsistent knowledge base $\mathcal{K} = (\mathcal{F}, \mathcal{R}, \mathcal{N})$. The soundness and completeness of the answer with respect to inconsistency tolerant semantics is ensured by the equivalence results presented in the next section.

The Logical Language: *Datalog*[±]

In this section we explain the logical language *Datalog*[±] used throughout the paper. We define the notion of *Datalog*[±] knowledge base, inconsistent knowledge base and explain the three inconsistency tolerant semantics mostly used in the literature. Its language \mathcal{L} is composed of formulas built with the usual quantifier (\exists, \forall) and *only* the connectors implication (\rightarrow) and conjunction (\wedge). We consider first-order vocabularies with constants but no other function symbol. An **atomic formula** (or atom) is of the form $p(t_1, \dots, t_n)$ where $p \in \mathcal{P}$ is an n -ary predicate, and t_1, \dots, t_n are terms. Classically, a fact is a ground atom. We denote by \vec{x} a vector of variables. An *existential rule* (or simply a rule) is a closed formula of the form $R = \forall \vec{x} \forall \vec{y} (B \rightarrow \exists \vec{z} H)$, where B and H are conjuncts, with $\text{vars}(B) = \vec{x} \cup \vec{y}$, and $\text{vars}(H) = \vec{x} \cup \vec{z}$. The variables \vec{z} are called the existential variables of the rule R . B and H are respectively called the *body* and the *head* of R . We denote them respectively $\text{body}(R)$ for B and $\text{head}(R)$ for H . We may sometimes omit quantifiers and write $R = B \rightarrow H$. A *negative constraint* (or simply a constraint) is a rule of the form $N = \forall \vec{x} (B \rightarrow \perp)$. A rule $R = B \rightarrow H$ is **applicable** to a fact F if there is a homomorphism σ from B to F . Let F be a fact and \mathcal{R} be a set of rules. A fact F' is called an **\mathcal{R} -derivation** of F if there is a finite sequence (called the **derivation sequence**) $(F_0 = F, \dots, F_n = F')$ such that for all $0 \leq i < n$ there is a rule R which is applicable to F_i and F_{i+1} is an immediate derivation from F_i . Given a fact F and a set of rules \mathcal{R} , the **chase** (or saturation) procedure starts from F and performs rule applications in a breadth-first manner. The chase computes the **closure** of F , i.e. $CL_{\mathcal{R}}(F)$, which is the smallest set that contains F and that is closed under R -derivation, i.e. for every \mathcal{R} -derivation F' of F we have $F' \in CL_{\mathcal{R}}(F)$. Given a chase variant C [4], we call C -finite the class of set of rules \mathcal{R} , such that the C -chase halts on any fact F , consequently produces a finite $CL_{\mathcal{R}}(F)$. We limit our work in this paper to these kind of classes.

Let F and F' be two facts. $F \models F'$ if and only if there is a homomorphism from F' to F . Given two facts F and F' and a set of rules \mathcal{R} we say $F, \mathcal{R} \models F'$ if and only if $CL_{\mathcal{R}}(F) \models F'$ where \models is the classical first-order entailment [16].

Knowledge base and inconsistency Let us denote by \mathcal{L} the language described so far, A knowledge base \mathcal{K} is a finite subset of \mathcal{L} . Precisely, \mathcal{K} is a tuple $(\mathcal{F}, \mathcal{R}, \mathcal{N})$ of a finite set of facts \mathcal{F} , rules \mathcal{R} and constraints \mathcal{N} . Saying that $\mathcal{K} \models F$ means $CL_{\mathcal{R}}(\mathcal{F}) \models F$. We say a set of facts \mathcal{F} is \mathcal{R} -inconsistent with respect to a set of constraints \mathcal{N} and rules \mathcal{R} if and only if there exists $N \in \mathcal{N}$ such that $CL_{\mathcal{R}}(\mathcal{F}) \models \text{body}(N)$, otherwise \mathcal{F} is \mathcal{R} -consistent. A knowledge base $\mathcal{K} = (\mathcal{F}, \mathcal{R}, \mathcal{N})$ is said to be inconsistent with respect

to \mathcal{R} and \mathcal{N} (inconsistent for short) if \mathcal{F} is \mathcal{R} -inconsistent. We may use the notation $CL_{\mathcal{R}}(\mathcal{F}) \models \perp$ to mean the same thing.

In the area of inconsistent ontological knowledge base query answering, we usually check what can be inferred from an inconsistent ontology. We usually begin by calculating all maximal consistent subsets of \mathcal{K} called *repairs*. Given a knowledge base $\mathcal{K} = (\mathcal{F}, \mathcal{R}, \mathcal{N})$, we call by $Repairs(\mathcal{K})$ the set of all repairs defined as:

$$Repairs(\mathcal{K}) = \{\mathcal{F}' \subseteq \mathcal{F} \mid \mathcal{F}' \text{ is maximal for } \subseteq \text{ and } \mathcal{R}\text{-consistent}\}$$

Different inconsistency tolerant semantics are used for inconsistent ontology knowledge base query answering (Intersection of All Repairs: IAR, All Repairs: AR, Intersection of Closed Repairs: ICR); these semantics can yield different results.

Definition 1 [cf [11]]. Let $\mathcal{K} = (\mathcal{F}, \mathcal{R}, \mathcal{N})$ be a knowledge base and α be a query.

- α is AR-entailed from \mathcal{K} : $\mathcal{K} \models_{AR} \alpha$ iff for $\forall r \in Repairs(\mathcal{K})$, $Cl_{\mathcal{R}}(r) \models \alpha$.
- α is ICR-entailed from \mathcal{K} , written $\mathcal{K} \models_{ICR} \alpha$ iff $\bigcap_{r \in Repairs(\mathcal{K})} Cl_{\mathcal{R}}(r) \models \alpha$
- α is IAR-entailed from \mathcal{K} , written $\mathcal{K} \models_{IAR} \alpha$ iff $CL_{\mathcal{R}}(\bigcap_{r \in Repairs(\mathcal{K})} r) \models \alpha$

Structured Argumentation for $Datalog^{\pm}$

In this section we address the problem of how to use structured argumentation for $Datalog^{\pm}$. We show how the $ASPIC^+$ framework can be instantiated to yield results equivalent to the state of the art in OBDA inconsistency tolerant semantics. We define the first instantiation in the literature of $ASPIC^+$ using $Datalog^{\pm}$.

$ASPIC^+$ [15] is a framework for obtaining logical based argumentation system using any logical language \mathcal{L} . It is meant to generate an abstract argumentation framework and was created because abstract argumentation does not specify the structure of arguments and the nature of attacks. $ASPIC^+$ is meant to provide guidance to those aspects without losing a large range of instantiating logics. Before going any further we will provide a few basic abstract argumentation notions needed later in this section. We consider $\mathcal{AS} = (\mathcal{A}, \mathcal{C})$, where \mathcal{A} is a set of arguments and $\mathcal{C} \subseteq \mathcal{A} \times \mathcal{A}$ is a binary attack relation on them. We say that the argument $a \in \mathcal{A}$ is *acceptable* w.r.t a set of arguments $\varepsilon \subseteq \mathcal{A}$ iff $\forall b \in \mathcal{A}$ such that $(b, a) \in \mathcal{C}$, $\exists c \in \varepsilon$ such that $(c, b) \in \mathcal{C}$. ε is *conflict-free* iff $\nexists a, b \in \varepsilon$ such that $(a, b) \in \mathcal{C}$. ε is *admissible* iff ε is conflict-free and all arguments of ε are acceptable w.r.t ε . ε is *preferred* iff it is maximal (for set inclusion) and admissible. ε is *stable* iff it is conflict-free and $\forall a \in \mathcal{A} \setminus \varepsilon$, $\exists b \in \varepsilon$ such that $(b, a) \in \mathcal{C}$. ε is *complete* iff it contains all arguments that are acceptable w.r.t ε . ε is *grounded* iff it is minimal (for set inclusion) and complete. Reasoning takes place on the various ε (also called extensions). Following [15], to use $ASPIC^+$, we need to choose a logical language \mathcal{L} closed under negation (\neg), provide a set of rules $\mathcal{R} = \mathcal{R}_d \cup \mathcal{R}_s$ composed of defeasible rules and strict rules with $\mathcal{R}_d \cap \mathcal{R}_s = \emptyset$, specify a contrariness function $cf : \mathcal{L} \rightarrow 2^{\mathcal{L}}$ and a partial naming function $n : \mathcal{R}_d \rightarrow \mathcal{L}$ that associates a well-formed formulas of \mathcal{L} to a defeasible rule. The function n will not be used in this instantiation. In $ASPIC^+$ an argumentation system is a triple $\mathcal{AS} = (\mathcal{L}, \mathcal{R}, n)$ and a knowledge base is $\mathcal{K} \subseteq \mathcal{L}$ consisting of two disjoint subsets \mathcal{K}_n (the axioms) and \mathcal{K}_p (the ordinary premises).

To instantiate $ASPIC^+$ for $Datalog^\pm$, we define \mathcal{L} as $Datalog^\pm$, rules in definition 2 and the contrariness function in definition 3. Please note that definition 4 and 7 are new w.r.t. state of art regarding $Datalog^\pm$ instantiations conform [15].

Definition 2 *Strict rules (resp. defeasible rules) are of the form $\forall \vec{x} \forall \vec{y} (B \rightarrow \exists \vec{z} H)$ (resp. $\forall \vec{x} \forall \vec{y} (B \Rightarrow \exists \vec{z} H)$) with B , the body and H , the head are atoms or conjunction of atoms with $\text{vars}(B) = \vec{x} \cup \vec{y}$, and $\text{vars}(H) = \vec{x} \cup \vec{z}$.*

Definition 3 [[15]]. *The function cf is a function from \mathcal{L} to $2^{\mathcal{L}}$ such that:*

- φ is the contrary of ψ if $\varphi \in cf(\psi)$, $\psi \notin cf(\varphi)$
- φ is the contradictory of ψ if $\varphi \in cf(\psi)$, $\psi \in cf(\varphi)$
- Each $\varphi \in \mathcal{L}$ has at least one contradictory.

We define our own contrariness function to instantiate $ASPIC^+$ for $Datalog^\pm$ ($\mathcal{L} = Datalog^\pm$). This contrariness function is necessary because it is used in the attack relation. It is worth noting that the idea that we want to capture (as also defined in [1]) is that x is the contrary of y iff they cannot be both true but they can be both false. They are contradictory if the truth of one implies the falsity of the other and vice versa.

Definition 4 ($Datalog^\pm$ **contrariness function**) *Let $a \in \mathcal{L}$ and b be a conjunction of atoms. $b \in cf(a)$ iff $\exists \psi$ an atom such that $a \models \psi$ and $\{b, \psi\}$ is \mathcal{R} -inconsistent.*

Here we recall that an $ASPIC^+$ argument can be built from axioms and ordinary premises or from rules and other arguments. The arguments are built once $\mathcal{R}_d, \mathcal{R}_s, cf$ and \mathcal{K} are known.

Definition 5 (Argument cf [15]) *Arguments in $ASPIC^+$ can be in two forms:*

- $\emptyset \rightarrow c$ (resp. $\emptyset \Rightarrow c$) with $c \in \mathcal{K}_n$ (resp. $c \in \mathcal{K}_p$ or $\emptyset \Rightarrow c \in \mathcal{R}_d$) such that $Prem(A) = \{c\}$, $Conc(A) = c$, $Sub(A) = \{A\}$ with $Prem$ returns premisses of A and $Conc$ returns its conclusion.
 $DefRules(A) = \emptyset$.
- $A_1, \dots, A_m \rightarrow c$ (resp. $A_1, \dots, A_m \Rightarrow c$), such that there exists a strict (resp. defeasible) rule $r = B \rightarrow H$ (resp. $r = B \Rightarrow H$) and a homomorphism σ from B to $X = Conc(A_1) \wedge Conc(A_2) \wedge \dots \wedge Conc(A_m)$.
 $Prem(A) = Prem(A_1) \cup \dots \cup Prem(A_m)$,
 $Conc(A) = c = \alpha(X, r, \sigma)$,
 $Sub(A) = Sub(A_1) \cup \dots \cup Sub(A_m) \cup \{A\}$,
 $TopRule(A) = \text{rule } r = B \rightarrow H$ (resp. $r = B \Rightarrow H$), such that there exists an homomorphism σ from B to X .
 $DefRules(A) = DefRules(A_1) \cup \dots \cup DefRules(A_m)$ (resp. $DefRules(A) = DefRules(A_1) \cup \dots \cup DefRules(A_m) \cup \{TopRule(A)\}$).

Attacks in $ASPIC^+$ are based on three notions (undercutting, undermining and rebutting). Each of those notions are useful as they capture different aspects of conflicts. In short, arguments can be attacked on a conclusion of a defeasible inference (rebutting attack), on a defeasible inference step itself (undercutting attack), or on an ordinary premise (undermining attack).

Definition 6 [cf [15]]. Let a and b be arguments, we say that a attacks b iff a undercuts, undermines or rebuts b , where:

- a undercuts argument b (on b') iff $\text{Conc}(a) \in \text{cf}(n(r))$ for some $b' \in \text{Sub}(b)$ such that b' 's top rule r is defeasible.
- a rebuts argument b (on b') iff $\text{Conc}(a) \in \text{cf}(\psi)$ for some $b' \in \text{Sub}(b)$ of the form $b'_0, \dots, b'_n \Rightarrow \psi$.
- a undermines b (on ψ) iff $\text{Conc}(a) \in \text{cf}(\psi)$ for an ordinary premise ψ of b .

We are now ready to define the mapping that allows the instantiation of ASPIC^+ with Datalog^\pm . The mapping will consider each fact as a defeasible rule because the inconsistency in the OBDA setting is assumed to come from the facts level. Therefore the only attack we consider in this instantiation is the undermine attack because we have simple defeasible rules. The rules of the ontology become strict rules.

Definition 7 (Mapping For ASPIC^+ Instantiation of Datalog^\pm) We denote by \mathcal{S} the set of all possible inconsistent knowledge bases of the form $\mathcal{K} = (\mathcal{F}, \mathcal{R}, \mathcal{N})$ and \mathcal{G} the set of all ASPIC^+ instantiation using Datalog^\pm language. The mapping $\tau : \mathcal{S} \rightarrow \mathcal{G}$ is defined as follows:

1. The mapping τ associates every \mathcal{R} -consistent subset $F_i \subseteq \mathcal{F}$ to its defeasible rule $\emptyset \Rightarrow \text{conjunct}(F_i)$ where $\text{conjunct}(F_i)$ denotes the conjunction of facts contained in F_i .
2. The mapping τ associates every rules $r_i \in \mathcal{R}$ to the same rule $r_i \in \mathcal{R}_s$.

We will considerate that if $\emptyset \Rightarrow c$, then c is an ordinary premise ($c \in \mathcal{K}_p$).

In order to give properties of the ASPIC^+ instantiation presented in this paper we remind few notions. ε is *admissible* iff ε is conflict-free and all arguments of ε are acceptable w.r.t ε . ε is *preferred* iff it is maximal (for set inclusion) and admissible. ε is *stable* iff it is conflict-free and $\forall a \in \mathcal{A} \setminus \varepsilon, \exists b \in \varepsilon$ such that $(b, a) \in \mathcal{C}$.

We denote by $\text{AF}_{\mathcal{K}}^A$ the ASPIC^+ argumentation framework constructed from \mathcal{K} using the mapping of definition 7. We restate that attacks in $\text{AF}_{\mathcal{K}}^A$ are composed only of undermining because we only have simple defeasible rules of the form $\emptyset \Rightarrow c$. The following lemma shows that stable extensions are closed under sub-arguments in the Datalog^\pm instantiation of ASPIC^+ .

Lemma 1 Let ε be an ASPIC^+ stable extension and $A \in \varepsilon$ an argument contained in ε . Then $\text{Sub}(A) \subseteq \varepsilon$.

Notation Let $c = \alpha_1 \wedge \alpha_2 \wedge \dots \wedge \alpha_n$ be a conjunction of facts. $\text{Elimination}(c) = \{\alpha_1, \alpha_2, \dots, \alpha_n\}$ is the set resulting from eliminating the conjunction of c . Let S be a set of facts. We denote by $\mathcal{P}(S)$ the superset of S which correspond to all subsets of S .

We can now define the set of arguments constructed on a consistent set of facts.

Definition 8 Let $\mathcal{K} = (\mathcal{F}, \mathcal{R}, \mathcal{N})$ be a knowledge base and $\text{AF}_{\mathcal{K}}^A$ be the corresponding ASPIC^+ instantiation and $S \subseteq \mathcal{F}$ a \mathcal{R} -consistent subset of \mathcal{F} . We denote by $\text{Arg}^A(S)$ the set of arguments such that their premises are contained in S . Formally:

$$Arg^A(S) = \{ASPIC^+ \text{ argument } a \mid \bigcup_{c \in Prem(a)} Elimination(c) \subseteq \mathcal{P}(S)\}$$

The main result shows that the set of stable extension coincides with the set of preferred one and it is obtained from the arguments built on repairs.

Theorem 1 (Repair Equivalence for $ASPIC^+$ Instantiation) *Let $\mathcal{K} = (\mathcal{F}, \mathcal{R}, \mathcal{N})$ be a knowledge base, $AF_{\mathcal{K}}^A$ be the corresponding $ASPIC^+$ instantiation and $\sigma \in \{\text{preferred}, \text{stable}\}$. Then:*

$$\{Arg^A(R) \mid R \in Repair(\mathcal{K})\} = Ext_{\sigma}(AF_{\mathcal{K}}^A)$$

The state of the art can also provide a structured argumentation framework of $Datalog^{\pm}$ [11,10]. Let $\mathcal{K} = (\mathcal{F}, \mathcal{R}, \mathcal{N})$ be a knowledge base, we denote by $AF_{\mathcal{K}}^M$ the instantiated logical argumentation framework $(\mathcal{A}, \mathcal{C})$ with $\mathcal{A} = Arg(\mathcal{F})$ and \mathcal{C} defined in [11,10]. According to [11] the arguments constructed on the set of repairs coincide with the arguments in the stable and preferred extension: $\{Arg(R) \mid R \in Repair(\mathcal{K})\} = Ext_{\sigma}(AF_{\mathcal{K}}^M)$. We can thus conclude that the preferred/stable extensions in the two instantiated frameworks are the same and that for each stable/preferred extension of one framework, there is a corresponding stable/preferred extension in the other and vice-versa. This is formalized in the theorem below.

Theorem 2 (Instantiations Equivalence) *Let $\mathcal{K} = (\mathcal{F}, \mathcal{R}, \mathcal{N})$ be a knowledge base, $AF_{\mathcal{K}}^M$ and $AF_{\mathcal{K}}^A$ be the two argumentation framework instantiations. Then if $\sigma \in \{\text{preferred}, \text{stable}\}$, $|Ext_{\sigma}(AF_{\mathcal{K}}^M)| = |Ext_{\sigma}(AF_{\mathcal{K}}^A)|$ and for each extension under semantics σ , $\varepsilon \in Ext_{\sigma}(AF_{\mathcal{K}}^M)$, there is a corresponding extension $\varepsilon_2 \in Ext_{\sigma}(AF_{\mathcal{K}}^A)$ and vice-versa (the corresponding extension can be found via repairs).*

Conclusions

In this paper we demonstrated how to benefit from structured argumentation frameworks and their implementations to provide for reasoning capabilities of OBDA systems under inconsistency tolerant semantics. More precisely, given an inconsistent $Datalog^{\pm}$ knowledge base we instantiated it using the $ASPIC^+$ framework and showed that the reasoning provided by $ASPIC^+$ is equivalent to the main inconsistent tolerant semantics in the literature. A workflow of interoperability between $ASPIC^+$ ACL framework and Graal $Datalog^{\pm}$ framework was thus formally underpinned. In future work we are interested in exploiting this workflow for the explanation capabilities of inconsistent tolerant semantics [2,3].

Acknowledgments

The authors acknowledge the support of ANR grant QUALINCA (ANR-12-0012).

References

- [1] L. Amgoud. Five weaknesses of ASPIC+. In *Advances in Computational Intelligence - 14th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, IPMU 2012, Catania, Italy, July 9-13, 2012, Proceedings, Part III*, pages 122–131, 2012.
- [2] A. Arioua and M. Croitoru. A dialectical proof theory for universal acceptance in coherent logic-based argumentation frameworks. *Proc of ECAI 2016 - to appear*.
- [3] A. Arioua and M. Croitoru. Dialectical characterization of consistent query explanation with existential rules. In *FLAIRS: Florida Artificial Intelligence Research Society*, 2016.
- [4] J.-F. Baget, F. Garreau, M.-L. Mugnier, and S. Rocher. Revisiting Chase Termination for Existential Rules and their Extension to Nonmonotonic Negation. *ArXiv e-prints*, May 2014.
- [5] M. Bienvenu. On the complexity of consistent query answering in the presence of simple ontologies. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence, July 22-26, 2012, Toronto, Ontario, Canada.*, 2012.
- [6] M. Bienvenu, C. Bourgaux, and F. Goasdoué. Querying inconsistent description logic knowledge bases under preferred repair semantics. In *AAAI*, pages 996–1002, 2014.
- [7] M. Bienvenu and R. Rosati. Tractable approximations of consistent query answering for robust ontology-based data access. In *IJCAI 2013, Proceedings of the 23rd International Joint Conference on Artificial Intelligence, Beijing, China, August 3-9, 2013*, 2013.
- [8] A. Cali, G. Gottlob, and T. Lukasiewicz. Datalog+/-: a unified approach to ontologies and integrity constraints. In *Proc of ICDT'09*, pages 14–30. ACM, 2009.
- [9] A. Cali, G. Gottlob, and T. Lukasiewicz. A general datalog-based framework for tractable query answering over ontologies. *J. Web Sem.*, 14:57–83, 2012.
- [10] M. Croitoru, R. Thomopoulos, and S. Vesic. Introducing preference-based argumentation to inconsistent ontological knowledge bases. In *PRIMA 2015: Principles and Practice of Multi-Agent Systems - 18th International Conference, Bertinoro, Italy, October 26-30, 2015, Proceedings*, pages 594–602, 2015.
- [11] M. Croitoru and S. Vesic. What can argumentation do for inconsistent ontology query answering? In *Scalable Uncertainty Management - 7th International Conference, SUM 2013, Washington, DC, USA, September 16-18, 2013. Proceedings*, pages 15–29, 2013.
- [12] P. M. Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artif. Intell.*, 77(2):321–358, 1995.
- [13] D. Lembo, M. Lenzerini, R. Rosati, M. Ruzzi, and D. F. Savo. Inconsistency-tolerant semantics for description logics. In *Web Reasoning and Rule Systems - Fourth International Conference, RR 2010, Bressanone/Brixen, Italy, September 22-24, 2010. Proceedings*, pages 103–117, 2010.
- [14] M. V. Martinez, C. A. D. Deagustini, M. A. Falappa, and G. R. Simari. Inconsistency-tolerant reasoning in datalog^{*}{\ pm} ontologies via an argumentative semantics. In *Ibero-American Conference on Artificial Intelligence*, pages 15–27. Springer, 2014.
- [15] S. Modgil and H. Prakken. The ASPIC⁺ framework for structured argumentation: a tutorial. *Argument & Computation*, 5(1):31–62, 2014.
- [16] M. Mugnier. Ontological query answering with existential rules. In *Web Reasoning and Rule Systems - 5th International Conference, RR 2011, Galway, Ireland, August 29-30, 2011. Proceedings*, pages 2–23, 2011.
- [17] A. Poggi, D. Lembo, D. Calvanese, G. De Giacomo, M. Lenzerini, and R. Rosati. Linking data to ontologies. *J. Data Semantics*, 10:133–173, 2008.
- [18] L. Popa, S. Abiteboul, and P. G. Kolaitis, editors. *Proceedings of the Twenty-first ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, June 3-5, Madison, Wisconsin, USA*. ACM, 2002.

Regular Papers

This page intentionally left blank

Argument Schemes for Reasoning About the Actions of Others

Katie ATKINSON^{a,1}, Trevor BENCH-CAPON^a

^a*Department of Computer Science, University of Liverpool, UK*

Abstract. In practical reasoning, it is important to take into consideration what other agents will do, since this will often influence the effect of actions performed by the agent concerned. In previous treatments, the actions of others must either be assumed, or argued for using a similar form of practical reasoning. Such arguments, however, will also depend on assumptions about the beliefs, values and preferences of the other agents, and so are difficult to justify. In this paper we capture, in the form of argumentation schemes, reasoning about what others will do, which depends not on assuming particular actions, but through consideration of the expected utility (based on the promotion and demotion of values) of particular actions and alternatives. Such arguments depend only on the values and preferences of the agent concerned, and do not require assumptions about the beliefs, values and preferences of the other relevant agents. We illustrate the approach with a running example based on Prisoner's Dilemma.

Keywords. practical reasoning, values, argumentation schemes, AATS

1. Introduction

In the method for value based practical reasoning proposed in [3] and later improved in [2], the reasoning goes through three stages. First there is a *problem formulation* stage in which states and actions allowing transition between them are modelled and the transitions labelled with the values they promote and demote. In [3] the modelling is done using an Alternation Action Based Transition system (AATS) [19]. Note that the transitions in an AATS are the *joint* actions of all the agents involved, since the state reached by a given action will often depend on what other agents choose to do. Next there is the *epistemic* stage in which the initial state must be determined (or assumed) and the particular joint action that will result from the agent's choice of action must be established or assumed. Finally conflicts between the various arguments that can be generated from this structure are resolved according to the preferences of the agent, using a Value Based Argumentation Framework (VAF) [7]. A significant problem with this method is the treatment of the actions of others. Although it is possible to justify the actions attributed to others, this does require assumptions to be made as to how they will formulate their part of the problem, the assumptions they themselves will make and the preferences they will use to resolve their VAF. All this can introduce rather more uncertainty than is desirable, and must be done for every other agent relevant to the scenario. An improved treatment,

¹Corresponding Author: tbc; E-mail: tbc@csc.liv.ac.uk

which reduces the need to make assumptions about others, was proposed in [4]. In this paper we will advance this initial work by expressing this proposal in the form of a set of argumentation schemes [16]. This will clarify the nature of the arguments, and how they can be deployed in dialogues.

Section 2 will give some essential background on the AATS and the well known *Prisoner's Dilemma* which will be used as the running example in this paper. Section 3 will summarise the proposal of [4], section 4 will give the schemes and their critical questions. Section 5 will show the use of the schemes in a dialogical setting and section 6 will offer some concluding remarks.

2. Background

2.1. Alternation Action Based Transition systems (AATS)

Based on Alternating Time Temporal Logic [1], AATS were originally presented in [19] as semantical structures for modelling game-like, dynamic, multi-agent systems in which the agents can perform actions in order to modify and attempt to control the system in some way. As such they provide an excellent basis for modelling situations in which a set of agents are required to make decisions. The definition in [19] is:

Definition 1: AATS.

An *Action-based Alternating Transition System* (AATS) is an $(n + 7)$ -tuple $S = \langle Q, q_0, Ag, Ac_1, \dots, Ac_n, \rho, \tau, \Phi, \pi \rangle$, where:

- Q is a finite, non-empty set of *states*;
- $q_0 \in Q$ is the *initial state*;
- $Ag = \{1, \dots, n\}$ is a finite, non-empty set of *agents*;
- Ac_i is a finite, non-empty set of actions, for each $ag_i \in Ag$ where $Ac_i \cap Ac_j = \emptyset$ for all $ag_i \neq ag_j \in Ag$;
- $\rho : Ac_{ag} \rightarrow 2^Q$ is an *action pre-condition function*, which for each action $\alpha \in Ac_{ag}$ defines the set of states $\rho(\alpha)$ from which α may be executed;
- $\tau : Q \times J_{Ag} \rightarrow Q$ is a partial *system transition function*, which defines the state $\tau(q, j)$ that would result by the performance of j from state q – note that, as this function is partial, not all joint actions are possible in all states (cf. the pre-condition function above);
- Φ is a finite, non-empty set of *atomic propositions*; and
- $\pi : Q \rightarrow 2^\Phi$ is an interpretation function, which gives the set of primitive propositions satisfied in each state: if $p \in \pi(q)$, then this means that the propositional variable p is satisfied (equivalently, true) in state q .

AATSs are particularly concerned with the joint actions of the set of agents Ag . j_{Ag} is the joint action of the set of n agents that make up Ag , and is a tuple $\langle \alpha_1, \dots, \alpha_n \rangle$, where for each α_j (where $j \leq n$) there is some $ag_i \in Ag$ such that $\alpha_j \in Ac_i$. Moreover, there are no two different actions α_j and $\alpha_{j'}$ in j_{Ag} that belong to the same Ac_i . The set of all joint actions for the set of agents Ag is denoted by J_{Ag} , so $J_{Ag} = \prod_{i \in Ag} Ac_i$. Given an element j of J_{Ag} and an agent $ag_i \in Ag$, ag_i 's action in j is denoted by j^i . This definition was extended in [3] to allow the transitions to be labelled with the values they promote.

Definition 2: AATS+V.

An AATS+V is defined by adding two more elements as follows:

- V is a finite, non-empty set of values.
- $\delta : Q \times Q \times V \rightarrow \{+, -, =\}$ is a *valuation function* which defines the status (promoted (+), demoted (-) or neutral (=)) of a value $v_u \in V$ ascribed to the transition between two states: $\delta(q_x, q_y, v_u)$ labels the transition between q_x and q_y with one of $\{+, -, =\}$ with respect to the value $v_u \in V$.

An *Action-based Alternating Transition System with Values* (AATS+V) is thus defined as a $(n + 9)$ tuple $S = \langle Q, q_0, Ag, Ac_1, \dots, Ac_n, \rho, \tau, \Phi, \pi, V, \delta \rangle$. The value may be ascribed on the basis of the source and target states, or in virtue of an action in the joint action, where that action has intrinsic value.

2.2. Prisoner's Dilemma

In this very well known game [13], both players may either cooperate or defect. Mutual cooperation results in a pay off of 3 to each player, mutual defection a payoff of 1 to each player, and if one cooperates and the other defects the defector receives 5 and the cooperator receives 0. Note first that the “correct” strategy is to defect since that gives a better payoff whichever move the other makes (is the *dominant* strategy), and second that it is not a zero-sum game: collective utility is maximised by mutual cooperation. Note also that, as in other situations empirically tested in behavioural economics (e.g. [12], [8] and [9]), the game-theoretic choice is rarely found in practice. As explained in [15] in many social situations conventions to encourage mutual cooperation emerge or are devised, and such conventions may be reinforced by defection being the subject of punishment [11]. In the example discussed in [15], in a military situation much effort is made to build up trust and loyalty to create an *esprit de corp* in a regiment so that members will cooperate rather than defect, feeling that they are able to rely on their comrades, and in turn reluctant to let their comrades down. The explanation for this deviation from game theoretic behaviour is that the participants have values other than the payoff to themselves, and they tend to import the values established in their culture into their behaviour in the game. Some other values therefore need to be considered. Here we will use the following values, suggested by the previous studies in experimental economics. Each value is relative to the player affected.

- Player Money (M1 and M2): promoted if player 1's (or 2's) payoff is greater than 1 (which is the least that can be ensured), and demoted if it is less than 1.
- Player Guilt (G1 and G2): demoted if player 1 (or 2) defects and player 2 (or 1) cooperates.
- Player Self-Esteem (S1 and S2): demoted if player 1 (or 2) cooperates and player 2 (or 1) defects: player 1 (or 2) may feel that they have allowed themselves to be taken advantage of and that they should have known better.

In this game there are four joint actions which promote and demote values as shown in Table 1. In the case of M1 and M2 we also show the relative extent of promotion and demotion of the values. Since a player can always ensure a payoff of 1, we consider money to be promoted only if it exceeds 1, and we take the degree of promotion as *payoff* – 1. Similarly the degree of demotion is taken as relative to the neutral situation of mutual defection.

Table 1. Value Promotion and Demotion in the Prisoner's Dilemma

Joint Action	Player 1	Player 2	Promoted	Demoted
j1	C	C	2M1,2M2	
j2	C	D	4M2	M1,S1,G2
j3	D	C	4M1	M2, S2,G1
j4	D	D		

3. Reasoning About Others with Expected Utilities

The current approach to reasoning about others' actions based on [3] is:

1. Select a desirable transition based on the values it promotes and demotes.
2. Argue for the individual action performed by the agent in the joint action corresponding to that transition.
3. Consider objections based on the other agents choosing different actions and so causing different joint actions to be performed.
4. Attempt to rebut these objections because:
 - (a) The values promoted and demoted by the alternative transition are acceptable.
 - (b) It is considered that the other agents will not act in this way.

Whereas 4a can be resolved on the basis of the agent concerned, 4b, which is very often needed, requires more assumptions about the other agents than can be really justified. To remedy this defect, [4] proposed that instead of a specific joint action, the *set* of joint actions that could result from the selected individual action should be considered. This is done by calculating the *expected utility* of performing the action, in terms of the probabilities of the joint actions containing that action. In order to facilitate this calculation it is necessary to express the various benefits of performing an action in a "common currency". Therefore as well as ordering values, the agent will provide weights expressing all the values in terms of the most preferred value (which will have a weight of 1). Thus given three values²: $V_1 \succ V_2 \succ V_3$, the agent may rate V_2 as $0.6V_1$ and V_3 as $0.3V_1$. How sensitive the arguments are to these relative weights is something which can be explored through objections and rebuttals, as we will see when we consider the argumentation schemes.

Definition 3: Agent Preferences

The preferences of an agent $ag \in Ag$ is the set $O_{ag} = \{\langle v_0 * w_0 \rangle, \langle v_1 * w_1 \rangle, \dots, \langle v_n * w_n \rangle\}$, where $v_0 \dots v_n$ are values and $w_0 \dots w_n$ are weights with $w_0 \geq w_1 \geq \dots \geq w_n$.

Using these weights we can calculate the expected utility of agent i performing α . We will assume that if the desired joint action (j_0) does not result from the performance of α the worst case alternative joint action (j_w) will be the one that does result (since this will represent a lower bound). Informally the expected utility of performing α will be the utility of j_0 multiplied by the probability of j_0 plus the utility of j_w (which will often be negative) multiplied by (1 minus the probability of j_0).

²Using VAF notation [7] where \succ denotes preference.

Definition 4: Expected Utility of ag performing α in state q_s

- Let $J_\alpha = \{j_0, j_1 \dots j_n\}$ be the set of joint actions in which ag performs α (i.e. $j^{ag} = \alpha$) available in the starting state, q_s .
- Let P_{ag_k} be the values for ag promoted by the performance of $j_k \in J_\alpha$ in q_s . Let D_{ag_k} be the values of ag demoted by the performance of $j_k \in J_\alpha$ in q_s .
- The positive utility for ag , $pu(ag, j_k)$, of the performance of $j_k \in J_\alpha$ in q_s is $\sum_{i=0}^{i=n} (v_i * w_i)$ where $v_1 \in P_{ag_k}$ and the negative utility for ag , $du(ag, j_k)$, of the performance of $j_k \in J_\alpha$ in q_s is $\sum_{i=0}^{i=n} (v_i * w_i)$ where $v_1 \in D_{ag_k}$. The utility, $u(ag, j_k)$, for ag of the performance of $j_k \in J_\alpha$ in q_s is $pu(ag, j_k) - du(ag, j_k)$.
- Let U_{ag} be the set of utilities for ag , $\{u_0, u_1 \dots u_n\}$, such that $u_i = u(ag, j_i)$ for $j_i \in J_\alpha$. Let u_w be such that for all $u_i \in U_{ag}$, $u_w \leq u_i$.
- Let $prob(j_0)$ be the probability of j_0 being the joint action performed when ag performs α in q_s .
- Now the expected utility, $eu_{ag}(\alpha)$ for ag of performing α in q_s is $(u(ag, j_0) * prob(j_0)) + (u(ag, j_w) * (1 - prob(j_0)))$

By taking j_w as the alternative to j_0 , we come up with the lower bound on the expected utility, which will always be “safe”. If we were able to assign actual probabilities to the other members of J_α , we could be exact, but in the kind of situations we wish to consider, this is rarely possible and so we will use the worst case. In PD the question as to which alternative joint action might result from performing α does not arise as there are only two joint actions for each of the actions available in the initial state.

In the traditional PD only the agent’s own payoff is recognised as having utility. The utility is the actual payoff minus the guaranteed payoff (i.e. the payoff from mutual defection). For cooperation the utility is 2 when the other cooperates and -1 when the other defects. For defection it is 4 when the other cooperates and 0 when the other defects. The expected utilities for ag cooperating (dark grey) and defecting (light grey) for the various probabilities of the other cooperating are shown in Figure 1.

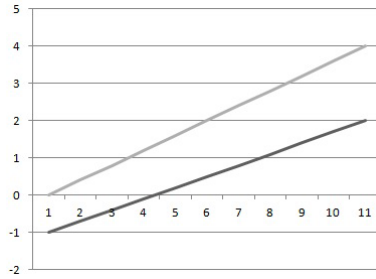


Figure 1. Expected Utilities for M1 only. Dark grey is ag cooperates, light grey is ag defects.

Suppose, however, that both the values $M1$ and $M2$ are recognised in PD, and $M2$ is weighted at $0.5M1$. Now the utility of cooperating when the other also cooperates will be $3M1$, and the utility of cooperating when the other defects $M1$. Similarly we can calculate the expected utility of defecting for the various probabilities of the other cooperating. Defecting when the other cooperates yields a utility of $3.5M1$, and mutual defection 0 (since this is the base line case, no values are considered promoted). Again

the desired joint action is performed when the other agent cooperates. This gives the graph shown as Figure 2a. The crossover is at $prob(j_0) = 0.67$.

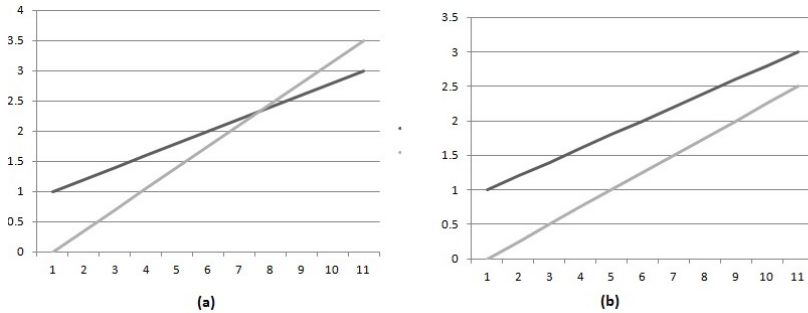


Figure 2. Expected Utilities for (a) $M2 = 0.5M1$ and (b) $M2 = 0.5M1$ and $G = M1$. Dark grey is *ag* cooperates, light grey is *ag* defects.

If we now add in the value of Guilt (with a weight of 1), which gives a negative utility when an agent defects and the other cooperates, we get the expected utilities shown in Figure 2b.

There are three possibilities, which correspond to these three figures. In Figure 1, which shows the traditional PD, we find that defection *dominates* cooperation: the expected utility is higher for every value of $prob(j_0)$. Therefore defection is the preferred action, whatever the probability of the other cooperating. In Figure 2b the reverse is true: the inclusion of additional values means that cooperation dominates defection. In Figure 2a, there is a crossover, at $prob(j_0) = 0.7$, so that for high probabilities of cooperation, defection is preferred, but for low levels, the utility afforded to the payoff received by the other makes cooperation preferred.

3.1. Arguments Using Expected Utilities

Several types of argument can be based on the expected utilities for PD.

1. With your value preferences, you should C (respectively, D) since the expected utility is always greater than any alternative
2. With your value preferences, you should C (respectively, D) since the expected utility is always positive
3. With your value preferences, you should C (respectively, D) since the expected utility is greater than the alternative when the probability of cooperation is greater (less) than P.

Of these (1) is appropriate when the action advocated is dominant, and is the strongest of the three. Argument (2) is rather weak: although the expected utility is always positive, the proposed action may be dominated by the alternative for some (or even all) values of $prob(j_0)$. It may, however, be useful if we wish to reach the target state in order to enable some more beneficial action, since it indicates that no harm is done, and so can be used to rebut objections. The argument shows that we suffer no loss, although there is an opportunity cost. Argument (3) can be effective provided we can give reasons to suppose that probability of cooperation is in the desired range.

4. The Argumentation Schemes

The above arguments (1)-(3) for PD can be generalised and presented as argumentation schemes in the manner of [16]. Note that the users of these schemes are not to be identified with the players in the PD. The dialogues below are supposed to represent one player being given advice (likely to be a persuasion situation), or two people acting as a team in the PD discussing their best course of action (likely to be a deliberation situation). The schemes have a number of premises, and the conclusion in common. These are the premises that set up the situation and identify the key elements. Then additionally there is one key premise for each scheme, characteristic of the scheme. All the schemes have

- **Conclusion:** ag should perform α

4.1. Common Premises

Each scheme will have four premises in common:

- **Values Premise:** V is the set of values considered to be relevant by ag
- **Weighting Premise:** The relative valuation of the members of V given by ag is S set of $\langle value, relativeweight \rangle$ pairs
- **Joint Action Premise:** $\{j_0, j_1, \dots, j_n\}$ is the set of joint action S in which ag performs α
- **Expected Utility Premise:** $eu_{ag}(\alpha, prob(j_0))$ returns the expected utilities of agent ag performing α for values of $prob(j_0)$ $0 \leq prob(j_0) \leq 1$ where j_0 is the desired joint action.

The first premise identifies the values which the agent will consider and the second weights them in terms of the most important value. The joint actions containing the advocated action α as the action of ag are then taken from the AATS to give the third premise. The fourth premise then establishes the expected utilities for the various probabilities of the desired joint action, j_0 , resulting from ag performing α .

4.2. Characteristic Premises

We have three schemes bases of the arguments (1)-(3) of section 3. We will name these as follows:

1. Argument from Dominance
2. Argument From Positive Expected Utility
3. Argument From Probable Compliance³

Each has its own characteristic premises. For Argument from Dominance:

- **Dominance Premise:** $eu_{ag}(\alpha, j_0) \geq eu_{ag}(\beta, j_0)$ for any alternative action β available to ag , for all values of $prob(j_0)$; where j_0 is the joint action compliant with the action of ag .

For Argument From Positive Expected Utility:

- **Positive Utility Premise:** $eu_{ag}(\alpha, j_0) \geq 0$ for all values of $prob(j_0)$

³We call the other agents acting so that j_0 results from ag performing α *compliance*.

Finally, for Argument From Probable Cooperation:

- **Probability Range Premise:** $eu_{ag}(\alpha, j_0) \geq eu_{ag}(\beta, j_0)$ for all values of $prob(j_0) \geq$ (respectively, \leq) *crossover*, where *crossover* is the point at which $eu_{ag}(\alpha, j_0)$ becomes greater (respectively, less) than $eu_{ag}(\beta, j_0)$

Here we are taking the joint action resulting from *ag* performing β to be the best alternative, namely the joint action containing β which yields *ag* the highest expected utility, i.e j_0 is the joint action compliant with the action of *ag*.

5. Critical Questions

These schemes can be associated with critical questions, as in [16]. Some will be common to all three schemes, while those associated with the characteristic premises will be applicable only to the particular scheme. We begin with those common to all schemes.

5.1. Critical Questions Applicable to All Schemes

- **CQ1** Are all the members of *V* relevant?
- **CQ2** Are any other Values (i.e values in the AATS+*V*, but not included in *V* for this argument) relevant?
- **CQ3** Are any members of *V* over weighted?
- **CQ4** Are any members of *V* under weighted?

CQ1 and CQ2 are directed at the Values Premise and CQ3 and CQ4 at the weighting premise. We have no CQs directed at the other two premises, which are taken directly from the AATS and so considered beyond challenge at this stage. If there are only two joint actions containing α , the Expected Utility Premise is fully determined by the labelling of transitions in the AATS, together with the Values and Weighting premises. If there are more than two such joint actions, the worst case should be used, as described in definition 4.

Once we have established which values we wish to consider, we can only challenge the characteristic premise of the Argument from Dominance by coming up with an alternative action γ for which $eu_{ag}(\gamma, j_0) > eu_{ag}(\alpha, j_0)$ for at least some probabilities of compliance. But if the dominance premise is indeed true, this would challenge the AATS, and so it is considered outside the scope of this stage of the argumentation. Therefore there are no CQs peculiar to the Argument from Dominance. Similarly the Argument From Positive Expected Utility has no individually applicable CQs. The Argument From Probable Cooperation does, however, have its own CQ:

- **CQ5** Can $prob(j_0)$ be assumed to be \geq (respectively, \leq) *crossover*?

5.2. Rebuttals

These critical questions will have their own typical rebuttals, but these may depend on the context supplied by the original scheme. For example CQ3 could be met by

*even if the relative weight of *v* is reduced to *n*%, $eu_{ag}(\alpha, j_0)$ remains greater than its alternatives for all values of $prob(j_0)$.*

in the context of the Argument from Dominance, but by

even if the relative weight of v is reduced to $n\%$, $eu_{ag}(\alpha, j_0)$ remains ≥ 0 for all values of $prob(j_0)$.

in the context of Argument From Positive Expected Utility. These rebuttals can be preempted by posing a more specific challenge: for example, to the Argument From Positive Expected Utility:

if the relative weight of v is reduced to $n\%$, $eu_{ag}(\alpha, j_0)$ becomes < 0 for values of $prob(j_0) < p$.

Perhaps a more natural way of making the last move in a dialogue is first to pose the appropriate CQ and then to put forward an argument of ones own. Thus the last challenge would be made using both CQ3, and an Argument from Probable Cooperation for an alternative to α .

6. Dialogue Based on These Schemes

These schemes, challenges based on the critical questions and rebuttals can be deployed in an adversarial discussion. As an example we will consider a dialogue between *Coop* and *Def*, concerning the action to take in the Prisoner's Dilemma.

In the dialogue, we will take it that the participants agree on the AATS, so that the schemes can be summarised in the form

Given *ListOfValueWeightPairs*, one should α because *CharacteristicPremise*.

Def begins the dialogue:

D1 Given $\langle M1, 1 \rangle$, one should defect because the expected value of defection is always greater than the expected value of cooperation.

Coop can now challenge this using CQ2. As there is only a single value, the other CQs cannot be used here. *Coop* needs to find a value demoted by defection. As Table 1 shows, there are three possibilities: the payoff of the other player, Guilt, or the self-esteem of the other player. *Coop* can make the challenge (here *Coop* uses the payoff of the other player) and then counter with an Argument From Probable Cooperation:

C1 You must take some account of the payoff to the other player.

C2 Given $\langle M1, 1 \rangle$, $\langle M2, 0.5 \rangle$, one should cooperate since the expected utility is greater for probability of the other cooperating less than 0.67.

At this point *Def* has several possibilities:

R1, based on CQ1: *There is no reason to care about the payoff of the other*. This simply refuses to modify the position of *D1*.

R2, based on CQ2: Introduce another value, demoted by cooperation. Self Esteem is a possibility. A weight of 1 for *S1* will restore *D* to dominance,

R3, based on CQ3: Argue that *M2* is overrated. For example, reducing the weight to 0.2 will restore defection to dominance. Any greater weight will give some value of $prob(j_0)$ at which cooperation is better.

R4. Since C2 expresses an Argument From Probable Cooperation, CQ5 is also available.

How *Coop* responds will depend on the move made by *Def*. For R1, much will depend on the context. If *Def* is trying to persuade *Coop*, *Coop* gets to choose the values [6], and so the move is not available to *Def*, since *Coop* has, in C1, already shown that M2 is, in his opinion, something to care about. In other situations, such as deliberation, they are in a different dialogue type, and a nested persuasion dialogue in which *Coop* will attempt to persuade *Def* that the value should be recognised must be entered. Unless *Coop* is trying to persuade *Def* (when *Def* has the last word on what values should be considered), R1 is probably best avoided at this point. R2 similarly depends on context. If it is *Coop* being persuaded, *Coop* can simply reject this challenge, but if *Def* is being persuaded, or in a deliberation it may be an effective move.

Probably the best tactic for *Def* is to use R3, since this explores the sensitivity of *Coop*'s challenge to the the weight used and so can establish the least weight that may be accorded to the payoff the other. Even if *Def* and *Coop* agree to compromise and accept a value for M2 between 0.2 and 0.5, then having made R3 means that R4 becomes more effective because of the reduction in the crossover point. For example, splitting the difference at 0.35 will reduce the crossover to 0.29.

Suppose, however, the dialogue in fact continues as follows (e.g. *Coop* is the persuadee, and so is able, in this context, to have the final say as to weights and values.)

D2 You have overrated M2. At 0.5, you would be happy for the other to defect when you cooperate⁴. Suppose we weight it at no more than 0.25M1.

D3 Given $\langle M1, 1 \rangle$ and $\langle M2, 0.25 \rangle$ one should defect because the expected value of defection is always greater than the expected value of cooperation.

C3 I think that 0.5 is the correct weight for M2.

Coop may now introduce a third value, say Guilt, which will enable the Argument from Dominance:

C4 Given $\langle M1, 1 \rangle$, $\langle M2, 0.5 \rangle$ and $\langle G1, 0.5 \rangle$, one should cooperate because the expected value of cooperation is always greater than the expected value of defection.

This will work well if *Coop* has the final say as to values. But if this is not so, *Coop* may still defend cooperation with the Argument From Positive Expected Utility:

C4a Given $\langle M1, 1 \rangle$ and $\langle M2, 0.5 \rangle$, I can cooperate because the expected value of cooperation is always greater than zero.

Suppose now that *Def* had responded to C2 with R4, arguing that there is no reason to think that the probability of cooperation will be below 0.67. Here *Coop* could try to argue why cooperation is unlikely (e.g. the game-theoretic dominance of defection) or reply with the Argument From Positive Expected Utility, which licenses the performance of the action while acknowledging that it may not be the best choice.

⁴This could be so in many concrete situations, depending on the relationship between the two players. A parent will often give preference to the needs of a child, or a cooperator may expect a present (or compensation) from one who defects. Normally, however, a player would be expected to wish to avoid the situation in which he cooperates and the other defects.

6.1. Discussion

As can be seen from the preceding section, the dialogue can take a variety of paths. The particular path taken will depend greatly on the context in which the dialogue is taking place, in particular the dialogue type [18]. If it is a persuasion dialogue, one participant (the person being persuaded) can decide on the values to be used, and the weights that they should be given. The other player can suggest additional values, and question the weights, and even present arguments for values to be recognised and for weights to be different, but is powerless to compel the acceptance of these suggestions.

In contrast in a deliberation dialogue (e.g [5], [17]), the participants need to agree on the values, and we would expect the values to be a union of the proposals of both participants, and the weights to represent some sort of compromise between them.

While studies of these sorts of game in behavioural economics such as [8], [12] and [9] make it clear that the best game theoretic choice is often not made since payoff seems rarely to be the only consideration, they make it equally clear that there is a great deal of inter-cultural (and even intra-cultural) variations in the additional values considered, and in the weights given to them. In deliberations, dialogues of this form are especially useful in refining proposals by including additional values so that the interests of the whole group are reflected, and the weights are such that the group as a whole considers them acceptable. Note that the arguments remain valid over a range of weights (and probabilities of success), so that the group can agree on a course of action without necessarily needing to reach full agreement on the weights and the probabilities, provided they can agree on a range acceptable to them all.

7. Concluding Remarks

We have provided a new way of capturing reasoning about the actions of others using argumentation and expected utilities. This account rectifies a serious defect in the account of practical reasoning procedure in [3] which required assumptions to be made about the beliefs, values and preferences of other agents whose choice of action affects the result of an agent's action. Modelling the other participant in a dialogue is difficult enough (e.g. [14] and [10]) and modelling several unseen agents is likely to be very much harder. In the proposed method here we avoid the need to make such assumptions, by considering not a particular joint action in which an agent performs α , but the set of joint actions which can result from the performance of α . Instead of the values promoted and demoted by a selected joint action, we considered the expected utility (with utility calculated in terms of the values promoted and demoted) of performing α .

We have presented this way of thinking about what the others might do in the form of a set of related argument schemes and critical questions, and considered how these schemes can be deployed in dialogues, both persuasion and deliberation dialogues. Possibly the most useful context is deliberation, as there these arguments provide a framework in which additional values can be introduced, the relative weights accorded to them discussed and possible compromises reached, and the range of probabilities of success for which the argument holds good to be established. Modelling other agents is a difficult and currently unresolved problem, and so the ability to take what others may do into account without making unfounded assumptions about their beliefs and preferences is

essential. The argumentation schemes presented here allow this to be done in the context of value-based practical reasoning based on an AATS in the manner of [3].

References

- [1] R. Alur, T. A. Henzinger, and O. Kupferman. Alternating-time temporal logic. *Journal of the ACM (JACM)*, 49(5):672–713, 2002.
- [2] K. Atkinson and T. Bench-Capon. Taking the long view: Looking ahead in practical reasoning. In *Computational Models of Argument - Proceedings of COMMA 2014*, pages 109–120.
- [3] K. Atkinson and T. Bench-Capon. Practical reasoning as presumptive argumentation using action based alternating transition systems. *Artif. Intell.*, 171(10-15):855–874, 2007.
- [4] K. Atkinson and T. Bench-Capon. Value-based reasoning and the actions of others. In *Proceedings of ECAI 2016*, in press, 2016.
- [5] K. Atkinson, T. Bench-Capon, and D. Walton. Distinctive features of persuasion and deliberation dialogues. *Argument and Computation*, 4(2):105–127, 2013.
- [6] T. Bench-Capon. Agreeing to differ: modelling persuasive dialogue between parties with different values. *Informal Logic*, 22:231–246, 2002.
- [7] T. Bench-Capon. Persuasion in practical argument using value-based argumentation frameworks. *Journal of Logic and Computation*, 13(3):429–448, 2003.
- [8] C. Engel. Dictator games: a meta study. *Experimental Economics*, 14(4):583–610, 2011.
- [9] J. Henrich, R. Boyd, S. Bowles, C. Camerer, E. Fehr, H. Gintis, and R. McElreath. In search of homo economicus: behavioral experiments in 15 small-scale societies. *The American Economic Review*, 91(2):73–78, 2001.
- [10] A. Hunter. Modelling the persuadee in asymmetric argumentation dialogues for persuasion. In *24th International Joint Conference on Artificial Intelligence*, pages 3055–3061, 2015.
- [11] S. Mahmoud, N. Griffiths, J. Keppens, A. Taweel, T. Bench-Capon, and M. Luck. Establishing norms with metanorms in distributed computational systems. *AI and Law*, 23(4):367–407, 2015.
- [12] H. Oosterbeek, R. Sloof, and G. Van De Kuilen. Cultural differences in ultimatum game experiments: Evidence from a meta-analysis. *Experimental Economics*, 7(2):171–188, 2004.
- [13] A. Rapoport and A. M. Chammah. *Prisoner's dilemma: A study in conflict and cooperation*. University of Michigan press, 1965.
- [14] T. Rienstra, M. Thimm, and N. Oren. Opponent models with uncertainty for strategic argumentation. In *IJCAI*, 2013.
- [15] E. Ullmann-Margalit. *The emergence of norms*. Clarendon Press Oxford, 1977.
- [16] D. Walton. *Argumentation Schemes for Presumptive Reasoning*. Lawrence Erlbaum Associates, Mahwah, NJ, USA, 1996.
- [17] D. Walton, K. Atkinson, T. Bench-Capon, A. Wyner, and D. Cartwright. Argumentation in the framework of deliberation dialogue. In *Arguing Global Governance: Agency, Lifeworld and Shared Reasoning*, pages 210–230. Taylor and Francis, 2010.
- [18] D. Walton and E. C. Krabbe. *Commitment in dialogue: Basic concepts of interpersonal reasoning*. SUNY press, 1995.
- [19] M. Wooldridge and W. van der Hoek. On obligations and normative ability: Towards a logical analysis of the social contract. *J. Applied Logic*, 3(3-4):396–420, 2005.

Verifiability of Argumentation Semantics

Ringo BAUMANN^a, Thomas LINSBICHLER^b, and Stefan WOLTRAN^b

^aComputer Science Institute, University of Leipzig, Germany

^bInstitute of Information Systems, TU Wien, Austria

Abstract. Dung’s abstract argumentation theory is a widely used formalism to model conflicting information and to draw conclusions in such situations. Hereby, the knowledge is represented by argumentation frameworks (AFs) and the reasoning is done via semantics extracting acceptable sets. All reasonable semantics are based on the notion of conflict-freeness which means that arguments are only jointly acceptable when they are not linked within the AF. In this paper, we study the question which information on top of conflict-free sets is needed to compute extensions of a semantics at hand. We introduce a hierarchy of verification classes specifying the required amount of information and show that well-known semantics are exactly verifiable through a certain such class. This also gives a means to study semantics lying between known semantics, thus contributing to a more abstract understanding of the different features argumentation semantics offer.

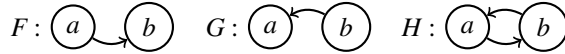
Keywords. abstract argumentation, argumentation semantics, verifiability, strong equivalence, intermediate semantics

1. Introduction

In the late 1980s the idea of using *argumentation* to model nonmonotonic reasoning emerged (see [1,2] as well as the survey [3]). Nowadays argumentation theory is a vibrant subfield of Artificial Intelligence, covering aspects of knowledge representation, multi-agent systems, and also philosophical questions. Among other approaches which have been proposed for capturing representative patterns of inference in argumentation theory [4], Dung’s abstract argumentation frameworks (AFs) [5] play an important role within this research area. At the heart of Dung’s approach lie *argumentation semantics* (cf. [6] for an excellent overview). Given an AF F , which is set-theoretically just a directed graph encoding arguments and attacks between them, a certain argumentation semantics σ returns acceptable sets of arguments $\sigma(F)$, so-called σ -extensions. Each of these sets represents a reasonable position w.r.t. F and σ .

Over the last 20 years a series of abstract argumentation semantics were introduced. The motivations of these semantics range from the desired treatment of specific examples to fulfilling a number of abstract principles. The comparison via abstract criteria of the different semantics available is a topic which emerged quite recently in the community ([7] can be seen as the first paper in this line). Our work takes a further step towards a comprehensive understanding of argumentation semantics. In particular, we study the following question: Do we really need the entire AF F to compute a certain argumentation semantics σ ? In other words, is it possible to unambiguously determine acceptable sets w.r.t. σ , given only partial information of the underlying framework F . In order to solve this problem let us start with the following reflections:

1. As a matter of fact, one basic requirement of almost all existing semantics (exemptions are [8,9,10]) is that of conflict-freeness¹, i.e. arguments within a reasonable position are not allowed to attack each other. Consequently, knowledge about conflict-free sets is an essential part for computing semantics.
2. The second step is to ask the following: Which information on top on conflict-free sets has to be added? Imagine the set of conflict-free sets given by $\{\emptyset, \{a\}, \{b\}\}$. Consequently, there has to be at least one attack between a and b . Unfortunately, this information is not sufficient to compute any standard semantics (except naive extensions, which are defined as \subseteq -maximal conflict-free sets) since we know nothing precise about the neighborhood of a and b . The following three AFs possess exactly the mentioned conflict-free sets, but differ with respect to other semantics.



3. The final step is to try to minimize the added information. In other words, which kind of knowledge about the neighborhood is somehow dispensable in the light of computation? Clearly, this will depend on the considered semantics. For instance, in case of stage semantics [12], which requests conflict-free sets of maximal range, we do not need any information about incoming attacks. This information can not be omitted in case of admissible-based semantics since incoming attacks require counterattacks.

The above considerations motivate the introduction of *verification classes* specifying a certain amount of information. In a first step, we study the relation of these classes to each other. We therefore introduce the notion of being *more informative*, capturing the intuition that a certain class can reproduce the information of another. We present a hierarchy w.r.t. this ordering, containing 15 different verification classes only. This is because many syntactically different classes collapse to the same amount of information.

We then formally define the essential property of a semantics σ being *verifiable* w.r.t. a certain verification class. We present a general theorem stating that any *rational* semantics is exactly verifiable w.r.t. one of the 15 different verification classes. Roughly speaking, a semantics is rational if attacks inbetween two self-loops can be omitted without affecting the set of extensions. An important aside hereby is that even the most informative class contains indeed less information than the entire framework by itself.

In this paper we consider a representative set of standard semantics. All of them satisfy rationality and thus, are exactly verifiable w.r.t. a certain class. Since the theorem does not provide an answer to which verification class perfectly matches a certain rational semantics we study this problem one by one for any considered semantics. As a result, only 6 different classes are essential to classify the considered standard semantics.

In the last part of the paper we study an application of the concept of verifiability. More precisely, we address the question of strong equivalence for semantics lying inbetween known semantics, called intermediate semantics in the following. Strong equivalence in nonmonotonic formalisms² is the natural counterpart to ordinary equivalence in monotonic logics. We provide characterization theorems relying on the notion of verifiability and thus, contributing to a more abstract understanding of the different features argumentation semantics offer. Besides these main results, we also give new characteri-

¹The alternative *labelling*-approach to argumentation semantics [11] does not explicitly exploit the notion of conflict-freeness; it still remains a basic property of all labelling semantics though.

²See [13,14] for abstract argumentation and [15,16,17,18] for other nonmonotonic theories.

zations for strong equivalence with respect to naive extensions and strongly admissible sets [7,19].

Due to limited space we have to refer to an extended version [20] for full proofs.

2. Preliminaries

An *argumentation framework* (AF) $F = (A, R)$ is a directed graph whose nodes $A \subseteq \mathcal{U}$ (with \mathcal{U} being an infinite set of arguments, the *universe*) are interpreted as *arguments* and whose edges $R \subseteq A \times A$ represent *conflicts* between them. We assume that all AFs possess finitely many arguments only and denote the collection of all AFs by \mathcal{A} . If $(a, b) \in R$ we say that a *attacks* b . An argument $a \in A$ is *defended* by a set $S \subseteq A$ if for each $b \in A$ with $(b, a) \in R$, $\exists c \in S$ s.t. $(c, b) \in R$. We define the *range* of S (in F) as $S_F^+ = S \cup \{a \mid \exists b \in S : (b, a) \in R\}$ and the *anti-range* of S (in F) as $S_F^- = S \cup \{a \mid \exists b \in S : (a, b) \in R\}$. A set S is *conflict-free* (in F) if there are no $a, b \in S$ with $(a, b) \in R$. The set of all conflict-free sets of F is denoted by $cf(F)$. For an AF $F = (B, S)$ we use $A(F)$ and $R(F)$ to refer to B and S , respectively. Finally, we introduce the union of AFs F and G as $F \cup G = (A(F) \cup A(G), R(F) \cup R(G))$.

A *semantics* σ assigns to each $F = (A, R)$ a set $\sigma(F) \subseteq 2^A$ where the elements are called σ -*extensions*. Numerous semantics are available. Each of them captures different intuitions about how to reason about conflicting knowledge. We consider $\sigma \in \{ad, na, stb, pr, co, gr, ss, stg, id, eg\}$ for admissible, naive, stable, preferred, complete, grounded, semi-stable, stage, ideal, and eager semantics [5,12,21,22,23].

Definition 1. Given an AF $F = (A, R)$ and let $S \subseteq A$.

1. $S \in ad(F)$ iff $S \in cf(F)$ and each $a \in S$ is defended by S ,
2. $S \in na(F)$ iff $S \in cf(F)$ and there is no $S' \in cf(F)$ s.t. $S \subsetneq S'$,
3. $S \in stb(F)$ iff $S \in cf(F)$ and $S_F^+ = A$,
4. $S \in pr(F)$ iff $S \in ad(F)$ and there is no $S' \in ad(F)$ s.t. $S \subsetneq S'$,
5. $S \in co(F)$ iff $S \in ad(F)$ and for any $a \in A$ defended by S , $a \in S$,
6. $S \in gr(F)$ iff $S \in co(F)$ and there is no $S' \in co(F)$ s.t. $S' \subsetneq S$,
7. $S \in ss(F)$ iff $S \in ad(F)$ and there is no $S' \in ad(F)$ s.t. $S_F^+ \subsetneq S_F'^+$,
8. $S \in stg(F)$ iff $S \in cf(F)$ and there is no $S' \in cf(F)$ s.t. $S_F^+ \subsetneq S_F'^+$,
9. $S \in id(F)$ iff $S \in ad(F)$, $S \subseteq \bigcap pr(F)$ and $\nexists S' \in ad(F)$ s.t. $S' \subseteq \bigcap pr(F) \wedge S \subsetneq S'$,
10. $S \in eg(F)$ iff $S \in ad(F)$, $S \subseteq \bigcap ss(F)$ and $\nexists S' \in ad(F)$ s.t. $S' \subseteq \bigcap ss(F) \wedge S \subsetneq S'$.

For two semantics σ, τ we use $\sigma \subseteq \tau$ to indicate that $\sigma(F) \subseteq \tau(F)$ for each AF $F \in \mathcal{A}$. If we have $\rho \subseteq \sigma$ and $\sigma \subseteq \tau$ for semantics ρ, σ, τ , we say that σ is ρ - τ -*intermediate*. Well-known relations between semantics are $stb \subseteq ss \subseteq pr \subseteq co \subseteq ad$, meaning, for instance, that ss is stb - pr -intermediate.

The role of self-attacking arguments is discussed quite controversially in the literature. If self-loops are allowed (and we do so to be as general as possible) we want to take the scepticism w.r.t. self-loops into account by calling a semantics rational if attacks between self-attacking arguments do not matter.

Definition 2. We call a semantics σ *rational* if for every AF F it holds that $\sigma(F) = \sigma(F^l)$, where $F^l = (A(F), R(F) \setminus \{(a, b) \in R(F) \mid (a, a), (b, b) \in R(F), a \neq b\})$.

Indeed, all semantics introduced in Definition 1 are rational. A prominent semantics that is based on conflict-free sets, but is not rational is the *cf2*-semantics [24], since here chains of self-loops can have an influence on the SCCs of an AF (see also [25]).

The main notions of equivalence available for non-monotonic formalisms are *ordinary* (or *standard*) *equivalence* and *strong* (or *expansion*) *equivalence*. A detailed overview of equivalence notions including their relations can be found in [26,27].

Definition 3. Given a semantics σ . Two AFs F and G are *standard equivalent* w.r.t. σ ($F \equiv^\sigma G$) iff $\sigma(F) = \sigma(G)$, and *expansion equivalent* w.r.t. σ ($F \equiv_E^\sigma G$) iff for each AF H : $F \cup H \equiv^\sigma G \cup H$.

Expansion equivalence can be decided syntactically via so-called *kernels* [13]. A kernel is a function $k : \mathcal{A} \mapsto \mathcal{A}$ mapping each AF F to another AF $k(F)$ (which we may also denote as F^k). Consider the following definitions.

Definition 4. Given an AF $F = (A, R)$ and a semantics σ . We define σ -kernels $F^{k(\sigma)} = (A, R^{k(\sigma)})$ whereby

- $R^{k(stb)} = R \setminus \{(a, b) \mid a \neq b, (a, a) \in R\}$,
- $R^{k(ad)} = R \setminus \{(a, b) \mid a \neq b, (a, a) \in R, \{(b, a), (b, b)\} \cap R \neq \emptyset\}$,
- $R^{k(gr)} = R \setminus \{(a, b) \mid a \neq b, (b, b) \in R, \{(a, a), (b, a)\} \cap R \neq \emptyset\}$,
- $R^{k(co)} = R \setminus \{(a, b) \mid a \neq b, (a, a), (b, b) \in R\}$.

A semantics σ is *compatible with a kernel* k if $F \equiv_E^\sigma G$ iff $F^k = G^k$. All semantics from Definition 1 (except *na*) are compatible with one of the kernels introduced above.

Theorem 1. [13,28] For any AFs F and G ,

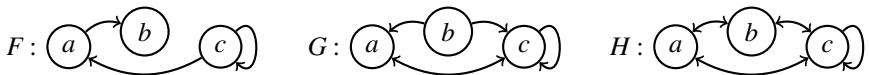
1. $F \equiv_E^\sigma G \Leftrightarrow F^{k(\sigma)} = G^{k(\sigma)}$ with $\sigma \in \{stb, ad, co, gr\}$,
2. $F \equiv_E^\tau G \Leftrightarrow F^{k(ad)} = G^{k(ad)}$ with $\tau \in \{pr, id, ss, eg\}$,
3. $F \equiv_E^{sg} G \Leftrightarrow F^{k(stb)} = G^{k(stb)}$.

3. Complementing Previous Results

In order to provide an exhaustive analysis of intermediate semantics (cf. Section 5) we provide missing kernels for naive semantics as well as strongly admissible sets. We start with the *naive kernel* characterizing expansion equivalence w.r.t. naive semantics. Note that the following kernel is the first one which adds attacks to the former attack relation.

Definition 5. Given an AF $F = (A, R)$. We define the *naive kernel* $F^{k(na)} = (A, R^{k(na)})$ whereby $R^{k(na)} = R \cup \{(a, b) \mid a \neq b, \{(a, a), (b, a), (b, b)\} \cap R \neq \emptyset\}$.

Example 1. Consider the AFs F and G . Note that $na(F) = na(G) = \{\{a\}, \{b\}\}$. Consequently, $F \equiv^{na} G$. In accordance with Definition 5 we observe that both AFs possess the same naive kernel $H = F^{k(na)} = G^{k(na)}$.



We can show that naive semantics is indeed compatible with this kernel.

Theorem 2. For all AFs F and G , it holds that $F \equiv_E^{na} G \Leftrightarrow F^{k(na)} = G^{k(na)}$.

We turn now to *strongly admissible sets* [7]. We will see that, besides grounded [13] and resolution-based grounded semantics [29,30], strongly admissible sets are compatible with the grounded kernel. Consider the following definition from [19].

Definition 6. Given an AF $F = (A, R)$. A set $S \subseteq A$ is *strongly admissible*, i.e. $S \in \text{sad}(F)$ iff any $a \in S$ is defended by a strongly admissible set $S' \subseteq S \setminus \{a\}$.

The following properties are needed to prove the characterization theorem. (1) and (2) are already shown in [7], (3) is an immediate consequence of the former.

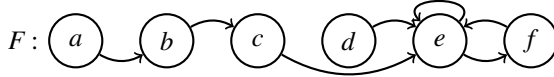
Proposition 3. Given two AFs F and G , it holds that

1. $\text{gr}(F) \subseteq \text{sad}(F) \subseteq \text{ad}(F)$,
2. if $S \in \text{gr}(F)$ we have: $S' \subseteq S$ for all $S' \in \text{sad}(F)$, and
3. $\text{sad}(F) = \text{sad}(G)$ implies $\text{gr}(F) = \text{gr}(G)$.

We now provide an alternative criterion for being a strongly admissible set. In contrast to the former it allows one to construct strongly admissible sets step by step. A proof that Definitions 6 and 7 are equivalent can be found in [20].

Definition 7. Given an AF $F = (A, R)$. A set $S \subseteq A$ is *strongly admissible*, i.e. $S \in \text{sad}(F)$ iff there are finitely many and pairwise disjoint sets A_1, \dots, A_n , s.t. $S = \bigcup_{1 \leq i \leq n} A_i$ and $A_1 \subseteq \Gamma_F(\emptyset)^3$ and furthermore, $\bigcup_{1 \leq i \leq j} A_i$ defends A_{j+1} for $1 \leq j \leq n-1$.

Example 2. Consider the following AF F .



We have $\Gamma_F(\emptyset) = \{a, d\}$. Hence, for all $S \subseteq \{a, d\}$, $S \in \text{sad}(F)$. Furthermore, $\Gamma_F(\{a\}) = \{a, c\}$, $\Gamma_F(\{d\}) = \{d, f\}$ and $\Gamma_F(\{a, d\}) = \{a, d, c, f\}$. This means, additionally $\{a, c\}, \{d, f\}, \{a, d, c\}, \{a, d, f\}, \{a, d, c, f\} \in \text{sad}(F)$. Finally, $\Gamma_F(\{a, c\}) = \{a, c, f\}$ justifying the last missing set $\{a, c, f\} \in \text{sad}(F)$.

The grounded kernel is insensitive w.r.t. strongly admissible sets, which then allows us to state the main result for strongly admissible sets.

Lemma 4. For any AF F , $\text{sad}(F) = \text{sad}(F^{k(\text{gr})})$.

Theorem 5. For any two AFs F and G , we have $F \equiv_E^{\text{sad}} G \Leftrightarrow F^{k(\text{gr})} = G^{k(\text{gr})}$.

Proof. (\Rightarrow) We show the contrapositive, i.e. $F^{k(\text{gr})} \neq G^{k(\text{gr})} \Rightarrow F \not\equiv_E^{\text{sad}} G$. Assuming $F^{k(\text{gr})} \neq G^{k(\text{gr})}$ implies $F \not\equiv_E^{\text{gr}} G$ (cf. Theorem 1). This means, there is an AF H , s.t. $\text{gr}(F \cup H) \neq \text{gr}(G \cup H)$. Due to statement 3 of Proposition 3, we deduce $\text{sad}(F \cup H) \neq \text{sad}(G \cup H)$ proving $F \not\equiv_E^{\text{sad}} G$. (\Leftarrow) Given $F^{k(\text{gr})} = G^{k(\text{gr})}$. Since expansion equivalence is a congruence w.r.t. \cup we obtain $(F \cup H)^{k(\text{gr})} = (G \cup H)^{k(\text{gr})}$ for any AF H . Consequently, $\text{sad}((F \cup H)^{k(\text{gr})}) = \text{sad}((G \cup H)^{k(\text{gr})})$. Due to Lemma 4 we deduce $\text{sad}(F \cup H) = \text{sad}(G \cup H)$, concluding the proof. \square

³Hereby, Γ is the so-called *characteristic function* [5] with $\Gamma_F(S) = \{a \in A \mid a \text{ is defended by } S \text{ in } F\}$. The term $\Gamma_F(\emptyset)$ can be equivalently replaced by $\{a \in A \mid a \text{ is unattacked}\}$.

4. Verifiability

In this section we study the question whether we really need the entire AF F to compute the extensions of a given semantics. Consider naive semantics. Obviously, in order to determine naive extensions it suffices to know all conflict-free sets. Conversely, knowing $cf(F)$ only does not allow to reconstruct F unambiguously. This means, knowledge about $cf(F)$ is indeed less information than the entire AF by itself. In fact, most of the existing semantics do not need information about the entire AF. We will categorize the amount of information by taking the conflict-free sets as a basis and distinguish between different amounts of knowledge about the neighborhood (range and anti-range) of these sets.

Definition 8. We call a function $\tau^x : 2^{\mathcal{U}} \times 2^{\mathcal{U}} \rightarrow (2^{\mathcal{U}})^n$ ($n > 0$), which is expressible via basic set operations only, *neighborhood function*. A neighborhood function τ^x induces the *verification class* mapping each AF F to $\tilde{F}^x = \{(S, \tau^x(S_F^+, S_F^-)) \mid S \in cf(F)\}$.

We coined the term neighborhood function because the induced verification classes apply these functions to the neighborhoods, i.e. range and anti-range of conflict-free sets. The notion of *expressible via basic set operations* simply means that (in case of $n = 1$) the expression $\tau^x(A, B)$ is in the language generated by the BNF $X ::= A \mid B \mid (X \cup X) \mid (X \cap X) \mid (X \setminus X)$. Consequently, in case of $n = 1$, we may distinguish eight set theoretically different neighborhood functions, namely

$$\begin{aligned} \tau^{\varepsilon}(S, S') &= \emptyset & \tau^{+}(S, S') &= S & \tau^{-}(S, S') &= S' & \tau^{\bar{+}}(S, S') &= S' \setminus S \\ \tau^{\pm}(S, S') &= S \setminus S' & \tau^{\cap}(S, S') &= S \cap S' & \tau^{\cup}(S, S') &= S \cup S' & \tau^{\Delta}(S, S') &= (S \cup S') \setminus (S \cap S') \end{aligned}$$

The names of the neighborhood functions are inspired by their usage in the verification classes they induce (cf. Definition 8). A verification class encapsulates a certain amount of information about an AF, as the following example illustrates.

Example 3. Consider the AF $F = (\{a, b, c\}, \{(a, b), (b, a), (b, b), (c, b)\})$. Now take, for instance, the verification class induced by τ^{+} , that is $\tilde{F}^{+} = \{(S, \tau^{+}(S_F^+, S_F^-)) \mid S \in cf(F)\} = \{(S, S_F^+) \mid S \in cf(F)\}$, storing information about conflict-free sets together with their associated ranges w.r.t. F . It contains the following tuples: (\emptyset, \emptyset) , $(\{a\}, \{a, b\})$, $(\{c\}, \{b, c\})$, and $(\{a, c\}, \{a, b, c\})$. For the verification class induced by τ^{\pm} , on the other hand, we have $\tilde{F}^{\pm} = \{(\emptyset, \emptyset), (\{a\}, \emptyset), (\{c\}, \{b\}), (\{a, c\}, \emptyset)\}$.

Intuitively, it should be clear that the set \tilde{F}^{+} suffices to compute stage extensions (i.e., range-maximal conflict-free sets) of F . This intuitive understanding of *verifiability* will be formally specified in Definition 10. Note that a neighborhood function τ^x may return n -tuples. Consequently, in consideration of the eight basic functions we obtain (modulo reordering, duplicates, empty set) $2^7 + 1$ syntactically different neighborhood functions and therefore the same number of verification classes. As usual, we denote the n -ary combination of basic functions $(\tau^{x_1}(S, S'), \dots, \tau^{x_n}(S, S'))$ as $\tau^x(S, S')$ by $x = x_1 \dots x_n$.

With the following definition we can put neighborhood functions into relation w.r.t. their information. This will help us to show that actually many of the induced classes collapse to the same amount of information.

Definition 9. Given neighborhood functions τ^x and τ^y returning n -tuples and m -tuples, respectively, we say that τ^x is *more informative* than τ^y , for short $\tau^x \succeq \tau^y$, iff there is a

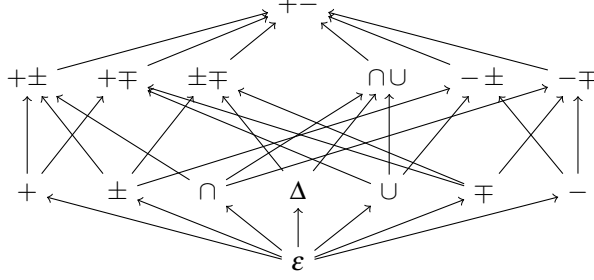


Figure 1. Representatives of neighborhood functions and their relation w.r.t. information; a node x stands for the neighborhood function τ^x ; an arrow from x to y means $\tau^x \prec \tau^y$.

function $\delta : (2^{\mathcal{U}})^n \rightarrow (2^{\mathcal{U}})^m$ such that for any two sets of arguments $S, S' \subseteq \mathcal{U}$, we have $\delta(\tau^x(S, S')) = \tau^y(S, S')$. We denote the strict part of \succeq by \succ , i.e. $\tau^x \succ \tau^y$ iff $\tau^x \succeq \tau^y$ and $\tau^y \not\succeq \tau^x$. Finally, $\tau^x \approx \tau^y$ (τ^x represents τ^y and vice versa) in case $\tau^x \succeq \tau^y$ and $\tau^y \succeq \tau^x$.

It turns out that many neighborhood functions yield the same amount of information. In particular, τ^{+-} represents all τ^{x_1, \dots, x_n} with $n > 2$.

Lemma 6. *All neighborhood functions are represented by the ones depicted in Figure 1 and the \prec -relation represented by arcs in Figure 1 holds.*

If the information provided by a neighborhood function is sufficient to compute the extensions under a semantics, we say that the semantics is verifiable by the class induced by the neighborhood function.

Definition 10. A semantics σ is *verifiable* by the verification class induced by the neighborhood function τ^x returning n -tuples (or simply, *x-verifiable*) iff there is a function (also called *criterion*) $\gamma_\sigma : (2^{\mathcal{U}})^n \times 2^{\mathcal{U}} \rightarrow 2^{2^{\mathcal{U}}}$ s.t. for every AF $F \in \mathcal{A}$ we have: $\gamma_\sigma(\tilde{F}^x, A(F)) = \sigma(F)$. Moreover, σ is *exactly x-verifiable* iff σ is x -verifiable and there is no verification class induced by τ^y with $\tau^y \prec \tau^x$ such that σ is y -verifiable.

We proceed with a list of criteria showing that any semantics mentioned in Definition 1 is verifiable by a verification class induced by a certain neighborhood function. In the following, we abbreviate the tuple $(\tilde{F}^x, A(F))$ by \tilde{F}_A^x .

$$\begin{aligned}
 \gamma_{na}(\tilde{F}_A^\varepsilon) &= \{S \mid S \in \tilde{F}, S \text{ is } \subseteq\text{-maximal in } \tilde{F}\}; \\
 \gamma_{stg}(\tilde{F}_A^{++}) &= \{S \mid (S, S^+) \in \tilde{F}^+, S^+ \text{ is } \subseteq\text{-maximal in } \{C^+ \mid (C, C^+) \in \tilde{F}^+\}\}; \\
 \gamma_{stb}(\tilde{F}_A^{++}) &= \{S \mid (S, S^+) \in \tilde{F}^+, S^+ = A\}; \\
 \gamma_{ad}(\tilde{F}_A^{+\mp}) &= \{S \mid (S, S^\mp) \in \tilde{F}^\mp, S^\mp = \emptyset\}; \\
 \gamma_{pr}(\tilde{F}_A^{+\mp}) &= \{S \mid S \in \gamma_{ad}(\tilde{F}_A^{+\mp}), S \text{ is } \subseteq\text{-maximal in } \gamma_{ad}(\tilde{F}_A^{+\mp})\}; \\
 \gamma_{ss}(\tilde{F}_A^{+\mp}) &= \{S \mid S \in \gamma_{ad}(\tilde{F}_A^{+\mp}), S^+ \text{ is } \subseteq\text{-maximal in } \{C^+ \mid (C, C^+, C^\mp) \in \tilde{F}^{+\mp}, C \in \gamma_{ad}(\tilde{F}_A^{+\mp})\}\}; \\
 \gamma_{id}(\tilde{F}_A^{+\mp}) &= \{S \mid S \text{ is } \subseteq\text{-maximal in } \{C \mid C \in \gamma_{ad}(\tilde{F}_A^{+\mp}), C \subseteq \bigcap \gamma_{pr}(\tilde{F}_A^{+\mp})\}\}; \\
 \gamma_{eg}(\tilde{F}_A^{+\mp}) &= \{S \mid S \text{ is } \subseteq\text{-maximal in } \{C \mid C \in \gamma_{ad}(\tilde{F}_A^{+\mp}), C \subseteq \bigcap \gamma_{ss}(\tilde{F}_A^{+\mp})\}\};
 \end{aligned}$$

$$\begin{aligned}
\gamma_{sad}(\tilde{F}_A^{\pm}) &= \{S \mid (S, S^-, S^\pm) \in \tilde{F}^{\pm}, \exists (S_0, S_0^-, S_0^\pm), \dots, (S_n, S_n^-, S_n^\pm) \in \tilde{F}^{\pm} : \\
&\quad (\emptyset = S_0 \subset \dots \subset S_n = S \wedge \forall i \in \{1, \dots, n\} : S_i^- \subseteq S_{i-1}^\pm)\}; \\
\gamma_{gr}(\tilde{F}_A^{\pm}) &= \{S \mid S \in \gamma_{sad}(\tilde{F}_A^{\pm}), \forall (\bar{S}, \bar{S}^-, \bar{S}^\pm) \in \tilde{F}^{\pm} : \bar{S} \supset S \Rightarrow (\bar{S}^- \setminus S^\pm) \neq \emptyset\}; \\
\gamma_{co}(\tilde{F}_A^{\pm}) &= \{S \mid (S, S^+, S^-) \in \tilde{F}^{+-}, (S^- \setminus S^+) = \emptyset, \forall (\bar{S}, \bar{S}^+, \bar{S}^-) \in \tilde{F}^{+-} : \bar{S} \supset S \Rightarrow (\bar{S}^- \setminus S^+) \neq \emptyset\}.
\end{aligned}$$

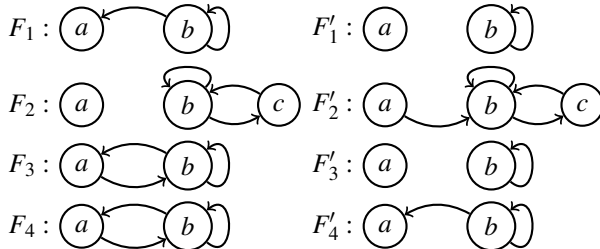
It is easy to see that the naive semantics is verifiable by the verification class induced by τ^e since the naive extensions can be determined by the conflict-free sets. Stable and stage semantics, on the other hand, utilize the range of each conflict-free set in addition. Hence they are verifiable by the verification class induced by τ^+ . Now consider admissible sets. Recall that a conflict-free S set is admissible if and only if it attacks all attackers. This is captured exactly by the condition $S^\mp = \emptyset$, hence admissible sets are verifiable by the verification class induced by τ^\mp . The same holds for preferred semantics, since we just have to determine the maximal conflict-free sets with $S^\mp = \emptyset$. Semi-stable semantics, however, needs the range of each conflict-free set in addition, see γ_{ss} , which makes it verifiable by the verification class induced by $\tau^{+\mp}$. Finally consider the criterion γ_{co} . The first two conditions for a set of arguments S stand for conflict-freeness and admissibility, respectively. Now assume the third condition does not hold, i.e., there exists a tuple $(\bar{S}, \bar{S}^+, \bar{S}^-) \in \tilde{F}^{+-}$ with $\bar{S} \supset S$ and $\bar{S}^- \setminus S^+ = \emptyset$. This means that every argument attacking \bar{S} is attacked by S , i.e., \bar{S} is defended by S . Hence S is not a complete extension, showing that $\gamma_{co}(\tilde{F}_A^{+-}) = co(F)$ for each $F \in \mathcal{A}$. One can verify that all criteria from the list are adequate in the sense that they describe the extensions of the corresponding semantics.

The concepts of verifiability and being more informative behave correctly insofar as more informative neighborhood functions do not lead to a loss of verification capacity.

Proposition 7. *If a semantics σ is x -verifiable, then σ is verifiable by all verification classes induced by some τ^y with $\tau^y \succeq \tau^x$.*

In order to prove unverifiability of a semantics σ w.r.t. a class induced by a certain τ^x it suffices to present two AFs F and G such that $\sigma(F) \neq \sigma(G)$ but, $\tilde{F}^x = \tilde{G}^x$ and $A(F) = A(G)$. Then the verification class induced by τ^x does not provide enough information to verify σ . In the following we will use this strategy to show exact verifiability. Consider a semantics σ which is verifiable by a class induced by τ^x . If σ is unverifiable by all verifiability classes induced by τ^y with $\tau^y \prec \tau^x$ we have that σ is exactly verifiable by τ^x . The following examples study this issue for the semantics under consideration.

Example 4. The complete semantics is $+-$ -verifiable as seen before. The following AFs show that it is even exactly verifiable by that class.



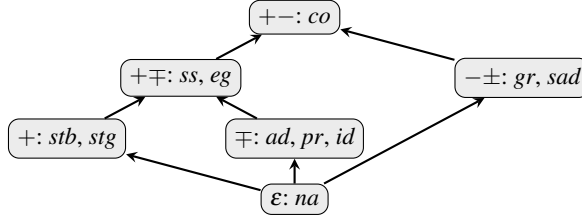
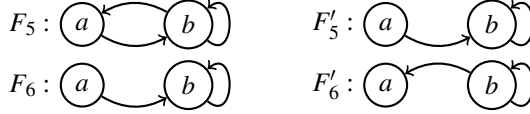


Figure 2. Semantics and their exact verification classes.



First consider the AFs F_1 and F'_1 , and observe that $\widetilde{F}_1^{+\pm} = \{(\emptyset, \emptyset, \emptyset), (\{a\}, \emptyset, \emptyset)\} = \widetilde{F}'_1^{+\pm}$. On the other hand F_1 and F'_1 differ in their complete extensions since $co(F_1) = \{\emptyset\}$ but $co(F'_1) = \{\{a\}\}$. Therefore complete semantics is unverifiable by the verification class induced by $\tau^{+\pm}$. Likewise, this can be shown for the classes induced by $\tau^{-\mp}$, $\tau^{\pm\mp}$, $\tau^{-\pm}$, $\tau^{+\mp}$, and τ^{\cup} , respectively:

- $\widetilde{F}_2^{-\mp} = \widetilde{F}'_2^{-\mp}$, but $co(F_2) = \{\{a\}, \{a, c\}\} \neq \{\{a, c\}\} = co(F'_2)$.
- $\widetilde{F}_3^{\pm\mp} = \widetilde{F}'_3^{\pm\mp}$, but $co(F_3) = \{\emptyset, \{a\}\} \neq \{\{a\}\} = co(F'_3)$.
- $\widetilde{F}_4^{-\pm} = \widetilde{F}'_4^{-\pm}$, but $co(F_4) = \{\emptyset, \{a\}\} \neq \{\emptyset\} = co(F'_4)$.
- $\widetilde{F}_5^{+\mp} = \widetilde{F}'_5^{+\mp}$, but $co(F_5) = \{\emptyset, \{a\}\} \neq \{\{a\}\} = co(F'_5)$.
- $\widetilde{F}_6^{\cup} = \widetilde{F}'_6^{\cup}$, but $co(F_6) = \{\{a\}\} \neq \{\emptyset\} = co(F'_6)$.

Hence complete semantics is exactly verifiable by the verification class induced by τ^{+-} .

Examples showing exact verifiability of the other semantics can be found in [20]. Figure 2 shows the resulting relation between the semantics under consideration with respect to their exact verification classes.

Theorem 8. *Every semantics which is rational is exactly verifiable by a verification class induced by one of the neighborhood functions presented in Figure 1.*

Proof. First of all note that by Lemma 6, τ^{ε} is the least informative neighborhood function and for every other neighborhood function τ^x it holds that $\tau^{\varepsilon} \preceq \tau^x$. Thus, if a semantics is verifiable by the class induced by any τ^x then it is exactly verifiable by a verification class induced by some τ^y with $\tau^{\varepsilon} \preceq \tau^y \preceq \tau^x$. Moreover, if a semantics is exactly verifiable by a class, then it is by definition also verifiable by this class. Hence it remains to show that every rational semantics is verifiable by a verification class of Figure 1.

We show the contrapositive. To this end, assume a semantics σ is not verifiable by one of the verification classes. This means σ is not verifiable by the verification class induced by τ^{+-} . Hence there exist two AFs F and G such that $\widetilde{F}^{+-} = \widetilde{G}^{+-}$ and $A(F) = A(G)$, but $\sigma(F) \neq \sigma(G)$. For every argument a which is not self-attacking, a tuple $(\{a\}, \{a\}^+, \{a\}^-)$ is contained in \widetilde{F}^{+-} (and in \widetilde{G}^{+-}). Hence F and G have the same not-self-attacking arguments, and moreover, these arguments have the same incoming and outgoing attacks in F and G . This, together with $A(F) = A(G)$ implies that $F^l = G^l$ (see Definition 2) holds. But since $\sigma(F) \neq \sigma(G)$ we get that σ is not rational. \square

Note that the criterion giving evidence for verifiability of a semantics by a certain class has access to the set of arguments of a given AF. In fact, only the criterion for stable semantics makes use of that – it can be omitted for the other semantics.

5. Intermediate Semantics

A type of semantics which has aroused quite some interest in the literature (see e.g. [31] and [32]) are intermediate semantics, i.e. semantics which yield results lying between two existing semantics. The introduction of σ - τ -intermediate semantics can be motivated by deleting *undesired* (or add *desired*) τ -extensions while guaranteeing all reasonable positions w.r.t. σ . In other words, σ - τ -intermediate semantics can be seen as sceptical or credulous acceptance shifts within the range of σ and τ .

A natural question is whether we can make any statements about compatible kernels of intermediate semantics. In particular, if semantics σ and τ are compatible with some kernel k , is then every σ - τ -intermediate semantics k -compatible? The following example answers this question negatively.

Example 5. Recall from Theorem 1 that both stable and stage semantics are compatible with $k(stb)$, i.e. $F \equiv_E^{stb} G \Leftrightarrow F \equiv_E^{stg} G \Leftrightarrow F^{k(stb)} = G^{k(stb)}$. Now we define the following stb - stg -intermediate semantics, say *stagle* semantics: Given an AF $F = (A, R)$, $S \in sta(F)$ iff $S \in cf(F)$, $S_F^+ \cup S_F^- = A$ and for every $T \in cf(F)$ we have $S_F^+ \not\subseteq T_F^+$. Obviously, it holds that $stb \subseteq sta \subseteq stg$ and $stb \neq sta$ as well as $sta \neq stg$, as witnessed by the AF F :



It is easy to verify that $stb(F) = \emptyset \subset sta(F) = \{\{b\}\} \subset stg(F) = \{\{b\}, \{c\}\}$. We proceed by showing that stagle semantics is not compatible with $k(stb)$. To this end consider $F^{k(stb)}$. Now, $sta(F^{k(stb)}) = \{\{b\}, \{c\}\}$ witnesses $F \not\equiv_E^{sta} F^{k(stb)}$ and therefore, $F \not\equiv_E^{stb} F^{k(stb)}$. Since $F^{k(stb)} = (F^{k(stb)})^{k(stb)}$ we are done, i.e. stagle semantics is indeed not compatible with the stable kernel.

It is the main result of this section that compatibility of intermediate semantics w.r.t. a certain kernel can be guaranteed if verifiability w.r.t. a certain class is presumed. The provided characterization theorems generalize former results presented in [13]. Moreover, due to the abstract character of the theorems the results are applicable to semantics which may be defined in the future.

Before turning to the characterization theorems we state some implications of verifiability. In particular, under the assumption that σ is verifiable by a certain class, equality of certain kernels implies expansion equivalence w.r.t. σ .

Proposition 9. *For a semantics σ it holds that*

- if σ is $+$ -verifiable then $F^{k(stb)} = G^{k(stb)} \Rightarrow F \equiv_E^\sigma G$.
- if σ is $+\mp$ -verifiable then $F^{k(ad)} = G^{k(ad)} \Rightarrow F \equiv_E^\sigma G$.
- if σ is $+-$ -verifiable then $F^{k(co)} = G^{k(co)} \Rightarrow F \equiv_E^\sigma G$.
- if σ is $-\pm$ -verifiable then $F^{k(gr)} = G^{k(gr)} \Rightarrow F \equiv_E^\sigma G$.
- if σ is ε -verifiable then $F^{k(na)} = G^{k(na)} \Rightarrow F \equiv_E^\sigma G$.

We proceed with general characterization theorems. The first one states that *stb-stg*-intermediate semantics are compatible with stable kernel if $+$ -verifiability is given.

Theorem 10. *Given a semantics σ which is $+$ -verifiable and *stb-stg*-intermediate, it holds that $F^{k(stb)} = G^{k(stb)} \Leftrightarrow F \equiv_E^\sigma G$.*

Proof. (\Rightarrow) Follows directly from Proposition 9. (\Leftarrow) We show the contrapositive, i.e. $F^{k(stb)} \neq G^{k(stb)} \Rightarrow F \not\equiv_E^\sigma G$. Assuming $F^{k(stb)} \neq G^{k(stb)}$ implies $F \not\equiv_E^{stg} G$, i.e. there exists an AF H such that $stg(F \cup H) \neq stg(G \cup H)$ and therefore, $stb(F \cup H) \neq stb(G \cup H)$. Let $B = A(F) \cup A(G) \cup A(H)$ and $H' = (B \cup \{a\}, \{(a, b), (b, a) \mid b \in B\})$. It is easy to see that $stb(F \cup H') = stb(F \cup H) \cup \{\{a\}\}$ and $stb(G \cup H') = stb(G \cup H) \cup \{\{a\}\}$. Since now both $stb(F \cup H') \neq \emptyset$ and $stb(G \cup H') \neq \emptyset$ it holds that $stb(F \cup H') = stg(F \cup H')$ and $stb(G \cup H') = stg(G \cup H')$. Hence $\sigma(F \cup H') \neq \sigma(G \cup H')$, showing that $F \not\equiv_E^{stb} G$. \square

The following theorems can be shown in a similar manner.

Theorem 11. *Given a semantics σ which is $-\pm$ -verifiable and *gr-sad*-intermediate, it holds that $F^{k(gr)} = G^{k(gr)} \Leftrightarrow F \equiv_E^\sigma G$.*

Theorem 12. *Given a semantics σ which is $+\mp$ -verifiable and ρ -ad-intermediate for any $\rho \in \{ss, id, eg\}$, it holds that $F^{k(ad)} = G^{k(ad)} \Leftrightarrow F \equiv_E^\sigma G$.*

Recall that complete semantics is a *ss-ad*-intermediate semantics. Furthermore, it is not compatible with the admissible kernel as already observed in [13]. Consequently, it is not $+\mp$ -verifiable (as we have shown in Example 4 with considerable effort).

6. Conclusions

In this work we have contributed to the analysis and comparison of abstract argumentation semantics. The main idea of our approach is to provide a novel categorization in terms of the amount of information required for testing whether a set of arguments is an extension of a certain semantics. The resulting notion of verification classes allows us to categorize any new semantics (given it is “rational”) with respect to the information needed and compare it to other semantics. Thus our work is in the tradition of the principle-based evaluation due to Baroni and Giacomin [7] and paves the way for a more general view on semantics, their common features, and their inherent differences.

Using our notion of verifiability, we were able to show kernel-compatibility for certain intermediate semantics. Concerning concrete semantics, our results yield the following observation: While preferred, semi-stable, ideal and eager semantics coincide w.r.t. strong equivalence, verifiability of these semantics differs. In fact, preferred and ideal semantics manage to be verifiable with strictly less information.

For future work we envisage extending the notion of verifiability classes in order to categorize semantics not captured by the approach followed in this paper, such as *cf2* [24] as well as labelling semantics [11]. Moreover, we want to study the link between containment in verification classes and the fulfillment of certain principles of [7].

Acknowledgements. This research has been supported by DFG (project BR 1817/7-1) and FWF (projects I1102, I2854 and P25521).

References

- [1] Loui, R.P.: Defeat among arguments: a system of defeasible inference. *Computational Intelligence* **14** (1987) 100–106
- [2] Pollock, J.L.: Defeasible reasoning. *Cognitive Science* **11** (1987) 481–518
- [3] Prakken, H., Vreeswijk, G.: Logics for defeasible argumentation. In: *Handbook of Philosophical Logic*. Dordrecht (2002) 219–318
- [4] Besnard, P., Garcia, A., Hunter, A., Modgil, S., Prakken, H., Simari, G., Toni, F.: Special issue: Tutorials on structured argumentation. *Argument and Computation* **5** (2014) 1–117
- [5] Dung, P.M.: On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artif. Intell.* **77** (1995) 321–357
- [6] Baroni, P., Caminada, M., Giacomin, M.: An introduction to argumentation semantics. *Knowledge Eng. Review* **26** (2011) 365–410
- [7] Baroni, P., Giacomin, M.: On principle-based evaluation of extension-based argumentation semantics. *Artif. Intell.* **171** (2007) 675–700
- [8] Jakobovits, H., Vermeir, D.: Robust semantics for argumentation frameworks. *JLC* **9** (1999) 215–261
- [9] Arieli, O.: Conflict-tolerant semantics for argumentation frameworks. In: *Proc. JELIA*. (2012) 28–40
- [10] Grossi, D., Modgil, S.: On the graded acceptability of arguments. In: *Proc. IJCAI*. (2015) 868–874
- [11] Caminada, M., Gabbay, D.M.: A logical account of formal argumentation. *Studia Logica* **93** (2009) 109–145
- [12] Verheij, B.: Two approaches to dialectical argumentation: admissible sets and argumentation stages. In: *Proc. NAIC*. (1996) 357–368
- [13] Oikarinen, E., Woltran, S.: Characterizing strong equivalence for argumentation frameworks. *Artif. Intell.* **175** (2011) 1985–2009
- [14] Baumann, R.: Characterizing equivalence notions for labelling-based semantics. In: *Proc. KR*. (2016) 22–32
- [15] Maher, M.J.: Equivalences of logic programs. In: *Proc. ICLP*. (1986) 410–424
- [16] Lifschitz, V., Pearce, D., Valverde, A.: Strongly equivalent logic programs. *ACM Trans. Comput. Log.* **2** (2001) 526–541
- [17] Turner, H.: Strong equivalence for causal theories. In: *Proc. LPNMR*. (2004) 289–301
- [18] Truszczyński, M.: Strong and uniform equivalence of nonmonotonic theories - an algebraic approach. *Ann. Math. Artif. Intell.* **48** (2006) 245–265
- [19] Caminada, M.: Strong admissibility revisited. In: *Proc. COMMA*. (2014) 197–208
- [20] Baumann, R., Linsbichler, T., Woltran, S.: Verifiability of argumentation semantics. In: *Proc. NMR*. (2016) Available at <http://arxiv.org/abs/1603.09502>.
- [21] Caminada, M., Carnielli, W.A., Dunne, P.E.: Semi-stable semantics. *JLC* **22** (2012) 1207–1254
- [22] Dung, P.M., Mancarella, P., Toni, F.: Computing ideal sceptical argumentation. *Artif. Intell.* **171** (2007) 642–674
- [23] Caminada, M.: Comparing two unique extension semantics for formal argumentation: Ideal and eager. In: *Proc. BNAIC*. (2007) 81–87
- [24] Baroni, P., Giacomin, M., Guida, G.: SCC-Recursiveness: A general schema for argumentation semantics. *Artif. Intell.* **168** (2005) 162–210
- [25] Gaggl, S.A., Woltran, S.: The cf2 argumentation semantics revisited. *JLC* **23** (2013) 925–949
- [26] Baumann, R., Brewka, G.: Analyzing the equivalence zoo in abstract argumentation. In: *Proc. CLIMA*. (2013) 18–33
- [27] Baumann, R., Brewka, G.: The equivalence zoo for Dung-style semantics. *JLC* (2015)
- [28] Baumann, R., Woltran, S.: The role of self-attacking arguments in characterizations of equivalence notions. *JLC: Special Issue on Loops in Argumentation* (2014)
- [29] Baroni, P., Dunne, P.E., Giacomin, M.: On the resolution-based family of abstract argumentation semantics and its grounded instance. *Artif. Intell.* **175** (2011) 791–813
- [30] Dvořák, W., Linsbichler, T., Oikarinen, E., Woltran, S.: Resolution-based grounded semantics revisited. In: *Proc. COMMA*. (2014) 269–280
- [31] Baroni, P., Giacomin, M.: Comparing argumentation semantics with respect to skepticism. In: *Proc. ECSQARU*. (2007) 210–221
- [32] Nieves, J.C., Osorio, M., Zepeda, C.: A schema for generating relevant logic programming semantics and its applications in argumentation theory. *Fundam. Inform.* **106** (2011) 295–319

From Arguments to Constraints on a Bayesian Network

Floris BEX^a, Silja RENOIJ^a

^a*Information and Computing Sciences, Utrecht University, The Netherlands*

Abstract. In this paper, we propose a way to derive constraints for a Bayesian Network from structured arguments. Argumentation and Bayesian networks can both be considered decision support techniques, but are typically used by experts with different backgrounds. Bayesian network experts have the mathematical skills to understand and construct such networks, but lack expertise in the application domain; domain experts may feel more comfortable with argumentation approaches. Our proposed method allows us to check Bayesian networks given arguments constructed for the same problem, and also allows for transforming arguments into a Bayesian network structure, thereby facilitating Bayesian network construction.

1. Introduction

Bayesian networks, graphical representations of probability distributions, are very well suited to epistemic reasoning because they capture the probabilistic (in)dependencies between variables in the domain of discourse. They have found a number of applications in domains such as medicine, forensics and the law [1]. However, constructing a Bayesian network and understanding the modelled influences between variables requires knowledge of the Bayesian network formalism, which means that domain experts (doctors, lawyers) can often only construct a network with the help of a modeller with the relevant mathematical background. In contrast, argumentation approaches can be said to more closely follow the reasoning of the domain experts, especially in legal or organizational contexts which are less mathematically inclined [2, 3]. Our aim is to bridge the communication gap between domain experts and Bayesian network analysts by developing a better understanding of the relation between the two kinds of reasoning.

The relations between arguments and Bayesian networks (BNs) can be considered from two directions. For the first direction, BNs are transformed into arguments or argument diagrams [4–6], which allows the knowledge captured in the BN to be understood more easily by domain experts accustomed to argumentative reasoning. This does not directly help domain experts to construct a BN based on argumentative reasoning, for which we need to transform in the other direction, that is, from arguments to BNs. A classic example is [3], who use Wigmore graphs, which are very similar to evidential arguments [7], as the basis for BNs. Schum and Kadane, however, do not provide a formal definition of their transformation. Such a definition is given in [8], in which Carneades argument evaluation structures are transformed into BNs, thus allowing existing BNs to be extended with the information contained in the arguments. The main focus in [8],

however, is on simulating the Carneades method of evaluating arguments through BNs. In contrast, the main aim of this article is to explore how domain knowledge expressed as argument structures can inform the construction of BNs.

One of the difficulties of interpreting generic argument structures as a BN is that theoretical assumptions have to be made about, for example, the causal direction [9] and strength of the inference rules [10, 11]. From a practical point of view, it is also an issue that different BNs can represent the same reasoning. Consider for example a case where a forensic specialist constructs a BN for a part of the case, while a judge constructs an argument about the same part of the case. If we directly translate this argument, we might not end up with the exact same network as the forensic specialist, but this does not mean that the judge and the forensic specialist disagree, as different BNs can represent the same probability distribution. Because a set of evidential arguments therefore do not – and cannot – uniquely define a BN about the same part of the case, it makes sense to transform the arguments to a set of constraints that can be met by multiple possible BNs.

We hence distinguish two cases for transforming arguments to (constraints on) BNs. In the first case, a network already exists – hand-crafted by experts or learned from data – for the same case as the arguments and we want to check if the BN complies with the arguments. Here, the arguments indicate a number of basic *constraints* a BN has to adhere to. In the second case, we use the identified constraints to construct a new BN on the basis of the available arguments. Because the constraints are not exhaustive (i.e., given a set of constraints based on an argument there will be multiple BNs that adhere to these constraints), we propose a general *heuristic* for transforming arguments into BNs.

Sections 2 and 3 briefly discuss the formal preliminaries: a structured argumentation framework and Bayesian networks. In Section 4.1, we then discuss possible constraints on the graph of a BN given structured arguments, and in Section 4.2 we propose a heuristic for transforming arguments to a BN. Note that initially, the focus is on constraints on the *structure* (i.e. the graph) of the Bayesian network. Deriving constraints on the (conditional) probabilities of a network from structured arguments is more difficult, because it depends on probabilistic interpretations of argumentation which are more contentious. In Section 5 we briefly discuss some possible constraints on the (conditional) probabilities in the Bayesian network, and how they can be used to expand our heuristic.

2. Structured Argumentation

In this section, we define a simple propositional system for argumentation based on the ASPIC⁺ framework [12], which captures the basic elements of structured argumentation. Because the idea is that domain experts have to be able to relatively easily construct arguments, we want to impose as few formal constraints as possible on these arguments. As has been shown [7], ASPIC⁺ allows for arguments that are very similar to the Wigmore graphs [3] with which many lawyers and judges are familiar.

Definition 2.1 [Argumentation systems] An *argumentation system* is a triple $AS = (\mathcal{L}, \mathcal{R}, n)$ where:

- \mathcal{L} is a logical language with contrariness function $\neg : \mathcal{L} \rightarrow 2^{\mathcal{L}}$.
- \mathcal{R}_s and \mathcal{R}_d are two disjoint sets of strict (\mathcal{R}_s) and defeasible (\mathcal{R}_d) inference rules of the form $\varphi_1, \dots, \varphi_n \rightarrow \varphi$ and $\varphi_1, \dots, \varphi_n \Rightarrow \varphi$ respectively (where φ_i, φ are meta-variables ranging over well-formed formulas in \mathcal{L}).

- n is a naming convention for defeasible rules, which to each rule r in \mathcal{R}_d assigns a well-formed formula φ from \mathcal{L} (written as $n(r) = \varphi$).

Here, \mathcal{L} is a propositional language, where φ is a *contrary* of ψ if $\varphi \in \overline{\psi}$ and $\psi \notin \overline{\varphi}$ (i.e. asymmetrical conflict) and φ is a *contradictory* of ψ , denoted by ' $\varphi = -\psi$ ', if $\varphi \in \overline{\psi}$ and $\psi \in \overline{\varphi}$ (i.e. symmetrical conflict). Furthermore, \mathcal{R}_s contains the inference rules of classical logic. Whereas in evidential reasoning we can distinguish between causal defeasible rules (fire causes smoke) and evidential defeasible rules (smoke is evidence for fire) [9], for now we make no assumptions as to the type of rule used. Finally, note that informally, $n(r)$ is a wff in \mathcal{L} which says that rule $r \in R$ is applicable.

Definition 2.2 [Knowledge bases] A *knowledge base* in an $AS = (\mathcal{L}, \mathcal{R}, n)$ is a set $\mathcal{K} \subseteq \mathcal{L}$ consisting of two disjoint subsets \mathcal{K}_e (the *evidence*) and \mathcal{K}_p (the *ordinary premises*).

The evidence in \mathcal{K}_e is similar to axiom premises [12], which cannot be denied or attacked. If, for example, a witness testimony is presented as evidence, then the existence of the testimony cannot be denied; of course we can still question its veracity. Ordinary premises can be undermined by other arguments (see Definition 2.4).

Arguments can be constructed from knowledge bases by chaining inference rules into trees. Here, for any argument A , the function Sub returns all sub-arguments of A ; Prop returns all the formulas in A ; Prem returns all the formulas of \mathcal{K} (called *premises*) used to build A , Conc returns A 's conclusion, Rules returns all inference rules in A and TopRule returns the last inference rule used in A .

Definition 2.3 [Arguments] An *argument* A on the basis of a knowledge base \mathcal{K} in an argumentation system $AS = (\mathcal{L}, \mathcal{R}, n)$ is:

1. φ if $\varphi \in \mathcal{K}$ with: $\text{Prem}(A) = \{\varphi\}$; $\text{Conc}(A) = \varphi$; $\text{TopRule}(A) = \text{undefined}$; $\text{Prop}(A) = \{\varphi\}$; $\text{Sub}(A) = \{\varphi\}$.
2. $A_1, \dots, A_n \rightarrow/\Rightarrow \psi$ if A_1, \dots, A_n are arguments such that there exists a strict/defeasible rule $\text{Conc}(A_1), \dots, \text{Conc}(A_n) \rightarrow/\Rightarrow \psi$ in $\mathcal{R}_s/\mathcal{R}_d$, with:
 $\text{Prem}(A) = \text{Prem}(A_1) \cup \dots \cup \text{Prem}(A_n)$;
 $\text{Conc}(A) = \psi$;
 $\text{TopRule}(A) = \text{Conc}(A_1), \dots, \text{Conc}(A_n) \rightarrow/\Rightarrow \psi$;
 $\text{Prop}(A) = \text{Prop}(A_1) \cup \dots \cup \text{Prop}(A_n) \cup \{\psi\}$;
 $\text{Sub}(A) = \text{Sub}(A_1) \cup \dots \cup \text{Sub}(A_n) \cup \{A\}$;
 $\text{Rules}(A) = \text{Rules}(A_1) \cup \dots \cup \text{Rules}(A_n) \cup \{\text{Conc}(A_1), \dots, \text{Conc}(A_n) \rightarrow/\Rightarrow \psi\}$

Arguments can be attacked in essentially two ways: on their conclusion (rebutting attack) or on a defeasible inference step (undercutting attack).

Definition 2.4 [Attack] A *attacks* B iff A *undercuts* or *rebuts*, where:

- A *undercuts* argument B (on r) iff $\text{Conc}(A) \in \overline{n(r)}$ for some $B' \in \text{Sub}(B)$ such that $\text{TopRule}(B) = r$ and r is defeasible.
- A *rebuts* argument B (on φ) iff $\text{Conc}(A) \in \overline{\varphi}$ and $\text{Conc}(B') = \varphi$ for some $B' \in \text{Sub}(B)$ where either $\text{TopRule}(B)$ is defeasible or $\varphi \in \mathcal{K}_p$.

Argumentation systems plus knowledge bases form argumentation theories, which induce structured argumentation frameworks. Note that we do not include an ordering on the arguments, as this is not needed for current purposes. Furthermore, because we only use the structure of arguments and not take the evaluation of arguments into account for our constraints on Bayesian networks, we will not discuss any of the possible argumentation semantics for ASPIC⁺ as given in [12].

Definition 2.5 [Structured Argumentation Frameworks] Let AT be an *argumentation theory* (AS, \mathcal{K}) . A *structured argumentation framework* (SAF) defined by AT , is a pair $\langle \mathcal{A}, \mathcal{C} \rangle$ where \mathcal{A} is the set of all finite arguments constructed from \mathcal{K} in AS and $(X, Y) \in \mathcal{C}$ iff X attacks Y .

Example 2.6 As an example of an argument, suppose that a burglary has taken place and that we are interested in whether some suspect is guilty of committing the burglary (Bur). Forensic analysis (For) shows a match between a pair of shoes owned by the suspect and footprints (Ftpr) found near the crime scene. However, there is also evidence that there may have been a mix up at the forensic lab (Mix): the exact history of the footprints from crime-scene to lab has not been properly documented. This suspect had a motive (Mot) to commit this burglary, which is confirmed by at least one reliable testimony (Tes1). Furthermore, it is argued that the suspect also had the opportunity (Opp) to commit this burglary, but this is denied (\neg Opp) by a further testimony of the suspect himself (Tes2). We can now build arguments based on the evidence, where $\mathcal{K}_e = \{\text{For}, \text{Mix}, \text{Tes1}, \text{Tes2}\}$ and $\mathcal{K}_p = \{\text{Opp}\}$. Furthermore, $\text{Mix} \in \overline{r_{for}}$, where $r_{for} : \text{For} \Rightarrow \text{Ftpr}$ is the rule applied in A'_1 .

$$\begin{array}{llll}
 A_1: \text{For} & A_3: \text{Tes1} & B_1: \text{Tes2} & C_1: \text{Mix} \\
 A'_1: A_1 \Rightarrow \text{Ftpr} & A'_3: A_3 \Rightarrow \text{Mot} & B'_1: B_1 \Rightarrow \neg \text{Opp} & \\
 A_2: \text{Opp} & A_4: A'_1, A_2, A'_3 \Rightarrow \text{Bur} & &
 \end{array}$$

Figure 1 shows the arguments, where inferences are modelled as dashed arrows (all the inferences in this example are defeasible) and attacks as thick black arrows. Grey propositions are in \mathcal{K}_e .

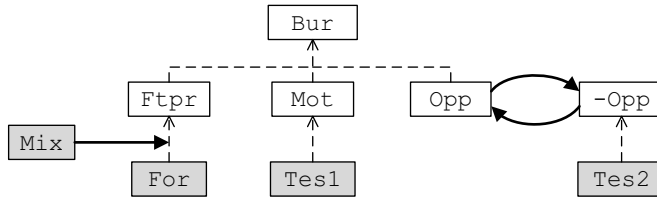


Figure 1. An argument in the example case

3. Bayesian Networks

In this section we briefly review Bayesian networks which combine a graph with conditional probability tables (CPTs) to compactly represent a joint probability distribution over a set of stochastic variables [13].

Definition 3.1 [Bayesian network] A BN is a triple $\mathcal{B} = (\mathbf{V}, \mathbf{A}, \mathcal{P})$ where:

- $G = (\mathbf{V}, \mathbf{A})$ is an acyclic directed graph with nodes \mathbf{V} and arcs $\mathbf{A} \subset \mathbf{V} \times \mathbf{V}$ (arc (V_i, V_j) is directed from V_i to V_j);
- $\mathcal{P} = \{\Pr_V \mid V \in \mathbf{V}\}$ where each \Pr_V is a set of (conditional) distributions $\Pr(V \mid \text{par}(V))$ over variables $V \in \mathbf{V}$, one for each combination of values for the parents $\text{par}(V)$ of V in graph G ; these distributions are typically represented as tables (CPTs).

Note from the above definition that in a BN there is a one-to-one correspondence between nodes and stochastic variables. Moreover, a BN allows for defining a joint probability distribution that respects the independences portrayed by its digraph (see below).

Proposition 3.2 The BN $\mathcal{B} = (\mathbf{V}, \mathbf{A}, \mathcal{P})$ uniquely defines the following joint probability distribution $\Pr(\mathbf{V})$:

$$\Pr(\mathbf{V}) = \prod_{V \in \mathbf{V}} \Pr(V \mid \text{par}(V))$$

A BN thus allows for computing any probability of interest over its variables; a typical query of interest is the probability $\Pr(h \mid \mathbf{e})$ of some hypothesis h given a combination of observations \mathbf{e} for a set of observed variables \mathbf{E} . The computation of probabilities from the network specification is called inference and the Bayesian network framework includes various algorithms to this end.

Although the directed arcs in a BN graph may suggest that the represented relations are causal, the arcs in fact only have meaning in combination with other arcs: together they capture the independences among the represented variables by means of the graphical *d-separation criterion* [13].

Definition 3.3 [d-separation] Consider three sets of nodes \mathbf{X} , \mathbf{Y} , and \mathbf{Z} in graph G . In addition, consider a simple chain s in G .

- Chain s is said to be *blocked*, or *inactive*, given \mathbf{Z} if the chain contains a node with two incoming arcs on the chain (a head-to-head node) which is not in \mathbf{Z} and has no descendants in \mathbf{Z} , or it contains a node in \mathbf{Z} that has at most one incoming arc on the chain; a chain that is not blocked by \mathbf{Z} is said to be *active* given \mathbf{Z} .
- \mathbf{X} and \mathbf{Y} are said to be *d-separated* by \mathbf{Z} if all possible chains s between nodes in \mathbf{X} and \mathbf{Y} are *inactive* given \mathbf{Z} .
- If \mathbf{X} and \mathbf{Y} are d-separated by \mathbf{Z} , then the two corresponding (sets of) variables \mathbf{X} and \mathbf{Y} are probabilistically independent given the third set \mathbf{Z} .

Note from the latter statement that the values of the variables in \mathbf{Z} are assumed to be actually observed and therefore known. We assume that two nodes that are directly connected by an arc are not d-separated.

Example 3.4 Figure 2 shows the graph of a Bayesian network, where observed variables have been shaded; the conditional probabilities are for now left implicit. Note that, compared to the argument graph in Figure 1 there is an additional variable, `Rel`, representing the reliability of the witness that gave testimony 1. Note that, for example, given evidence

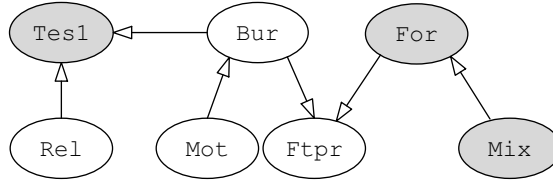


Figure 2. An Bayesian network graph for the example case

for *Tes1*, nodes *Rel* and *Bur* are not d-separated: given the witness testimony, their reliability will influence the probability of the suspect having committed the burglary. On the other hand, *Mix* and *Ftpr* are d-separated due to the observation for *For*.

4. From Argumentation to (Constraints on) Bayesian Network Graphs

A structured argumentation framework in ASPIC^+ can be used as the basis for (constraints on) the structure of a Bayesian network. First, we will discuss a number of generic constraints that can be put on a Bayesian network given an argumentation framework. After this, we will provide a heuristic for building a Bayesian network graph given these constraints based on arguments.

4.1. Constraints on the Structure of a Bayesian Network

First, we need to interpret all the elements of an argumentation framework in terms of the structure of a BN graph. The graph of a BN conveys the (in)dependencies between variables represented by the nodes in the graph. With structural constraints we denote both properties of nodes (which variables must be represented, and what are their values) and chains of arcs. In this paper we assume that all variables V are binary-valued, with possible values $\text{Val}(V) = \{\text{true}, \text{false}\}$. Since (in)dependencies are dynamic and can change depending on the observed variables, we will also include observations for variables in the structural constraints.

Constraint 4.1 [Nodes and values] Given a $\text{SAF} = \langle \mathcal{A}, \mathcal{C} \rangle$, the following constraints can be put on the nodes \mathbf{V} of a Bayesian network.

- For every atomic proposition v or $\neg v$ in $\text{Prop}(A)$, where $A \in \mathcal{A}$, there exists a single node $V \in \mathbf{V}$ such that $v \equiv V = \text{true}$ and $\neg v \equiv V = \text{false}$;
- Every atomic proposition e_i or $\neg e_i$ in \mathcal{K}_e is taken to represent the observed value of node E_i from the set of observed variables $\mathbf{E} \subset \mathbf{V}$.

We say that propositions v and $\neg v$ are *associated with* node V .

We now address the (chains of) arcs \mathbf{A} that capture the (in)dependencies between variables in a BN. Whenever a rule is applied to infer some conclusion from a set of premises, we should be able to reason from each node associated with the premises to the node associated with the consequence in the BN. That is, given the context of the evidence upon which an argument is built, there should exist active chains between the nodes associated with the applied rules.

Constraint 4.2 [Inference chains] Given a $SAF = \langle \mathcal{A}, \mathcal{C} \rangle$, the following constraints based on inferences can be put on the chains of arcs \mathbf{A} of a Bayesian network.

- For every rule $r : \varphi_1, \dots, \varphi_n \rightarrow \psi$ or $r : \varphi_1, \dots, \varphi_n \Rightarrow \psi$ such that $r = \text{Rules}(A)$ and $A \in \mathcal{A}$, there exist active chains between each of the nodes associated with $\varphi_1, \dots, \varphi_n$ and the node associated with ψ , given the observed nodes \mathbf{E} associated with evidence in \mathcal{K}_e .

An argumentation framework also includes a set of attack relations \mathcal{C} . As for inference, it makes sense to assume that whenever two propositions in arguments are in conflict (i.e. contrary or contradictory), that there is an influence between the values representing these propositions in the BN. Because contradictory propositions φ and $-\varphi$ are captured as two values of a single variable (see Constraint 4.1), not all of the propositions will translate to separate nodes in the BN, and hence the influences between the values associated with φ and $-\varphi$ will be captured in the (conditional) probabilities involving these values. In the case of φ being contrary to ψ – which occurs in rebutting attacks – the influence should be captured as an active chain between the two associated nodes given the evidence. Similarly, if a proposition undercuts the inference from one proposition to another then, given the evidence, there should be active chains between the nodes associated with this undercutting attack.

Constraint 4.3 [Attack chains] Given a $SAF = \langle \mathcal{A}, \mathcal{C} \rangle$, the following constraints based on attacks can be put on the chains of arcs \mathbf{A} of a Bayesian network. For every attack relation $(A, B) \in \mathcal{C}$:

- if $\text{Conc}(A) = \varphi$ is a *contrary* of $\text{Conc}(B) = \psi$, then there exists an active chain between the node associated with φ and the node associated with ψ , given the observed nodes \mathbf{E} associated with evidence in \mathcal{K}_e .
- if A *undercuts* B on r , where $\text{Conc}(A) = \chi$ and $r : \varphi_1, \dots, \varphi_n \Rightarrow \psi$, then there exist active chains between the node associated with χ and the nodes associated with $\varphi_1, \dots, \varphi_n, \psi$, given the observed nodes \mathbf{E} associated with evidence in \mathcal{K}_e .

Example 4.4 In our case, a Bayesian network expert with limited domain knowledge builds the network in Figure 2, and a judge builds the argument in Figure 1. The question is now whether, given the evidence, the network conforms to the constraints imposed by the arguments. In the argumentation framework of the case, there is an argument about the suspect having the opportunity to commit the burglary (Opp), and a counterargument based on the suspect's testimony (Tes2). The BN, however, does not include variables representing Opp or Tes2 and it thus violates some of the node constraints posed by the argumentation framework. Note that the BN also includes additional information that is not captured in the argumentation framework: there is a variable representing the reliability of a witness testimony (Rel) whereas none of the arguments includes a proposition about witness reliability. Remember that our current aim is to use the argumentation framework to put constraints on the BN and not vice versa, so any additional knowledge or reasoning in the BN is fine as long as it does not lead to violation of the constraints.

In order to determine the active chains in the BN of Figure 2, we have to consider the set of observations \mathbf{E} based on the evidence in \mathcal{K}_e and the nodes in the BN, so $\mathbf{E} = \{\text{For} = \text{true}, \text{Mix} = \text{true}, \text{Tes1} = \text{true}\}$. With respect to inference chains,

we now see that there are active chains from Ftpr to For (inference in A'_1), Tes1 to Mot (inference in A'_3), Ftpr to Bur and Mot to Bur (both part of the inference in A'_4). Because the node representing Opp is not in the network, there is no chain from Opp to Bur as per the constraint based on the top rule application in A_4 . Furthermore, because Tes2 is not a node in the network, there is no active chain from Tes2 to Opp , which is required by both the inference chain constraint and the first attack chain constraints. Finally, the second attack chain constraint says that, because Mix undercuts the application of $r_{\text{for}} : \text{For} \Rightarrow \text{Ftpr}$ in A'_1 , there should be active chains from Mix to For and from Mix to Ftpr . While these chains are both in the BN, the latter one, from Mix to Ftpr is blocked by the observation $\text{For} = \text{true}$, so the constraint is violated.

4.2. Transforming Arguments into Bayesian Network Graphs

In this section we propose a heuristic for the construction of a BN from arguments. Our BN graph construction heuristic starts by building an undirected skeleton in which all nodes that should be connected by active chains, according to the constraints from section 4.1, are connected directly by an edge. The resulting undirected graph will typically be densely connected and represent only few independencies. After this, it has to be decided which connections to retain and how to direct the edges. Recall that arguments are typically constructed by performing successive inference steps based on evidence. The rules behind these inferences can be either causal (fire causes smoke) or evidential rules (smoke is evidence for fire) [9]. In BNs – though the graph is just a representation of an independence relation – people tend to attach a causal interpretation to the arcs, and the notion of causality is typically used as a heuristic to guide the construction of the graph with the help of domain experts [13]. Hence, we propose to use the same heuristic for choosing the direction of the arcs. Furthermore, undercutting can be seen as a form of intercausal reasoning called *explaining away* (see Section 5). In BNs head-to-head nodes (see Definition 3.3) are explicitly used to model induced intercausal dependencies, such as exploited in explaining away, and we therefore propose to use head-to-head connections among the nodes involved in an undercutting attack. Our heuristic is then as follows:

BN graph construction heuristic

- 1) construct nodes according to Constraint 4.1;
- 2) each active chain dictated by Constraints 4.2 and 4.3 is implemented as a direct, undirected edge;
- 3) for all undercutting attacks: if a rule $r : \varphi_1, \dots, \varphi_n \Rightarrow \psi$ is undercut by $\chi \in \overline{n(r)}$ then remove edge (χ, ψ) (or (ψ, χ)) and turn the other edges involved into the following directed ones: (χ, φ_i) and (ψ, φ_i) for all φ_i ;
- 4) for all remaining undirected edges, choose the causal direction if a causal interpretation is possible, and an arbitrary direction, otherwise;
- 5) verify that the graph is acyclic and that all chains that should be active indeed are; otherwise remove or reverse appropriate arcs in consultation with a domain expert.

With respect to directing the edges we remark that BN graphs which share the same skeleton (undirected edges) and the same immoralities (head-to-head nodes for which the parents are not directly connected) are said to be Markov equivalent and capture

the same independence relation. Arc reversal therefore does not change the modelled independence relation as long as the resulting graph is Markov equivalent.

Example 4.5 Using the heuristic described above on the arguments depicted in Figure 1 we can first construct an undirected graph G' (Figure 3). We then apply the third step of the heuristic: *Mix* – which in the argumentation framework is modelled as an under-cutter of the rule $\text{For} \Rightarrow \text{Ftpr}$ – is now captured in the BN as a cause of *For*, together with *Ftpr*. The observation of head-to-head node *For* makes the chain between *Mix* and *Ftpr* active, so evidence for a mix-up will now change the probability of the footprints being the cause of the forensic results – further probabilistic constraints (see Section 5) could ensure that this intercausal effect is indeed “explaining away”. For the fourth step of the heuristic, we can quite sensibly interpret the edges causally: having a motive and opportunity cause the suspect to commit the burglary, which in turn causes the footprints to be found. Furthermore, the evidential observations are caused by the events that led to them. Now it can be verified that all chains that should be active given $\mathbf{E} = \{\text{Mix}, \text{For}, \text{Tes1}, \text{Tes2}\}$ indeed are, and we end up with graph G in Figure 3.

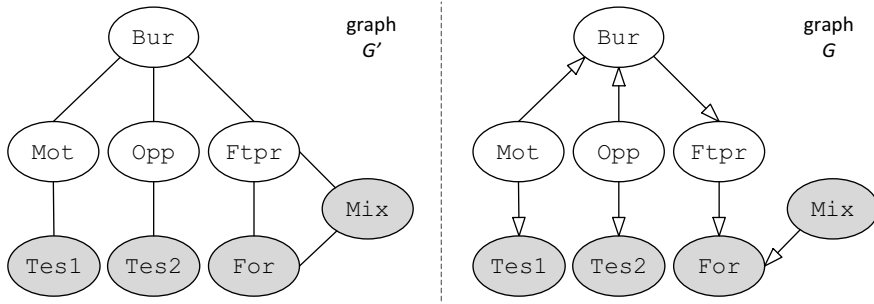


Figure 3. Undirected and directed graphs based on the argument in Figure 1

The structure of the final graph G is quite close to the structure of the original arguments in Figure 1. This structural similarity can also be seen in [3]: in a Wigmore diagram, the arrows denote evidential influences, which are then reversed in the causal direction to form a BN graph. A difference between the informal heuristic of [3] and ours is that we do not assume all influences in the argument to be evidential.

5. Constraints on the Probabilities of a Bayesian Network

Since a BN represents a probability distribution with its independence relation, an important part of the BN is formed by the (conditional) probabilities over the variables in a network. Putting constraints on these probabilities based on the evidential arguments in a case requires a probabilistic interpretation of such arguments, which is not straightforward (see, a.o., [8, 10, 11, 14]). A full exploration of probabilistic constraints is therefore beyond the scope of this paper. Rather, we will show a few possibilities of transforming argument structures into probabilistic constraints, and briefly discuss some of the issues concerning a probabilistic interpretation of structured arguments.

As for structural constraints, conditional probability constraints can be derived from the inferences and attacks in an argumentation framework. The interpretation of a strict

rule, for example, is fairly straightforward: the conclusion is necessarily true given the premises. However, with respect to defeasible rules different ideas exist on how they should be modelled probabilistically.

Constraint 5.1 [Inference probabilities] Given a $SAF = \langle \mathcal{A}, \mathcal{C} \rangle$, the following constraints can be put on the distribution \Pr defined by a Bayesian network.

- For every strict rule $r : \varphi_1, \dots, \varphi_n \rightarrow \psi$ such that $r = \text{Rules}(A)$, $A \in \mathcal{A}$, we have that $\Pr(\psi \mid \varphi_1, \dots, \varphi_n) = 1$;
- For every defeasible rule $r : \varphi_1, \dots, \varphi_n \Rightarrow \psi$ such that $r = \text{Rules}(A)$, $A \in \mathcal{A}$, we have that $\Pr(\psi \mid \varphi_1, \dots, \varphi_n) > 0$.

Note that all probabilities are taken to hold in the context of the available evidence, that is, each probability should also be conditioned on the subset of evidence for the nodes not associated with the rule. The above constraint for defeasible rules, which has been adapted from [11], is fairly weak. It can, for example, be argued that the premise should make it more likely than not that the conclusion is true ($\Pr(\psi \mid \varphi) > 0.5$), or that the premises are only a relevant reason for the conclusion if they increase the belief in the conclusion (e.g. $\Pr(\psi \mid \varphi) > \Pr(\psi)$ [10]). The point is that there are many measures of support [14], and choosing exactly which ones to use to interpret inference is not trivial.

As for attacks, it makes sense to interpret rebuttal using classical negation, as rebutting propositions φ and $-\varphi$ are values of one variable. Undercutting, where the application of a rule is attacked, can be seen as a form of *explaining away*, where the influence of a certain observation should be removed, or at least weakened, given an additional observation [13] (for more on the relations between undercutting evidential arguments and alternative causal explanations also see [9]). In BNs, explaining away is considered to be a type of intercausal reasoning often found between two causes (say ψ and χ) of a common effect (say φ): if the effect is observed, both causes become more likely; however, if we subsequently know which cause actually occurred (e.g. χ), the probability of the other cause (ψ) being present as well decreases.

Constraint 5.2 [Attack probabilities] Given a $SAF = \langle \mathcal{A}, \mathcal{C} \rangle$, the following constraints based on attacks can be put on the conditional probabilities of a BN. For every attack relation $(A, B) \in \mathcal{C}$:

- if $\text{Conc}(A) = \varphi$ and $\text{Conc}(B) = \psi$ and $\varphi \in \bar{\psi}$, then $\Pr(\psi \mid \varphi) = 0$.
- if $\text{Conc}(A) = \chi \in \bar{n}(r)$ and $r : \varphi_1, \dots, \varphi_n \Rightarrow \psi$ such that $r = \text{Rules}(B)$, then $\Pr(\psi \mid \varphi_1, \dots, \varphi_n, \chi) < \Pr(\psi \mid \varphi_1, \dots, \varphi_n)$.

A stronger version of the second constraint says that $\Pr(\psi \mid \varphi, \chi) = 0$ [11], which fits with the idea that undercutting arguments always defeat the argument they attack [12]. Similarly, in [8] an argument is not applicable given an exception.

Example 5.3 Again consider the arguments in Figure 1 and the BN in Figure 2. Consider the two defeasible rules $r_{for} : \text{For} \Rightarrow \text{Ftpr}$ and $r_{mot} : \text{Mot} \Rightarrow \text{Bur}$, and the undercutting attack (C_1, A'_1) . In the context of the relevant evidence \mathbf{e}' , our choice of probabilistic interpretation now gives the following constraints:

$$\Pr(\text{Ftpr} = \text{true} \mid \text{For} = \text{true}, \mathbf{e}') > 0$$

$$\Pr(\text{Ftpr} = \text{true} \mid \text{For} = \text{true}, \text{Mix} = \text{true}, \mathbf{e}') < \Pr(\text{Ftpr} = \text{true} \mid \text{For} = \text{true}, \mathbf{e}')$$

From Figure 2 we see that given *For*, the effect of a mix-up on footprints is blocked, which means that the last constraint is violated since due to the blocking influence of *Mix* it holds that $\Pr(\text{Ftpr} = \text{true} \mid \text{For} = \text{true}, \text{Mix} = \text{true}, \mathbf{e}') = \Pr(\text{Ftpr} = \text{true} \mid \text{For} = \text{true}, \mathbf{e}')$.

In addition to rules and attacks, there are also other aspects of argumentation frameworks that can be interpreted probabilistically. For example, ASPIC⁺ includes the option to define preferences over ordinary premises in \mathcal{K}_p and defeasible rules in \mathcal{R}_d , which translate more or less directly to constraints on probabilities: if premise φ is preferred over premise ψ , then $\Pr(\varphi) > \Pr(\psi)$. Furthermore, it is also possible to capture the evaluation of arguments probabilistically, as in [8], where the (conditional) probabilities depend on the status of propositions in the Carneades argument evaluation structure.

Once we have a probabilistic interpretation of an argumentation framework, we can use the probability constraints in an elicitation procedure for obtaining the required local probability distributions [15]. However, the above discussion shows that there are different ways to interpret arguments probabilistically, and that different aspects of argumentation frameworks can be incorporated as constraints on probabilities. The difficulty here lies in the fact that the exact probabilistic interpretation of arguments and evidence, and hence the various types of constraints on a BN, is a contentious issue. In fact, one way to deal with discussions surrounding constraints on probabilities is to allow arguments *about* the various probabilistic constraints in a BN [16].

6. Conclusion

In this paper we proposed a method for establishing constraints on Bayesian networks from structured arguments. Using our method, the type of arguments typically constructed by, for example, legal experts can be used to check BNs and in particular the knowledge contained in them. Observed differences between the arguments and a BN may carry useful information, as they point to possible differences in the reasoning of a Bayesian network expert and the domain expert. Hence, the communication gap between these two types of experts can be bridged.

In addition to deriving constraints for BNs from a set of arguments, we have also designed a heuristic for constructing a BN graph based on these constraints. In this way, the constraints can help with the construction of a BN: domain experts can build evidential arguments, which are then transformed into BNs or BN skeletons, which can in turn be interpreted and further extended by Bayesian network experts in conjunction with the domain experts. The proposed heuristic thus expands the toolbox for building a BN relatively easily, and can be used together with existing techniques for building BNs such as idioms, recurrent BN structures based on standard arguments such as inductive arguments or arguments based on evidence [1].

One of the key issues when transforming arguments into BNs is the probabilistic interpretation of the elements of argumentation. It turns out that there are numerous possible interpretations of evidential support and argument strength [10, 11, 14], which obviously lead to different constraints on the probability distribution represented by a BN.

It is at present not clear if and how the choice of constraints influences the outcomes of the reasoning in BNs. Furthermore, the choice of interpretation also depends on the (implicit) assumptions regarding evidential argument strength the domain experts that build structured arguments have. We would like to address these questions in future research.

References

- [1] N. Fenton and M. Neil. *Risk assessment and decision analysis with Bayesian networks*. CRC Press, 2012.
- [2] V. Franqueira, T.T. Tun, Y. Yu, R. Wieringa, and B. Nuseibeh. Risk and argument: a risk-based argumentation method for practical security. In *19th IEEE International Requirements Engineering Conference*, pages 239–248. IEEE, 2011.
- [3] J.B. Kadane and D.A. Schum. *A probabilistic analysis of the Sacco and Vanzetti evidence*, volume 773. John Wiley & Sons, 2011.
- [4] J. Keppens. Argument diagram extraction from evidential Bayesian networks. *Artificial Intelligence and Law*, 20(2):109–143, 2012.
- [5] S.T. Timmer, J.-J. Ch. Meyer, H. Prakken, S. Renooij, and B. Verheij. Extracting legal arguments from forensic Bayesian networks. In *JURIX 2014: The Twenty-seventh Annual Conference*, volume 217, pages 71–80, 2014.
- [6] G.A.W. Vreeswijk. Argumentation in Bayesian belief networks. In *Argumentation in Multi-Agent Systems*, volume 3366 of *Lecture Notes in Computer Science*, pages 111–129, 2005.
- [7] F.J. Bex, H. Prakken, C.A. Reed, and D.N. Walton. Towards a formal account of reasoning about evidence: Argumentation schemes and generalisations. *Artificial Intelligence and Law*, 11(2/3):125–165, 2003.
- [8] M. Grabmair, T.F. Gordon, and D. Walton. Probabilistic semantics for the Carneades argument model using Bayesian networks. In *Proceedings of COMMA*, pages 255–266, 2010.
- [9] F.J. Bex. An integrated theory of causal stories and evidential arguments. In *Proceedings of the 15th International Conference on Artificial Intelligence and Law*, pages 13–22. ACM, 2015.
- [10] U. Hahn and J. Hornikx. A normative framework for argument quality: argumentation schemes with a Bayesian foundation. *Synthese*, pages 1–41.
- [11] B. Verheij. Arguments and their strength: Revisiting Pollock’s anti-probabilistic starting points. In *Proceedings of COMMA 2014, Computational Models of Argument*, pages 433–444. IOS Press, 2014.
- [12] H. Prakken. An abstract framework for argumentation with structured arguments. *Argument and Computation*, 1:93–124, 2010.
- [13] F. V. Jensen and T. D. Nielsen. *Bayesian Networks and Decision Graphs*. Springer, Berlin, 2007.
- [14] V. Crupi, K. Tentori, and M. Gonzalez. On Bayesian measures of evidential support: Theoretical and empirical issues. *Philosophy of Science*, 74(2):229–252, 2007.
- [15] M.J. Druzdzel and L.C. van der Gaag. Elicitation of probabilities for belief networks: Combining qualitative and quantitative information. In *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*, pages 141–148, 1995.
- [16] J. Keppens. On modelling non-probabilistic uncertainty in the likelihood ratio approach to evidential reasoning. *Artificial Intelligence and Law*, 22(3):239–290, 2014.

On Efficiently Enumerating Semi-Stable Extensions via Dynamic Programming on Tree Decompositions

Bernhard BLIEM^a Markus HECHER^a, Stefan WOLTRAN^a

^a*Institute of Information Systems, TU Wien, Austria*

Abstract. Many computational problems in the area of abstract argumentation are intractable. For some semantics like preferred and semi-stable, important decision problems can even be hard for classes of the second level of the polynomial hierarchy. One approach to deal with this inherent difficulty is to exploit structure of argumentation frameworks. In particular, algorithms that run in linear time for argumentation frameworks of bounded treewidth have been proposed for several semantics. In this paper, we contribute to this line of research and propose a novel algorithm for the semi-stable semantics. We also present an implementation of the algorithm and report on some experimental results.

Keywords. Abstract Argumentation, Fixed-Parameter Tractability, Dynamic Programming on Tree Decompositions

1. Introduction

Dung's abstract argumentation frameworks [5] are a central concept in many argumentation formalisms and systems. Efficient and versatile methods for abstract argumentation are therefore important for further advances in the field. For some important semantics like the preferred and semi-stable extensions (see [7, 12]), the high worst-case complexity is a major obstacle to finding algorithms that evaluate argumentation frameworks (AFs) of real-world size in reasonable time. In fact, standard algorithms tend to be problematic for larger instances even if the inherent structure of the frameworks remains simple; a situation that is likely to appear when frameworks are obtained during some instantiation process. It is thus valuable to design alternative algorithms, where the size of the framework has less influence on the runtime.

The field of parameterized complexity theory [4] formally captures this intuition and is based on the following observation: Many hard problems become tractable if some problem parameter is bounded by a constant. This property is referred to as *fixed-parameter tractability* (FPT). One important parameter of graphs is the treewidth, which measures the “tree-likeness” of a graph and is thus also applicable to AFs. In the field of argumentation, research in this direction was initiated by Dunne [6] who showed that many intractable problems can be solved in linear time for argumentation frameworks of bounded treewidth. Later these results were extended to the more general structural parameter of clique-width [9]. Further parameterized complexity results include [8, 15].

Showing that a problem parameterized by treewidth is FPT often does not yield a practically useful algorithm automatically. For obtaining such algorithms, dynamic programming (DP) algorithms that operate on a tree decomposition (TD) of the input usually have to be designed. For admissible, preferred, and ideal semantics, such algorithms have been presented in [11], and the system DYNPARTIX [3] implements algorithms for admissible, preferred, stable and complete semantics. However, semi-stable semantics has not been considered so far. This semantics is challenging, as it is among the most complex ones for abstract argumentation, with credulous acceptance being Σ_2^P -complete and skeptical acceptance being Π_2^P -complete [12].

In this work, we present a DP algorithm that computes semi-stable extensions in linear time on AFs of bounded treewidth. We briefly report on an implementation of our algorithms and some experimental evaluation.

2. Background

Abstract Argumentation. We first review the Dung argumentation framework [5].

Definition 2.1. An argumentation framework (AF) is a pair $F = (A_F, R_F)$, where A_F is a set of arguments and $R_F \subseteq A \times A$ is a set of attacks. Instead of $(a, b) \in R_F$, we write $a \rightarrow^{R_F} b$, and we sometimes omit R_F if it is clear from the context. For any set $S \subseteq A_F$, we write $S \rightarrow^{R_F} b$ if there is some $a \in S$ s.t. $a \rightarrow^{R_F} b$. We say that a is defended by S if $S \rightarrow^{R_F} b$ for each $b \in A_F$ with $b \rightarrow^{R_F} a$. We call $S_{R_F}^+ = S \cup \{b \mid S \rightarrow^{R_F} b\}$ the range of S .

A semantics characterizes so-called *extensions* of an AF, i.e., sets of arguments that are acceptable. For a semantics $\psi \in \{\text{conflict-free, admissible, preferred, semi-stable, stable}\}$ and an AF F , we write $\psi(F)$ to denote the set of ψ -extensions in F .

Definition 2.2. Let F be an AF and $S \subseteq A_F$. We define (a) $S \in \text{conflict-free}(F)$ if there are no $a, b \in S$ with $a \rightarrow^{R_F} b$; (b) $S \in \text{admissible}(F)$ if $S \in \text{conflict-free}(F)$ and each $a \in S$ is defended by S ; (c) $S \in \text{preferred}(F)$ if $S \in \text{admissible}(F)$ and $S' \not\supseteq S$ holds for each $S' \in \text{admissible}(F)$; (d) $S \in \text{semi-stable}(F)$ if $S \in \text{admissible}(F)$ and $S_{R_F}^+ \not\supseteq S_{R_F}^+$ holds for each $S' \in \text{admissible}(F)$; (e) $S \in \text{stable}(F)$ if $S \in \text{conflict-free}(F)$ and $A_F = S_{R_F}^+$.

One can show that for every AF F it holds that $\text{stable}(F) \subseteq \text{semi-stable}(F) \subseteq \text{preferred}(F) \subseteq \text{admissible}(F) \subseteq \text{conflict-free}(F)$.

Tree decompositions (TDs). A parameterized problem is a problem whose instances are accompanied by an integer that represents a certain parameter of the instance. Such a problem is called *fixed-parameter tractable* (FPT) if it is solvable in time $f(k) \cdot n^{\mathcal{O}(1)}$, where n is the input size and f is a function that only depends on the value k of the parameter [4]. The parameter we consider is *treewidth*, which is defined by means of tree decompositions, originally introduced in [18]. The intuition behind TDs is to obtain a tree from a (potentially cyclic) graph by subsuming multiple vertices under one node and thereby isolating the parts responsible for cyclicity.

Definition 2.3. A tree decomposition of a graph $G = (V, E)$ is a pair $T = (\mathcal{T}, \mathcal{X})$ where $\mathcal{T} = (V_{\mathcal{T}}, E_{\mathcal{T}})$ is a (rooted) tree and $\mathcal{X} = \{X_{t_1}, \dots, X_{t_n}\}$ assigns to each node $t \in V_{\mathcal{T}}$ a subset X_t of V (called the bag of t) as follows: (1) For each vertex $v \in V$, there is a node

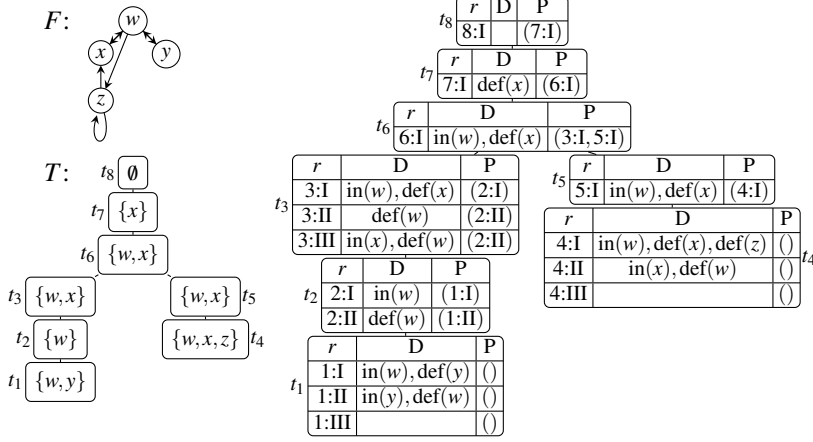


Figure 1: DP computation of $\text{stable}(F)$ w.r.t. normalized TD $T = (\mathcal{T}, \mathcal{X})$.

$t \in V_{\mathcal{T}}$ such that $v \in X_t$; (2) For each edge $e \in E$, there is a node $t \in V_{\mathcal{T}}$ such that $e \subseteq X_t$; (3) For each $v \in V$, the subgraph of \mathcal{T} induced by $\{t \in V_{\mathcal{T}} \mid v \in X_t\}$ is connected. We call $\max_{t \in V_{\mathcal{T}}} |X_t| - 1$ the width of T . The treewidth of G is the minimum width over all its TDs.

For $\mathcal{T} = (V_{\mathcal{T}}, E_{\mathcal{T}})$ we often write $t \in \mathcal{T}$ instead of $t \in V_{\mathcal{T}}$. We only consider TDs of the following form that can be achieved in linear time without increasing the width [16].

Definition 2.4. We call a TD $(\mathcal{T}, \mathcal{X})$ normalized if its root has an empty bag and each node $t \in \mathcal{T}$ is of one of the following types. LEAF: t is a leaf of \mathcal{T} . FORGET: t has only one child t' and $X_t = X_{t'} \setminus \{v\}$ for some $v \in X_{t'}$. INSERT: t has only one child t' and $X_t = X_{t'} \cup \{v\}$ for some $v \notin X_{t'}$. JOIN: t has two children t', t'' and $X_t = X_{t'} = X_{t''}$.

Example 2.5. Figure 1 depicts a normalized TD T (having width 2) of the AF F .

Dynamic programming (DP) on Tree decompositions. Algorithms for dynamic programming on TDs generally traverse the TD in bottom-up order. At each node, partial solutions for the subgraph induced by the vertices encountered so far are computed and stored in a table associated with the node (cf. [17]). The size of each table is typically bounded by the width of the TD, and the number of TD nodes is linear in the input size. Hence, if the width is bounded by a constant, the search space for the subproblem is constant as well, and the number of subproblems only grows by a linear factor for larger instances. Each row in such a table corresponds to partial solutions that take only a part of the instance into account, namely the part of the instance that has been “encountered” during the bottom-up traversal:

Definition 2.6. Let F be an AF, $(\mathcal{T}, \mathcal{X})$ be a TD of F , and $t \in \mathcal{T}$. We use $X_{\geq t}$ to denote the union of all bags $X_s \in \mathcal{X}$ such that s occurs in the subtree of \mathcal{T} rooted at t . Moreover, $X_{> t}$ denotes $X_{\geq t} \setminus X_t$. We define F_t as the AF $(A_F \cap X_t, R_F \cap X_t^2)$ and call it the subframework in t . Finally, we define $F_{\geq t}$ as the AF $(A_F \cap X_{\geq t}, R_F \cap X_{\geq t}^2)$ and call it the subframework induced by t . (Note that $F_{\geq t} = F$ holds if t is the root of the TD.)

Example 2.7. We illustrate the DP for stable extensions of the example AF F in Figure 1. The tables in that figure are computed as follows. For a TD node t , each table row r

consists of data $D(r)$, which may assign a status to arguments in X_t . For any argument a , $D(r)$ contains $\text{in}(a)$ or $\text{def}(a)$ if for each set S of arguments represented by r it holds that $a \in S$ or $S \rightarrow a$, respectively. The set $P(r)$ contains so-called extension pointer tuples (EPTs) that denote the rows in the children of t where r was constructed from. By following these pointers, we can obtain the sets represented by r .

We make sure that all sets represented by a row are conflict-free. For instance, at node t_1 , $X_{t_1} = \{w, y\}$ holds and these arguments attack each other. Hence the table at t_1 contains a row for each of the three conflict-free subsets of X_{t_1} . At t_2 , the child rows are extended and the status assignment is updated by removing the status of arguments that are not contained in X_{t_2} . Observe that row 1:III is not extended by any row at t_2 . This is because the removed argument y (i.e., $y \in X_{t_1} \setminus X_{t_2}$) has neither status “in” nor “def”: Any solution S that is constructed using 1:III would satisfy neither $y \in S$ nor $S \rightarrow y$, so S would not be a stable extension. At t_3 , we extend child rows and guess a status for the introduced argument x . Note that we must discard rows containing both $\text{in}(w)$ and $\text{in}(x)$ because otherwise the partial solutions represented by such rows would not be conflict-free. At t_6 , only rows that agree on the status of the common arguments may be joined. We continue this procedure recursively until we reach the TD’s root.

To decide whether there is a stable extension, it suffices to check if the table in the root node is non-empty. The overall procedure is in FPT time because the number of nodes in the TD is bounded by the input size, and each node t is associated with a table of size at most $\mathcal{O}(3^{|X_t|})$ (i.e., the number of possible status assignments). The AF F has a stable extension due to existence of row 8:I. Solutions (stable extensions of F) can be enumerated with linear delay by starting at the root and following the EPTs while collecting arguments with status “in” according to the extended rows. Solution $\{w\}$ is constructed by starting at 8:I and following EPTs (7:I), (6:I), (3:I, 5:I), (2:I), (1:I) and (4:I). The union of the arguments having status “in” according to these rows is $\{w\}$. It is easy to see that $\{w\}$ is the only stable extension of F .

3. Algorithm for Admissible Semantics

We first provide an algorithm for admissible semantics that uses DP on TDs, modifying concepts from [11], and then extend it to semi-stable semantics in Section 4. The adaption is needed, since we require to distinguish partial solutions not only with respect to the status of already processed arguments, but have to guess whether arguments might be attacked or not by arguments appearing later. As we will see in Section 4, this allows us to relate partial solutions to those which possess a larger range. For space reasons, we only provide proof sketches and refer to [14] for full proofs. First we adapt the concept of restricted-admissible sets from [11] for our purposes.

Definition 3.1. Let $F = (A_F, R_F)$ be an AF and B a set of arguments. A tuple (S, D) satisfying $S, D \subseteq A_F$ and $S \cap D = \emptyset$ is a B -restricted admissible tuple for F if (1) S is conflict-free in F and S defends itself in F against all elements of $A_F \cap B$, and (2) for each $a \in A_F$, whenever $S \rightarrow^{R_F} a$ or $a \rightarrow^{R_F} S$, then $a \in D$. We call S a B -restricted admissible set for F if there is a set D such that (S, D) is a B -restricted admissible tuple for F .

Note that for $A_F \subseteq B$, B -restricted admissible sets of AF $F = (A_F, R_F)$ are just admissible sets for F . For $A_F \cap B = \emptyset$, B -restricted admissible sets are just the conflict-free

sets for F . Intuitively, if (S, D) is a B -restricted admissible tuple, then D consists of at least those arguments (different from S) that are defeated or still require defeating by S .

Example 3.2. Consider the framework F and TD T given in Figure 1. Let $F' = (A_{F'}, R_{F'})$ be the subframework induced by node t_3 of T minus the attack (y, w) , i.e., $A_{F'} = \{w, x, y\}$, $R_{F'} = \{(w, x), (x, w), (w, y)\}$. The $\{x, y\}$ -restricted admissible sets are \emptyset , $\{w\}$, $\{x\}$, $\{y\}$ and $\{x, y\}$. The set $\{y\}$, however, is not $\{w\}$ -restricted admissible, since $w \rightarrow^{R_{F'}} y$ but y does not defend itself against w . From the stated $\{x, y\}$ -restricted admissible sets we obtain $\{x, y\}$ -restricted admissible tuples by adding the required second component. Since condition (2) of Definition 3.1 is always trivially satisfied if D contains all arguments, we only state the smallest sets D (w.r.t. subset inclusion) that satisfy condition (2). These are (\emptyset, \emptyset) , $(\{w\}, \{x, y\})$, $(\{x\}, \{w\})$, $(\{y\}, \{w\})$ and $(\{x, y\}, \{w\})$.

The following concept of (valid) colorings helps to prove correctness of our algorithm.

Definition 3.3. Let $(\mathcal{T}, \mathcal{X})$ be a TD of an AF $F = (A_F, R_F)$ and $t \in \mathcal{T}$. We call $C : X_t \rightarrow \{\text{in}, \text{attc}, \text{def}, \text{out}\}$ a coloring and define $[C] = \{a \mid C(a) = \text{in}\}$ and $[[C]] = \{a \mid C(a) = \text{def} \text{ or } C(a) = \text{attc}\}$. Moreover, we define $e_t(C)$ as the collection of $X_{>t}$ -restricted admissible tuples (S, D) for $F_{\geq t}$ that satisfy the following properties for each $a \in X_t$.

- (i) $C(a) = \text{in} \iff a \in S$
- (ii) $C(a) = \text{def} \iff S \rightarrow^{R_F} a$
- (iii) $C(a) = \text{attc} \iff S \not\rightarrow^{R_F} a \text{ and } a \rightarrow^{R_F} S$
- (iv) $C(a) = \text{out} \implies S \not\rightarrow^{R_F} a \text{ and } a \not\rightarrow^{R_F} S$
- (v) $C(a) \in \{\text{def}, \text{attc}\} \iff a \in D$

If $e_t(C) \neq \emptyset$, C is called a valid coloring for t ; \mathcal{C}_t denotes the set of valid colorings for t . For convenience we use $e'_t(C) := \{S \mid (S, D) \in e_t(C)\}$.

Intuitively, the color “in” means that the argument a is in the $X_{>t}$ -restricted admissible set S and “def” means a is defeated by S . Arguments that attack S without being defeated (yet) have color “attc”, but this color may also be assigned to other arguments that don’t need to be colored “in” or “def”. The color “attc” means that the argument is expected to be defeated in the future. Finally, all remaining arguments are assigned color “out”. Any valid coloring C_t for a TD node t forms exactly one admissible tuple (S, D) , with S being the $X_{>t}$ -restricted admissible set and D being those arguments that are either defeated by S (color def) or attack S or shall be defeated by it (both color attc).

Example 3.4. Let F and T be the AF and TD, respectively, from Figure 1. Observe that $X_{t_3} = \{w, x\}$, $X_{>t_3} = \{y\}$ and $F_{\geq t_3} = (\{w, x, y\}, \{(w, x), (x, w), (w, y), (y, w)\})$. Let C be the coloring for t_3 s.t. $C(w) = \text{in}$, $C(x) = \text{def}$. The only tuple in $e_{t_3}(C)$ that is $X_{>t_3}$ -restricted admissible for $F_{\geq t_3}$ and satisfies the conditions of Definition 3.3 is $(\{w\}, \{x, y\})$.

We now show that valid colorings indeed correspond to admissible extensions.

Proposition 3.5. Assuming that r is the root of a normalized TD of an AF F and ε is the valid coloring for r , we have that $e'_r(\varepsilon) = \text{admissible}(F)$.

Proof sketch. Let r be the root of the TD (recall that $X_r = \emptyset$) and ε the (trivial) coloring for r . For $A_F \subseteq B$, B -restricted admissible sets of AF $F = (A_F, R_F)$ are just admissible sets for F . Hence $e'_r(\varepsilon) = \text{admissible}(F)$ follows immediately from Definitions 3.1 and 3.3. \square

Because of this correspondence, our goal is to efficiently compute $e'_r(\varepsilon)$ for the root r of a TD T of an AF F using coloring ε for r , as we have seen that $e'_r(\varepsilon) = \text{admissible}(F)$. However, to guarantee FPT w.r.t. treewidth, we cannot afford to compute $e_t(\cdot)$ explicitly. In the following, we show how we can compute compact representations of $e_t(\cdot)$. For this, we first define the following operations that we will use in our algorithm.

Definition 3.6. Let $(\mathcal{T}, \mathcal{X})$ be a TD of an AF $F = (A_F, R_F)$, and let C' and C'' be colorings for nodes t' and t'' , respectively, in \mathcal{T} . We define the following operations.

$$\begin{aligned}
 (C' - a)(b) &= \begin{cases} C'(b) & \text{for each } b \in X_{t'} \setminus \{a\} \end{cases} \\
 (C' +_{\text{attc}} a)(b) &= \begin{cases} C'(b) & \text{if } b \in X_{t'} \\ \text{def} & \text{if } a = b \text{ and } [C'] \xrightarrow{R_F} a \\ \text{attc} & \text{otherwise} \end{cases} \\
 (C' \hat{+}_{\text{out}} a)(b) &= \begin{cases} C'(b) & \text{if } b \in X_{t'} \\ \text{out} & \text{otherwise} \end{cases} \\
 (C' \dot{+}_{\text{in}} a)(b) &= \begin{cases} \text{in} & \text{if } a = b \text{ or } C'(b) = \text{in} \\ \text{out} & \text{if } a \neq b, (a, b) \notin R_F, (b, a) \notin R_F, C'(b) = \text{out} \\ \text{def} & \text{if } a \neq b \text{ and } ((C'(b) = \text{attc} \text{ and } (a, b) \in R_F) \text{ or } C'(b) = \text{def}) \\ \text{attc} & \text{otherwise} \end{cases} \\
 (C' \bowtie C'')(b) &= \begin{cases} \text{in} & \text{if } C'(b) = C''(b) = \text{in} \\ \text{out} & \text{if } C'(b) = C''(b) = \text{out} \\ \text{def} & \text{if } C'(b) = \text{def} \text{ or } C''(b) = \text{def} \\ \text{attc} & \text{otherwise} \end{cases}
 \end{aligned}$$

In our algorithm, we will use the “−” operation at FORGET nodes, the three “+” operations at INSERT nodes and “ \bowtie ” at JOIN nodes. The intuitions behind the three different “+” operations are the following: The operation $C +_{\text{attc}} a$ causes that a newly introduced atom a is considered an attacking candidate (attc); hence a will have to be defeated (def) by a resulting extension S (i.e., $S \xrightarrow{R_F} a$). The operation $C \dot{+}_{\text{in}} a$ puts the new atom a in the resulting extension. Finally, $C \hat{+}_{\text{out}} a$ results in a being neither in the extension nor being an attacking candidate of it. Using these operations, we now define our algorithm that traverses a TD in a bottom-up manner to compute *vcolorings*, which serve as compact representations of valid colorings.

Definition 3.7. Let $(\mathcal{T}, \mathcal{X})$ be a TD of an AF F , and let $t \in \mathcal{T}$. If t has exactly one child, let t' denote this child; if t has two children, let them be denoted by t' and t'' . We define the set of *vcolorings* for t depending on the node type of t :

LEAF: A coloring $C' : X_{t'} \rightarrow \{\text{in}, \text{out}, \text{def}, \text{attc}\}$ is a *vcoloring* for t if the following conditions hold for all $x \in X_{t'}$: (i) $C'(x) = \text{in} \implies C'(y) \in \{\text{def}, \text{attc}\}$ for all $y \xrightarrow{R_F} x$;
(ii) $C'(x) = \text{def} \iff \exists y : C'(y) = \text{in} \text{ and } y \xrightarrow{R_F} x$

FORGET ($X_t = X_{t'} \setminus \{a\}$ for some argument a): If C' is a *vcoloring* for t' and $C'(a) \neq \text{attc}$, then $C' - a$ is a *vcoloring* for t .

INSERT ($X_t = X_{t'} \cup \{a\}$ for some argument a):

(i) If C' is a *vcoloring* for t' , then $C' +_{\text{attc}} a$ is a *vcoloring* for t .

(ii) If C' is a *vcoloring* for t' , $[C'] \not\xrightarrow{R_F} a$ and $a \not\xrightarrow{R_F} [C']$, then $C' \hat{+}_{\text{out}} a$ is a *vcoloring* for t .

(iii) If C' is a vcoloring for t' , $a \not\vdash a$, $[C'] \not\vdash a$, $a \not\vdash [C']$ and $[[C']] = [[C' \dot{+}_{in} a]]$, then $C' \dot{+}_{in} a$ is a vcoloring for t .

JOIN: If C' and C'' are vcolorings for t' and t'' , respectively, and $[C'] = [C'']$ as well as $[[C']] = [[C'']]$ hold, then $C' \bowtie C''$ is a vcoloring for t .

Example 3.8. Let F and $T = (\mathcal{T}, \mathcal{X})$ be the AF and TD, respectively, from Figure 1. Figure 2 illustrates the bottom-up computation of the vcolorings for the AF and TD of Figure 1. Each row r contains a vcoloring in the set $D(r)$, and $P(r)$ contains the EPTs as described in Section 2. By following the EPTs from the row 8:I in the root table, we obtain the sets \emptyset , $\{w\}$ and $\{y\}$, which are indeed exactly the admissible sets.

We now illustrate how to compute the tables in Figure 2 from the bottom up. Consider LEAF node t_1 with bag $\{w, y\}$ and $F_{\geq t_1} = (\{w, y\}, \{(w, y), (y, w)\})$. Its table contains six vcolorings, which correspond to \emptyset -restricted admissible tuples (that encode conflict-free sets) for $F_{\geq t_1}$, namely (\emptyset, \emptyset) , $(\emptyset, \{w\})$, $(\emptyset, \{y\})$, $(\emptyset, \{w, y\})$, $(\{w\}, \{y\})$ and $(\{y\}, \{w\})$.

The next node t_2 is of type FORGET and removes y ($X_{> t_2} = \{y\}$). The vcolorings for t_2 are obtained from vcolorings for t_1 except for C' with $C'(y) = \text{attc}$. Intuitively, such colorings are not extended further because y is still an undefeated attacking candidate (i.e., y requires defeating). By properties (2) and (3) of TDs, y is not attacked by any argument outside $X_{\geq t_2}$, i.e., y will not be defeated in nodes further toward the root. The colorings for t_2 correspond to the $X_{> t_2}$ -restricted admissible tuples for $F_{\geq t_2} = F_{\geq t_1}$.

Node t_3 introduces argument x . Consider the coloring C' of t_2 with $C'(w) = \text{attc}$. We have now three possibilities for argument x (corresponding to our “+” operations).

- $C = C' \dot{+}_{\text{attc}} x$: This results in $C(x) = \text{attc}$ and $C(w) = \text{attc}$.
- $C = C' \dot{+}_{in} x$: If we set $C(x) = \text{in}$, this leads to $C(w) = \text{def}$ since $x \succ^{R_F} w$.
- $C = C' \dot{+}_{out} x$: This leads to $C(x) = \text{out}$ and $C(w) = \text{attc}$.

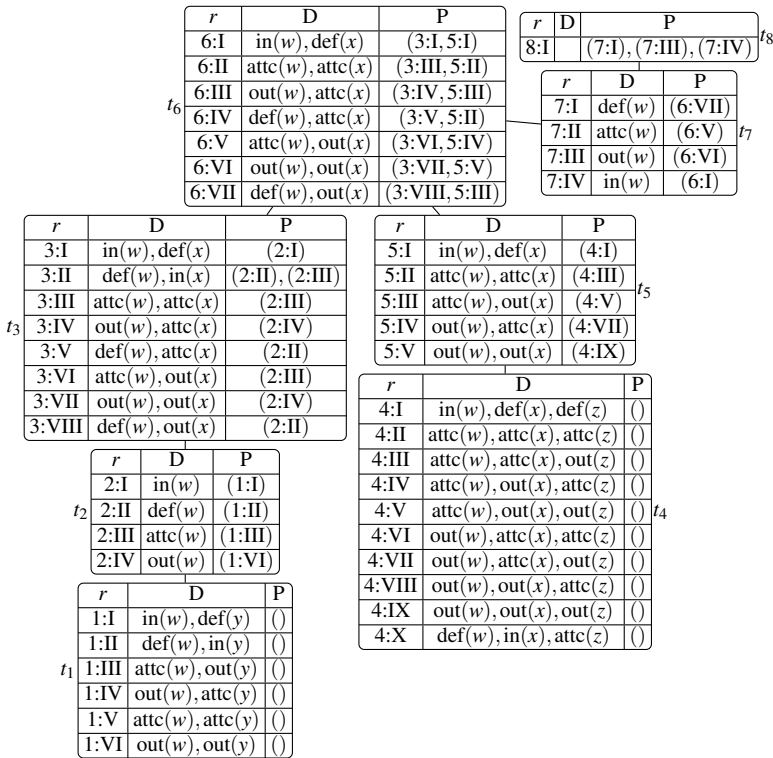
The only JOIN node is t_6 , which combines subframeworks $F_{\geq t_3}$ and $F_{\geq t_5}$. Let C' and C'' be colorings for t_3 and t_5 , respectively, and let $C'(w) = \text{in} = C''(w)$ and $C'(x) = \text{def} = C''(x)$. Since $[C'] = [C'']$ and $[[C']] = [[C'']]$, the colorings coincide on $X_{\geq t_3} \cap X_{\geq t_5}$ and we can join these colorings without any conflict, leading to $C = C' \bowtie C''$ with $C(x) = C'(x) = C''(x)$ and $C(w) = C'(w) = C''(w)$ for node t_6 . Now consider coloring C^* for node t_3 with $C^*(w) = \text{def}$ and $C^*(x) = \text{in}$. It holds that $[C''] \neq [C^*]$, and $[C''] \cup [C^*] = \{w, x\}$ are in conflict, leading to the fact that C'' and C^* do not result in a vcoloring for node t_6 . In fact, there is no resulting vcoloring for node t_7 originating from C^* .

Together with Proposition 3.5, the following theorem proves that the algorithm described in Definition 3.7 is sound.

Theorem 3.9. Let $(\mathcal{T}, \mathcal{X})$ be a TD of an AF F . For each coloring C for a node $t \in \mathcal{T}$, C is a valid coloring for t if and only if C is a vcoloring for t .

Proof sketch. The proof is by structural induction over the given TD and shows equivalence between valid colorings and vcolorings for all node types, see [14]. \square

Recall that the A_F -restricted admissible sets for an AF $F = (A_F, R_F)$ are the admissible sets for F . Because of Theorem 3.9 and Proposition 3.5, we can construct a valid coloring ε for the root r of any TD T by computing vcolorings in a bottom-up manner. This allows us to enumerate admissible sets via $e'_r(\varepsilon)$. Observe that \emptyset is always an admissible extension, so ε trivially exists, but for enumerating $e'_r(\varepsilon)$, the vcolorings for

Figure 2: DP computation of vcolorings for $F = (A_F, R_F)$ w.r.t. T (see Figure 1).

all the nodes of T are required. Enumeration can be done with linear delay by combining vcolorings from the different nodes of T according to the EPTs [1].

4. Algorithm for Semi-Stable Semantics

We now present our algorithm for semi-stable semantics by re-using concepts from the algorithm for admissible semantics from Section 3. First we define the counterparts of valid colorings and vcolorings for semi-stable semantics, namely *valid pairs* and *vpairs*.

Definition 4.1. Let $(\mathcal{T}, \mathcal{X})$ be a TD of an AF F , $t \in \mathcal{T}$, and (C, Γ) a pair with C being a coloring for t and Γ being a set of colorings for t . We call (C, Γ) simply a pair for t and define $e_t(C, \Gamma)$ as the collection of tuples (S, D) that satisfy the following conditions.

- (i) $(S, D) \in e_t(C)$;
- (ii) For all $C' \in \Gamma$, there is an $(E, U) \in e_t(C')$ such that $S \cup D \subset E \cup U$;
- (iii) For all tuples (E, U) that are $X_{>t}$ -restricted admissible for $F_{\geq t}$ s.t. $S \cup D \subset E \cup U$, there is a $C' \in \Gamma$ with $(E, U) \in e_t(C')$.

If $e_t(C, \Gamma) \neq \emptyset$, we call (C, Γ) a valid pair for t . We define $e'_t(C, \Gamma) = \{S \mid (S, D) \in e_t(C, \Gamma)\}$.

Given a pair (C, Γ) for a TD node t , the coloring C again represents admissible sets. Recall that an admissible set S is a semi-stable extension if there is no admissible set S'

whose range is a proper superset of the range of S . The colorings in Γ represent exactly such sets S' . Thus, in our algorithm for semi-stable semantics, we will again compute all colorings that represent admissible sets, but for each such coloring C we store colorings in Γ that cause all solution candidates represented by C to be rejected.

Definition 4.2. Let $(\mathcal{T}, \mathcal{X})$ be a TD of an AF $F = (A_F, R_F)$, and let Γ and Δ be sets of colorings for nodes t' and t'' , respectively, in \mathcal{T} . We define (similar to Definition 3.6):

$$\begin{aligned} \Gamma - a &= \{C' - a \mid C' \in \Gamma, C'(a) \neq \text{attc}\} \\ \Gamma +_{\text{attc}} a &= \{C' +_{\text{attc}} a \mid C' \in \Gamma, [C'] \not\vdash^{R_F} a, a \not\vdash^{R_F} [C']\} \\ \Gamma \dot{+}_{\text{in}} a &= \{C' \dot{+}_{\text{in}} a \mid C' \in \Gamma, [C'] \not\vdash^{R_F} a, a \not\vdash^{R_F} [C'], a \not\vdash^{R_F} a \text{ and } [[C']] = [[C' \dot{+}_{\text{in}} a]]\} \\ \Gamma \dot{+}_{\text{out}} a &= \{C' \dot{+}_{\text{out}} a \mid C' \in \Gamma\} \\ \Gamma \bowtie \Delta &= \{C' \bowtie C'' \mid C' \in \Gamma, C'' \in \Delta, [C'] = [C''] \text{ and } [[C']] = [[C'']]\} \end{aligned}$$

Definition 4.3. Let $(\mathcal{T}, \mathcal{X})$ be a TD of an AF F and let $t \in \mathcal{T}$ be a node with t' , t'' its possible children. Depending on the node type of t , we define a vpair for t .

LEAF: Each (C, Γ) with $C \in \mathcal{C}_t$ and $\Gamma = \{C' \in \mathcal{C}_t \mid [C] \cup [[C]] \subset [C'] \cup [[C']]\}$ is a vpair for t .

FORGET ($X_t = X_{t'} \setminus \{a\}$ for some argument a): If (C', Γ') is a vpair for t' and $C'(a) \neq \text{attc}$, then $(C' - a, \Gamma' - a)$ is a vpair for t .

INSERT ($X_t = X_{t'} \cup \{a\}$ for some argument a): If (C', Γ') is a vpair for t' and $C' \dot{+}_{\text{in}} a$ is a vcoloring for t , then $(C' \dot{+}_{\text{in}} a, (\Gamma' +_{\text{attc}} a) \cup (\Gamma' \dot{+}_{\text{in}} a))$ is a vpair for t ; if moreover $C' \dot{+}_{\text{out}} a$ is a vcoloring for t , then $(C' \dot{+}_{\text{out}} a, (\{C'\} +_{\text{attc}} a) \cup (\{C'\} \dot{+}_{\text{in}} a) \cup (\Gamma' +_{\text{attc}} a) \cup (\Gamma' \dot{+}_{\text{in}} a) \cup (\Gamma' \dot{+}_{\text{out}} a))$ is a vpair for t ; $(C' +_{\text{attc}} a, (\Gamma' +_{\text{attc}} a) \cup (\Gamma' \dot{+}_{\text{in}} a))$ is a vpair for t .

JOIN: If (C', Γ') is a vpair for t' , (C'', Γ'') is a vpair for t'' , $[C'] = [C'']$ and $[[C']] = [[C'']]$, then $(C' \bowtie C'', (\Gamma' \bowtie \Gamma'') \cup (\{C'\} \bowtie \Gamma'') \cup (\Gamma' \bowtie \{C''\}))$ is a vpair for t .

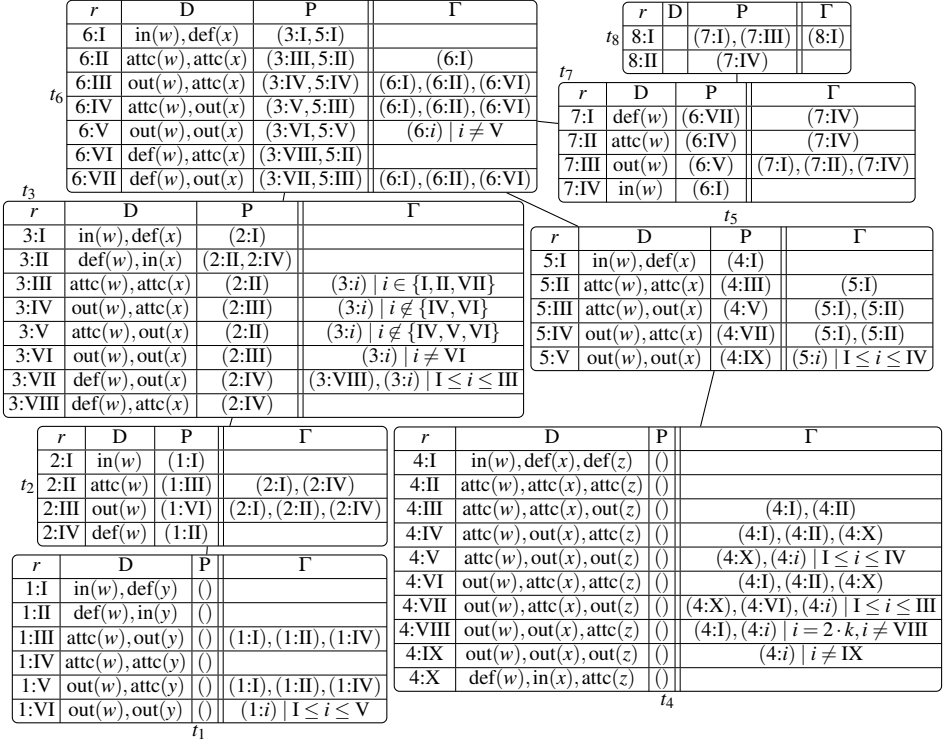
Example 4.4. Figure 3 illustrates the computation of the vpairs for the AF F and the TD from Figure 1. For each row r , we store a set Γ of references to rows r' from the same table such that r' represents admissible sets whose range is a proper superset of the range of each admissible set represented by r . By following the EPTs, we obtain exactly one set, namely $\{w\}$, which is in fact the only semi-stable extension.

For our correctness proof, we need another lemma and a proposition.

Lemma 4.5. Let $(\mathcal{T}, \mathcal{X})$ be a TD of an AF F and $t \in \mathcal{T}$. For each $X_{>t}$ -restricted admissible tuple (S, D) for $F_{\geq t}$, there is a coloring $C \in \mathcal{C}_t$ s.t. $(S, D) \in e_t(C)$.

Proposition 4.6. Let r be the root of a TD $(\mathcal{T}, \mathcal{X})$ of an AF $F = (A_F, R_F)$. It holds that $e'_r(\varepsilon, \emptyset) = \text{semi-stable}(F)$.

Proof sketch. Recall that $e'_r(\varepsilon) = \text{admissible}(F)$ (see Proposition 3.5, Definitions 3.1 and 3.3). To show $e'_r(\varepsilon, \emptyset) \subseteq \text{semi-stable}(F)$, let (S, D) be an arbitrary tuple s.t. $(S, D) \in e_r(\varepsilon, \emptyset)$. By condition (i) from Definition 4.1, S is admissible for $F_{\geq r} = F$. Furthermore, by (iii) and the fact that $\Gamma = \emptyset$ we conclude that there is no admissible tuple (E, U) for F with $E \cup U$ being a proper superset of $S \cup D$, i.e., S is a semi-stable extension of F . It remains to show that $e'_r(\varepsilon, \emptyset) \supseteq \text{semi-stable}(F)$. Let $S \in \text{semi-stable}(F)$ be an arbitrary semi-stable

Figure 3: DP computation of vpairs for $F = (A_F, R_F)$ w.r.t. T (see Figure 1).

extension of F with range $S_{R_F}^+$. We set $D = S_{R_F}^+ \setminus S$ to get the arguments that require defeating. It can be shown [14] that there exists a pair (C, Γ) such that $(S, D) \in e_r(C, \Gamma)$. Since the root node has an empty bag, $C = \varepsilon$, and furthermore, by condition (ii) from Definition 4.1 and the fact that $S \cup D$ is maximal (w.r.t. \subseteq) in F , $\Gamma = \emptyset$ holds as well. \square

Finally, we state the main theorem of this section, which analogously to Theorem 3.9 can be proved by structural induction [14].

Theorem 4.7. *Let $(\mathcal{T}, \mathcal{X})$ be a TD of an AF F . For each pair (C, Γ) for a node t , it holds that (C, Γ) is a valid pair for t if and only if (C, Γ) is a vpair for t .*

Together with Proposition 4.6, this guarantees that we can compute semi-stable extensions via vpairs. For the root r of a TD T of a framework F , we can compute vpairs in a bottom-up manner along T and thus obtain valid pairs. For enumerating $e_r'(\varepsilon, \emptyset)$ (i.e., semi-stable extensions of $F_{\geq r} = F$), we combine vpairs from all nodes of T .

Proposition 4.8. *Let T be a TD of width w for an AF F of size n . The DP computation for semi-stable extensions according to Definition 4.3 is feasible in FPT time, i.e., in time $f(w) \cdot n^{\mathcal{O}(1)}$, for some function f that depends only on w .*

Proof sketch. For the induction base, let t be a LEAF node. There are up to $\mathcal{O}(4^w)$ many vcolorings and vpairs for t , which can be computed in time $g(w) \cdot n^{\mathcal{O}(1)}$, for some function g . For the induction step, let t be a FORGET, INSERT or JOIN node and k be the number

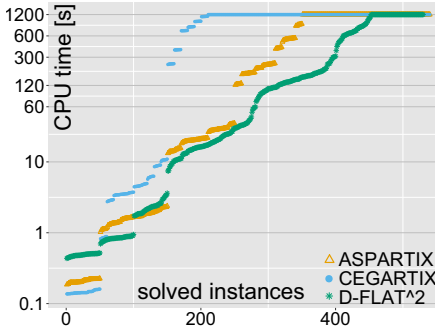


Figure 4: Plot on grids, TW ≤ 5, P 0.9.

	ASPARTIX		CEGARTIX		D-FLAT ²	
	t[s]	t outs	t[s]	t outs	t[s]	t outs
TW 4, P 0.9	443.9 (100)		712(160)		232.9	(49)
TW 4, P 0.7	280 (40)		555.4 (90)		34.4	(0)
TW 4, P 0.5	17.3 (0)		127 (20)		2.1	(0)
TW 5, P 0.9	551.4 (90)		938.5(170)		358.2	(29)
TW 5, P 0.7	365.9 (60)		611.6(110)		174.2	(4)
TW 5, P 0.5	41.6	(0)	278.7 (50)		88.2	(0)
TW 6, P 0.9	705.1	(100)	1200(200)		1168.5	(190)
TW 6, P 0.7	394.1	(57)	640.2 (70)		1137.3	(176)
TW 6, P 0.5	31.1	(0)	200.9 (30)		1076.7	(157)

Table 1.: Tabular results on grids.

of children of t , and assume that the time required for computing the tables of the children is $g_i(w) \cdot n^{\mathcal{O}(1)}$, for $1 \leq i \leq k$ and some function g_i . There are at most $\mathcal{O}(4^w)$ many vcolorings for t , and there can be at most $\mathcal{O}(4^w \cdot 2^{4^w})$ vpairs, which can be computed in time $g(w) \cdot n^{\mathcal{O}(1)} \cdot \prod_{i=1}^k g_i(w)$, for some function g , as described in Definition 4.3. Since we may assume that T has size $\mathcal{O}(n)$, the claim holds. \square

Our algorithm for semi-stable semantics can be easily turned into an algorithm for preferred semantics (as an alternative to [11]) by simplifying Definitions 4.1 and 4.3.

5. Preliminary Evaluation

We implemented the algorithm of Section 4 as an ASP encoding for the D-FLAT² system¹. This is an extended version of the D-FLAT system [1] and capable of efficiently solving problems from the second level of the polynomial hierarchy if the treewidth is small. We compared D-FLAT² 1.0.3 with CEGARTIX 0.4 [10] and ASPARTIX [13]. D-FLAT² internally uses ASP systems Gringo 4.5.4 and Clasp 3.1.4; we also used these versions for ASPARTIX. DYNPARTIX [3] cannot compute semi-stable extensions yet.

DP on TDs makes sense on instances with small treewidth, but usually yields poor performance if the treewidth is very large. For instances of the International Competition on Computational Models of Argumentation (ICCMA), we observed widths between 60 and 200, which is too much for our system. Hence we used randomly generated instances obtained from grids: Vertices are arranged on an $n \times m$ matrix, and edges connect horizontally, vertically and diagonally neighboring vertices with a certain probability (P).

We considered the problem of enumerating semi-stable extensions and compared the systems on instances with ≤ 70 nodes and treewidth (TW) ≥ 4 ; the observed widths of the TDs are ≤ 11 . Each D-FLAT² instance was run ten times with different TDs, and every run was limited to twenty minutes and three GB of memory. Figure 4 shows a cactus plot, and Table 1 lists running times in seconds and the number of timeouts. D-FLAT² exhibited the best performance, while CEGARTIX and ASPARTIX often time out, especially on larger instances. On the other hand, the performance of D-FLAT² becomes worse with increasing treewidth, thus reflecting our runtime estimation from Proposition 4.8. For detailed results, we refer to [14].

¹The system [2] is open source and available at <https://github.com/hmarkus/dflat-2>

6. Conclusion

We presented a new algorithm for computing semi-stable semantics using dynamic programming on tree decompositions that runs in linear time on AFs of bounded treewidth. For this purpose, we extended the concept of restricted-admissible sets [11]. Our experimental results indicate performance advantages over existing systems in case of bounded treewidth. It should be noted that such DP algorithms should not be seen as general solvers that outperform standard techniques on average. Instead, DP algorithms qualify as an alternative approach when instances are structurally rather “close” to trees.

Acknowledgements This research has been supported by the Austrian Science Fund (FWF) through projects Y698, P25607, I1102 and I2854.

References

- [1] M. Abseher, B. Bliem, G. Charwat, F. Dusberger, M. Hecher, and S. Woltran. D-FLAT: Progress report. Technical Report DBAI-TR-2014-86, TU Wien, 2014.
- [2] B. Bliem, G. Charwat, M. Hecher, and S. Woltran. D-FLAT²: Subset minimization in dynamic programming on tree decompositions made easy. In *ASPOCP*, 2015.
- [3] G. Charwat and W. Dvořák. dynPARTIX 2.0 - Dynamic programming argumentation reasoning tool. In *COMMA*, volume 245 of *FAIA*, pages 507–508. IOS Press, 2012.
- [4] R. G. Downey and M. R. Fellows. *Parameterized Complexity*. Monographs in Computer Science. Springer, 1999.
- [5] P. M. Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artif. Intell.*, 77(2):321–357, 1995.
- [6] P. E. Dunne. Computational properties of argument systems satisfying graph-theoretic constraints. *Artif. Intell.*, 171(10-15):701–729, 2007.
- [7] P. E. Dunne and T. J. M. Bench-Capon. Coherence in finite argument systems. *Artif. Intell.*, 141(1/2):187–203, 2002.
- [8] W. Dvořák, S. Ordyniak, and S. Szeider. Augmenting tractable fragments of abstract argumentation. *Artif. Intell.*, 186:157–173, 2012.
- [9] W. Dvořák, S. Szeider, and S. Woltran. Abstract argumentation via monadic second order logic. In *SUM*, volume 7520 of *LNCIS*, pages 85–98. Springer, 2012.
- [10] W. Dvořák, M. Järvisalo, J. P. Wallner, and S. Woltran. Complexity-sensitive decision procedures for abstract argumentation. *Artif. Intell.*, 206:53 – 78, 2014.
- [11] W. Dvořák, R. Pichler, and S. Woltran. Towards fixed-parameter tractable algorithms for abstract argumentation. *Artif. Intell.*, 186:1–37, 2012.
- [12] W. Dvořák and S. Woltran. Complexity of semi-stable and stage semantics in argumentation frameworks. *Inf. Process. Lett.*, 110:425–430, 2010.
- [13] U. Egly, S. A. Gaggl, and S. Woltran. Answer-set programming encodings for argumentation frameworks. *Argument and Computation*, 1(2):147–177, 2010.
- [14] M. Hecher. Optimizing Second-Level Dynamic Programming Algorithms; The D-FLAT² System: Encodings and Experimental Evaluation. Master’s thesis, Vienna University of Technology, 2015.
- [15] E. J. Kim, S. Ordyniak, and S. Szeider. Algorithms and complexity results for persuasive argumentation. *Artif. Intell.*, 175(9-10):1722–1736, 2011.
- [16] T. Kloks. *Treewidth: Computations and Approximations*, volume 842 of *LNCIS*. Springer, 1994.
- [17] R. Niedermeier. *Invitation to Fixed-Parameter Algorithms*. OUP, 2006.
- [18] N. Robertson and P. D. Seymour. Graph minors. III. Planar tree-width. *J. Comb. Theory, Ser. B*, 36(1):49–64, 1984.

An Ontology for Argumentation on the Social Web: Rhetorical Extensions to the AIF

Tom BLOUNT ^a, David E. MILLARD ^a and Mark J. WEAL ^a

^a*Web and Internet Science, University of Southampton, UK*

Abstract. A key area in the research agenda of modelling argumentation is to accurately model argumentation on the social web. In this paper we propose additional extensions to our ontology for argumentation on the social web (which integrates elements of the Argument Interchange Format and the Semantically Interlinked Online Communities project) for the purposes of modelling social and rhetorical tactics used in eristic or irrational arguments. We then present a review of these extensions from a panel of experts in the fields of argumentation modelling, web science, philosophy and open and linked data and discuss the value of modelling social argument, the challenges faced to create usable and accurate models and the completeness, clarity and consistency of our proposed additions.

Keywords. argumentation, rhetoric, social web, social media, ASWO, AIF, SIOC

1. Introduction

The social web and social media describe the relationships and communities that form over the world wide web, and the way in which people share content, ideas and information. As the social web becomes more and more ubiquitous, the potential for using it to investigate how truly massive communities interact, communicate and argue increases dramatically. A key area in the research agenda of modelling argumentation is to accurately model argumentation on the social web [1].

Currently, the majority of argumentation modelling tools and ontologies are primarily geared towards formal, rather than informal, argumentation. This approach is highly suited towards AI-based methods and can allow for reasoning over arguments to determine the final outcome, or the correct course of action. However, it neglects the set of informal social argumentation that, while virtually impossible to reason over, represents an equally valuable area of argumentation, particularly on the web. Rising levels of e-bile make understanding how otherwise civil discussions can evolve to turn abusive and toxic an important topic to consider [2].

In this work we build on our previous work of bringing together the Semantically Interlinked Online Communities (SIOC) project and the Argument Interchange Format (AIF) [3] and our extensions of this model in the Argumentation on the Social Web Ontology (ASWO) to incorporate rhetorical attacks and declarations of support [4] with additional features to capture some of the extra-logical tactics used in informal argumen-

tation. We then conducted an expert review of these additions to determine how they affected the clarity, completeness and consistency of the ontology, and the overall inherent value in attempting to model this form of social argumentation.

2. Background

Argumentation can, very broadly, be separated into two categories: dialectic, and eristic. The terms dialectic and eristic were coined in Ancient Greece to describe modes of argumentation with different goals and were popularised in Plato's *Republic* [5], and more recently by Walton and Krabbe [6,7]. A dialectic argument takes the form of two or more parties engaged in rational discourse with the aim of either discovering the truth behind a particular matter, or formulating a solution or resolution for a set of circumstances [8]. For example, an academic presenting their findings to an audience of their peers and rationalising that they are indeed valid, is an example of a dialectic argument, but so too can be a group of friends trying to decide on the best place to have lunch. These arguments tend to rely heavily on the weight of facts and evidence, although personal preference can still hold some sway (for example, a free market vs. protectionism or take-away vs. a restaurant). In contrast, an eristic argument is an argument in which there is no clear resolution in the minds of the participants: they are not motivated by solving a problem, or convincing their opponent [8]. Instead, they may be quarrelling for its own sake as a form of catharsis [9], or to be seen to "win" the argument in the eyes of any spectators [10]. As a result, these arguments favour more emotive language and facts may be deliberately distorted to serve a participant's agenda.

Many theoretical models of argumentation are based on the assumption of a dialectic argument, which is useful when building systems to aid automated reasoning to discover the final resolution to a discussion. However, on the social web there is a clear proliferation of eristic argumentation that often will not have a resolution. Nonetheless, this style of argumentation is also important to consider.

The social web presents a number of challenges for extracting and analysing arguments, particularly due to the lack of clear "indicators" of argument or structure. This problem is compounded by the type of language used; often consisting of highly informal language, incorporating quickly evolving slang and irregular punctuation and grammar [11].

3. Existing Models

3.1. *Argument Interchange Format*

The Argument Interchange Format (AIF) is a framework for representing argumentation as a directed graph [12], modelling information "nodes" and the relationships (such as inference or conflict) between them. In their work on an extension to the AIF, dubbed AIF+, Reed et al. differentiate between these logical relations and the actual words spoken during the debate [13] and introduce a web-based tool, Online Visualisation of Argument (OVA+), to annotate, display and share argumentation on the web [14].

Information nodes (*I-nodes*) represent a (purported) piece of information, data, or claim. Scheme nodes denote a logical connection between information nodes,

whether an inference (*RA-node*), a conflict (*CA-node*), or a value preference (*PA-node*). Illocutionary-Anchor nodes (*YA-nodes*) tie the information and logical structure of an argument with the spoken or written locution. Locution nodes (*L-nodes*) represent the actual words that are spoken or written by participants. Transition nodes (*TA-nodes*) represent links between locutions. However, this is adapted by the ASWO to instead denote locutions that do not add information nodes, but still further the debate (such as prompting for more details, evidence, etc.).

3.2. Semantically Interlinked Online Communities

The Semantically Interlinked Online Communities project (SIOC), a semantic-web vocabulary for representation social media, aims to enable the cross-platform, cross-service representation of data from the social web [15]. This allows for semantic representations of Sites, which hold Forums, which contain Posts, authored by a UserAccount (explicitly *not* a person, as a person can own and manage more than one UserAccount). SIOC also allows the modelling of replies between posts.

4. Proposed Additions

Previously, we examined how to link the AIF and SIOC to provide further contextual information about arguments on social media [3]. We now propose several additional nodes to aid modelling rhetorical or “extra-logical” argument with the ASWO.

One of the key additions is the Persona node (*P-nodes*): this represents the “character” that a person assumes during the discussion. For example, a person may argue in a different fashion in a debate about music than they would about technical expertise. This allows one UserAccount to have many Personas where necessary. The inverse, linking one Persona to multiple UserAccounts, is also possible and can be used to represent a participant attempting to artificially solidify their position by creating multiple accounts.

Faction and Audience nodes (*F-* and *A-nodes*) represent abstract groups of Personas; a Faction is any grouping of Personas and can potentially include those outside the Thread, whereas the Audience represents all Personas currently participating in, or observing, the discussion.

Personal Support and Personal Conflict nodes (*PS-* and *PC-nodes*) allow a means of representing support and attack that does not rely on logic and instead uses rhetorical force, social pressure or some other form of “extra-logical” tactic.

Implication nodes (*Im-nodes*) allow analysts to represent a participant implying a relationship between two (or more) nodes, such as Personas. These can be combined with the Personal Support/Conflict nodes to indicate whether the implication is positive or negative.

5. Expert Review

Six experts, from the fields of argumentation systems, web science, philosophy, and linked data, were chosen to review these proposed additions to the model. Experts A and B have a background in argumentations systems and modelling argumentation, and are familiar with the AIF. Expert A is a computer science lecturer whose research is con-

1. **User 1:** *The tech industry is often biased against women*
User 2: *@User1 You would say that, you're a woman*
User 3: *@User1 **** off and die you ***** nazi before I come and **** you up*

2. **User 1:** *Guns killed 33,000 people last year, they need to be banned*
User 2: *@User1 And a lot of those were minors*
User 3: *@User2 According to who?*

3. **User 1:** *What does Barack Obama call illegal aliens? Undocumented democrats!*
User 2: *@User1 You're so stupid you probably went to the library to find Facebook*

Figure 1. The three argumentation samples the experts were asked to model

cerned with argumentation-based models of communication and formal reasoning, with interests in AI and behaviour change. Expert B is a post-doctoral researcher with degrees in library and information science, mathematics, and liberal arts whose thesis focused on the problem of analysing, integrating, and reconciling information in online discussions. Expert C is a web-science graduate student, researching the relation between social structures in virtual worlds and the real world, with a focus on practices of gender and power. Expert D is a philosophy graduate student, specialising in ethics, moral obligations and with a background in argumentation and formal logic. Experts E and F are specialists in the area of open and linked data working in web and data innovation and development. Expert E is an institutional open data specialist and Expert F is a senior technical specialist.

Each expert was provided with a document describing the background of this area and an overview of the existing models. They were then asked to model three argumentation samples shown in Figure 1, illustrating a variety of different rhetorical structures, by speaking aloud and/or sketching with pen and paper. They were then shown the additions to the model, and asked to model the three argumentation samples again. They were then asked a series of semi-structured question aimed to evaluate their thoughts on how best (and whether) to model social (and anti-social) argumentation, the completeness of the ontology, the clarity of the ontology and the consistency of the ontology.

5.1. Results and Analysis

Table 1 shows an overview of the key points discussed by the experts along the themes of modelling social argumentation, completeness, clarity and consistency (and relevant sub-themes).

5.1.1. Social Argumentation

Each of the experts agreed that there was value in modelling social argumentation, Expert F going so far as to say they believed there was no argument that didn't have social components. Expert D discussed how understanding the nuances of how people argue socially could lead to ways of helping or encouraging them to argue "better", in a more cooperative or polite manner.

The challenges of modelling social argumentation the experts foresaw were mostly a question of scale. In part, the sheer volume of data in a social media discussion can be

Table 1. Summary of experts' opinions on key aspects of ASWO

Theme	Sub-theme	Comments
Social Argumentation	<i>Value</i>	<p>"...if we're going to have a realistic model of how people argue, we've got to look at how people really argue rather than how our "ideal reasoner" would argue" – Expert A</p> <p>"I think modelling social argumentation is very important...I want to say it's useful in trying to help people argue 'better'." – Expert D</p>
	<i>Challenges</i>	<p>"Even in quite a simple back-and-forth argument, there's quite a lot going on...scale is a challenge" – Expert C</p> <p>"...enthymemes, humour, there's lots of missing information, there's lots of playing to particular audiences...there are lots of things that are current events or would only make sense to a particular group" – Expert B</p>
	<i>Abuse/Threats</i>	<p>"I, personally, tend to ignore all of those because I'm...focusing on the informal proof structures" – Expert A</p> <p>"...it's hard to exclude them...if you think about what you're going to do with the model...do you want to retrieve threatening and abusive comments? Well you might want to exclude them from being retrieved, which also makes it relevant to model that" – Expert B</p>
Completeness	<i>Implicit/Explicit Premises</i>	<p>"I think when people model arguments it's pretty common to infer the reading, and what's interesting is that there can be multiple readings. So it wouldn't be wrong to...put in some interpretation, as long as it's clear it's an interpretation and there can be others. " – Expert B</p>
	<i>Social Meta-Data</i>	<p>"One other thing... is other people's opinions of statements. A lot of systems have thumbs up and thumbs down...what you need is, I think, an audience response" – Expert F</p>
Clarity	<i>Generalisation</i>	<p>"If anything I think maybe your default conflict is a superclass - everything is a conflict, and one of the subclasses is a...rational argument. But then you've also got personal attack, ad hominem...these are all alternatives to rational argument, but at the default it might be worth allowing modelling of a conflict. Not a conflict as it is in the original model, but as a superclass of interaction." – Expert F</p>
Consistency	<i>Internal consistency</i>	<p>"whenever you try to model anything in a formalised system...if you give two people the same thing...unless it's something really simple, they will always find two different ways of modelling it" – Expert E</p> <p>"...rather than having the minimal number of nodes and encouraging people to just misuse them, I would rather say 'Here's a definite type of argumentation we want to capture and share...'" – Expert A</p>
	<i>External consistency</i>	<p>"Consistent with [the AIF], maybe not, but building on? Definitely" – Expert C</p>

overwhelming, particularly when considering the speed with which it can grow, but also in terms of the variety of information, which is often contextual, such as references to current events, or cultural “in-jokes”.

Experts A and D explained that they would not consider abusive argumentation as a valid when modelling an argumentation structure (as they focused broadly on dialectic arguments and that was the current standard for their domain), although they agreed it was a potentially valuable area to explore. Expert B explained that it depended very much on the purpose of the model — in some cases it may be important to model threatening and abusive attacks specifically so they can be excluded when presenting the model to users. Expert E also noted that excluding this type of argument can lead to confusion if a particular abusive comment changes the course of the argument, or causes the quality of the rest of the discussion to degenerate.

5.1.2. Completeness

Experts A and B both made explicit mention of the ability to mark certain posts as being in direct response to other participants in the discussion as a useful addition to argumentation frameworks.

Expert B noted that as many annotations have the potential to be subjective, it would be possible to extend this to include further subjective annotations such as an analyst’s confidence in a particular reading of an inference. Expert C had similar views and discussed including mappings of a participant’s agreement or disagreement with key positions in the dialogue as well.

Expert F discussed the potential for an “activity” score for each location, derived from the social meta-data of each post (e.g. number of replies, number of up- or down-votes or number of retweets); this metric could be derived on a per-purpose basis to allow analysts to correctly categorise different platforms for their own needs, and to highlight key areas of the discussion that had solicited or stimulated large amounts of discussion.

Broadly, all experts agreed that to adequately model social argument that it was necessary to include further context about the participants, such as demographic information where available, such as by linking the SIOC UserAccount to a FOAF Agent, or additional information about key events related to the discussion to maintain relevance of the model for future analysis, and to limit the number of assumptions needed to be made by analysts.

5.1.3. Clarity

Expert D was concerned that, when faced as an analyst with a statement that appeared ambiguous (for example, a statement of support that could be interpreted as genuine or sarcastic) they may struggle to accurately and objectively model it, and proposed a means of allowing analysts to mark such relations as existing without committing to associating them with either a support or an attack.

Expert F proposed a similar solution, by means of generalising the model to include super-classes of Support and Conflict. “Personal” conflict, for example, is perhaps too specific a name for all non-logical conflicts: there are rhetorical attacks that can target institutions or accounts run by software, but also, importantly, positions and information. These Support and Conflict super-classes would encompass Logical Support/Conflict and Rhetorical Support/Conflict and could then be further sub-classed to provide more specific instances of each, where apparent, allowing analysts to defer when unsure.

5.1.4. Consistency

The majority of experts felt that these additions to the ASWO were consistent with the nodes used in the AIF. However, Experts C and F disagreed, pointing to the fact that the ASWO was intentionally inconsistent with the AIF because they were developed for different purposes.

In terms of inter-rater reliability — whether two analysts attempting to model the same argument would reach the same result — the experts were much more divided. While they agreed that the objective parts of the model (i.e. the locutions, user account and, in most circumstances, the persona) could be modelled identically (and in most cases, automated), Experts C and B felt that both analysts would reach the same conclusion overall with minor deviations, whereas Experts A, D and E disagreed, stating there was too much subjective information to model identically. Expert A felt that the analyst would naturally perceive the argument through their own lens of cultural and social context and Expert D noted the different levels of detail an analyst may choose to use, whether focusing only on premises that have been explicitly stated, or including additional implicit information.

How important this is was also a matter of some debate: Experts B and C felt that it was likely there would (and should) be one “correct” representation of an argument. Experts D and F agreed to an extent, citing their proposals for handling ambiguous content being able to aid annotators in this regard, so that if the model could not be complete, it could be consistent. Expert A felt that ideally analysts should reach the same conclusion but in practice, the subjective nature of the task might make this impossible. Expert E felt the consistency of annotators would, in practice, be less important and would be a factor of the intended purpose of the model.

6. Conclusions

In this paper we provide further extensions to the ASWO to incorporate other modes of rhetorical persuasion that contrast with logical argument. We conducted an expert review that highlighted some key strengths of this model, such as the ability to model directed replies, the ability to model the audience and the ability to model instances of irrational and eristic argument that were previously difficult or impossible to achieve with the AIF alone.

This review highlights some current limitations of the ASWO framework as it stands that will need to be addressed to further improve the model. Firstly, the issue of scalability: annotating web-based argumentation in this manner remains a high-cost affair in terms of knowledge and time. Future work will examine how suitable a crowd-sourced annotation approach is, with respect to accuracy and inter-rater reliability. Secondly, automation: because social argumentation can rely heavily on nuanced contextual information (such as the ability to recognise humour, sarcasm or references to current events) it is likely impossible to model it in such a way that it could be automatically reasoned over. However, because the ASWO provides additional information about rhetorical tactics in use, it provides human analysts the means to explore the resulting structure in greater detail and context. This can also potentially be used to highlight areas of particular interest, or assist in community decision-making environments.

The review also highlights useful directions of further work, such as including further contextual information such as participant demographics or social meta-data, or generalising the ontology further. It also lays groundwork for an investigation into how such rhetorical structures are used on different social web platforms. By using this extended framework, we aim to determine if and how the perceived contribution and value of a comment correlates to the dialectic and eristic content.

Our hope is that these developments lead to a means of more accurately modelling social argumentation which in turn provides a path to creating tools to allow social media users to critically analyse discussions in progress and to encourage them to engage with debates in good faith.

Acknowledgements

This work is funded by a scholarship provided by the UK Engineering and Physical Sciences Research Council. Thanks also go to the reviewers for their detailed, constructive and helpful reviews.

References

- [1] J. Schneider, T. Groza, and A. Passant, "A review of argumentation for the Social Semantic Web," *Semantic Web*, vol. 4, no. 2, pp. 159–218, 2013.
- [2] E. A. Jane, "Your a Ugly, Whorish, Slut" Understanding E-bile," *Feminist Media Studies*, vol. 14, no. 4, pp. 531–546, 2014.
- [3] T. Blount, D. E. Millard, and M. J. Weal, "Towards Modelling Dialectic and Eristic Argumentation on the Social Web," in *14th workshop on Computational Models of Natural Argument*, 2014.
- [4] T. Blount, D. E. Millard, and M. J. Weal, "An Investigation into the Use of Logical and Rhetorical Tactics within Eristic Argumentation on the Social Web," in *ACM Conference on Hypertext and Social Media*, 2015.
- [5] Plato, *Book V. The Republic*, Basic Books, 380BC. (Bloom, A.D. Trans. 1991).
- [6] D. Walton and E. C. Krabbe, *Commitment in Dialogue: Basic Concepts of Interpersonal Reasoning*. Albany, NY: State University of New York Press, 1995.
- [7] D. N. Walton, *The New Dialectic: Conversational Contexts of Argument*. University of Toronto Press, 1998.
- [8] G. B. Kerferd, *The Sophistic Movement*. Cambridge University Press, 1981.
- [9] J. Schneider, S. Villata, and E. Cabrio, "Why did they post that argument? Communicative Intentions of Web 2.0 Arguments," in *Arguing on the Web 2.0*, (Amsterdam), SINTELNET, European Network for Social Intelligence, 2014.
- [10] C. Jørgensen, "Public Debate – An Act of Hostility?," *Argumentation*, vol. 12, no. 4, pp. 431–443, 1998.
- [11] J. Schneider, B. Davis, and A. Wyner, "Dimensions of argumentation in social media," *Lecture Notes in Computer Science*, vol. 7603, pp. 21–25, 2012.
- [12] C. Chesnevar, J. McGinnis, S. Modgil, I. Rahwan, C. Reed, G. Simari, M. South, G. Vreeswijk, and S. Willmott, "Towards an argument interchange format," *Knowledge Engineering Review*, vol. 21, no. 4, pp. 293–316, 2006.
- [13] C. Reed, S. Wells, J. Devereux, and G. Rowe, "AIF+: Dialogue in the Argument Interchange Format.," *FRONTIERS IN ARTIFICIAL INTELLIGENCE AND APPLICATIONS*, vol. 172, p. 311, 2008.
- [14] M. J. J. L. C. REED, "Ova+: An argument analysis interface," in *Computational Models of Argument: Proceedings of COMMA*, vol. 266, p. 463, 2014.
- [15] J. G. Breslin, S. Decker, A. Harth, and U. Bojars, "SIOC: An Approach to Connect Web-Based Communities," *International Journal of Web Based Communities*, vol. 2, no. 2, pp. 133–142, 2006.

Abstract Dialectical Argumentation Among Close Relatives

ALEXANDER BOCHMAN

*Computer Science Department,
Holon Institute of Technology, Israel*

Abstract. We establish a uniform modular translation of Abstract Dialectical Frameworks into the formalism of the causal calculus, and discuss the correspondences this translation creates between a number of semantics suggested for ADFs and their causal counterparts.

Keywords. formal argumentation, abstract dialectical frameworks, causal reasoning

1. Introduction

Abstract Dialectical Frameworks (ADFs) have been introduced in [8,7] as an abstract argumentation formalism purported to capture more general forms of argument interaction than just attacks among arguments, which form the basis of the original, Dung's argumentation frameworks. To achieve this, each argument in an ADF is associated with an *acceptance condition*, which is some propositional function determined by arguments that are linked to it. Using such acceptance conditions, ADFs allow to express that arguments may jointly support another argument, or that two arguments may jointly attack a third one, and so on. Dung's argumentation frameworks are recovered in this setting by acceptance condition saying that an argument is accepted if none of its parents is.

The authors of ADFs have repeatedly stressed that they primarily see their formalism not as a knowledge representation tool, but rather as a convenient and conceptually neutral abstraction tool ('argumentation middleware') that is intended to encompass a broad range of more specific argumentation and other nonmonotonic formalisms. On the other hand, [16] has considered ADFs as a particular knowledge representation formalism. In our study also, we will view ADFs as a specific knowledge representation formalism and show its close conceptual connections with the formalism of causal reasoning. This will also help us to single out some of the basic principles behind the construction of ADFs and their semantics, as well as to situate the latter in the range of closely related KR formalisms.

The plan of the paper is as follows. We present first a brief description of the formalism of ADF and the relevant parts of the causal calculus. Then we will establish a simple modular translation of ADFs into the causal calculus, and explore the counterparts of the main semantics introduced for ADFs under this translation. It will be shown, in particular, that the basic operator Γ of ADFs can be significantly enhanced by taking into account disjunctive information. This translation will also suggest a natural gener-

alization of ADFs to a general rule-based formalism that will already subsume Logic Programming.

2. Preliminaries I: Abstract Dialectical Frameworks

An abstract dialectical framework (ADF) is a directed graph whose nodes represent statements or positions which can be accepted or not. The links represent dependencies: the status of a node s only depends on the status of its parents (denoted $par(s)$), that is, the nodes with a direct link to s . In addition, each node s has an associated acceptance condition C_s specifying the exact conditions under which s is accepted. C_s is a function assigning to each subset of $par(s)$ one of the truth values \mathbf{t}, \mathbf{f} . Intuitively, if for some $R \subseteq par(s)$ we have $C_s(R) = \mathbf{t}$, then s will be accepted provided the nodes in R are accepted and those in $par(s) \setminus R$ are not accepted.

Definition 1. An abstract dialectical framework is a tuple $D = (S, L, C)$ where

- S is a set of statements (positions, nodes),
- $L \subseteq S \times S$ is a set of links,
- $C = \{C_s\}_{s \in S}$ is a set of total functions $C_s : 2^{par(s)} \rightarrow \{\mathbf{t}, \mathbf{f}\}$, one for each statement s . C_s is called acceptance condition of s .

A more ‘logical’ representation of ADFs can be obtained simply by assigning each node s a *classical* propositional formula corresponding to its acceptance condition C_s (see [11]). In this case we can tacitly assume that the acceptance formulas implicitly specify the parents a node depends on. It is then not necessary to give the links L , so an ADF D amounts to a tuple (S, C) where S is a set of statements, and C is a set of propositional formulas, one for each statement from S . The notation $s[C_s]$ is used by the authors to denote the fact that C_s is the acceptance condition of s .

A two-valued interpretation v is a (two-valued) *model* of an ADF (S, C) whenever for all statements $s \in S$ we have $v(s) = v(C_s)$, that is, v maps exactly those statements to true whose acceptance conditions are satisfied under v . This notion of a model provides a natural semantics for ADFs. In addition to this semantics, however, the authors define appropriate generalizations for all the major semantics of Dung’s argumentation frameworks. In [7], all these semantics are defined by generalizing the two-valued interpretations to three-valued ones. All of them are formulated using the basic operator Γ_D over three-valued interpretations that was introduced, in effect, already in [8]. In the formulation of [7], for an ADF D and a three-valued interpretation v , the interpretation $\Gamma_D(v)$ is given by the mapping

$$s \mapsto \prod \{w(C_s) \mid w \in [v]_2\},$$

where \prod is the product operator on interpretations, while $[v]_2$ is the set of all two-valued interpretations that extend v .

For each statement s , the operator Γ_D returns the consensus truth value for its acceptance formula C_s , where the consensus takes into account all possible two-valued interpretations w that extend the input valuation v . If v is two-valued, we get $\Gamma_D(v)(s) = v(C_s)$, so v is a two-valued model for D iff $\Gamma_D(v) = v$. In other words, two-valued models of D are precisely those classical interpretations that are fixed points of Γ_D .

The *grounded model* of an ADF D can now be defined as the least fixpoint of Γ_D . This fixpoint is in general three-valued, and it always exists since the operator Γ_D is monotone in the information ordering \leq_i , as shown in [8]. This grounded semantics is viewed by the authors as the greatest possible consensus between all acceptable ways of interpreting the ADF at hand¹.

The operator Γ_D also provides a proper basis for defining admissible, complete and preferred semantics for arbitrary ADFs.

Definition 2. A three-valued interpretation v for an ADF D is

- admissible iff $v \leq_i \Gamma_D(v)$;
- complete iff $\Gamma_D(v) = v$;
- preferred iff it is \leq_i -maximal admissible.

As can be shown, the above definitions provide proper generalizations of the corresponding semantics for Dung's argumentation frameworks and, moreover, preserve much of the properties and relations of the latter. Thus, the grounded semantics is always a complete model, and each complete model is admissible. In addition, as it is the case for AFs, all preferred models are complete, the grounded model is the \leq_i -least complete model, and the set of all complete models forms a complete meet-semilattice with respect to the information ordering \leq_i .

In [8], the standard Dung semantics of stable extensions was generalized only to a restricted type of ADFs called bipolar, but [7] has suggested a new definition that avoids unintended features of the original definition, and covers arbitrary ADFs, not only bipolar ones (see also [16]). This new definition is based on the notion of a *reduct* of an ADF, similar to the Gelfond-Lifschitz transformation of logic programs. We will discuss the representation of the stable semantics in ADFs later in this study.

3. Preliminaries II: Causal Reasoning

The causal calculus has been introduced in [14] as a nonmonotonic formalism purported to serve as a logical basis for reasoning about action and change. This line of research has led to the action description language $C+$, which is based on this calculus [12]. A logical basis of the causal calculus was described in [1], and it has been argued in [2] that this calculus is not necessarily restricted to temporal domains, but has actually a vast potential and representation capabilities for serving as a general-purpose nonmonotonic formalism (see also [3,4,5]).

We will assume in this section that our underlying language is an ordinary classical propositional language with the usual connectives and constants $\{\wedge, \vee, \neg, \rightarrow, \mathbf{t}, \mathbf{f}\}$. \models and Th will stand, respectively, for the classical entailment and the associated logical closure operator. We will reserve also the letters p, g, r, \dots for denoting propositional atoms, while A, B, C, \dots will denote arbitrary classical propositions of the language.

By a *causal rule* we will mean an expression of the form $A \Rightarrow B$ (" A causes B "), where A and B are propositional formulas. A *causal theory* is a set of causal rules. A causal rule $A \Rightarrow B$ is *determinate*, if B is a literal. A determinate causal theory is a set of determinate causal rules.

¹We will qualify this claim in what follows.

We will begin with a general notion of production inference which is actually just a slight modification of the input-output logic from [13].

Definition 3. A *production inference relation* is a binary relation \Rightarrow on the set of classical propositions satisfying the following conditions:

(Strengthening) If $A \models B$ and $B \Rightarrow C$, then $A \Rightarrow C$;

(Weakening) If $A \Rightarrow B$ and $B \models C$, then $A \Rightarrow C$;

(And) If $A \Rightarrow B$ and $A \Rightarrow C$, then $A \Rightarrow B \wedge C$;

(Truth) $t \Rightarrow t$;

(Falsity) $f \Rightarrow f$.

A characteristic property of production inference is that the reflexivity postulate $A \Rightarrow A$ does not hold for it.

We extend causal rules to rules having arbitrary sets of propositions as premises using the familiar compactness recipe: for any set u of propositions, we define

$$u \Rightarrow A \equiv \bigwedge a \Rightarrow A, \text{ for some finite } a \subseteq u$$

$\mathbb{C}(u)$ will denote the set of propositions caused by u , that is

$$\mathbb{C}(u) = \{A \mid u \Rightarrow A\}$$

As could be expected, the causal operator \mathbb{C} plays much the same role as the usual derivability operator for consequence relations. Note that $\mathbb{C}(u)$ is always a deductively closed set (due to And, Weakening, and Truth). Also, it satisfies monotonicity:

Monotonicity If $u \subseteq v$, then $\mathbb{C}(u) \subseteq \mathbb{C}(v)$.

Actually, due to compactness, \mathbb{C} is not only monotonic, but also a continuous operator. Still, it is not inclusive, that is, $u \subseteq \mathbb{C}(u)$ does not always hold. Also, it is not idempotent, that is, $\mathbb{C}(\mathbb{C}(u))$ can be distinct from $\mathbb{C}(u)$.

3.1. Regular, basic and causal inference

A production inference relation is *regular* if it satisfies the following well-known rule:

(Cut) If $A \Rightarrow B$ and $A \wedge B \Rightarrow C$, then $A \Rightarrow C$.

Cut is one of the basic rules for ordinary consequence relations. In the context of production inference it plays the same role, namely, allows for a reuse of produced propositions as premises in further productions². It corresponds to the following characteristic condition on the causal operator:

$$\mathbb{C}(u \cup \mathbb{C}(u)) \subseteq \mathbb{C}(u).$$

Following [13], a production inference relation will be called *basic* if it satisfies

²Such production relations correspond to input-output logics with reusable output in [13].

(Or) If $A \Rightarrow C$ and $B \Rightarrow C$, then $A \vee B \Rightarrow C$.

For basic production inference, the set of propositions caused by a theory u coincides with the set of propositions that are caused by every world containing u :

$$\mathbb{C}(u) = \bigcap \{ \mathbb{C}(\alpha) \mid u \subseteq \alpha \text{ \& } \alpha \text{ is a world} \}$$

Another important fact about basic production inference is that any causal rule is reducible to a set of *clausal* rules of the form $\bigwedge l_i \Rightarrow \bigvee l_j$, where l_i, l_j are classical literals.

Finally, a production inference relation will be called *causal* if it is both basic and regular.

3.2. General nonmonotonic semantics

Production inference determines a natural nonmonotonic semantics, and provides thereby a logical basis for a particular form of nonmonotonic reasoning.

Definition 4. • A set u of propositions is an *exact theory* of a production inference relation if it is consistent, and $u = \mathbb{C}(u)$.

- A set u of propositions is an *exact theory of a causal theory* Δ , if it is an exact theory of the least production relation \Rightarrow_Δ that includes Δ .
- A *general nonmonotonic semantics* of a causal theory is the set of all its exact theories.
- A *causal nonmonotonic semantics* of a causal theory is the set of its exact theories that are worlds (complete deductively closed sets).

An exact theory describes an information state in which every proposition is caused, or *explained*, by other propositions accepted in this state. Accordingly, restricting our universe of discourse to exact theories amounts to imposing a kind of an *explanatory closure assumption*. Namely, it amounts to requiring that any accepted proposition should also have an explanation, or justification, for its acceptance.

The general nonmonotonic semantics is indeed nonmonotonic in the sense that adding new causal rules to a causal theory may lead to a nonmonotonic change of the associated semantics, and thereby to a nonmonotonic change in the derived information. This happens even though the causal rules themselves are monotonic, since they satisfy Strengthening (the Antecedent).

Exact theories are consistent fixed points of the operator \mathbb{C} . Since the latter operator is monotonic and continuous, exact theories (and hence the nonmonotonic semantics) always exist. Moreover, there always exists a least exact theory. In addition, the union of any chain of exact theories (with respect to set inclusion) is an exact theory, so any exact theory is included in a maximal such theory.

It has been shown in [2] (using an appropriate strong equivalence theorem) that regular production inference provides an adequate and maximal logical framework for reasoning with general exact theories.

As an interesting application of this result for our present study, it can be shown that the least exact theory of a regular inference relation coincides with the set of propositions that are caused by truth \mathbf{t} . Thus, we obtain the following

Lemma 1. *The least exact theory of a causal theory Δ coincides with the set of propositions that are provable from Δ using the postulates of regular production inference.*

Finally, it has been shown that the *causal* nonmonotonic semantics, as defined above, is equivalent to the original semantics described in [14]. In addition, as a consequence of the corresponding strong equivalence theorem, it has been shown that the full system of causal inference relations (that is both regular and basic) constitutes an adequate logical basis for reasoning with respect to this semantics.

4. The Causal Representation of ADFs

Now we are going to provide a uniform and modular translation of ADFs into the causal calculus. An essential precondition of this causal representation, however, will consist in transforming the underlying semantic interpretations of ADFs in terms of three-valued models (used, e.g., in [7]) into ordinary classical logical descriptions. This latter transformation will also allow us to clarify to what extent the various semantics suggested for ADFs admit a classical logical reading. In fact, the very possibility of such a classical reformulation stems from the crucial fact that the basic operator Γ of an ADF, described earlier, is defined, ultimately, in terms of ordinary classical interpretations extending a given three-valued one. Nevertheless, our reformulation will also reveal a significant discrepancy between these semantics and their immediate causal counterparts.

4.1. Three-valued interpretations versus classical theories

To begin with, any three-valued interpretation v on the set of statements S can be faithfully encoded using an associated set of literals $[v] = S_0 \cup \neg S_1$ such that $S_0 = \{p \in S \mid v(p) = \mathbf{t}\}$ and $S_1 = \{p \in S \mid v(p) = \mathbf{f}\}$. Moreover, this set of literals generates a unique deductively closed theory $\text{Th}([v])$ that corresponds in this sense to the source three-valued interpretation v . Conversely, let us say that a deductively closed set u is a *literal theory*, if it is a deductive (classical) closure of some set of literals. Then the latter set of literals will correspond to a unique three-valued interpretation v such that $u = \text{Th}([v])$. These simple facts establish a precise bi-directional correspondence between three-valued interpretations and classical literal theories. Moreover, we will see in what follows that the main operator Γ of ADFs will correspond under this reformulation to a ‘literal’ restriction of the causal operator \mathbb{C} of basic production inference.

4.2. Acceptance conditions as causal rules

As our starting point, we note a striking similarity between the official definition of an ADF and the notion of a *causal model*, used by Judea Pearl in [15].

According to [15, Chapter 7], a causal model is a triple $M = \langle U, V, F \rangle$, where

- (i) U is a set of *background* (or *exogenous*) variables.
- (ii) V is a set $\{V_1, V_2, \dots, V_n\}$ of *endogenous* variables that are determined by variables in $U \cup V$.
- (iii) F is a set of functions $\{f_1, f_2, \dots, f_n\}$ such that each f_i is a mapping from $U \cup (V \setminus V_i)$ to V_i , and the entire set, F , forms a mapping from U to V .

Symbolically, F is represented as a set of equations

$$v_i = f_i(pa_i, u_i) \quad i = 1, \dots, n$$

where pa_i is any realization of the unique minimal set of variables PA_i in $V \setminus \{V_i\}$ (parents) sufficient for representing f_i , and similarly for $U_i \subseteq U$.

In Pearl's account, every instantiation $U = u$ of the exogenous variables determines a particular "causal world" of the causal model. Such worlds stand in one-to-one correspondence with the solutions to the above equations in the ordinary mathematical sense. However, structural equations also encode causal information in their very syntax by treating the variable on the left-hand side of $=$ as the effect and treating those on the right as causes. Accordingly, the equality signs in structural equations convey the asymmetrical relation of "is determined by".

Being restricted to the classical propositional language, Pearl's notion of a causal model can be reduced to the following notion of a Boolean causal model, used in [6]:

Definition 5. Assume that the set of propositional atoms is partitioned into a set of *background* (or *exogenous*) atoms and a finite set of *explainable* (or *endogenous*) atoms.

- A *Boolean structural equation* is an expression of the form $p = F$, where p is an endogenous atom and F is a propositional formula in which p does not appear.
- A *Boolean causal model* is a set of Boolean structural equations $p = F$, one for each endogenous atom p .

As can be seen, the above definition is much similar to the logical reformulation of ADFs, with equations $p = F$ playing essentially the same role as the acceptance conditions $p[F]$. The differences are that only endogenous atoms are determined by their associated conditions in causal models, but on the other hand, there are no restrictions on appearances of atoms on both sides in ADF's acceptance conditions. Furthermore, plain (two-valued) models of ADFs correspond precisely to causal worlds of the causal model, as defined in [6]:

Definition 6. A *solution* (or a *causal world*) of a Boolean causal model M is any propositional interpretation satisfying the equivalences $p \leftrightarrow F$ for all equations $p = F$ in M .

Now, a modular representation of Boolean causal models as causal theories of the causal calculus has been given in [6], and it can now be seamlessly transformed into the following causal representation of ADFs:

Definition 7 (*Causal representation of an ADF*). For any ADF D , Δ_D is the causal theory consisting of the rules

$$F \Rightarrow p \text{ and } \neg F \Rightarrow \neg p$$

for all acceptance conditions $p[F]$ in D .

The above representation is fully modular, and it will be taken as a uniform basis for the correspondences described in this study.

To begin with, based on the correspondence results from [6], we immediately establish

Theorem 2. *The two-valued semantics of an ADF D corresponds precisely to the causal nonmonotonic semantics of Δ_D .*

As a consequence, the full system of causal inference provides a precise logical basis for this nonmonotonic semantics.

4.3. General correspondences

Now we are going to show that the above causal representation also survives the transition to three-valued models of ADFs. To this end, however, we will have to retreat from the system of causal inference to a weaker system of basic production inference.

A broader correspondence between various semantics of ADFs and general nonmonotonic semantics of the causal calculus arises from the fact that the operator Γ of an ADF naturally corresponds to a particular causal operator of the associated causal theory.

Let L denote the set of classical literals of the underlying language. We will denote by \mathbb{C}^L the restriction of a causal operator \mathbb{C} to literals, that is, $\mathbb{C}^L(u) = \mathbb{C}(u) \cap L$. As we are going to show, the operator Γ of ADFs corresponds precisely to this ‘literal restriction’ of the causal operator associated with a basic production inference. As before, $[v]$ will denote the set of literals corresponding to a three-valued interpretation v .

Lemma 3. *For any three-valued interpretation v ,*

$$[\Gamma_D(v)] = \mathbb{C}_D^L([v]),$$

where \mathbb{C}_D is a basic causal operator corresponding to Δ_D .

The above equation has immediate consequences for the broad correspondence between the semantics of ADFs that are defined in terms of the operator Γ_D and natural sets of propositions definable wrt associated causal theory. Thus, we have

Theorem 4. *Complete models of an ADF D correspond precisely to the fixed points of \mathbb{C}_D^L :*

$$v = \Gamma_D(v) \quad \text{iff} \quad [v] = \mathbb{C}_D^L([v])$$

As a result, we immediately conclude that preferred models of an ADF correspond to maximal fixpoints of \mathbb{C}_D^L (with respect to set inclusion), while the grounded model corresponds to the least fixpoint of \mathbb{C}_D^L .

It turns out, however, that when viewed in a classical logical setting, the restriction of the causal operator to literals inadvertently leads to an information loss. More precisely, though disjunctive formulas can appear in acceptance conditions used by Γ in an ADF, the operator itself records, in effect, only literals that are produced, and thereby disregards all other information that can be obtained from its output. The following example illustrates this.

Example 1. Let us consider the following ADF D :

$$q[p] \quad r[\neg p] \quad s[q \vee r]$$

The grounded model of this ADF is empty (all atoms are unknown). However, the associated causal theory Δ_D comprises the following rules:

$$\begin{array}{lll} p \Rightarrow q & \neg p \Rightarrow r & q \vee r \Rightarrow s \\ \neg p \Rightarrow \neg q & p \Rightarrow \neg r & \neg q \wedge \neg r \Rightarrow \neg s \end{array}$$

In view of Lemma 1, the least exact theory of \mathbb{C}_D is precisely the set of propositions that are provable from the above theory using the postulates of *causal* inference (since it is both basic and regular). Now, the first two rules imply $\mathbf{t} \Rightarrow q \vee r$ (by Or), and hence $\mathbf{t} \Rightarrow s$ by Cut. Similarly, the fourth and fifth rule imply $\mathbf{t} \Rightarrow \neg q \vee \neg r$. As result, the least exact theory of \mathbb{C}_D is much more informative, namely $\text{Th}(\{q \leftrightarrow \neg r, s\})$.

It can also be seen from the above example that the restriction of exact theories to literals does not necessary produce fixed points of the corresponding literal operator \mathbb{C}^L . Still, it can be shown that for any fixpoint of the latter (that is, for any complete model an ADF) there exists a least exact theory that contains it. The latter theory may contain, however, more information than its literal source.

5. Justification Frames, Logic Programs and Generalized ADFs

A revised definition of a stable model has been given in [7], generalized already to arbitrary ADFs. Roughly, a two-valued model v of an ADF D is a *stable model* of D if the set of statements that are true in it coincides with the grounded extension of the reduced ADF D^v obtained from D by replacement of all false statements in v by their truth value in each acceptance formula. As has been shown by the authors, this definition properly generalizes stable extensions of Dung's argumentation frameworks.

It should be noted, however, that from the 'non-abstract' knowledge representation view of ADFs that we pursue in the present study, the above definition of a stable semantics constitutes a certain departure from the original formulation of ADFs that was based on *classical* acceptance conditions. Indeed, the above definition of a stable model implicitly breaks the classical symmetry between positive and negative statements, so the acceptance conditions cannot already be viewed as classical formulas. Instead, they acquire a non-classical reading that is quite familiar from logic programming.

It is well-known that the formalism of ADFs, taken in its original sense, does not capture all the semantic distinctions that are expressible in the language of Logic Programming (see, e.g., [16]). Still, the causal representation of ADFs, described in the preceding section, can also suggest a proper generalization of ADFs that would cover Logic Programming under its various semantics, while still preserving the original classical reading of their acceptance conditions (and even their original two-valued semantics). Due to space limitations, however, we can only be brief here.

To begin with, the causal representation of ADFs, described earlier, transforms them into rule-based causal theories, while the latter constitute, in turn, a very special, 'classical' case of *justification frames*, introduced in [9]. In particular, the justification rules of the latter have the general form $x \leftarrow S$, where x is a literal, while S is a set of literals. In the case of the classical negation, such justification frames correspond precisely to determinate causal theories under basic production inference.

However, the causal rules of the causal calculus have an additional expressivity in that they allow arbitrary classical formulas not only in the bodies, but also in the heads of the rules. It turns out that this expressive capability is already sufficient for representing logic programming rules and their semantics.

A causal representation of logic programming rules under various semantics for the latter has been described in [3]. It was defined for general program rules of the form

$$\mathbf{not} \, d, c \leftarrow a, \mathbf{not} \, b \quad (*)$$

where a, b, c, d are finite sets of atoms.

A general understanding of logic programs presupposes an asymmetric treatment of negative information, which is reflected in viewing the negation **not** as denoting *negation as failure*. This understanding can be formally captured in the causal calculus by accepting the following additional postulate:

(Default Negation) $\neg p \Rightarrow \neg p$, for any propositional atom p .

The above postulate makes negations of propositional atoms self-explainable propositions (or abducibles), so it expresses, in effect, the *Closed World Assumption* (CWA).

Given this postulate, a causal representation of logic programs under the stable semantics is provided by interpreting a program rule (*) as the following causal rule:

$$d, \neg b \Rightarrow \bigwedge a \rightarrow \bigvee c$$

This interpretation provides a classical understanding for **not**, so its non-classicality amounts solely to the non-classicality of \Rightarrow . Nevertheless, it has been shown that the stable semantics of logic programs corresponds precisely to the causal nonmonotonic semantics of the resulting causal theories, that is, to the exact worlds of the latter. Furthermore, the same causal nonmonotonic semantics has turned out to be appropriate also for logic programs under the *supported semantics*, provided we interpret the program rule (*) differently, namely as the following causal rule:

$$a, d, \neg b \Rightarrow \bigvee c$$

The only difference with the previous stable interpretation amounts to treating positive premises in a as explanations rather than as part of what is explained. Note that a normal program rule $p \leftarrow a, \mathbf{not} \, b$ corresponds under this interpretation to the causal rule

$$a, \neg b \Rightarrow p$$

which can be directly transformed into (part of) an acceptance condition for p in ADFs.

The above considerations and results suggest a natural generalization of an ADF to acceptance conditions of the form $A[B]$, where both A and B are classical formulas. This would supply the Abstract Argumentation Frameworks with further representation capabilities, and thereby even contribute to the original aim of the authors of providing a powerful and widely applicable abstraction tool for Argumentation and Reasoning.

6. Summary, Related Work and Conclusions

It has been shown in this study that Abstract Dialectical Frameworks can be uniformly translated into the causal calculus in a way that creates a broad correspondence between the main semantics for ADFs and their causal counterparts.

Among many other things, the suggested translation can be used for determining the place of ADFs (viewed as a specific KR formalism) in the broad range of formalisms for argumentation and reasoning. Thus, it has been shown in [4] that a great number of key systems for argumentation and nonmonotonic reasoning, including the causal calculus, can be viewed as direct instantiations of the original Dung's argumentation frameworks in different logical languages. Due to the results of the present study, the Abstract Dialectical Frameworks also find their natural place in this larger picture. This topic deserves, however, a separate discussion that goes beyond the scope of the present study. Still, a couple of general comments are in order here.

The field of formal argumentation is abundant with different formalisms, which creates a fertile ground for extensive and rapid development. But there is also a lot of conceptual affinity among these argumentation formalisms, as well as between the latter and the major KR representation languages. It is this affinity that allows us to use many of them for basically the same reasoning tasks. This situation creates, however, an obvious incentive for unification, namely for constructing a general theory of argumentation and reasoning where these formalisms could find their proper and hospitable place.

An algebraic approach to unification of different KR formalisms has been suggested in [10], which describes a general method for deriving approximations of operators associated with particular knowledge representation systems. This approach has been successfully applied to ADFs in [16], which also contains comparisons with Logic Programming.

The above approximation theory can be viewed as a paradigmatic *abstraction approach*, in which a general algebraic formalism is shown to be capable of encompassing many particular KR systems. In contrast, our present study can be seen as an instance of a somewhat more specific *generalization approach*, which aims to single out conceptual principles common to a number of formalisms for argumentation and reasoning³. For instance, we take it to be a virtue of the original ADFs that they employ classical descriptions in the acceptance conditions. This makes an ADF a natural extension of classical reasoning (instead of being a replacement for the latter), an extension that incorporates, however, some key features of our commonsense reasoning that go beyond pure logical inference.

There is a number of concepts and features that are pervasive in commonsense reasoning, though they escape a purely logical description. The general field of nonmonotonic reasoning has considerably advanced our understanding of these features, which include concepts like explanation, justification, causation, and even definition. The key notions of the modern formal argumentation theory such as support, defeat and attack also belong to this class. It could even be argued that the main contribution of Dung's abstract argumentation theory has consisted not so much in suggesting a new abstract framework for argumentation, but rather in incorporating these notions as the main conceptual ingredients of argumentation. It is this conceptual advancement that has given

³The formal theory of justifications [9] could also be seen as a step in this direction.

the formal argumentation theory its current impetus. Accordingly, a systematic study of these novel features of argumentation should be viewed as one of the principal tasks of argumentation theory in general.

Finally, it is an undeniable fact that all the above mentioned notions are also intimately related, which could be seen as the ultimate reason why there are mutual translations between the associated formalisms, as well as why they are so often interchangeable in specific reasoning and argumentation settings. Accordingly, a large part of the task of studying and clarifying the scope of the main building blocks of argumentation consists in determining the relationships and translations among these diverse concepts (often formulated in entirely different formalisms). The correspondence between acceptance conditions of ADFs and causal rules of the causal calculus, established in this study, should hopefully facilitate this general effort.

References

- [1] A. Bochman. A logic for causal reasoning. In *IJCAI-03, Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence, Acapulco, Mexico, August 9-15, 2003*, pages 141–146, Acapulco, 2003. Morgan Kaufmann.
- [2] A. Bochman. A causal approach to nonmonotonic reasoning. *Artificial Intelligence*, 160:105–143, 2004.
- [3] A. Bochman. A causal logic of logic programming. In D. Dubois, C. Welty, and M.-A. Williams, editors, *Proc. Ninth Conference on Principles of Knowledge Representation and Reasoning, KR'04*, pages 427–437, Whistler, 2004.
- [4] A. Bochman. Propositional argumentation and causal reasoning. In *IJCAI-05, Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence, Edinburgh, Scotland, UK, July 30-August 5, 2005*, pages 388–393, 2005.
- [5] A. Bochman. A causal theory of abduction. *Journal of Logic and Computation*, 17:851–869, 2007.
- [6] A. Bochman and V. Lifschitz. Pearl's causality in a logical setting. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA.*, pages 1446–1452. AAAI Press, 2015.
- [7] G. Brewka, H. Strass, S. Ellmauthaler, J. P. Wallner, and S. Woltran. Abstract dialectical frameworks revisited. In *IJCAI 2013, Proceedings of the 23rd International Joint Conference on Artificial Intelligence, Beijing, China, August 3-9, 2013*, 2013.
- [8] G. Brewka and S. Woltran. Abstract dialectical frameworks. In *Principles of Knowledge Representation and Reasoning: Proceedings of the Twelfth International Conference, KR 2010, Toronto, Ontario, Canada, May 9-13, 2010*, 2010.
- [9] M. Denecker, G. Brewka, and H. Strass. A formal theory of justifications. In *Logic Programming and Nonmonotonic Reasoning - 13th International Conference, LPNMR 2015, Lexington, KY, USA, September 27-30, 2015. Proceedings*, pages 250–264, 2015.
- [10] M. Denecker, V. W. Marek, and M. Truszczynski. Ultimate approximation and its application in non-monotonic knowledge representation systems. *Inf. Comput.*, 192(1):84–121, 2004.
- [11] S. Ellmauthaler. Abstract Dialectical Frameworks: Properties, Complexity, and Implementation. Master's thesis, Technische Universität Wien, Institut für Informationssysteme, 2012.
- [12] E. Giunchiglia, J. Lee, V. Lifschitz, N. McCain, and H. Turner. Nonmonotonic causal theories. *Artificial Intelligence*, 153:49–104, 2004.
- [13] D. Makinson and L. van der Torre. Input/Output logics. *Journal of Philosophical Logic*, 29:383–408, 2000.
- [14] N. McCain and H. Turner. Causal theories of action and change. In *Proceedings AAAI-97*, pages 460–465, 1997.
- [15] J. Pearl. *Causality: Models, Reasoning and Inference*. Cambridge UP, 2000. 2nd ed., 2009.
- [16] H. Strass. Approximating operators and semantics for abstract dialectical frameworks. *Artif. Intell.*, 205:39–70, 2013.

Argumentation Ranking Semantics Based on Propagation¹

Elise BONZON^a, Jérôme DELOBELLE^b, Sébastien KONIECZNY^b and
Nicolas MAUDET^c

^a*LIPADE, Université Paris Descartes, Paris, France - bonzon@parisdescartes.fr*

^b*CRIL, Université d'Artois, Lens, France - {delobelle,konieczny}@cril.fr*

^c*LIP6, Université Pierre et Marie Curie, Paris, France - nicolas.maudet@lip6.fr*

Abstract. Argumentation is based on the exchange and the evaluation of interacting arguments. Unlike Dung's theory where arguments are either accepted or rejected, ranking-based semantics rank-order arguments from the most to the least acceptable ones. We propose in this work six new ranking-based semantics. We argue that, contrarily to existing ranking semantics in the literature, that focus on evaluating attacks and defenses only, it is reasonable to give a prominent role to non-attacked arguments, as it is the case in standard Dung's semantics. Our six semantics are based on the propagation of the weight of each argument to its neighbors, where the weight of non-attacked arguments is greater than the attacked ones.

Keywords. Argumentation, Ranking-based semantics, Propagation

1. Introduction

Argumentation is a very natural framework for representing conflicting information. A proof of its appeal is the recent development of online platforms where people participate in debates using argumentation graphs (e.g. *debategraph.org* or *arguman.org*) such representation tools become more and more popular.

The question now goes towards the reasoning part: how to automatically use these argumentation graphs that are constructed this way? Argumentation has been a very active topic in Artificial Intelligence since more than two decades now, and Dung's work on abstract argumentation framework [1] can be used to represent the graphs (even if some additional information should be also represented, like the number of people that agree/disagree with an argument and/or an attack, or a support between arguments, ...). But the main issue is about the semantics that one should use in this case. In fact classical Dung's semantics, using extensions [1] (or equivalently labellings [2]), with their dichotomous evaluation of arguments (accepted/rejected) do not seem very well suited for such applications. As discussed in [3], on such online platforms, with a big number of arguments, and a lot of individuals participating, it can be problematic (in particular quite unintuitive for the participants) to have such a drastic evaluation, that is not that in-

¹This work benefited from the support of the project AMANDE ANR-13-BS02-0004 of the French National Research Agency (ANR).

formative (since there are only two levels of acceptability), or to propose several possible results (different extensions). So in this case a finer evaluation of arguments seems to be more adequate. The idea is then to have ranking-based semantics, that allow to produce a full ranking of the arguments, from the most to the least acceptable ones. This kind of semantics seems very natural, and it is then quite surprising that they have received little attention until recently [3,4,5,6,7,8,9,10]. These semantics basically rely on the attacks and defenses of each argument in order to evaluate its acceptability rank.

In this work we propose a new family of semantics, that relies on attacks and defenses, like previous semantics, but that also puts a strong emphasis on non-attacked arguments. While many principles remain discussed and controversial, all semantics agree on the fact that non-attacked arguments should have the highest rank. The idea of our semantics is that these arguments should also have a greater impact on the evaluation of the other ones.

In this paper, we propose six new semantics based on the idea of propagation. Each argument has an initial weight that depends on its status (non-attacked arguments have a greater value than attacked ones), and then these weights are progressively propagated to their neighbors. Of course at each propagation the polarity of the weight changes in order to comply with the attack relation meaning. The difference between these semantics lies in the method that is chosen to differentiate non-attacked arguments and attacked ones, and in the choice of considering one or all paths between arguments.

In the next section, we define the notions and notations we will need to define our six propagation semantics in Section 3. Section 4 recalls the logical properties proposed in the literature for ranking-based semantics, and studies which ones are satisfied by our semantics. In Section 5, we study the links between our semantics and previous ones. Section 6 provides an example in order to illustrate the behaviour of the propagation semantics and to relate them to other semantics.

2. Background

Following [1], an argumentation system is a (finite) set of arguments together with the binary conflicts among them.

Definition 1. An **argumentation framework (AF)** is a directed graph $\langle A, \hookrightarrow \rangle$ where the set of nodes A is a finite set of **arguments**, and the set of edges $\hookrightarrow \subseteq A \times A$ is an **attack relation** between arguments. A set of arguments $C \subseteq A$ **attacks** an argument $b \in A$, if there exists $a \in C$, such that $(a, b) \in \hookrightarrow$. C **defends** a iff $\forall b \in A$ such that $(b, a) \in \hookrightarrow$, $\exists c \in C$ such that $(c, b) \in \hookrightarrow$.

In the following, \mathbb{AF} will represent the set of all argumentation frameworks.

Definition 2. Let $AF = \langle A, \hookrightarrow \rangle$ and $a, b \in A$. A **path** from a to b , denoted by $a \rightsquigarrow b$, is a sequence of nodes $s = \langle a_0, \dots, a_n \rangle$ such that from each node there is an edge to the next node in the sequence : $a_0 = a$, $a_n = b$ and $\forall i < n, (a_i, a_{i+1}) \in \hookrightarrow$. Its **length** is noted $|a \rightsquigarrow b|$ and is equal to the number of edges it is composed of.

In order to encode the fact that there are several possible paths between two arguments, we introduce the notion of multiset of attackers and defenders of an argument.

Definition 3. Let $AF = \langle A, \hookrightarrow \rangle$ and $a, b \in A$. Let $\oplus \in \{M, S\}$, where M (resp. S) stands for multiset (resp. set). The (multi)set of arguments such that there exists a path to a with a length of n is denoted by $\downarrow_n^\oplus(a) = \{b \mid \exists b \rightsquigarrow a, \text{ with } |b \rightsquigarrow a| = n\}$. An argument $b \in \downarrow_n^\oplus(a)$ is a **defender** (resp. **attacker**) of a if $n \in 2\mathbb{N}$ (resp. $n \in 2\mathbb{N} + 1$). Let us note $\downarrow^\oplus(a) = \bigcup_n \downarrow_n^\oplus(a)$.

Note that the direct attackers of an argument a belong to $\downarrow_1^\oplus(a)$.

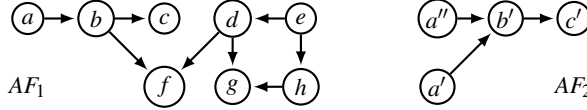


Figure 1. Two argumentation frameworks

Let us discuss how using sets or multisets can influence the result of the ranking-based semantics. Obviously there is no change for the direct attackers because an argument cannot be directly attacked several times by the same argument. However, several paths of length greater than one between two arguments can exist: on AF_1 (Figure 1) there are two paths of length 2 from e to g : $\langle e, d, g \rangle$ and $\langle e, h, g \rangle$. So with sets $\downarrow_2^S(g) = \{e\}$, whereas with multisets the result is $\downarrow_2^M(g) = \{e, e\}$.

The use of sets can be seen as a focus on the arguments at the end of the path without taking into account the number of possible paths, whereas the multisets encodes the fact that there are several possible paths.

Definition 4. Let $\oplus \in \{M, S\}$. A path from b to a is a **branch** if b is not attacked, that is if $\downarrow_1^\oplus(b) = \emptyset$. It is a **defense branch** (resp. **attack branch**) if b is a defender (resp. attacker) of a .

One of the main goals of argumentation theory is to identify which arguments are rationally *acceptable* according to different notions of acceptability. In [1], the acceptability of an argument depends on its membership to some extensions, whereas ranking-based semantics aim to rank arguments from the most to the least acceptable ones.

Definition 5. A **ranking-based semantics** σ is a function that transforms any argumentation framework $AF = \langle A, \hookrightarrow \rangle$ into a ranking \succeq_{AF}^σ on A , where \succeq_{AF}^σ is a preorder (a reflexive and transitive relation) on A . $a \succeq_{AF}^\sigma b$ means that a is at least as acceptable as b ($a \simeq_{AF}^\sigma b$ is a shortcut for $a \succeq_{AF}^\sigma b$ and $b \succeq_{AF}^\sigma a$, and $a \succ_{AF}^\sigma b$ is a shortcut for $a \succeq_{AF}^\sigma b$ and $b \not\succeq_{AF}^\sigma a$).

When there is no ambiguity about the argumentation framework in question, we will use \succeq^σ instead of \succeq_{AF}^σ .

Finally, we need to introduce the notion of lexicographical order and a shuffle operation between vectors of real number in order to define our new ranking-based semantics.

Definition 6. A **lexicographical order** between two vectors of real numbers $V = \langle V_1, \dots, V_n \rangle$ and $V' = \langle V'_1, \dots, V'_n \rangle$ is defined as $V \succ_{lex} V'$ iff $\exists i \leq n$ s.t. $V_i > V'_i$ and $\forall j < i, V_j = V'_j$. $V \simeq_{lex} V'$ means that $V \not\succ_{lex} V'$ and $V' \not\succ_{lex} V$; and $V \succeq_{lex} V'$ means that $V' \not\succ_{lex} V$.

Definition 7. The **shuffle** \cup_s between two vectors of real numbers $V = \langle V_1, \dots, V_n \rangle$ and $V' = \langle V'_1, \dots, V'_n \rangle$ is defined as $V \cup_s V' = \langle V_1, V'_1, V_2, V'_2, \dots, V_n, V'_n \rangle$.

3. Propagation Semantics

A standard principle of existing ranking-based semantics is to base the evaluation of an argument on the number of its attackers and of its defenders: the less attackers and the more defenders an argument, the more acceptable the argument. For example, if we compare AF_1 with AF_2 in Figure 1, b is better than b' because b has less attackers than b' (only one for b against two for b'). Inversely, c' has more defenders than c (two for c' against one for c) and the same number of attackers so c' can be considered better than c .

Another important principle to take into account is the role and impact of non-attacked arguments. For example, in AF_1 , as a is non-attacked and c is defended by a against the attack of b , a and c are both accepted with respect to Dung's semantics. However, one could go further and say that a is better than c because a is not attacked, whereas c is attacked and is accepted only because of the defense of a . So, we can clearly see that a , and more generally the non-attacked arguments, play a key role in the (classical) acceptability of an argument. Thus, our propagation method will allow those non-attacked arguments to play a key role in the ranking of arguments.

Our approach is based on these two principles. The propagation methods are defined in two steps. The first step consists in assigning a positive initial weight to each argument. The score of 1 attached to non-attacked arguments is set to be higher than the score of attacked arguments, which is an ε between 0 and 1. The value of this ε is chosen accordingly to the degree of influence of the non-attacked arguments that we want: the smaller the value of ε is, the more important the influence of non-attacked arguments on the order prevails. Then, during the second step, we propagate the weights into the graph in changing their polarities in order to comply with the attack relation meaning (attack or defense). For each argument, we accumulate and store the weights from its attackers and defenders in the argumentation framework.

Definition 8. Let $F = \langle A, \hookrightarrow \rangle$ be an AF. The valuation P of $a \in A$, at step i , is given by:

$$P_i^{\varepsilon, \oplus}(a) = \begin{cases} v_\varepsilon(a) & \text{if } i = 0 \\ P_{i-1}^{\varepsilon, \oplus}(a) + (-1)^i \sum_{b \in \downarrow_i^{\oplus}(a)} v_\varepsilon(b) & \text{otherwise} \end{cases}$$

where $v : A \rightarrow \mathbb{R}^+$ is a valuation function giving an initial weight to each argument, with $\varepsilon \in [0, 1]$ such that $\forall b \in A$, $v_\varepsilon(b) = 1$ if $\downarrow_1^{\oplus}(b) = \emptyset$; $v_\varepsilon(b) = \varepsilon$ otherwise.

The **Propagation vector** of a is denoted $P^{\varepsilon, \oplus}(a) = \langle P_0^{\varepsilon, \oplus}(a), P_1^{\varepsilon, \oplus}(a), \dots \rangle$.

Example 1. Let us calculate the value of P when $\varepsilon = 0.75$ for AF_1 (Figure 1). If no distinction exists between set and multiset then the value is put in the same cell (Table 1). Otherwise, the cell is divided into two parts (valuation for set at left and for multiset at right). For instance, when $i = 2$, $P_2^{0.75, S}(c) = P_2^{0.75, M}(c) = 1$ but $P_2^{0.75, S}(g) = 0.25$ whereas $P_2^{0.75, M}(g) = 1.25$.

In Table 1, argument f begins with an initial weight of 0.75 ($P_0^{0.75, \oplus}(f) = 0.75$) because it is attacked. Then, during the second step, the direct attackers (b and d which are also attacked) propagate negatively their weights of 0.75 to f , so $P_1^{0.75, \oplus}(f) = P_0^{0.75, \oplus}(f) - (v_{0.75}(d) + v_{0.75}(b)) = -0.75$. Finally, during the third step, f receives

$P_i^{0.75, \oplus}$	a, e		b, d, h		c		f		g	
	S	M	S	M	S	M	S	M	S	M
0		1		0.75		0.75		0.75		0.75
1		1		-0.25		0		-0.75		-0.75
2		1		-0.25		1		1.25	0.25	1.25

Table 1. Computation of propagation vector when $\varepsilon = 0.75$

positively the weights of 1 from a and e which are non-attacked, so $P_2^{0.75, \oplus}(f) = P_1^{0.75, \oplus}(f) + (v_{0.75}(a) + v_{0.75}(e)) = 1.25$. As there exists no path to f with a length higher than 2, this score remains the same, and $P^{0.75, \oplus}(f) = \langle 0.75, -0.75, 1.25, 1.25, \dots \rangle$.

It is important to note that $P^{\varepsilon, \oplus}(a)$ may be infinite (this may occur when an argument is involved in at least one cycle). Moreover, the valuation $P_n^{\varepsilon, \oplus}(x)$ of an argument x is not even necessarily bounded as $n \rightarrow \infty$. After a finite number of steps though, an argument is bound to receive the influence of exactly the same arguments than in a previous step of the vector (which means that the vector can be finitely encoded). More precisely, this number of steps is in the order of the least common multiplier of the cycle lengths occurring in the argumentation graph. As a ranking-based semantics is not concerned with the exact values of arguments, but only in their relative ordering, this is sufficient for our purpose.

3.1. Propa_ε

Once the propagation vector is calculated for each argument in the argumentation framework, we can compare the different vectors in order to obtain an order between all the arguments. We want the influence of arguments to quickly decrease with the length of a path, so an option is to use a lexicographical comparison for comparing these vectors. For the first semantics we just compare the propagation vectors for a given ε .

Definition 9. Let $\oplus \in \{M, S\}$. The ranking-based semantics $\text{Propa}_\varepsilon^{\varepsilon, \oplus}$ associates to any $AF = \langle A, \hookrightarrow \rangle$ a ranking $\succeq_{AF}^{P_\oplus}$ on A such that $\forall a, b \in A$, $a \succeq_{AF}^{P_\oplus} b$ iff $P^{\varepsilon, \oplus}(a) \succeq_{lex} P^{\varepsilon, \oplus}(b)$.

So this defines two semantics, one using sets and one using multisets for the attack and defense branches computations.

Example 1 (cont.). In Table 1, if $\oplus = S$, we obtain the ranking $a \simeq^{P_S} e \succ^{P_S} c \succ^{P_S} b \simeq^{P_S} d \simeq^{P_S} h \succ^{P_S} f \succ^{P_S} g$. If $\oplus = M$, we have $a \simeq^{P_M} e \succ^{P_M} c \succ^{P_M} b \simeq^{P_M} d \simeq^{P_M} h \succ^{P_M} f \simeq^{P_M} g$.

These semantics focus mainly on the attackers and defenders in adding the fact that if there exists non-attacked arguments among them, these ones will be more influential than attacked arguments. But this influence depends also on the value of ε . Indeed, two values of ε can lead to different orders. On Example 1, with $\varepsilon = 0.75$, if we focus on f , which is defended twice, and h , which is attacked (and not defended), we can see that h is better than f because $P_1^{0.75, \oplus}(f) < P_1^{0.75, \oplus}(h)$. But if we take $\varepsilon < 0.5$, we obtain the opposite case. For example, with $\varepsilon = 0.3$, we have $P^{0.3, \oplus}(f) = \langle 0.3, -0.3, 1.7, \dots \rangle$ and $P^{0.3, \oplus}(h) = \langle 0.3, -0.7, \dots \rangle$. With the lexicographical order, f is now better than h because $P_1^{0.3, \oplus}(f) > P_1^{0.3, \oplus}(h)$.

So, with Propa_ε semantics, an argument with only (but numerous) defense branches can be worse than an argument only attacked by one non-attacked argument. It is a pos-

sible point of view to focus only on the attackers in saying that the more an argument is directly attacked, the less acceptable the argument. It is the case, for instance, with the semantics proposed by Amgoud and Ben-Naim [8]. But other approaches are possible, as we shall see now.

3.2. $Propa_{1+\varepsilon}$

If we do not want the influence of non-attacked arguments to be drown in by the influence of attacked arguments, we have to split and lexicographically compare the influence of the two kinds of arguments.

Definition 10. Let $\oplus \in \{M, S\}$. The ranking-based semantics $Propa_{1+\varepsilon}^{\varepsilon, \oplus}$ associates to any $AF = \langle A, \hookrightarrow \rangle$ a ranking $\succeq_{AF}^{\hat{P}_\oplus}$ on A such that $\forall a, b \in A$,

$$a \succeq_{AF}^{\hat{P}_\oplus} b \text{ iff } P^{0, \oplus}(a) \cup_s P^{\varepsilon, \oplus}(a) \succeq_{lex} P^{0, \oplus}(b) \cup_s P^{\varepsilon, \oplus}(b)$$

With these semantics, we simultaneously look at the result of the two propagation vectors $P^{0, \oplus}$ and $P^{\varepsilon, \oplus}$ step by step, using the shuffle operation, starting with the first value of the propagation vector $P^{0, \oplus}$ (i.e. the one that takes into account non-attacked arguments only). In the case where two arguments are still equally acceptable, we compare the first value of the propagation vector $P^{\varepsilon, \oplus}$. Then, in case of equality, we move to the second step and so on.

Example 1 (cont.). Let us focus on f with the two propagation vectors: $P^{0, \oplus}(f) = \langle 0, 0, 2, \dots \rangle$ (see Table 2 where $\varepsilon = 0$) and $P^{0.75, \oplus}(f) = \langle 0.75, -0.75, 1.25, \dots \rangle$ (see Table 1 where $\varepsilon = 0.75$). We use the shuffle \cup_s to combine the previous propagation vectors: $P^{0, \oplus}(f) \cup_s P^{0.75, \oplus}(f) = \langle 0, 0.75, 0, -0.75, 2, 1.25, \dots \rangle$. We apply the same method for all the others arguments and we use the lexicographical order to compare them. So if $\oplus = S$, we obtain the ranking $a \simeq^{\hat{P}_S} e \succ^{\hat{P}_S} c \succ^{\hat{P}_S} f \succ^{\hat{P}_S} g \succ^{\hat{P}_S} b \simeq^{\hat{P}_S} d \simeq^{\hat{P}_S} h$ whereas if $\oplus = M$, we have $a \simeq^{\hat{P}_M} e \succ^{\hat{P}_M} c \succ^{\hat{P}_M} f \simeq^{\hat{P}_M} g \succ^{\hat{P}_M} b \simeq^{\hat{P}_M} d \simeq^{\hat{P}_M} h$.

$P_i^{0, \oplus}$	a, e		b, d, h		c		f		g	
	S	M	S	M	S	M	S	M	S	M
0		1		0		0		0		0
1		1		-1		0		0		0
2		1		-1		1		2		1

Table 2. Computation of propagation vector when $\varepsilon = 0$

It is also important to notice that $Propa_{1+\varepsilon}$, conversely to $Propa_\varepsilon$, returns the same order whatever the value of ε , that removes the problem of choosing “a good” ε :

Proposition 1. Let $\oplus \in \{M, S\}$, $\forall AF \in \mathbb{AF}$ and $\forall \varepsilon, \varepsilon' \in]0, 1]$,

$$Propa_{1+\varepsilon}^{\varepsilon, \oplus}(AF) = Propa_{1+\varepsilon}^{\varepsilon', \oplus}(AF)$$

3.3. $Propa_{1 \rightarrow \varepsilon}$

A last possibility is to give a higher priority to the non-attacked arguments, by propagating only their weights in the graph. And if two arguments are equivalent for this comparison, they are compared using the $Propa_\varepsilon$ method. Technically, the priority to the non-attacked arguments is given by using $\varepsilon = 0$. So we compare first the propagation vector $P^{0, \oplus}$. And if the two propagation vectors are identical, we restart with a non-zero ε :

Definition 11. Let $\oplus \in \{M, S\}$. The ranking-based semantics $Propa_{1 \rightarrow \epsilon}^{\epsilon, \oplus}$ associates to any $AF = \langle A, \hookrightarrow \rangle$ a ranking $\succeq_{AF}^{\bar{P}_{\oplus}}$ on A such that $\forall a, b \in A$,

$$a \succeq_{AF}^{\bar{P}_{\oplus}} b \text{ iff } P^{0, \oplus}(a) \succ_{lex} P^{0, \oplus}(b) \text{ or, } (P^{0, \oplus}(a) \simeq_{lex} P^{0, \oplus}(b) \text{ and } P^{\epsilon, \oplus}(a) \succeq_{lex} P^{\epsilon, \oplus}(b))$$

Example 1 (cont.). $Propa_{1 \rightarrow \epsilon}$ focuses first in Table 2, where $\epsilon = 0$, and then in Table 1, where $\epsilon = 0.75$. If $\oplus = S$, we obtain the ranking $a \simeq^{\bar{P}_S} e \succ^{\bar{P}_S} f \succ^{\bar{P}_S} c \succ^{\bar{P}_S} g \succ^{\bar{P}_S} b \simeq^{\bar{P}_S} d \simeq^{\bar{P}_S} h$ whereas if $\oplus = M$, we have $a \simeq^{\bar{P}_M} e \succ^{\bar{P}_M} f \simeq^{\bar{P}_M} g \succ^{\bar{P}_M} c \succ^{\bar{P}_M} b \simeq^{\bar{P}_M} d \simeq^{\bar{P}_M} h$.

With these semantics, an argument with a lot of defense branches will receive a lot of positive weights, and conversely, an argument with a lot of attack branches, will receive a lot of negative weights. Thus, as f and g have one more defense branch than c (with the multiset), which has also one more defense branch than b , d and h , we have that f and g are better than c which is better than b , d and h . However, focusing only on $\epsilon = 0$, we cannot distinguish the arguments with the same number of defense/attack branches. This is why we use the propagation vector with $\epsilon \neq 0$ to decide between those.

It is also important to notice that $Propa_{1 \rightarrow \epsilon}$, like $Propa_{1+\epsilon}$, returns the same order whatever the value of ϵ :

Proposition 2. Let $\oplus \in \{M, S\}$, $\forall AF \in \mathbb{AF}$ and $\forall \epsilon, \epsilon' \in]0, 1]$,

$$Propa_{1 \rightarrow \epsilon}^{\epsilon, \oplus}(AF) = Propa_{1 \rightarrow \epsilon}^{\epsilon', \oplus}(AF)$$

4. Properties

4.1. Properties for ranking-based semantics

In the last few years, a set of properties have been proposed in different papers, allowing to better understand the behavior of the different ranking-based semantics. We adopt the recent catalogue of properties listed in [11] (for space reasons we point out to this paper for their formal definitions).

One can find the properties Abs, In, VP, DP, CT, SCT, CP, QP and DDP in [8], the properties In, VP and SC in [6], the properties $\oplus DB$, $+DB$, $\uparrow AB$, $+AB$, $\uparrow DB$ in [5,11], and the properties Tot and AvsFD in [11]. We do not claim that all these properties are mandatory (in particular some of them are incompatible and we do not necessarily endorse all of them). Let a and b two arguments in an AF.

Abstraction (Abs) The arguments' ranking should only depend on the attack relation.

Independence (In) The ranking between two arguments should be independent of arguments that are not connected to either of them.

Void Precedence (VP) A non-attacked argument should be strictly more acceptable than an attacked argument.

Self-Contradiction (SC) An argument that attacks itself should be strictly less acceptable than an argument that does not.

Cardinality Precedence (CP) If a has strictly more direct attackers than b , then b should be strictly more acceptable than a .

Quality Precedence (QP) If a has a direct attacker strictly more acceptable than any direct attacker of b , then a should be strictly more acceptable than b .

Counter-Transitivity (CT) If the direct attackers of b are at least as numerous and acceptable as those of a , then a should be at least as acceptable as b .

Strict Counter-Transitivity (SCT) If CT is satisfied and if the direct attackers of b are either strictly more numerous, or strictly more acceptable than those of a , then a should be strictly more acceptable than b .

Defense Precedence (DP) If arguments a and b have the same number of direct attackers, and if a is defended at least once whereas b is not, a should be ranked higher than b .

Distributed-Defense Precedence (DDP) A defense where each defender attacks a distinct attacker is better than any other.

In order to introduce the next properties, let us define the notion of ancestor's graph:

Definition 12. Let $AF = \langle A, \hookrightarrow \rangle$ and $a \in A$. The **ancestor's graph** of a is denoted by $Anc(a) = \langle A', \hookrightarrow' \rangle$ with $A' = \downarrow^S(a)$ and $\hookrightarrow' = \{(a_1, a_2) \in \hookrightarrow \mid a_1 \in A' \text{ and } a_2 \in A'\}$.

Strict addition of Defense Branch (\oplus DB) If a and b have the same ancestor's graph, except that a has an additional defense branch, then a should be strictly more acceptable than b .

Addition of Defense Branch (+DB) If a and b have the same ancestor's graph, which is not empty, except that a has an additional defense branch, then a should be strictly more acceptable than b .

Increase of Attack Branch (\uparrow AB) If a and b have the same ancestor's graph, except that one attack branch of a is longer than for b , then a should be strictly more acceptable than b .

Addition of Attack Branch (+AB) If a and b have the same ancestor's graph, except that a has an additional attack branch, then a should be strictly less acceptable than b .

Increase of Defense Branch (\uparrow DB) If a and b have the same ancestor's graph, except that one defense branch of a is longer than for b , then a should be strictly less acceptable than b .

Total (Tot) All arguments can be compared.

Attack vs Full Defense (AvsFD) A fully defended argument (without any attack branch) should be strictly more acceptable than an argument attacked once by a non-attacked argument.

Finally, we introduce a new property, which generalizes the property Non-attacked Equivalence [11], that says that all the non-attacked arguments are equally acceptable.

Argument Equivalence (AE) If two arguments have the same ancestor's graph, then they are equally acceptable.

It is important to note that the reverse is not always true. Indeed, on Example 1, f and g have different ancestor's graphs, *but* are equally acceptable when taking the multiset.

4.2. Properties Satisfied by Propagation

We are now in a position to check which of these properties are satisfied by our six ranking-based semantics based on propagation. A first remark is that if we choose $\varepsilon = 0$, the three kinds of semantics return exactly the same order.

Proposition 3. Let $\oplus \in \{M, S\}$, for all $AF \in \mathbb{AF}$,

$$Propa_{\varepsilon}^{0, \oplus}(AF) = Propa_{1+\varepsilon}^{0, \oplus}(AF) = Propa_{1 \rightarrow \varepsilon}^{0, \oplus}(AF)$$

In this case, only the weights propagated by the non-attacked arguments are taken into account. This can make sense, but when there is no non-attacked argument in the AF, all the arguments have a propagation vector composed only of 0, therefore trivializing the result. So we remove this possibility for studying the properties.

Let us begin to check which properties are satisfied by $Propa_\varepsilon$

Proposition 4. *Let $\oplus \in \{M, S\}$ and $\varepsilon \in]0, 1]$. $Propa_\varepsilon^{\varepsilon, \oplus}$ satisfies Abs, In, VP, DP, CT, SCT, $\uparrow AB$, $\uparrow DB$, +AB, Tot and AE. The other properties are not satisfied.*

$Propa_{1+\varepsilon}$ satisfies more properties.

Proposition 5. *Let $\oplus \in \{M, S\}$ and $\varepsilon \in]0, 1]$. $Propa_{1+\varepsilon}^{\varepsilon, \oplus}$ satisfies Abs, In, VP, DP, CT, SCT, DDP, $\uparrow AB$, $\uparrow DB$, +AB, Tot, AE and AvsFD. The other properties are not satisfied.*

Finally, let us list the properties satisfied by $Propa_{1 \rightarrow \varepsilon}$.

Proposition 6. *Let $\oplus \in \{M, S\}$ and $\varepsilon \in]0, 1]$. $Propa_{1 \rightarrow \varepsilon}^{\varepsilon, \oplus}$ satisfies Abs, In, VP, DP, DDP, +DB, $\uparrow AB$, $\uparrow DB$, +AB, Tot, AE and AvsFD. The other properties are not satisfied.*

First it is interesting to remark than choosing sets or multisets for the definitions, although clearly leading to different semantics, do not have any impact on the satisfaction of these properties. This may suggest that some properties allowing to make such a distinction are still missing.

Without surprise, the semantics satisfy the properties Abs, In, VP, DP, +AB and Tot, which are expected according to [11]. Contrastingly, CP, QP, SC and $\oplus DB$ are not satisfied by our semantics, but they describe very specific ranking-based semantics behaviors, which differ from the ones designed here. First of all, $\oplus DB$ is incompatible with VP [11] which is satisfied by all our semantics. CP and QP focus only on the direct attackers whereas our semantics look also at the impact of the non-attacked arguments in the entire graph. Finally, concerning the property SC, our semantics consider that an argument that attacks itself is a cycle with a length equal to 1. So an argument which attacks itself remains better than an argument which is attacked by a non-attacked argument.

Now, comparing the three families, $Propa_\varepsilon$, $Propa_{1+\varepsilon}$, and $Propa_{1 \rightarrow \varepsilon}$; $Propa_\varepsilon$ is the only one that does not satisfy DDP and AvsFD. Finally, we can see that $Propa_{1 \rightarrow \varepsilon}$ is the only one to satisfy +DB and to not satisfy CT and SCT because it is the only one to consider the defense as a reinforcement for the defended argument.

5. Links between Semantics

In this section, we establish the links between the six ranking based-semantics based on propagation, but also between these semantics and some ranking-based semantics existing in the literature. A first important remark is that $Propa_{1+\varepsilon}$ can be seen as a special case of $Propa_\varepsilon$. Let us make this link more formal.

Let $AF = \langle A, \hookrightarrow \rangle \in \mathbb{AF}$ be an argumentation framework, and $\oplus \in \{M, S\}$. We define $\maxdeg(AF)$, the maximal indegree of AF , as $\maxdeg(AF) = \max_{a \in A} |\downarrow_1^\oplus(a)|$.

Proposition 7. *Let $AF = \langle A, \hookrightarrow \rangle \in \mathbb{AF}$. For any $\varepsilon < \frac{1}{\maxdeg(AF)}$,*

$$Propa_{1+\varepsilon}^{\varepsilon, \oplus}(AF) = Propa_\varepsilon^{\varepsilon, \oplus}(AF)$$

But as we saw previously with the satisfied properties, even if, in the light of the above result, $Propa_{1+\varepsilon}$ could be considered as particular case of $Propa_\varepsilon$ for a given ε , it forms a sufficiently interesting subclass for being defined and studied on its own right.

In addition to the case where $\varepsilon = 0$, there is another particular situation where all the propagation semantics return the same order: when there exists no non-attacked argument in the argumentation framework.

Proposition 8. *Let $\oplus \in \{M, S\}$, $\forall AF = \langle A, \hookrightarrow \rangle \in \mathbb{AF}$, $\forall \varepsilon \in]0, 1]$, if $\nexists a \in A$ s.t. $\downarrow_1^\oplus(a) = \emptyset$ then $Propa_\varepsilon^{\varepsilon, \oplus}(AF) = Propa_{1+\varepsilon}^{\varepsilon, \oplus}(AF) = Propa_{1 \rightarrow \varepsilon}^{\varepsilon, \oplus}(AF)$.*

Indeed, if there is no non-attacked argument, for $Propa_{1+\varepsilon}$ and $Propa_{1 \rightarrow \varepsilon}$, the first case where $\varepsilon = 0$ returns the same propagation vector for all the arguments (for all argument a , $P^{0, \oplus}(a) = \langle 0, 0, \dots \rangle$). Consequently the only way to make a difference between arguments is to look at the case where $\varepsilon \neq 0$ exactly like $Propa_\varepsilon$. In other words, when there is no non-attacked argument, the semantics compare the arguments only on the number of attackers/defenders.

In this case, a link can be established between our ranking-based semantics and one semantics of the literature. The Discussion-based semantics [8] compares arguments by counting the number of direct attackers. If this number is the same for some arguments, the size of paths is recursively increased until a difference is found:

Definition 13. [8] *Let $AF = \langle A, \hookrightarrow \rangle$, $a \in A$, and $i \in \mathbb{N}$. Let $Dis_i(a) = (-1)^{i+1} |\downarrow_i^M(a)|$, and $Dis(a) = \langle Dis_1(a), Dis_2(a), \dots \rangle$. The **ranking-based semantics Dbs** associates to AF a ranking \succeq_{AF}^{Dbs} on A such that $\forall a, b \in A$, $a \succeq_{AF}^{Dbs} b$ iff $Dis(b) \succeq_{lex} Dis(a)$.*

Dbs and the propagation semantics share similar principles regarding the way paths are counting and use the lexicographical comparison. However, let us recall that, in the general case, our semantics also take into account the role of the non-attacked arguments which has consequences on the order between arguments. But in the case where there is no non-attacked argument, the order returned by both semantics is the same.

Proposition 9. *$\forall AF = \langle A, \hookrightarrow \rangle \in \mathbb{AF}$, $\forall \varepsilon \in]0, 1]$, if $\nexists a \in A$ such that $\downarrow_1^M(a) = \emptyset$, then $Propa_\varepsilon^{\varepsilon, M}(AF) = Propa_{1+\varepsilon}^{\varepsilon, M}(AF) = Propa_{1 \rightarrow \varepsilon}^{\varepsilon, M}(AF) = Dbs(AF)$.*

Note that this result is obtained with the multiset version of the three kinds of semantics. The set versions are not equivalent.

6. Example

In this section, we apply the different existing ranking-based semantics and the six semantics based on propagation on an example with few arguments. The objective is to illustrate their behaviors with regard to some particular situations. We consider the semantics based on Social Argumentation Frameworks *SAF* [3], the semantics Categoriser *Cat* [4,9], the semantics based on tuple² values *Tuples** [5], the semantics proposed by Matt and Toni *M&T* [6], the semantics proposed by Grossi and Modgil *G&M* [10], the Discussion-based semantics *Dbs* and the Burden-based semantics *Bbs* [8].

²In order to avoid infinite tuples, we consider this approach for acyclic graph only. See [11] for details.

Example 1 (cont.). For a better visibility of the obtained orders, we do not consider the argument e which is similar to a (both are non-attacked) and the arguments d and h which are similar to b (all attacked once by one non-attacked argument) in the final pre-order because they are always equally acceptable.

Semantics	Order between arguments	$\text{Propa}_e^{0.75,S}$	$a \succ^{P_S} c \succ^{P_S} b \succ^{P_S} f \succ^{P_S} g$
Cat	$a \succ^{\text{Cat}} c \succ^{\text{Cat}} b \simeq^{\text{Cat}} f \simeq^{\text{Cat}} g$	$\text{Propa}_e^{0.3,S}$	$a \succ^{P_S} c \succ^{P_S} f \succ^{P_S} g \succ^{P_S} b$
SAF	$a \succ^{\text{SAF}} c \succ^{\text{SAF}} f \simeq^{\text{SAF}} g \succ^{\text{SAF}} b$	$\text{Propa}_e^{0.75,M}$	$a \succ^{P_M} c \succ^{P_M} b \succ^{P_M} f \simeq^{P_M} g$
Tuples*	$a \succ^T f \simeq^T g \succ^T c \succ^T b$	$\text{Propa}_e^{0.3,M}$	$a \succ^{P_M} c \succ^{P_M} f \simeq^{P_M} g \succ^{P_M} b$
M&T	$a \succ^{\text{MT}} c \simeq^{\text{MT}} f \simeq^{\text{MT}} g \succ^{\text{MT}} b$	$\text{Propa}_{1 \rightarrow \epsilon}^{\epsilon,S}$	$a \succ^{\bar{P}_S} f \succ^{\bar{P}_S} c \succ^{\bar{P}_S} g \succ^{\bar{P}_S} b$
G&M	$a \succ^{\text{GM}} c \simeq^{\text{GM}} f \simeq^{\text{GM}} g \succ^{\text{GM}} b$	$\text{Propa}_{1 \rightarrow \epsilon}^{\epsilon,M}$	$a \succ^{\bar{P}_M} f \simeq^{\bar{P}_M} g \succ^{\bar{P}_M} c \succ^{\bar{P}_M} b$
Dbs	$a \succ^{\text{Dbs}} c \succ^{\text{Dbs}} b \succ^{\text{Dbs}} f \simeq^{\text{Dbs}} g$	$\text{Propa}_{1+\epsilon}^{\epsilon,S}$	$a \succ^{\hat{P}_S} c \succ^{\hat{P}_S} f \succ^{\hat{P}_S} g \succ^{\hat{P}_S} b$
Bbs	$a \succ^{\text{Bbs}} c \succ^{\text{Bbs}} b \succ^{\text{Bbs}} f \simeq^{\text{Bbs}} g$	$\text{Propa}_{1+\epsilon}^{\epsilon,M}$	$a \succ^{\hat{P}_M} c \succ^{\hat{P}_M} f \simeq^{\hat{P}_M} g \succ^{\hat{P}_M} b$

Table 3. Order obtained with the different semantics on AF_1 (Figure 1).

First of all, all semantics consider a (and e) as the best argument because it is non-attacked (see property VP). On the contrary, b (but also d and h) is most of the time the worst argument because it is attacked by the better argument. It is not the case with Dbs and Bbs because they satisfy the property Cardinality Precedence, where the greater the number of direct attackers for an argument, the weaker the level of acceptability of this argument. It is why, for these two semantics, b is better than f and g which are both defended. Note that it is also the case with Propa_e when $\epsilon > 0.5$.

It is interesting to note that almost all semantics make no distinction between f and g , both defended twice (by non-attacked arguments). Only our Propagation semantics using sets make a distinction between the two, preferring f that is defended by two arguments, whereas g is defended twice but by the same argument e .

Finally, concerning the three defended arguments (c, f and g), the order reflects the position of the semantics about the notion of defense. We can see that, for $\text{Propa}_{1 \rightarrow \epsilon}$ and Tuples*, f and g are better than c because they consider a defense as a reinforcement, contrary to all the others semantics.

7. Conclusion

In this work we proposed six new ranking-based semantics based on the propagation of the weights of arguments, that give a higher weight to non-attacked arguments. The differences between the six semantics lie in the choice of the interaction between attacked and non-attacked arguments (i.e. how much priority do we give to non-attacked arguments), and in the choice of sets or multi-sets as tracking of attacking and defending arguments.

The basic motivating idea behind these semantics, and one of the main contributions of this work, is that one can not take into account only information on attacks and defenses of an argument, but also has to take into account the impact of non-attacked arguments. This idea follows the principle of classical Dung's semantics. However it should be noted that full compatibility of rankings with extensions (sets of mutually acceptable arguments) is a difficult to reach objective, as these semantics do not capture the inter-

action between arguments and remain at the level of the acceptability of single arguments. For instance, two arguments may be highly ranked but mutually incompatible: ranking-based semantics are blind to this.

We show that these semantics have interesting properties. In particular they satisfy the properties that should be satisfied by any ranking semantics according to [11]. In particular the semantics $Propa_{1+\varepsilon}$ and $Propa_{1\rightarrow\varepsilon}$ satisfy the very natural AvsFD property, that is not satisfied by most of previously proposed ranked-based semantics.

We also show some relationships between these semantics and other ones: all the propagation semantics based on multisets coincide with the semantics Dbs when there is no non-attacked arguments in the AF. So they can be viewed as improvement of Dbs allowing to take into account the impact of non-attacked arguments.

Also, by many respect semantics $Propa_{1\rightarrow\varepsilon}$ is close to the Tuples* semantics [5]. The Tuples* semantics does not necessarily provide a total pre-order, and it cannot be applied (easily) if there is a cycle in the AF. So in a sense $Propa_{1\rightarrow\varepsilon}$ could be seen as an improvement of the ideas of Tuples* that allows to overcome these limitations.

This work on ranked-based semantics is motivated by applications for online debates platforms. On these platforms people can usually vote on arguments and/or on attacks. So this provides weights on the arguments and on the attacks. The SAF framework [3] allows to take these information into account. We started with the basic framework, without any weights. Now the plan is to study the full framework, with weights on attacks and on arguments. We want to study how to generalize these semantics with weights, and to study which are the adaptations of the properties, or the missing ones, in this case.

References

- [1] P. M. Dung, "On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games," *Artificial Intelligence*, vol. 77, no. 2, pp. 321–358, 1995.
- [2] M. Caminada, "On the issue of reinstatement in argumentation," in *Proc. of the 10th European Conference on Logics in Artificial Intelligence, (JELIA'06)*, pp. 111–123, 2006.
- [3] J. Leite and J. Martins, "Social abstract argumentation," in *Proc. of the 22nd International Joint Conference on Artificial Intelligence, (IJCAI'11)*, pp. 2287–2292, 2011.
- [4] P. Besnard and A. Hunter, "A logic-based theory of deductive arguments," *Artificial Intelligence*, vol. 128, no. 1-2, pp. 203–235, 2001.
- [5] C. Cayrol and M. Lagasquie-Schiek, "Graduality in argumentation," *Journal of Artificial Intelligence Research*, vol. 23, pp. 245–297, 2005.
- [6] P. Matt and F. Toni, "A game-theoretic measure of argument strength for abstract argumentation," in *Proc. of the 11th European Conference on Logics in Artificial Intelligence, (JELIA'08)*, pp. 285–297, 2008.
- [7] C. da Costa Pereira, A. Tettamanzi, and S. Villata, "Changing one's mind: Erase or rewind?," in *Proc. of the 22nd International Joint Conference on Artificial Intelligence, (IJCAI'11)*, pp. 164–171, 2011.
- [8] L. Amgoud and J. Ben-Naim, "Ranking-based semantics for argumentation frameworks," in *Proc. of the 7th International Conference on Scalable Uncertainty Management, (SUM'13)*, pp. 134–147, 2013.
- [9] F. Pu, J. Luo, Y. Zhang, and G. Luo, "Argument ranking with categoriser function," in *Proc. of the 7th International Conference on Knowledge Science, Engineering and Management, (KSEM'14)*, pp. 290–301, 2014.
- [10] D. Grossi and S. Modgil, "On the graded acceptability of arguments," in *Proc. of the Twenty-Fourth International Joint Conference on Artificial Intelligence, (IJCAI'15)*, pp. 868–874, 2015.
- [11] E. Bonzon, J. Delobelle, S. Konieczny, and N. Maudet, "A Comparative Study of Ranking-based Semantics for Abstract Argumentation," in *Proc. of the 30th AAAI Conference on Artificial Intelligence (AAAI'16)*, pp. 914–920, 2016.

Using Argument Features to Improve the Argumentation Process

Maximiliano C. D. Budán ^{a,b,c} Gerardo I. Simari ^{a,b} Guillermo R. Simari ^a

^aArtificial Intelligence R&D Laboratory (LIDIA), ICIC

Universidad Nacional del Sur – Alem 1253, (8000) Bahía Blanca, Argentina

^bConsejo Nacional de Investigaciones Científicas y Técnicas

Av. Rivadavia 1917, (1000-1499) Buenos Aires, Argentina

^cDepartment of Mathematics – Universidad Nacional de Santiago del Estero

Belgrano(s) 1912, (4200) Capital, Sgo. del Estero, Argentina

Abstract. Argumentation has become an important topic in artificial intelligence; the basic idea is to identify arguments in favor and against a statement, select the acceptable ones, and determine whether the original statement can be accepted or not. However, the arguments involved in an argumentative discussion may have different relevance degrees; for this reason, argumentation frameworks need to represent the qualities that describe the soundness of an argument in order to refine the acceptability process performed over the argumentation model.

Keywords. Abstract Argumentation, Bipolar Models, Quantitative Analysis

1. Introduction and Preliminary Concepts

Argumentation aims towards formalizing reasoning mechanisms with the capability of handling contradictory, incomplete and/or uncertain information [13], taking as inspiration commonsense reasoning and the human-like mechanism of defending a given statement by giving reasons for its acceptance. In this process, both the original statement and its support are subject to scrutiny, since reasons supporting conflicting conclusions can also be advanced. Argumentation theories have been proposed for applications in many different domains, such as legal reasoning [12], dialogue and persuasion [16], recommender systems [6], agents and MAS [9], cyber security [14], and others [11,15]. Several argument-based formalisms have emerged to study the different relations among arguments. In [7], Dung proposes *Abstract Argumentation Frameworks (AF)* to model real-world situations representing the attack relations between abstract entities called *arguments*, and providing different acceptability semantics to determine which sets of arguments are acceptable. Subsequently, Cayrol and Lagasquie-Schiex [5] extended Dung's framework taking into account two independent relations between arguments: *attack* and *support*. In this formalism, called *Bipolar Argumentation Frameworks (BAF)*, the authors allow to model situations in which one argument reinforces another, giving more reasons to believe in it. In addition, they adapt Dung's acceptability semantics taking into account the support relationship between arguments.

Although these formalizations model certain aspects of real-world situations, they do not provide tools to represent the particular features of arguments that affect the relations (support and attack) between the arguments involved in an argumentative discussion. However, in certain applications it is necessary to provide further details about the arguments, considering their features in order to refine the analysis and provide extra information about their acceptance [1]. In this work, we extend *BAF* by taking into account the properties associated with the arguments in the form of *labels*, increasing the representational capabilities of this formalization. These labels can be combined and propagated through the bipolar argumentation graph in accordance with the arguments' interaction. Then, using the extra information provided by these labels, we can improve the semantics offered by *BAF* by: (i) deriving more information regarding argument acceptability, (ii) determining special coefficients associated with the argumentation model that represent the *effectiveness* of the support and conflict relation, (iii) refining the argumentation model in order to improve the argumentative discussion excluding the less relevant arguments, and (iv) defining new acceptability extensions.

Algebra of Argumentation Labels

The use of labels gives us the possibility of representing *distinctive features* of arguments; these labels change according to the existing relations between arguments. Following this idea, we use an algebrization that consists of a set of labels equipped with a collection of operators to be used in combining and propagating the labels according argument interactions [2]. The algebra is based on an ordered set, allowing the comparison of labels; this set is also characterized in an abstract way to allow the adaptation to different applications. A natural way of representing this information is to use a scale that measures a particular *feature* of the argument. We will consider valuations ranging between two distinguished elements: \top and \perp , where \top represents the least possible degree in which an argument may possess a certain attribute, and \perp the maximum.

Definition 1. An algebra of argumentation labels is a 6-tuple of the form $A = \langle \mathcal{L}, \leq, \oplus, \ominus, \top, \perp \rangle$, where:

- \mathcal{L} is a set of labels called the domain of labels, where \top and \perp are two distinguished elements. \top is the last label with respect to \leq , while \perp is the first.
- \leq is a partial order over \mathcal{L} (that is, a reflexive, antisymmetric, and transitive relation).
- $\oplus : \mathcal{L} \times \mathcal{L} \rightarrow \mathcal{L}$ is called a support operation, and it satisfies: (i) commutativity; (ii) associativity; (iii) monotonicity; and (iv) \perp is the identity element.
- $\ominus : \mathcal{L} \times \mathcal{L} \rightarrow \mathcal{L}$ is called a conflict operation, and it satisfies: (i) for all $\alpha, \beta \in \mathcal{L}$, $\alpha \ominus \beta \leq \alpha$ if $\beta < \alpha$; (ii) for all $\alpha, \beta \in \mathcal{L}$, $\alpha \ominus \beta = \perp$ if $\beta \geq \alpha$; (iii) for all $\alpha, \beta \in \mathcal{L}$, if $\alpha \ominus \beta = \perp$ and $\beta \ominus \alpha = \perp$, then $\alpha = \beta$; (iv) monotonicity; and (v) \perp is the identity element.

The support operation, denoted with \oplus , is used to determine the strengthened valuation of an argument based on the weakened valuations of the arguments that support it. It is clear that one wants this operation to be invariant with respect to the order in which the supporting arguments are considered, and therefore the operation is both commuta-

tive and associative, with \perp as the neutral element. The conflict operation, denoted with \ominus , determines the valuation of an argument after considering the weakened valuations associated with the attacker arguments. The \ominus operation enjoys the monotonicity property, ensuring that the valuation of an argument decreases if the valuation of its attackers increases. Also, \perp is the neutral element for \ominus , specifying that the valuation associated with an argument is not affected by counterarguments with the least possible valuation.

Bipolar Abstract Argumentation

In the argumentation domain, arguments have different roles with respect to each other; one might then say that arguments are presented in a *bipolar* way since those in favor of a conclusion can be considered as positive while those against the conclusion as negative. Abstracting away from the inner structure of the arguments, the *Abstract Bipolar Argumentation Framework* proposed in [5] extends Dung's notion of acceptability by distinguishing two independent forms of interaction between arguments: support and attack.

Definition 2. A *Bipolar Argumentation Framework (BAF)* is a 3-tuple $\Theta = \langle \text{Arg}, R_a, R_s \rangle$, where Arg is a set of arguments, and R_a, R_s are disjoint binary relations on Arg called *attack* and *support*, respectively.

In *BAF*, the notion of graph presented by Dung [7] is extended by adding the representation of support between arguments. This argumentation model provides a starting point for enriching the analysis of discussions with the bipolarity of human reasoning.

Definition 3. Let $\Theta = \langle \text{Arg}, R_a, R_s \rangle$ be a *BAF*. The *directed graph* for Θ , denoted as G_Θ , is defined by taking as nodes the elements in Arg , and two types of arcs: one for the *attack* relation, and one for the *support* relation.

In addition, we have the notions of *supported* and *secondary* defeat, which combine a sequence of supports with a direct defeat in order to consider the interaction between supporting and defeating arguments. In this work we will be considering *well-founded BAFs*, which are those *BAFs* with no infinite path (therefore no cycles, no self-attacking nor self-supporting arguments).

Definition 4. Let $\Theta = \langle \text{Arg}, R_a, R_s \rangle$ be a *BAF*, and $A, B \in \text{Arg}$ two arguments. Then: (i) a *supported defeat* from A to B is a sequence $A_1 R_1 \dots R_{n-1} A_n$, with $n \geq 3$, where $A_1 = A$ and $A_n = B$, such that $\forall i = 1 \dots n-2, R_i = R_s$ and $R_{n-1} = R_a$; and (ii) a *secondary defeat* from A to B is a sequence $A_1 R_1 \dots R_n A_n$, with $n \geq 3$, where $A_1 = A$ and $A_n = B$, such that $R_1 = R_a$ and $\forall i = 2 \dots n-1, R_i = R_s$.

In [5], Cayrol and Lagasque-Schiex argued that a set of arguments must be in some sense coherent to model one side of an intelligent dispute. The coherence of a set of arguments is analyzed *internally* (a set of arguments in which an argument attacks another in the same set is not acceptable), and *externally* (a set of arguments that contains both a supporter and an attacker for the same argument is not acceptable). The internal coherence is captured by extending the definition of *conflict free set* proposed in [7], and external coherence is captured with the notion of *safe set*.

Definition 5. Let $\Phi = \langle \text{Arg}, R_a, R_s \rangle$ be a BAF and $S \subseteq \text{Arg}$. Then: (i) S is Conflict-free iff $\nexists A, B \in S$ s.t. there is a supported or secondary defeat from A to B ; and (ii) S is Safe iff $\nexists A \in \text{Arg}$ and $\nexists B, C \in S$ s.t. there is a supported or secondary defeat from B to A , and either there is a sequence of support from C to A , or $A \in S$.

The notion of conflict-freeness requires to take supported and secondary defeats into account, becoming a more restrictive definition than the classical version of conflict-freeness proposed by Dung. In addition, the notion of safety was shown to be powerful enough to encompass conflict-freeness. The closure under R_s , which concerns only the support relation, was also considered in [5].

Definition 6. Let $\Phi = \langle \text{Arg}, R_a, R_s \rangle$ be a BAF and $S \subseteq \text{Arg}$. S is closed under R_s iff $\forall A \in S, \forall B \in \text{Arg}$, if $A R_s B$ then $B \in S$.

Based on the previous concepts, the notion of defense for an argument with respect to a set of arguments is extended by taking into account the relations of support and conflict.

Definition 7. Let $\Phi = \langle \text{Arg}, R_a, R_s \rangle$ be a BAF, $S \subseteq \text{Arg}$ be a set of arguments and $A \in \text{Arg}$ be an argument. S defends collectively A iff $\forall B \in \text{Arg}$, if B is a supported or secondary defeat of A then $\exists C \in S$ such that C is a supported or secondary defeat of B . In this case, it can be interpreted that C defends A from B .

Three admissibility notions were proposed, from the most general (based on Dung's definition) to the most specific (considering external coherence and closure under R_s).

Definition 8. Let $\Phi = \langle \text{Arg}, R_a, R_s \rangle$ be a BAF and $S \subseteq \text{Arg}$. The admissibility of a set S is defined as follows: (i) S is cf-admissible if S is conflict-free and defends all its elements; (ii) S is s-admissible if S is safe and defends all its elements; and (iii) S is c-admissible if S conflict-free, closed for R_s , and defends all its elements.

The following new semantics were proposed in [5] based on the notions of coherence and admissibility, and by extending the propositions introduced in [7].

Definition 9. Let $\Phi = \langle \text{Arg}, R_a, R_s \rangle$ be a BAF and $S \subseteq \text{Arg}$. S is a stable extension of Φ if S is conflict-free and for all $A \notin S$, there is a supported or a secondary defeat of A in S . S is a cf-preferred (resp. s-preferred, c-preferred) extension if S is maximal (for set-inclusion) among the cf-admissible (resp. s-admissible, c-admissible) subsets of Arg .

Note that in [5] all attacking arguments are considered as defeating the attacked argument, while in the extension that we propose in the next section attacks can be successful or not. We adapted some of the definitions presented in this section to reflect this, and we propose how we refine the argumentation model to consider only the relevant arguments.

2. Labeled Bipolar Argumentation Frameworks

In previous work [3], we presented an early version of this formalism; the novel aspects in the present paper include (i) the refinement of the support and conflict coefficients into

particular and general for a finer-grained analysis of argument impact, (ii) the proposal of nine kinds of extensions resulting from combining the classical bipolar extensions with the results of the labeling process, and (iii) the analysis of underlying principles for the labeling process.

Definition 10. A Labeled Bipolar Argumentation Framework (L-BAF) is a 5-tuple $\Psi = \langle \text{Arg}, R_a, R_s, A_s, F_v \rangle$, where $\langle \text{Arg}, R_a, R_s \rangle$ is a Bipolar Argumentation Framework, A_s is a set of Algebras of Argumentation Labels A_1, A_2, \dots, A_n (one for each feature represented by the labels), and F_v is a function that assigns to each element of Arg an n -tuple of elements in the algebras $A_i, i = 1, \dots, n$. That is, $F_v : \text{Arg} \longrightarrow A_1 \times A_2 \times \dots \times A_n$.

Notation: Given $A \in \text{Arg}$, the set $\{A_i \in \text{Arg} \mid A_i R_a A\}$ is denoted with $S^{\rightarrow}(A)$, and the set $\{A_i \in \text{Arg} \mid A_i R_s A\}$ is denoted with $S^{\neg\rightarrow}(A)$.

Definition 11. Let Ψ be an L-BAF, Θ the underlying BAF, G_Θ the associated bipolar graph, and A_i be an algebra in A_s . A labeled bipolar graph is an assignment of three valuations in each of the algebras to each argument A defined in Θ , denoted with $\langle \alpha_i^A, \mu_i^A, \delta_i^A \rangle$, where α_i^A is the original value of the attribute assigned to the argument by F_v , μ_i^A accounts for the aggregation of the attributes of arguments supporting A , and δ_i^A is obtained after taking the attacks into account. If A is an argument defined in Θ , its valuations are determined as follows: i) $\alpha_i^A = F_v(A)$; ii) If $S^{\neg\rightarrow}(A) = \emptyset$, then $\mu_i^A = \alpha_i^A$; iii) If $S^{\rightarrow}(A) = \emptyset$, then $\delta_i^A = \mu_i^A$; iv) If $S^{\neg\rightarrow}(A) \neq \emptyset$, then $\mu_i^A = \alpha_i^A \oplus (\oplus_{j=1}^n \delta_i^{A_j})$, with $A_j \in S^{\neg\rightarrow}(A)$; and v) If $S^{\rightarrow}(A) \neq \emptyset$, then $\delta_i^A = \mu_i^A \ominus (\oplus_{j=1}^m \delta_i^{B_j})$, with $B_j \in S^{\rightarrow}(A)$.

For each $A \in \text{Arg}$ and for each algebra A_i in A_s representing a feature to be associated with A , the triple $\langle \alpha_i^A, \mu_i^A, \delta_i^A \rangle$ is called the *label* of A with respect to A_i . The following proposition describes the relationship between these valuations.

Proposition 1. Let Ψ be an L-BAF, Θ the underlying BAF, G_Θ the associated graph, A_i be one of the algebras in A_s , and A be an argument in G_Θ . Then, the labels $\langle \alpha_i^A, \mu_i^A, \delta_i^A \rangle$ related to algebra A_i satisfy: (i) $\mu_i^A \geq \delta_i^A$; (ii) $\mu_i^A \geq \alpha_i^A$; and (iii) If $\mu_i^A = \perp_i$, then $\delta_i^A = \alpha_i^A = \perp_i$.

The following underlying principles are satisfied by all the valuations defined according to the labeling process. In general, these principles describe the behavior of valuations associated with arguments in our framework.

Property. The valuations given by Definition 11 respect the following principles:

The Weakened and Strengthened Valuations (The weakened valuation is equal to the strengthened valuation for the arguments without attackers; for an attacked but undefeated argument, the weakened valuation is less than the strengthened valuation, whenever the attacking arguments are strong enough to weaken it);

Attack Strength (The weakened valuation for an argument depends, in a non-increasing manner, on the weakened valuation of the attacking argument); and

Support Strength (The weakened valuations of the supporting arguments contribute to increase the strengthened valuation of the supported argument).

Unlike the original BAF, in our proposal an attack does not always mean a defeat. Based on the conflict operator defined in the algebra of argumentation labels, attacks between arguments can produce a weakening of the valuations associated with the attacked argument that can result in a defeat, a weakening, or a strengthening depending on the attacking arguments' strength, while no effect over the valuations associated with the attacked argument is produced when all the attacking arguments have the least possible feature degree.

Definition 12. Let $\Psi = \langle \text{Arg}, R_a, R_s, A_s, F_v \rangle$ be an L-BAF, G_Ψ be the corresponding labeled bipolar argumentation graph, and A be an argument (node) in G_Ψ . For each of the algebras A_i in A_s , A has assigned one of four possible statuses: (i) Defeated iff $\delta_i^A = \perp_i$; (ii) Weakened iff $\perp_i < \delta_i^A < \mu_i^A$; (iii) Strengthened iff $\alpha_i^A < \delta_i^A$; and (iv) Unchallenged iff $\mu_i^A = \delta_i^A \neq \perp_i$. Finally, for each argument, we form a vector with the acceptability of that argument with respect to each of the attributes, and take the least degree of those that appear in the vector as the acceptability degree for the argument as a whole. We denote with S_g the gradual status assignment to the bipolar graph G_Ψ .

Based on the status assigned to each argument, it is possible to partition Arg into four categories: *defeated*, denoted as S^d ; *weakened*, denoted as S^w ; *strengthened*, denoted as S^s ; and *unchallenged*, denoted as S^u .

Proposition 2. The gradual status assignment S_g to the bipolar graph G_Ψ is unique.

Proposition 3. Let S_g be the gradual status assignment to the bipolar graph G_Ψ , and S^d , S^w , S^s , and S^u be the set of defeated, weakened, strengthened and unchallenged arguments in Ψ . Then: $\{S^d, S^w, S^s, S^u\}$ is a partition of Arg .

We generally wish to determine a consistent set of arguments in favor or against certain conclusions, commonly referred to as the *semantics of acceptability*. In this work, we will use the status associated with the arguments to define our semantics. Clearly, these sets may not always be conflict-free or safe. Here, we use a preference relation, denoted as \succ , defined over Arg that uses the extra information contained in the labels in order to obtain the different conflict-free subsets of arguments corresponding to each particular set of arguments.

Definition 13. Let $\Psi = \langle \text{Arg}, R_a, R_s, A_s, F_v \rangle$ be an L-BAF, \succ a preference relation over Arg , and $S \subseteq \text{Arg}$. Then, S is conflict-free iff for all $A, B \in \text{Arg}$ s.t. A attacks B , if $A \succ B$ then $A \in S$ (and thus $B \notin S$).

Definition 14. Let $\Psi = \langle \text{Arg}, R_a, R_s, A_s, F_v \rangle$ be an L-BAF, and S^w , S^s , and S^u the sets of weakened, strengthened, and unchallenged arguments associated with Ψ . Then: (i) $S \subseteq S^w$ is a CF-W extension of Φ if S is a maximal conflict-free set; (ii) $S \subseteq S^w$ is an S-W extension of Φ if S is a maximal safe set; (iii) $S \subseteq S^w$ is a C-W extension of Φ if S is a maximal conflict-free and closed under R_s ; (iv) $S \subseteq S^u$ is a CF-U extension of Φ if S is a maximal conflict-free set; (v) $S \subseteq S^u$ is an S-U extension of Φ if S is a maximal safe set; (vi) $S \subseteq S^u$ is a C-U extension of Φ if S is maximal conflict-free and closed under R_s ; (vii) $S \subseteq S^s$ is a CF-S extension of Φ if S is a maximal conflict-free set; (viii) $S \subseteq S^s$ is an S-S extension of Φ if S is a maximal safe set; and (ix) $S \subseteq S^s$ is a C-S extension of Φ if S is a maximal conflict-free and closed under R_s .

Proposition 4. Let $\Psi = \langle \text{Arg}, R_a, R_s, A_s, F_v \rangle$ be an L-BAF, and S^w , S^s , and S^u the sets of weakened, strengthened, and unchallenged arguments associated with Ψ . Then: (i) For every CF-W (resp. CF-S, CF-U) extension S_1 there exists S-W (resp. S-S, S-U) extension S_2 such that $S_1 \subseteq S_2$; (ii) For every S-W (resp. S-S, S-U) extension S_1 there exists C-W (resp. C-S, C-U) extension S_2 such that $S_1 \subseteq S_2$; and (iii) Any S-W (resp. S-S, S-U) extension closed under R_s is also a C-W (resp. C-S, C-U) extension.

In certain applications, it is necessary to improve the argumentative model considering only those arguments that satisfy the constraints imposed by the domain. Having extra information associated with arguments introduces the possibility of analyzing and refining the argumentation model to improve the argumentative discussion by excluding the least relevant arguments. To do this, we calculate the ‘coefficients of conflict and support’ of the model, indicating the efficiency of the relations among the arguments.

Definition 15. Let $\Psi = \langle \text{Arg}, R_a, R_s, A_s, F_v \rangle$ be an L-BAF, G_Ψ be the corresponding labeled bipolar argumentation graph, $A_s = \{A_1, \dots, A_n\}$ the set of algebras labels, and $|R_a|$ and $|R_s|$ be the cardinalities of the attack and support relations, respectively. Then, the effectiveness degree associated with the argument roles in Ψ is defined (denominators assumed to be non-zero, otherwise the coefficient is 0):

Particular Conflict Cf.	Particular Support Cf.	General Conflict Cf.	General Support Cf.
$\omega_i^a = \frac{\sum_{A \in \text{Arg}} \mu_i^A - \delta_i^A}{ R_a }$	$\omega_i^s = \frac{\sum_{A \in \text{Arg}} \mu_i^A - \alpha_i^A}{ R_s }$	$\Omega^a = \frac{\sum_{A_i \in A_s} \omega_i^a}{n}$	$\Omega^s = \frac{\sum_{A_i \in A_s} \omega_i^s}{n}$

Remark 1. If $\Omega^s = 0$ and $\Omega^a = 1$, then the labeled bipolar argumentation framework is equivalent to a Dung framework.

Remark 2. If $\Omega^s = 1$ and $\Omega^a = 1$, then the label bipolar argumentation framework is equivalent to a Bipolar argumentation framework.

In particular, if $\omega_i^a = 0$ and $\omega_i^s = 1$ for only a subset of attributes, then the same property holds – this can be applied, for instance, in a simplified analysis (for those attributes) based on Dung frameworks instead of labeled BAF frameworks. The same observation applies in the case in which $\omega_i^a = 1$ and $\omega_i^s = 1$.

3. Related Work and Conclusions

In [1], Bench-Capon persuasively posits that in situations involving practical reasoning, it is impossible to demonstrate conclusively that either party is wrong; thus, in such cases the role of argumentation is to persuade rather than to prove, demonstrate, or refute. In his own words: “*The point is that in many contexts the soundness of an argument is not the only consideration: arguments also have a force which derives from the value they advance or protect.*”. In a similar way, in [10] Pollock points out the fact that, in defeasible reasoning, most semantics ignore the issue of the *inner force* of arguments, i.e., that some arguments support their conclusions more strongly. But once we acknowledge that arguments can differ in strength and conclusions can differ in their degree of justification, things become more complicated. In particular, Pollock introduces the notion of

diminishers, which are defeaters that cannot completely defeat their target, but instead lower the degree of justification of that argument. Both of these ideas motivate our work. In a similar vein, in [8] Dunne *et al.* explore a natural extension of Dung's well-known model of argument systems in which attacks are associated a *weight* indicating the relative strength of the attack. In [4], Cayrol and Lagasquie-Schiex argue that argumentation is based on the exchange and valuation of interacting arguments, followed by the selection of the most acceptable of them. They propose the notion of "graduality" in the selection of the best arguments in order to represent different levels of acceptability.

Based on the intuitions of these research lines, we combine *Bipolar Argumentation Frameworks* with *Algebras of Argumentation Labels* in order to extend the representation capability of argument structures; in this system the labels represent argument features, generalizing the notion of *value* and *weight*. Moreover, the interaction between arguments can affect their labels, causing strengthening and weakening among arguments. Thus, the information contained in the labels allows us to improve the analysis performed over the argumentation model and refine it using only the set of relevant arguments. In particular, in this expanded framework it is possible to determine the acceptability of sets of arguments, as well as additional information justifying their acceptability status.

Acknowledgments. This work was supported by funds from Universidad Nacional del Sur in Bahía Blanca and CONICET, Argentina.

References

- [1] T. Bench-Capon. Persuasion in practical argument using value-based argumentation frameworks. *Logic and Computation*, 13(3):429–448, 2003.
- [2] M. C. D. Budán, M. Lucero Gómez, I. Viglizzo, and G. R. Simari. A labeled argumentation framework. *Applied Logic*, 13(4):534–553, 2015.
- [3] M. C. D. Budán, I. Viglizzo, and G. R. Simari. A labeled abstract bipolar argumentation framework. In *Advances in Artificial Intelligence*, pages 28–40. Springer, 2014.
- [4] C. Cayrol and M. C. Lagasquie-Schiex. Graduality in argumentation. *Journal of Artificial Intelligence Research (JAIR)*, 23:245–297, 2005.
- [5] C. Cayrol and M. C. Lagasquie-Schiex. On the acceptability of arguments in bipolar argumentation frameworks. In *Proc. of ECSQARU*, pages 378–389. Springer, 2005.
- [6] C. I. Chesñevar and A. G. Maguitman. Arguenet: An argument-based recommender system for solving web search queries. In *Intelligent Systems*, volume 1, pages 282–287. IEEE, 2004.
- [7] P. M. Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning and logic programming and *n*-person games. *Artificial Intelligence*, 77:321–357, 1995.
- [8] P. E. Dunne, A. Hunter, P. McBurney, S. Parsons, and M. Wooldridge. Weighted argument systems: Basic definitions, algorithms, and complexity results. *Artif. Intell.*, 175(2):457–486, 2011.
- [9] S. Kraus. Negotiation and cooperation in multi-agent environments. *Artif. Intell.*, 94(1):79–97, 1997.
- [10] J. L. Pollock. Defeasible reasoning and degrees of justification. *Arg. and Comp.*, 1(1):7–22, 2010.
- [11] D. L. Poole, A. K. Mackworth, and R. Goebel. *Computational intelligence: a logical approach*, volume 79. Oxford University Press New York, 1998.
- [12] H. Prakken and G. Sartor. *Logical models of legal argumentation*. Springer, 1997.
- [13] I. Rahwan and G. R. Simari. *Argumentation in Artificial Intelligence*. Springer Verlag, 2009.
- [14] P. Shakarian, G.I. Simari, G. Moores, D. Paulo, S. Parsons, M. Falappa, and A. Aleali. Belief revision in structured probabilistic argumentation: Model and application to cyber security. *AMAI, In Press*.
- [15] G. A. W. Vreeswijk. Abstract argumentation systems. *Artificial intelligence*, 90(1):225–279, 1997.
- [16] D. Walton and D. M. Godden. Persuasion dialogue in online dispute resolution. *Artificial Intelligence and Law*, 13(2):273–295, 2005.

Minimal Cost Semantics in Argumentation Framework on Semiring Cost Assignment

Shuya BUNDO¹, and Kazunori YAMAGUCHI

The University of Tokyo

Abstract. Cost based methods have been proposed as ways to extend Argumentation frameworks proposed by Dung but have been found to contain an anomaly. In this paper, we introduce a new cost based method that is free from the anomaly. We also describe a preliminary algorithm to calculate the cost.

Keywords. Argumentation Frameworks, Argumentation Semantics, C-semiring, Weighted Argumentation Frameworks

1. Introduction

An argumentation framework (AF) reported in [5] is a widely accepted framework for argumentation. In AFs, complete extensions are considered suitable to represent the extensions of the subjective standpoint of the decision maker [4]. It is often the case that AF has multiple complete extensions. In [2], c-semiring values are used to select the “best” complete extension. However, in their formalization, one can decrease the cost to accept arguments by adding a critical argument to them as detailed in Example 5.7. This is contrary to intuition. In this paper, we introduce a formalism that uses an AF structure and is free from such counter-intuitive behavior.

This paper is organized as follows. In Section 2 we present preliminary definitions. In Section 3 we introduce a model to accept arguments incrementally in order to determine the set of arguments we must accept explicitly when we have to accept given arguments in a complete extension. Then in Section 4 we explore properties and generalize definitions. We define the cost for accepting given arguments in Section 5. Section 6 concludes the paper with a brief summary.

2. Preliminary Definitions

We call $F = \langle A, \rightarrow \rangle$ an *argumentation framework* (AF). Here, A is a finite set of arguments and $\rightarrow \subseteq A \times A$ is a binary relation on A representing attack relations. We denote A of $F = \langle A, \rightarrow \rangle$ as $A(F)$ and \rightarrow as \rightarrow_F . The relation $a \rightarrow_F b$ means that the argument a attacks the argument b in F .

¹bundo@graco.e.u-tokyo.ac.jp

For $S, T \subseteq A(F)$, $S \rightarrow_F T$ iff there exist $s \in S$ and $t \in T$ such that $s \rightarrow_F t$. We write $s \rightarrow_F T$ instead of $\{s\} \rightarrow_F T$ and $S \rightarrow_F t$ instead of $S \rightarrow_F \{t\}$. For $S \subseteq A(F)$, if $S \rightarrow_F S$ does not hold, we call S *conflict-free*. For $S \subseteq A(F)$ and $a \in A(F)$, if $b \rightarrow a$ for $b \in A(F)$ then $S \rightarrow b$, we say that S *defends* a . Let $\text{Def}_F(S) = \{a \mid S \text{ defends } a\}$. If S is conflict-free and $S \subseteq \text{Def}_F(S)$, we call S *admissible*. In particular, if S is admissible and $S = \text{Def}_F(S)$, we call S *complete*. We denote the composition of Def_F by n times as Def_F^n . $\text{Def}_F^0(S) = S$. If S is admissible, $\text{Def}_F(S)$ is also admissible. The increasing sequence of $\text{Def}_F^n(S)$ converges at sufficiently large n and the limit is complete. We denote the limit as $\text{Comp}_F(S)$. We define $\text{Atk}_F(S) = \{a \in A(F) \mid S \rightarrow a\}$ and $\text{Undec}_F(S) = A(F) - (S \cup \text{Atk}_F(S))$. If S is conflict-free, S , $\text{Atk}_F(S)$, and $\text{Undec}_F(S)$ are the partition of $A(F)$. For $A' \subseteq A(F)$, we define $F \downarrow_{A'} = \langle A', \rightarrow_F \cap A' \times A' \rangle$, which we call the *restriction* of F to A' .

3. Progress of Argumentation

Complete extensions are possible and reasonable positions one can take because one cannot point out internal inconsistency even though one can disagree on them [4]. Here, complete extensions are used instead of admissible extensions on the basis of the greedy principle that if one accepts admissible E , one has to accept $\text{Def}_F(E)$ that are all the arguments that E defends and its closure $\text{Comp}_F(E)$. Here, we distinguish two acceptance by calling accepting E “we agree on E ” and accepting $\text{Comp}_F(E) - E$ by *justifying* $\text{Comp}_F(E) - E$. An argument that is attacked by some accepted argument is called *defeated*.

As an example, we represent the arguments on bacteria in [4] as AF $F = \langle \{0, 1, 2, 3\}, \{0 \rightarrow 1, 1 \rightarrow 0, 0 \rightarrow 2, 1 \rightarrow 2, 2 \rightarrow 3\} \rangle$. In this F , the arguments 0 and 1 attack each other and no other argument attacks them. Therefore, we cannot justify either argument objectively. If we agree on the admissible set $\{0\}$, the argument 3 is automatically justified because $\text{Comp}(\{0\}) = \{0, 3\}$.

Up to now, we restricted what we agree on to an admissible extension because we cannot agree on an argument set that has contradiction or that cannot defend itself. However, if we consider argumentation as a process of agreeing on arguments in a step-by-step manner, such restriction is not necessary. Let us consider arguments 4, 5, and 6 and attacking relations $3 \rightarrow 4$, $4 \rightarrow 5$, $5 \rightarrow 6$, and $6 \rightarrow 5$ are added to the previous F as shown in Figure 1.

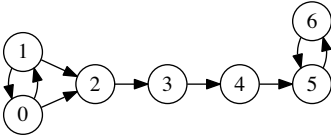


Figure 1. Example for Incremental Acceptance

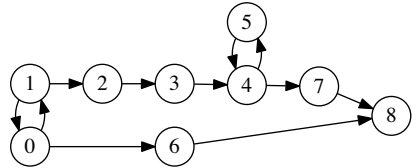


Figure 2. Example for Cond_F

As before, if we agree on 0, 0 and 3 are accepted. Then, 1, 2, and 4, which are attacked by 0 and 3, are defeated. The remaining arguments are 5 and 6. Because they are attacking each other there is no objective choice. Let us agree on 5 and accept it explicitly. As a result, we accepted $\{0, 3, 5\}$ by agreeing on $B = \{0, 5\}$ and justifying $\{3\}$. Note that B is not admissible.

Now let us formalize this step-by-step process.

Definition 3.1. We call (E_1, E_2, \dots, E_n) “progress of argumentation” of AF F if there exist AF $F_i (1 \leq i \leq n)$ such that $F_1 = F$, E_i is admissible in $F_i (1 \leq i \leq n)$, and $F_{i+1} = F_i \downarrow_{\text{Undec}_{F_i}(\text{Comp}_{F_i}(E_i))} (1 \leq i < n)$. We define $\text{Cond}_F(C) = \{B \mid B \text{ is conflict-free and } \exists (E_1, \dots, E_n): \text{progress of argumentation such that } \cup_{i=1}^n E_i \subseteq B \text{ and } C = \cup_{i=1}^n \text{Comp}_{F_i}(E_i)\}$ for $C \subseteq A(F)$.

In this expression, B is a set of arguments that we may agree on and C is the set justified by the agreement.

Example 3.2. Let $F = \langle \{0, 1, 2, 3, 4, 5, 6, 7, 8\}, \{0 \rightarrow 1, 1 \rightarrow 0, 1 \rightarrow 2, 2 \rightarrow 3, 3 \rightarrow 4, 4 \rightarrow 5, 5 \rightarrow 4, 0 \rightarrow 6, 4 \rightarrow 7, 6 \rightarrow 8, 7 \rightarrow 8\} \rangle$ as shown in Figure 2. Then, (E_1, E_2) for $E_1 = \{0\}$ and $E_2 = \{4\}$ is a progress of argumentation. In fact, from $\text{Comp}_F(E_1) = \{0, 2\}$, we get $\text{Undec}_F(\text{Comp}_F(E_1)) = \{4, 5, 7, 8\}$, $F_2 = \langle \{4, 5, 7, 8\}, \{4 \rightarrow 5, 5 \rightarrow 4, 4 \rightarrow 7, 4 \rightarrow 8\} \rangle$. In F_2 , $E_2 = \{4\}$ is admissible and $\text{Comp}_{F_2}(E_2) = \{4, 8\}$. Thus, we get $C = \text{Comp}(E_1) \cup \text{Comp}(E_2) = \{0, 2, 4, 8\}$. Therefore, $B = \{0, 4\} \in \text{Cond}_F(C)$.

4. Property of Incremental Acceptance Semantics

In this section, we show properties of $\text{Cond}_F(C)$.

Lemma 4.1. Let F be an AF, E_0 be an admissible extension and $F' = F \downarrow_{\text{Undec}_F(E_0)}$, $E' \subseteq A(F')$ be conflict-free. Then, the following holds.

1. $\text{Def}_F(E_0 \cup E') = E_0 \cup \text{Def}_{F'}(E')$.
2. $E - E_0$ is admissible in F' iff E is admissible in F and $E \cup E_0$ is conflict-free.

Proof. 1. First, for $a \in \text{Def}_F(E_0 \cup E') - E_0$, let us show $a \in \text{Def}_{F'}(E')$. $a \in \text{Undec}_F(E_0)$ holds. If $b \rightarrow_{F'} a$, $E_0 \cup E' \rightarrow_F b$ holds, and from $b \in \text{Undec}_F(E_0)$, $E' \rightarrow_{F'} b$ holds. Thus, $a \in \text{Def}_{F'}(E')$ holds. Conversely, for $a \in E_0 \cup \text{Def}_{F'}(E')$, we show $a \in \text{Def}_F(E_0 \cup E')$. We show the case that $a \in \text{Def}_{F'}(E')$. Suppose that $b \rightarrow_F a$. From $a \in \text{Undec}_F(E_0)$, $b \in \text{Atk}_F(E_0) \cup \text{Undec}_F(E_0)$ holds. If $b \in \text{Atk}_F(E_0)$, $E_0 \rightarrow_F b$ holds, otherwise if $b \in \text{Undec}_F(E_0)$, $b \rightarrow_{F'} a$ holds, so $E' \rightarrow_{F'} b$ holds, that is, $E' \rightarrow_F b$ holds. From the above, in either case $E_0 \cup E' \rightarrow_F b$ holds, and $a \in \text{Def}_F(E_0 \cup E')$ is proved.

2. follows from 1 directly. ($\text{Def}_F(E) = E_0 \cup \text{Def}_{F'}(E - E_0)$, $E_0 \cap \text{Def}_{F'}(E - E_0) = \emptyset$)

□

Lemma 4.2. Let F be an AF, E be admissible in F , $F' = F \downarrow_{\text{Undec}_F(E)}$, and E' be admissible in F' . Then,

1. $E \cup E'$ is admissible in F , and
2. $\text{Comp}_F(E \cup E') = E \cup \text{Comp}_{F'}(E')$.

Proof. 1. was proved in the proof of 2. in 4.1 and 2. follows from 1. of it immediately. \square

Lemma 4.3. Let C be complete in F and C' be complete in $F' = F \downarrow_{\text{Undec}_F(C)}$. Then, $C \cup C'$ is complete in F and $\text{Undec}_F(C \cup C') = \text{Undec}_{F'}(C')$.

Proof. From Lemma 4.2, $C \cup C'$ is admissible in F and from $\text{Def}_F(C \cup C') = C \cup \text{Def}_{F'}(C') = C \cup C'$, $C \cup C'$ is complete in F . Because the latter is equivalent to $C \cup C' \cup \text{Atk}_F(C \cup C') = C \cup \text{Atk}_F(C) \cup C' \cup \text{Atk}_{F'}(C')$ by taking the complement set of $A(F)$, it suffices to show that $\text{Atk}_F(C) \cup \text{Atk}_{F'}(C') = \text{Atk}_F(C \cup C')$. For $a \in \text{Atk}_F(C \cup C') - \text{Atk}_F(C)$, $C' \rightarrow_F a$. Because $C \cup C'$ is conflict-free, we have $a \notin C$. Then, $a \in \text{Undec}_F(C)$, $C' \rightarrow_{F'} a$, and $a \in \text{Atk}_{F'}(C')$. From these, we have $\text{Atk}_F(C \cup C') \subseteq \text{Atk}_F(C) \cup \text{Atk}_{F'}(C')$. Obviously, the reverse inclusion holds. \square

Next, we generalize $\text{Comp}_F(B)$ to $\Gamma_F(B)$ so that we can apply it to any conflict-free B .

Definition 4.4. Let $B \subseteq A(F)$ be conflict-free, and let $\Gamma_F(B)$ be the minimum set satisfying the condition that $\forall E \subseteq B \cup \Gamma_F(B), (E \text{ is admissible} \Rightarrow \text{Def}_F(E) \subseteq \Gamma_F(B))$.

$\Gamma = A(F)$ satisfies the expression and if $\Gamma_1, \dots, \Gamma_n$ satisfies the expression $\bigcap_{i=1}^n \Gamma_i$ satisfies the condition also. Hence the minimum exists and $\Gamma_F(B)$ is well-defined.

Lemma 4.5. For conflict-free B , $\Gamma_F(B) = \Gamma_F(B \cap \Gamma_F(B))$.

Proof. Let $E \subseteq B \cup \Gamma_F(B \cap \Gamma_F(B))$ be an admissible extension. For $E \subseteq B \cup \Gamma_F(B)$, $\text{Def}_F(E) \subseteq \Gamma_F(B)$. In particular, $E \subseteq \Gamma_F(B)$. Therefore, $E \subseteq (B \cap \Gamma_F(B)) \cup \Gamma_F(B \cap \Gamma_F(B))$ and $\text{Def}_F(E) \subseteq \Gamma_F(B \cap \Gamma_F(B))$. Thus, $\Gamma_F(B) = \Gamma_F(B \cap \Gamma_F(B))$. \square

By using Γ_F , we can have a simple characterization of Cond_F .

Theorem 4.6. Let $B \subseteq C \subseteq A(F)$ and let B be conflict-free. Then $B \in \text{Cond}_F(C)$ iff $\Gamma_F(B) = C$.

Proof. Let $B \in \text{Cond}_F(C)$. Then there exists progress of argumentation (E_1, \dots, E_n) such that $\bigcup_i E_i \subseteq B$. Let F_i be an AF in the definition of progress of argumentation and let $C_i = \text{Comp}_{F_i}(E_i)$. We show that $C_i \subseteq \Gamma_F(B)$ by induction. For $i = 1$, as $E_1 \subseteq B$, $C_1 = \text{Comp}_F(E_1) \subseteq \Gamma_F(B)$. Let $i > 1$, and suppose that the proposition holds for $1, \dots, i-1$. $C_1, \dots, C_{i-1}, E_i \subseteq \Gamma_F(B)$. From Lemma 4.3, $C_1 \cup \dots \cup C_{i-1}$ is complete in F . Therefore, from Lemma 4.2, $C_i = \text{Comp}_{F_i}(E_i) \subseteq \text{Comp}_F(C_1 \cup \dots \cup C_{i-1} \cup E_i)$. Thus, $C_i \subseteq \Gamma_F(B)$ and finally $C = \bigcup_i C_i \subseteq \Gamma_F(B)$. From Lemma 4.3, C is complete. Because $B \subseteq C$ and Γ_F is monotonous, $\Gamma_F(B) \subseteq \Gamma_F(C) = C$. Therefore, $C = \Gamma_F(B)$.

Conversely, Let $C = \Gamma_F(B)$. We can take admissible extensions E_1, \dots, E_n in F such that $E_i \subseteq B \cup \text{Comp}_F(E_1) \cup \dots \cup \text{Comp}_F(E_{i-1})$ for $1 \leq i \leq n$, and $\text{Comp}_F(E_1) \cup \dots \cup \text{Comp}_F(E_n) = C$. Let $E'_1 = E_1$, $E'_{i+1} = E_{i+1} - \text{Comp}_F(E_1 \cup \dots \cup E_n)$ for $i < n$, and $E'_{n+1} = C - \text{Comp}_F(E_1 \cup \dots \cup E_n)$, $C = \text{Comp}_F(E_1 \cup E_2 \cup \dots \cup E_n \cup C) = \text{Comp}_{F_1}(E'_1) \cup \dots \cup \text{Comp}_{F_n}(E'_n) \cup \text{Comp}_{F_{n+1}}(E'_{n+1})$. Thus, (E'_1, \dots, E'_n) is a progress of argumentation, and $C = \bigcup_i \text{Comp}_{F_i}(E'_i)$. Therefore, $B \in \text{Cond}_F(C)$. \square

Corollary 4.7. Let F be an AF, $B, C \subseteq A(F)$ and B be conflict-free.

1. $\text{Cond}_F(C) = \emptyset$ if C is not complete.
2. $\Gamma_F(B)$ is complete.

Proof. 1. was already shown in the proof of Theorem 4.6.

2. follows from Lemma 4.5 and Theorem 4.6. \square

Corollary 4.8. $\text{Cond}_F(C) = \{B \subseteq A(F) \mid B \text{ is conflict-free, } C = \Gamma_F(B \cap C)\}.$

Proof. For conflict-free B , $B \in \text{Cond}_F(C)$ iff $B \cap C \in \text{Cond}_F(C)$. The conclusion follows from this and Theorem 4.6. \square

Now let us find out what we have to agree on in order to justify arguments in general. For this purpose, we define $\text{Guar}_F(S)$ representing sets that “guarantee” S .

Definition 4.9. For conflict-free S , $\text{Guar}_F(S) = \bigcup_{C: \text{complete, } S \subseteq C} \text{Cond}_F(C).$

Theorem 4.10. For conflict-free S , $\text{Guar}_F(S) = \{B \mid B \text{ is conflict-free, } \Gamma_F(B) \supseteq S\}.$

Proof. If $B \in \text{Guar}_F(S)$, there exists complete $C \supseteq S$ such that $\Gamma_F(B \cap C) = C$ from Corollary 4.8. Thus, $\Gamma_F(B) \supseteq S$. Conversely, if $\Gamma_F(B) \supseteq S$, $B \in \text{Cond}_F(\Gamma_F(B))$. \square

Adding an argument x may make it more difficult to accept S or easier. We formally discuss this effect in the following for an AF G and an AF F , which is G with x and attacking relations involving x added. The difficulty is measured by the sets in Guar .

We say *an argument x defends an argument y indirectly* if $x = y$ or there exist a_1, \dots, a_{2n-1} (n is a positive integer) such that $x \rightarrow_F a_1 \rightarrow_F \dots \rightarrow_F a_{2n-1} \rightarrow_F y$.

Theorem 4.11. Let $x \in A(F)$ and $S \subseteq A(F)$, and let X be the set of arguments defended by x indirectly in F . For $G = F \downarrow_{A(F) - \{x\}}$ and $B \in \text{Guar}_F(S)$, $B - X \in \text{Guar}_G(S - X)$ holds.

Proof. Let $B \in \text{Guar}_F(S)$. Then, we can find admissible extensions E_1, \dots, E_n in F such that $E_i \subseteq B \cup \text{Def}_F(E_1) \cup \dots \cup \text{Def}_F(E_{i-1})$ and $\text{Def}_F(E_1) \cup \dots \cup \text{Def}_F(E_n) \supseteq S$ for $1 \leq i \leq n$. We show that for each E_i , $E_i - X$ is admissible in F . Let $a \rightarrow_F E_i - X$. Then there exists $b \in E_i$ such that $b \rightarrow_F a$. Assume that $b \in X$. Then there exists an argument in $E_i - X$ that is defended indirectly by x and contradicts. Therefore, $b \in E_i - X$ and $E_i - X$ is admissible in F . Because $x \notin E_i - X$, $E_i - X$ is admissible in G . Now we show $\text{Def}_F(E_i) - X \subseteq \text{Def}_G(E_i - X)$. Let $a \in \text{Def}_F(E_i) - X$. If $b \rightarrow_G a$, there exists $c \in E_i$ such that $c \rightarrow_F b$. If $c \in X$, then $a \in X$ holds and contradicts. Therefore, $c \in E_i - X$. Now we have $E_i - X \rightarrow_G b$ and $a \in \text{Def}_G(E_i - X)$. Therefore, for each i , $E_i - X \subseteq (B - X) \cup \text{Def}_G(E_1 - X) \cup \dots \cup \text{Def}_G(E_{i-1} - X)$ and $\text{Def}_F(E_1 - X) \cup \dots \cup \text{Def}_F(E_n - X) \supseteq S - X$. This means that $B - X \in \text{Guar}_G(S - X)$. \square

For arguments x and y , we say “ x attacks y indirectly” if there exists an argument b such that x defends b indirectly and $b \rightarrow y$.

Theorem 4.12. Let $x \in A(F)$ and $S \subseteq A(F)$, and let X be the set of arguments attacked by x indirectly in F . For $G = F \downarrow_{A(F) - \{x\}}$ and $B \in \text{Guar}_G(S)$, $B - X \in \text{Guar}_F(S - X)$.

Proof. Let $B \in \text{Guar}_G(S)$. We can take admissible extensions E_1, \dots, E_n in F such that for $1 \leq i \leq n$, $E_i \subseteq B \cup \text{Def}_G(E_1) \cup \dots \cup \text{Def}_G(E_{i-1})$ and $\text{Def}_G(E_1) \cup \dots \cup \text{Def}_G(E_n) \supseteq S$. Let X be the set of arguments attacked indirectly by x in G . We show that for each E_i , $E_i - X$ is admissible in F . For $a \rightarrow_F E_i - X$, there exists $b \in E_i$ such that $b \rightarrow_F a$. If $b \in X$ there exists an argument in $E_i - X$ that is attacked by x indirectly and contradicts. Therefore, $b \in E_i - X$ and $E_i - X$ is admissible in G . As x does not attack any argument in $E_i - X$, $E_i - X$ is admissible in F holds. We also show that $\text{Def}_G(E_i) - X \subseteq \text{Def}_F(E_i - X)$. For $a \in \text{Def}_G(E_i) - X$ and $b \rightarrow_F a$, as $b \neq x$, $b \rightarrow_G a$. Therefore, there exists $c \in E_i$ such that $c \rightarrow_G b$. If $c \in X$, then $a \in X$ and contradicts. Therefore, $c \in E_i - X$. Now, $E_i - X \rightarrow_F b$, and $a \in \text{Def}_F(E_i - X)$. Finally, for each i , $E_i - X \subseteq (B - X) \cup \text{Def}_F(E_1 - X) \cup \dots \cup \text{Def}_F(E_{i-1} - X)$ and $\text{Def}_G(E_1 - X) \cup \dots \cup \text{Def}_G(E_n - X) \supseteq S - X$. This completes the proof that $B - X \in \text{Guar}_F(S - X)$. \square

Theorems 4.11 and 4.12 are the generalization of the desirable properties stated in Section 1. For more limited situations, they can be understood as: if x does not defend any argument in S indirectly, adding argument x to an AF makes it more difficult to justify S ; and If x does not attack any argument in S indirectly, adding argument x to an AF makes it easier to justify S .

5. Minimal Cost Semantics

In this section, we introduce a new cost model of AFs. The cost model is based on the semiring-based AFs introduced in [2]. The semiring-based AFs use c-semiring defined below as the generalized domain of costs.

Definition 5.1. c-semiring is a semiring $\langle R, +_R, \times_R, \mathbf{0}_R, \mathbf{1}_R \rangle$ with $\Sigma: 2^R \rightarrow R$ defined as $\Sigma(\emptyset) = \mathbf{0}_R$, $\Sigma(\{a\}) = a$, $\Sigma(A \cup B) = \Sigma(A) + \Sigma(B)$ for $A, B \subseteq R$ satisfies that $\Sigma(R) = \mathbf{1}_R$, $\Sigma(\cup_{i \in I} A_i) = \Sigma(\{\Sigma(A_i) \mid i \in I\})$, and for $a \in R$ and $B \subseteq R$, $a \times \Sigma(B) = \Sigma(\{a \times b \mid b \in B\})$.

We define a partial order \preceq_R on R by $a + b = b \Leftrightarrow a \preceq_R b$. In the partial order, the minimum is $\mathbf{0}_R$ and the maximum is $\mathbf{1}_R$. We interpret $a \preceq_R b$ as that b is better than or equal to a . $a \preceq_R a'$ implies $a +_R b \preceq_R a' +_R b$ and $a \times_R b \preceq_R a' \times_R b$. $a \times_R b \preceq_R a$ always holds. This means that more constrained is less preferable. For more details, please refer to [1].

We show an example of c-semiring.

Example 5.2. Let R be positive real and ∞ . Having $+_R = \min$, $\times_R = +$, $\mathbf{0}_R = \infty$, and $\mathbf{1}_R = 0$, $\langle R, +_R, \times_R, \mathbf{0}_R, \mathbf{1}_R \rangle$ is a c-semiring called a *weighted semiring* in [1]. Note that $a \preceq b$ implies $b \leq a$ for a and b in the weighted semiring.

Definition 5.3. For AF F , c-semiring R , a function $W: A(F) \rightarrow R$ is called a *weighted function*. We extend W to $S \subseteq A(F)$ by $W(S) = \prod_{s \in S} W(s)$. $W(\emptyset) = \mathbf{1}_R$.

$B_1 \subseteq B_2$ implies $W(B_2) \preceq W(B_1)$ in general. In the following, we consider multiple semiring-based AFs with common c-semiring R and weighted function W . Here, for semiring-based AFs $\langle F, R, W_1 \rangle$ and $\langle G, R, W_2 \rangle$, we say that the weighted functions are common in $\langle F, R, W_1 \rangle$ and $\langle G, R, W_2 \rangle$, iff there exists W such that $W_1 = W|_{A(F)}$ and $W_2 = W|_{A(G)}$.

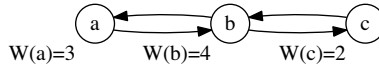


Figure 3. Example for Anomaly of Using $W(S)$ as Cost

Now, we assume that persons are rational. Therefore, if persons agree on a , then they also justify $\Gamma_F(\{a\})$. Taking this into consideration, we define the cost to justify arguments S as follows.

Definition 5.4. $\text{cost}_F(S) = \sum \{W(B) \mid B \in \text{Guar}_F(S)\}$.

The next theorem follows from Theorem 4.11.

Theorem 5.5. Let $x \in A(F)$ and $S \subseteq A(F)$, and let X be the set of arguments defended by x indirectly in F . For $G = F \downarrow_{A(F) - \{x\}}$, $\text{cost}_F(S) \preceq \text{cost}_G(S - X)$.

Similarly, the next theorem follows from Theorem 4.12.

Theorem 5.6. Let $x \in A(F)$ and $S \subseteq A(F)$, and let X be the set of arguments attacked by x indirectly in F . For $G = F \downarrow_{A(F) - \{x\}}$, $\text{cost}_G(S - X) \preceq \text{cost}_F(S)$.

For the special case that $S - X = S$, that is, $S \cap X = \emptyset$, the above theorems can be restated as that if an argument that does not attack S indirectly is added, the cost of S does not increase and if an argument that does not defend S indirectly is added, the cost of S does not decrease.

Example 5.7. We show an example of the above definitions. Let R be the weighted semiring defined in Example 5.2. Let us take a semiring-based AF using R as follows. Let $F = \langle \{a, b, c\}, \{a \rightarrow b, b \rightarrow a, b \rightarrow c, c \rightarrow b\} \rangle$, $W(a) = 3$, $W(b) = 4$, and $W(c) = 2$ as shown in Figure 3. Here, as $W(a)$, for example, we can use the number of persons who do not agree on argument a .

There are two nonempty complete extensions $S_1 = \{a, c\}$ and $S_2 = \{b\}$ in F . Let us consider which of the two is preferable.

In [2], W is used as the cost of S . Then, $W(S_1) = 3 + 2 = 5$, $W(S_2) = 4$, and $W(S_1) > W(S_2)$, resulting in the decision that S_2 is preferable. However, this has the following anomaly. Consider an AF $G = \langle \{a, b\}, \{a \rightarrow b, b \rightarrow a\} \rangle$, which is F with c removed. In G there exist two nonempty complete extensions $S_3 = \{a\}$ and $S_2 = \{b\}$ and in this case $W(S_3)$ is preferable because $W(S_3) = 3 < W(S_2) = 4$. However, it is counter-intuitive that adding an argument that attacks S_2 and does not attack S_3 makes the attacked S_2 preferable to the unattacked S_3 .

Here, we considered only nonempty complete extensions S_1 and S_2 . What happens if we consider other sets of arguments in F ? For example, if we consider S_3 that is admissible but not complete in F , $W(S_3) \neq W(S_1)$ holds even though $\text{Def}_F(S_3) = \text{Def}_F(S_1)$ resulting in another anomaly. This suggests that considering only complete extensions is not the cause of anomalies.

In order to avoid such anomaly, in this paper, we propose to use $cost_F(S)$ defined in Definition 5.4 instead of $W(S)$ as cost. Using $cost$, $cost_F(S_1)(=cost_F(S_3)) = 2$, $cost_F(S_2) = cost_G(S_2) = 4$, and $cost_G(S_3) = 3$. Therefore, S_1 is preferable to S_2 in F and S_3 is preferable to S_2 in G , and the anomaly is resolved. As we see in Theorem 5.5, $cost_F(S_3) \leq cost_G(S_3)$ and $cost_G(S_2) \leq cost_F(S_2)$.

c-semiring is also employed in [3]. In the paper, they gave weight to attacks as well as to arguments, and the aforementioned anomaly does not occur. However, the use of the weight on attacks is for relaxing the condition of conflict-free and it can not be used to decide which of S_1 and S_2 is preferable for S_1 and S_2 are already conflict-free.

6. Conclusion

In this paper, we introduced $Guar_F(S)$ as a basis of acceptance of arguments. This definition satisfies the desirable property that adding an argument *against* the arguments increases the cost while adding an argument *for* the arguments reduces the cost.

For accepting S , we need the minimal $B \in Guar_F(S)$. B is not admissible but each argument in B is in some simple cycle of even length. We are planning to explore more properties of B , construct an efficient algorithm to find the minimal B , and analyze the average-case performance of the algorithm under reasonable assumptions on AFs.

References

- [1] Stefano Bistarelli, Ugo Montanari, and Francesca Rossi. Semiring-based constraint satisfaction and optimization. *J. ACM*, 44(2):201–236, March 1997.
- [2] Stefano Bistarelli, Daniele Pirolandi, and Francesco Santini. Solving weighted argumentation frameworks with soft constraints. In Javier Larrosa and Barry O’ Sullivan, editors, *Recent Advances in Constraints*, volume 6384 of *Lecture Notes in Computer Science*, pages 1–18. Springer Berlin Heidelberg, 2011.
- [3] Stefano Bistarelli and Francesco Santini. A common computational framework for semiring-based argumentation systems. In *Proceedings of the 2010 Conference on ECAI 2010: 19th European Conference on Artificial Intelligence*, pages 131–136, Amsterdam, The Netherlands, The Netherlands, 2010. IOS Press.
- [4] Martin Caminada. On the issue of reinstatement in argumentation. In *Proceedings of the 10th European Conference on Logics in Artificial Intelligence*, JELIA’06, pages 111–123, Berlin, Heidelberg, 2006. Springer-Verlag.
- [5] Phan Minh Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77(2):321 – 357, 1995.

Spectral Techniques in Argumentation Framework Analysis

James BUTTERWORTH Paul E. DUNNE

Dept. of Computer Science, University of Liverpool, Liverpool, L69 7ZF, UK

Abstract. Spectral analysis – the study of the properties of the eigenvalues associated with some matrix derived from an underlying graph form – has proven to offer valuable insights in many domains where graph-theoretic models are prevalent. Abstract argumentation frameworks (AFs) are, of course, one such model and have provided a unifying basis for defining semantic properties related to concepts of “argument acceptability”. In this paper we consider the possible benefits of adopting spectral methods as a tool for analysing argumentation structures, presenting a preliminary empirical study of semantics in AFs and properties of the associated spectrum.

Keywords. abstract argumentation frameworks; directed graph spectrum; extension-based semantics

Introduction

A notable feature of formal analytic treatments of Dung’s seminal model of abstract argumentation from [16] is the focus on *discrete* methodologies. Typical of such directions has been the exploitation of graph-theoretic structures in defining semantics, e.g. Dung [16], Baroni *et al.* [4,2], and Caminada [12]. Developments seeking to alleviate issues with the highly abstracted form of Dung’s approach – such as Amgoud and Cayrol [1], Bench-Capon [7], Brewka and Woltran [9] – similarly embrace discrete mechanisms. While there are exceptions in which continuous measures are, in principle, permitted, e.g. within divers forms of so-called “weighted” frameworks, e.g. Dunne *et al.* [20], Barringer *et al.* [5], and, more directly, in models of *probabilistic* frameworks, such as Li *et al.* [29], it could be argued that the presence of continuous numerical quantities in such is more a consequence of the problems addressed than a direct analytic tool.

The aim of this article is to consider what scope for determining argumentation framework properties may be provided by considering the *spectrum* of the $(0,1)$ -matrix¹ defined through the directed graph describing the framework. We review the formal definition of “*graph spectrum*” subsequently, but for the purpose of this introduction it suffices to note that the spectrum of an $n \times n$ matrix

¹Although the $(0,1)$ structure is a natural choice it is often useful – especially within *directed* graph forms to make use of “transformed” $n \times n$ matrix definitions, one of the most widely used of these being the so-called Laplace operator, see e.g. Bauer [6].

is given by an n -tuple $\langle \lambda_1, \lambda_2, \dots, \lambda_n \rangle$ of (possibly) complex values² corresponding to the *eigenvalues* of the matrix. That is to say for $\chi(A, \lambda)$ the polynomial of degree n in λ , (thus $\chi(A, \lambda) = \sum_{i=0}^n c_i \lambda^i$) given³ by $|\lambda I - A|$ the n (not necessarily distinct) solutions of $\chi(A, \lambda) = 0$. If λ_A is an eigenvalue of the matrix A , then one may find n -tuples, \underline{x} , with at least one non-zero component, for which $A\underline{x} = \lambda_A \underline{x}$. Such n -tuples being referred to as *eigenvectors*.

At first inspection it may seem that there is little connection between the rather abstruse notion of eigenvalue (especially when these lie in the complex plane) and the, apparently, more practically grounded concepts offered within established ideas of abstract argumentation semantics. In order to motivate our subsequent study, it is worth reviewing contexts both within computational and other domains where their analysis is known to provide important benefits.

Undoubtedly one of the best known such applications is found in Web search-engines and the mechanisms used to rank pages, see Bryan and Leise [11] for further discussions. Other computational applications building on properties of eigenvalues within a supporting graph structure include pattern matching, e.g. Kirby and Shilovich [28], Shi and Malik [32], power control in communication networks, see e.g. Bertoni [8]. Similarly within empirical studies from the physical sciences concepts such as the *Estrada index* – an invariant defined via the eigenvalues arising from a graph introduced in Estrada [22] – have been argued to have important properties with respect to models of molecular structures, see e.g. Gutman and Graovac [26], Ilić and Stevanović [27]. Finally the use of graph spectra to inform reasoning about combinatorial structures within graphs is well established, e.g. Brouwer and Haemers [10].

The exploitation of what are often referred to as “*spectral techniques*”, in the light of their use in other domains, may provide some useful insight into aspects of argumentation frameworks. The principal aim of the current paper is to explore this potential. Our approach is empirical rather than analytic in nature. In particular, we consider evidence for links between divers argumentation structures, e.g. acceptability of arguments with respect to given semantics, existence of extensions containing some number of arguments, etc. and various spectral measures defined on the underlying framework, amongst which are invariants such as the Estrada index, the spectral spread – i.e. the difference between largest and smallest eigenvalue, etc.

Before proceeding with the technical presentation we elaborate on what the aims of our empirical investigations are and, of equal importance, what is *not* being asserted.

The central conceit motivating this paper may, informally, be expressed in the following question: do spectral techniques offer a possible basis for studying structural, especially semantic, properties within abstract argumentation frameworks? In support of a positive answer to this question, we have noted the numerous examples in other computational domains, particularly those wherein di-

²In special cases, in particular when the underlying matrix $[a_{ij}]$ is symmetric, its spectrum consists of values drawn from \mathbb{R} .

³For an $n \times n$ real-valued matrix, A we use $|A|$ to denotes its *determinant*, recalling that a matrix B for which $A \times B = B \times A = I$ exists if and only if $|A| \neq 0$.

rected graphs provide a natural modelling formalism, of spectral analysis providing insight.

Of course, the fact that a given formalism has proven helpful in one arena of study does not imply it will also prove useful within different but superficially similar fields. Nevertheless, it would seem reasonable prior to rejecting outright the notion that “spectral methods have a rôle within the analysis of argumentation frameworks”, to consider evidence in its support. In addition, we note recent studies of argumentation frameworks have explored operations on the matrix representation as an approach to capturing particular semantics in terms of matrix properties. Notable here is the recent work of Xu and Cayrol [33].

Thus, our principal aim is not to provide a *full* analytic or even empirical study of the relationships between spectra and argumentation but rather to consider connections between one specialized class of AFs and its spectra. For the class of AFs examined, its behaviour with respect to one argumentation semantics is well-characterized this characterization does not, however, assist computationally: that is to say, the canonical decision questions become no more tractable. In principle, however, given what is already known regarding the structural properties of this class, one might reasonably hope that this *could* in turn be tied with spectral properties. We develop this idea in fuller detail within Section 2 below.

We present background to Dung’s abstract AF model and review some elements regarding linear algebra and matrices in Section 1. In Section 2 we outline the basis and motivation underlying the structure of the experimental studies, and report on preliminary findings from these. Conclusions are presented in Section 3.

1. Preliminaries

We begin by recalling the concept of abstract argumentation frameworks and terminology from Dung [16]

Definition 1 We use \mathcal{X} to denote a finite set of arguments with $\mathcal{A} \subseteq \mathcal{X} \times \mathcal{X}$ the so-called attack relationship over these. An argumentation framework (AF) is a pair $\mathcal{H} = \langle \mathcal{X}, \mathcal{A} \rangle$. A pair $\langle x, y \rangle \in \mathcal{A}$ is referred to as ‘ y is attacked by x ’ or ‘ x attacks y ’. Using S to denote an arbitrary subset of arguments for $S \subseteq \mathcal{X}$,

$$\begin{aligned} S^- &=_{\text{def}} \{ p : \exists q \in S \text{ such that } \langle p, q \rangle \in \mathcal{A} \} \\ S^+ &=_{\text{def}} \{ p : \exists q \in S \text{ such that } \langle q, p \rangle \in \mathcal{A} \} \end{aligned}$$

We say that: $x \in \mathcal{X}$ is acceptable with respect to S if for every $y \in \mathcal{X}$ that attacks x there is some $z \in S$ that attacks y . Given $S \subseteq \mathcal{X}$, $\mathcal{F}(S) \subseteq \mathcal{X}$ is the set of all arguments that are acceptable with respect to S , i.e.

$$\mathcal{F}(S) = \{ x \in \mathcal{X} : \forall y \text{ such that } \langle y, x \rangle \in \mathcal{A}, \exists z \in S \text{ s.t. } \langle z, y \rangle \in \mathcal{A} \}$$

A subset, S , is conflict-free if no argument in S is attacked by any other argument in S . The \subseteq -maximal conflict-free sets are referred to as naive extensions. A conflict-free set S is admissible if every $y \in S$ is acceptable w.r.t S . S is a complete extension if S is conflict-free and should $x \in \mathcal{F}(S)$ then $x \in S$, i.e. every

argument that is acceptable to S is a member of S , so that $\mathcal{F}(S) = S$. The set of \subseteq -maximal complete extensions coincide with the set of \subseteq -maximal admissible sets these being termed preferred extensions. The set S is a stable extension if S is conflict free and $S^+ = \mathcal{X} \setminus S$. It is a semi-stable extension (Caminada [12]) if admissible and has $S \cup S^+ \subseteq$ -maximal among all admissible sets.

The grounded extension of $\langle \mathcal{X}, \mathcal{A} \rangle$ is defined as the \subseteq -minimal complete extension.

We use σ to denote an arbitrary semantics and for a given semantics σ and AF, $\mathcal{H}(\mathcal{X}, \mathcal{A})$, $\mathcal{E}_\sigma(\mathcal{H})$ denotes the set of all subsets of \mathcal{X} that satisfy the conditions specified by σ . We say that σ is a *unique status* semantics if $|\mathcal{E}_\sigma(\mathcal{H})| = 1$ for every AF, \mathcal{H} , denoting the unique extension by $E_\sigma(\mathcal{H})$.

We complete this, brief, overview by describing the three canonical decision problems that may be instantiated for a given semantics: *Verification* (VER), *Credulous Acceptance* (CA) and *Sceptical Acceptance* (SA). Formal definitions of these problems for AFs are presented in Table 1.

Table 1. Decision Problems in AFs

Problem Name	Instance	Question
<i>Verification</i> (VER_σ)	$\mathcal{H}(\mathcal{X}, \mathcal{A}); S \subseteq \mathcal{X}$	Is $S \in \mathcal{E}_\sigma(\mathcal{H})$?
<i>Credulous Acceptance</i> (CA_σ)	$\mathcal{H}(\mathcal{X}, \mathcal{A}); x \in \mathcal{X}$	$\exists S \in \mathcal{E}_\sigma(\mathcal{H})$ for which $x \in S$?
<i>Sceptical Acceptance</i> (SA_σ)	$\mathcal{H}(\mathcal{X}, \mathcal{A}); x \in \mathcal{X}$	$\forall T \in \mathcal{E}_\sigma(\mathcal{H})$ is $x \in T$?

Similarly we have two *function* problems – CONSTRUCT and COUNT –

Table 2. Function Problems in AFs

Problem Name	Instance	Computation
<i>Construction</i> (CONSTRUCT_σ)	$\mathcal{H}(\mathcal{X}, \mathcal{A})$	Return some $S \in \mathcal{E}_\sigma(\mathcal{H})$
<i>Count</i> (COUNT_σ)	$\mathcal{H}(\mathcal{X}, \mathcal{A})$	Return $ \mathcal{E}_\sigma(\mathcal{H}) $

Both of the function problems of Table 2 may be qualified so that instances specify a given argument $x \in \mathcal{X}$. In such cases, one is asked to construct a representative (resp. to count the number of subsets) in $\mathcal{E}_\sigma(\mathcal{H})$ containing the given argument x .

1.1. Review of Matrix Algebra

For an AF, $\mathcal{H} = \langle \mathcal{X}, \mathcal{A} \rangle$ with $|\mathcal{X}| = n$ we denote by $\mathbf{M}^\mathcal{H}$ the $n \times n$ $(0, 1)$ -matrix with entries m_{ij} defined via $m_{ij} = 1$ if and only if $\langle x_i, x_j \rangle \in \mathcal{A}$. With \mathbb{C} denoting the complex plane,⁴ $\lambda \in \mathbb{C}$ is said to be an *eigenvalue* of $\mathbf{M}^\mathcal{H}$ if there is some $n \times 1$ vector \underline{v} (with \underline{v} having at least one non-zero component) for which $\mathbf{M}^\mathcal{H} \underline{v} = \lambda \underline{v}$. A witnessing vector \underline{v} for λ is referred to as an *eigenvector* with respect to $\langle \mathbf{M}^\mathcal{H}, \lambda \rangle$.⁵

The tuple

⁴That is pairs $(a, b) \in \mathbb{R}$ defining the complex number $z = a + ib$, $i^2 = -1$.

⁵It is, on occasion, useful to distinguish so-called *right* eigenvectors w.r.t. $\langle \mathbf{M}^\mathcal{H}, \lambda \rangle$ from *left* eigenvectors w.r.t. $\langle \mathbf{M}^\mathcal{H}, \lambda \rangle$: the former being $n \times 1$ vectors, \underline{v} with $\mathbf{M}^\mathcal{H} \underline{v} = \lambda \underline{v}$, the latter $1 \times n$ vectors \underline{w} for which $\underline{w} \mathbf{M}^\mathcal{H} = \lambda \underline{w}$.

$$\sigma(\mathcal{H}) = \langle \lambda_1, \lambda_2, \dots, \lambda_n \rangle$$

formed by the n eigenvalues of $\mathbf{M}^{\mathcal{H}}$ is called the *spectrum* (of \mathcal{H}). The *spectral radius* of $\mathbf{M}^{\mathcal{H}}$, denoted $\rho(\mathbf{M}^{\mathcal{H}})$ is

$$\max \{ |\lambda| : \lambda \text{ is an eigenvalue of } \mathbf{M}^{\mathcal{H}} \}$$

where for $\lambda = a + ib \in \mathbb{C}$, $|\lambda| = +\sqrt{a^2 + b^2}$. We assume an ordering of the spectrum for \mathcal{H} such that whenever $i \leq j$ it holds that $|\lambda_i| - |\lambda_j| \geq 0$ so that the eigenvalues are considered in a non-decreasing order and $|\lambda_1| = \rho(\mathbf{M}^{\mathcal{H}})$. The largest eigenvalue (that is to say, λ_1) will be termed the *dominant* eigenvalue. This (and on occasion its successor λ_2) are the typical focus of spectral treatments.

The *Estrada Index* of \mathcal{H} ([22]), $\mathbf{E}(\mathcal{H})$ is given as

$$\mathbf{E}(\mathcal{H}) = \sum_{\lambda \in \sigma(\mathcal{H})} e^{\lambda}$$

We briefly recall some well known properties of eigenvalues in,

Fact 1

- a. For an $n \times n$ -matrix, \mathbf{A} , not necessarily $(0, 1)$, let $|\mathbf{A}|$ denote its determinant, and $\chi_{\mathbf{A}}(x)$ the polynomial of degree n in x defined through $|x\mathbf{I} - \mathbf{A}|$ (\mathbf{I} being the $(0, 1)$ identity matrix with (i, j) entries equal to 1 if and only if $i = j$). The quantity $\lambda \in \mathbb{C}$ is an eigenvalue of \mathbf{A} if and only if λ is a root of $\chi_{\mathbf{A}}(x)$, i.e. $\chi_{\mathbf{A}}(\lambda) = 0$.⁶
- b. For $(a, b) \in \mathbb{R}^2$, $\lambda = a + ib$ is an eigenvalue of \mathbf{A} if and only if $\bar{\lambda} = a - ib$ is an eigenvalue of \mathbf{A} .
- c. If \mathbf{A} is a symmetric matrix ($a_{ij} = a_{ji}$ for all $1 \leq i, j \leq n$) then all eigenvalues of \mathbf{A} lie in \mathbb{R} .

The concepts of eigenvalue and eigenvectors arise with respect to $n \times n$ real-valued matrices: of particular interest are the class of *non-negative* matrices and the subset of these defined by *positive* matrices.

Definition 2 Let $\mathbf{A} = [a_{ij}]$ be an $n \times n$ real-valued matrix. We say that \mathbf{A} is non-negative if for each i and j ($1 \leq i, j \leq n$) $a_{ij} \geq 0$. It is a *positive matrix* if every a_{ij} satisfies $a_{ij} > 0$.

It is obvious for the mapping described that $\mathbf{M}^{\mathcal{H}}$ is always a non-negative matrix, however, an apparent difficulty with this representation is that there is *exactly one* AF, \mathcal{H} , that gives rise to a positive matrix: namely the AF in which *every* attack between arguments is present (including self-attacks). There are, however, a large class of \mathcal{H} whose structural properties allow $\mathbf{M}^{\mathcal{H}}$ to be related to positive matrices with consequential benefits.

⁶Eigenvalues corresponding to unique roots of $\chi_{\mathbf{A}}(x)$ are referred to as *simple*, e.g. $\lambda = 1$ is a simple eigenvalue (root) of $(x - 1)(x + 1)$ but not of $(x - 1)(x - 1)$.

Definition 3 Let \mathbf{A} be a non-negative $n \times n$ matrix. If, for some $k \in \mathbb{N}$, \mathbf{A}^k is a positive matrix, then \mathbf{A} is said to be primitive.

If for each i, j ($1 \leq i, j \leq n$) there is some $k_{ij} \in \mathbb{N}$ for which $[\mathbf{A}^{k_{ij}}]_{ij} > 0$ then \mathbf{A} is said to be irreducible.

With regards to irreducible matrices we have the following classic theorem, which has been widely applied in many of the applications described in the introduction.

Theorem 1 (*Perron-Frobenius Theorem [30,24]*)

If \mathbf{A} is an irreducible $n \times n$ matrix then,

PF1. There is (at least one) positive real eigenvalue, λ^A , of \mathbf{A} with positive eigenvectors, that is for which there are associated eigenvectors \underline{x} all of whose components are strictly greater than 0.

PF2. There is a unique positive and dominant eigenvalue λ_{pf}^A , i.e. $\lambda_{pf}^A = \rho(\mathbf{A})$, and simple.

PF3. If $\mathbf{A}\underline{x} = \lambda\underline{x}$ and \underline{x} is positive then $\lambda = \lambda_{pf}^A$.

PF4. If $\mathbf{B} \geq \mathbf{A}$ and $\mathbf{B} \neq \mathbf{A}$ then $\rho(\mathbf{B}) > \lambda_{pf}^A$.⁷

PF5. If $\mathbf{B} \leq \mathbf{A}$ and $\mathbf{B} \neq \mathbf{A}$ then $\rho(\mathbf{B}) < \lambda_{pf}^A$.

The eigenvector associated with λ_{pf}^{www} where (informally) \mathbf{www} is the matrix corresponding to web-page connectivity, is central to many web search page-ranking algorithms, cf. the discussion in Bryan and Leise [11].

Thm. 1 applies to $\mathbf{M}^{\mathcal{H}}$ for a wide-ranging class of AFs, whose importance has earlier been demonstrated in Baroni *et al.* [4] and in connection with algorithmic study of the semantics considered in [2].

Fact 2 If $\mathcal{H} = \langle \mathcal{X}, \mathcal{A} \rangle$ is strongly-connected⁸ then $\mathbf{M}^{\mathcal{H}}$ is irreducible.

To conclude this overview we note that the property stated in Fact 1(b), allows us to show,

Fact 3 For all \mathcal{H} , $\mathbf{E}(\mathcal{H}) \in \mathbb{R}$.

2. Experiment Structure and Motivation

The experimental framework in essence uses *randomly generated* AFs constructed so that the AF's density (that is the ratio $|\mathcal{A}|/|\mathcal{X}|$) varies. For each randomly constructed AF within a given class a specific semantic property is assessed and comparative figures accumulated over all (generated test instances of relevant size) sharing the property and the average of specific spectral parameters.

For the basis of our empirical overview we focus on three measures: the dominant eigenvalue i.e. $\lambda_1 = \rho(\mathbf{M}^{\mathcal{H}})$; the second largest such eigenvalue (λ_2); and,

⁷For $n \times n$ real matrices \mathbf{A}, \mathbf{B} we say $\mathbf{B} \geq \mathbf{A}$ if and only if $b_{ij} \geq a_{ij}$ for $1 \leq i, j \leq n$, i.e. the comparison is component-wise.

⁸A directed graph, $\langle X, E \rangle$ is said to be strongly-connected if for all $\langle x_i, x_j \rangle \in X \times X$ there is a directed path of links from E starting in x_i and ending in x_j .

in order, to glean some indication of effects arising from the entire range of $\sigma(\mathcal{H})$, its Estrada index $\mathbf{E}(\mathcal{H})$.

The frameworks of interest are characterized by three parameters, $\langle n, m, k \rangle$ ($\mathcal{F}^{(n,m,k)}$) denoting those AFs with the structure referred to and having these parameters set to $\langle n, m, k \rangle$ so that the entire space of interest is

$$\mathbb{S} = \bigcup_{k=1}^{\infty} \bigcup_{n=1}^{\infty} \bigcup_{m=0}^{2^k \binom{n}{k}} \mathcal{F}^{(n,m,k)}$$

The class of AFs examined have a number of important properties which we first summarize in terms of their relationship to argumentation semantics.

- A1. There is a polynomial-time computable mapping τ that associates an AF, $\tau(\varphi) \in \mathcal{F}^{(n,m,k)}$ for φ an n variable, m clause, k -CNF formula. For such formulae, the framework, $\tau(\varphi)$ has exactly $2n + m + 2$ arguments and $4n + (k + 1)m + 1$ attacks.
- A2. The AF $\tau(\varphi)$ has a *non-empty preferred extension* (which is also a *stable extension*) if and only if its source CNF formula is *satisfiable*.

Regarding properties of $\mathbf{M}^{\tau(\varphi)}$ the important one of interest (in the light of Thm 1) is that the non-negative matrix $\mathbf{M}^{\tau(\varphi)}$ is *irreducible*: the AF $\tau(\varphi)$ being strongly-connected.

Before describing the structure of $\tau(\varphi)$ in greater depth, we emphasize that the random variable involved is not drawn from the space of *all* AFs *per se* but rather a *subset* of these, namely for $|\mathcal{X}| = 2n + m + 2$, $|\mathcal{A}| = 4n + (k + 1)m + 1$,

$$\mathbb{P}[\langle \mathcal{X}, \mathcal{A} \rangle \text{ is chosen}] = \begin{cases} 0 & \text{if } \langle \mathcal{X}, \mathcal{A} \rangle \notin \mathcal{F}^{(n,m,k)} \\ > 0 & \text{if } \langle \mathcal{X}, \mathcal{A} \rangle \in \mathcal{F}^{(n,m,k)} \end{cases}$$

with these likelihoods being essentially uniformly distributed over eligible AFs, i.e. those in $\mathcal{F}^{(n,m,k)}$.⁹

Now, although in general one cannot make inferences about the behaviour of one class of random combinatorial structures (for example, directed graphs) *as a whole* via mappings from a different class of random structures (e.g. k -CNF formulae), this, of course, is *not* what we claim to be the focus of our experiments. It is, rather the case that *should* there be any observable link between spectral aspects of the AFs considered and semantic properties then it may well be the case that such behaviour is evident when the source formulae exhibit specific characteristics.

In the case of random k -CNF formulae, such characteristics have been validated (from initial experimental studies) analytically. For further background we refer the reader to, among others, Chao and Franco [13], Freeman [23], Dunne *et al.* [19].

In particular we have,

⁹Describing the distribution as “uniform” is a slight over-simplification, however, the difference between “true” uniform and that pertinent to the experiments themselves is insignificant.

Fact 4 Let ψ be drawn uniformly at random from the space of n variable, m clause k -CNF formulae where $k \geq 2$. For each k , there are constants $\langle \theta_k^l, \theta_k^u \rangle \in \mathbb{R}^+$ (with $\theta_k^l \leq \theta_k^u$) such that

Letting $r = m/n$,

$$\begin{aligned} P[\psi \text{ is satisfiable}] &\rightarrow 1 \text{ if } r < \theta_k^l \\ P[\psi \text{ is satisfiable}] &\rightarrow 0 \text{ if } r > \theta_k^u \end{aligned}$$

The behaviour indicated becoming increasingly pronounced as the sample space induced by n increases in size. When $k = 2$, that $\theta_2^l = \theta_2^u = 1$ has been proven analytically by Goerdt [25].

The “threshold” behaviours observed in random k -CNF formulae together with the properties of the AF constructed by τ as described in (A2), suggest investigating the following as an initial stage regarding putative connections between spectra and semantics:

“Is the pattern whereby random k -CNF with few clauses (relative to n) are almost certainly satisfiable whilst those with many clauses are not (the transition from “few” to “many” being witnessed by a constant multiple (θ_k) of n), reflected in spectral properties of the AF defined through τ ?”

The cases reported below consider a range of randomly generated 3-CNF using clause-to-variable ratios ranging from almost certainly satisfiable ($r \leq 4$) to almost surely unsatisfiable ($r \geq 5$). Before proceeding to describe these in detail, we conclude this overview by recalling the transformation from k -CNF formulae, φ , to AFS $\tau(\varphi)$.

Definition 4 Given a k -CNF, φ over propositional variables $Z = \{z_1, \dots, z_n\}$ and clause set $\{C_1, C_2, \dots, C_m\}$ the standard translation of φ is the AF, $\mathcal{H}_\varphi = \langle \mathcal{X}_\varphi, \mathcal{A}_\varphi \rangle$

$$\begin{aligned} \mathcal{X}_\varphi &= \{\varphi\} \cup \{C_1, \dots, C_m\} \cup \{z_1, \dots, z_n\} \cup \{\neg z_1, \dots, \neg z_n\} \\ \mathcal{A}_\varphi &= \{\langle C_j, \varphi \rangle : 1 \leq j \leq m\} \cup \{\langle z_i, \neg z_i \rangle, \langle \neg z_i, z_i \rangle : 1 \leq i \leq n\} \\ &\quad \cup \{\langle y_i, C_j \rangle : y_i \text{ is a literal (i.e., } z_i \text{ or } \neg z_i) \text{ of the clause } C_j\} \end{aligned}$$

The AF, $\tau(\varphi)$ is formed from \mathcal{H}_φ by adding a new argument, ψ to \mathcal{X}_φ with \mathcal{A}_φ extended with attacks

$$\{\langle \varphi, \psi \rangle\} \cup \bigcup_{i=1}^n \{\langle \psi, z_i \rangle, \langle \psi, \neg z_i \rangle\}$$

The standard translation (and its variants such as τ) has formed an important device in the complexity analysis of decision problems in argumentation semantics since its introduction by Dimopoulos and Torres [15], e.g. Dunne and Bench-Capon [18], Dunne [17], Dvořák and Woltran [21], etc. For our purposes the important property of $\tau(\varphi)$, demonstrated in [15] is,

Fact 5 Let φ be any CNF formula. The following are equivalent properties respecting φ :

- a. The formula φ is satisfiable.
- b. The argument φ in both \mathcal{H}_φ and $\tau(\varphi)$ is credulously accepted w.r.t. admissible semantics.
- c. The AF $\tau(\varphi)$ has a non-empty preferred extension.
- d. The AF $\tau(\varphi)$ has a stable extension.

A series of trials involving the following steps were carried out:

- S1. Set n the number of propositional variables.
- S2. Set m the number of clauses.
- S3. Generate a random m -clause, 3-CNF formula, φ .
- S4. Form the AF, $\tau(\varphi)$.
- S5. Determine, for the (irreducible) matrix $\mathbf{M}^{\tau(\varphi)}$,
 - L1 The dominant eigenvalue, $\lambda_1 = \rho(\mathbf{M}^{\tau(\varphi)})$.
 - L2 The second largest eigenvalue, λ_2 .
 - EE The Estrada index, $\mathbf{E}(\tau(\varphi))$

For reasons of space we focus on the experimental outcomes arising from the behaviour of the dominant eigenvalue.

Fig. 1 shows (x -axis) varying clause-to-variable ratio from $r = 3$ (predominantly satisfiable cases) to $r = 8$ (unsatisfiable) and n ranging from 6 to 16. The 24 specific cases result in $\mathbf{M}^{\tau(\varphi)}$ of dimensions 62×62 , 72×72 and 82×82 corresponding to the three curves indicated.

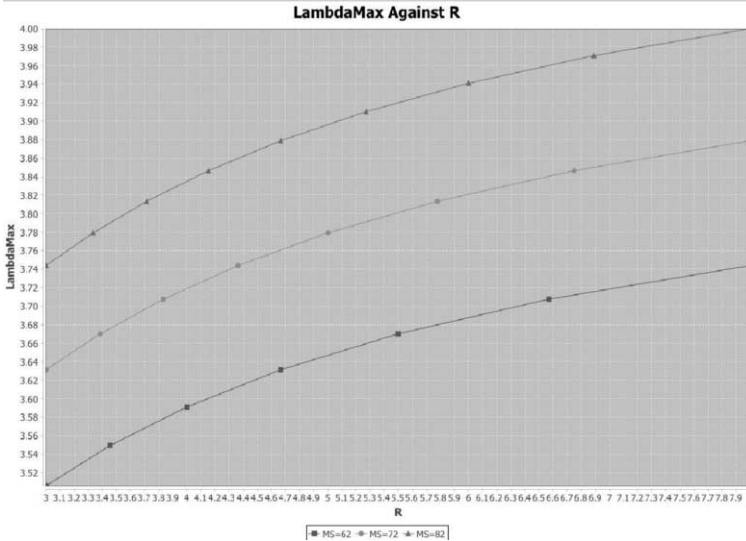


Figure 1. Clause Variable Ratio R vs. Dominant Eigenvalue

In Fig. 2, these ratios are compared against the Estrada Index of the corresponding AF.

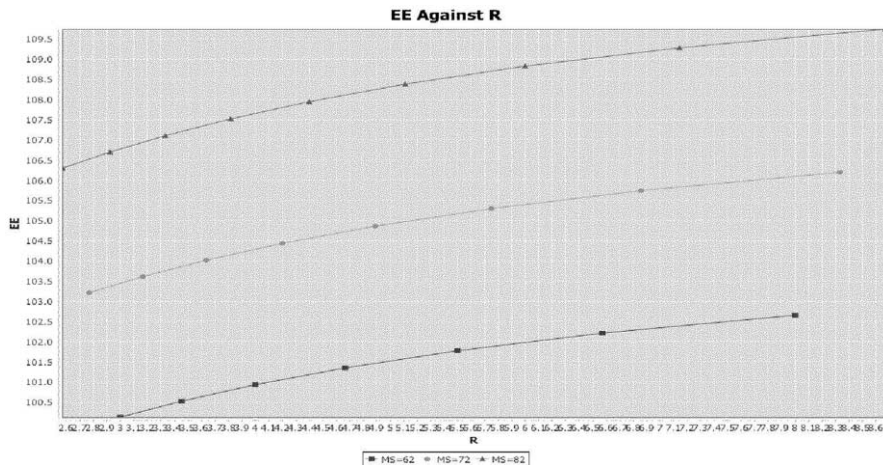


Figure 2. Clause Variable Ratio R vs. Estrada Index

There is some indication that the dominant eigenvalue is dependent on r . The close similarity between the outcomes shown for dominant eigenvalue (Fig. 1) and Estrada index¹⁰ (Fig. 2) could be accounted for by the presence of a large number of very small values in the relevant spectra, so that a significant contribution to $\mathbf{E}(\tau(\varphi))$ is from $(e^{\lambda_1} + e^{\lambda_2})$. We note, however, terms defining $\mathbf{E}(\tau(\varphi))$ that arise from smaller eigenvalues are also important so that estimating $\mathbf{E}(\tau(\varphi))$ as $(e^{\lambda_1} + e^{\lambda_2})$ fails to be accurate.

Of course, these outcomes are very far from being able to argue that $\rho(\mathbf{M}^{\mathcal{H}}) > \alpha$ allows some semantic properties of \mathcal{H} to be deduced. The behaviour, however, does suggest (on the basis of established properties of random CNF formulae) a possible continuation, namely: rather than mapping random 3-CNF to AFS via the standard translation, construct AFS with varying dominant eigenvalues (a non-trivial task) and consider semantic properties of the given AF. This direction is the subject of current work.

3. Conclusions

The use of spectral techniques, while widespread in many fields exploiting graph models, has had comparatively little attention with respect to potential use in studying argumentation frameworks. The primary thesis of this article is that a deeper analysis of the relationship between AF spectra and argumentation properties, such as extension-based semantics, offers possible insights into (among others), algorithm synthesis. In this regard, spectral techniques provide directions well-suited to the consideration of *weighted* frameworks.

We conclude by outlining two (out of many) directions for further research.

¹⁰Although not shown here, in fact R vs. λ_2 exhibits very similar behaviour.

- D1. Cyclic structures in AFS. Several researchers, e.g. Baroni and Giacomin [3], Coste-Marquis *et al.* [14], have observed that directed cycles among arguments (and the parity of such cycles) has a significant influence on argument acceptability and algorithmic behaviour. A well-known relationship between the spectrum of a directed graph, D , and the number of “cyclic paths of length k in D ” is that the latter is $\sum_{i=1}^n \lambda_i^k$. (Note that this counts *non-simple* cycles). Thus, the spectrum of \mathcal{H} provides information about cycles in \mathcal{H} . Notice that, as a consequence, returning to the expression of eigenvalues as roots of a polynomial, it follows that the governing polynomial for acyclic AFS is simply x^n , i.e all eigenvalues are 0.
- D2. Argument ranking. A growing area of interest within argumentation has been capturing concepts of argument “strength” and defining “rankings” of arguments, e.g. Pu *et al.* [31], Zhao *et al.* [34]. Many of the problems with “naïve” approaches (e.g. quantifying weakness by the *number* of attackers ignoring the nature of the attack itself) have parallels with naïve approaches to web page-ranking (e.g. using the number of links to a page to determine its importance). Pursuing this analogy suggests that applying consequences of Thm. 1 (the Perron-Frobenius Theorem) – the mechanism underpinning Google’s page ranking – offers one technique for exploring argument strength.

In total these and other possibilities suggest that spectral techniques offer, as these have been found to provide in other graph based arenas, a rich potential for effective exploitation applied to abstract argumentation frameworks.

References

- [1] L. Amgoud and C. Cayrol. A reasoning model based on the production of acceptable arguments. *Annals of Math. and AI*, 34:197–215, 2002.
- [2] P. Baroni, P. E. Dunne, and M. Giacomin. On the resolution-based family of abstract argumentation semantics and its grounded instance. *Artificial Intelligence*, 175:791–813, 2011.
- [3] P. Baroni and M. Giacomin. Solving semantic problems with odd-length cycles in argumentation. In *Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, pages 440–451. Springer, 2003.
- [4] P. Baroni, M. Giacomin, and G. Guida. SCC-recursiveness: a general schema for argumentation semantics. *Artificial Intelligence*, 168(1–2):162–210, 2005.
- [5] H. Barringer, D. M. Gabbay, and J. Woods. Temporal dynamics of support and attack networks: From argumentation to zoology. In *Mechanizing Mathematical Reasoning (LNCS Volume 2605)*, pages 59–98, 2005.
- [6] F. Bauer. Normalized graph Laplacians for directed graphs. *Linear Algebra and its Applications*, 436(11):4193–4222, 2012.
- [7] T. J. M. Bench-Capon. Persuasion in Practical Argument Using Value-based Argumentation Frameworks. *Journal of Logic and Computation*, 13(3):429–448, 2003.
- [8] H. L. Bertoni. *Radio propagation for modern wireless systems*. Prentice-Hall, 2000.
- [9] G. Brewka and S. Woltran. Abstract dialectical frameworks. In *Proc. 12th KR*, pages 102–111. AAAI Press, 2010.
- [10] A. E. Brouwer and W. H. Haemers. *Spectra of graphs*. Springer Science & Business Media, 2011.
- [11] K. Bryan and T. Leise. The \$25,000,000,000 eigenvector: The linear algebra behind google. *Siam Review*, 48(3):569–581, 2006.

- [12] M. Caminada. Semi-stable semantics. In *Proc. 1st COMMA*, volume 144 of *FAIA*, pages 121–130. IOS Press, 2006.
- [13] M. T. Chao and J. Franco. Probabilistic analysis of a generalization of the unit-clause literal selection heuristics for the k -satisfiable problem. *Inform. Sci.*, 51:289–314, 1990.
- [14] S. Coste-Marquis, C. Devred, and P. Marquis. Prudent semantics for argumentation frameworks. In *Proc. 17th ICTAI*, pages 5–9, 2005.
- [15] Y. Dimopoulos and A. Torres. Graph theoretical structures in logic programs and default theories. *Th. Comp. Sci.*, 170:209–244, 1996.
- [16] P. M. Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming, and N -person games. *Artificial Intelligence*, 77:321–357, 1995.
- [17] P. E. Dunne. The Computational Complexity of Ideal Semantics. *Artificial Intelligence*, 173(18):1559–1591, 2009.
- [18] P. E. Dunne and T. J. M. Bench-Capon. Coherence in finite argument systems. *Artificial Intelligence*, 141:187–203, 2002.
- [19] P. E. Dunne, A. Gibbons, and M. Zito. Complexity-theoretic models of phase transitions in search problems. *Theoretical Computer Science*, 249:243–263, 2000.
- [20] P. E. Dunne, A. Hunter, P. McBurney, S. Parsons, and M. Wooldridge. Weighted argument systems: Basic definitions, algorithms, and complexity results. *Artificial Intelligence*, 175(2):457–486, 2011.
- [21] W. Dvořák and S. Woltran. Complexity of semi-stable and stage semantics in argumentation frameworks. *Inf. Process. Lett.*, 110:425–430, 2010.
- [22] E. Estrada. Characterization of 3d molecular structure. *Chemical Physics Letters*, 319(5):713–718, 2000.
- [23] J. W. Freeman. Hard random 3-SAT problems and the Davis–Putnam procedure. *Artificial Intelligence*, 81:183–198, 1996.
- [24] G. Frobenius. Über matrizen aus nicht negativen elementen. *Sitz. Königl. Preuss. Akad. Wiss.*, pages 456–477, 1912.
- [25] A. Goerd. A threshold for unsatisfiability. *Journal of Computer and System Sciences*, 53(3):469–486, 1996.
- [26] I. Gutman and A. Graovac. Estrada index of cycles and paths. *Chemical physics letters*, 436(1):294–296, 2007.
- [27] A. Ilić and D. Stevanović. The Estrada index of chemical trees. *Journal of mathematical chemistry*, 47(1):305–314, 2010.
- [28] M. Kirby and L. Sirovich. Application of the Karhunen-Loeve procedure for the characterization of human faces. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 12(1):103–108, 1990.
- [29] H. Li, N. Oren, and T. J. Norman. Probabilistic argumentation frameworks. In *Proc. 1st TAFE*, pages 1–16, 2011.
- [30] O. Perron. Zur theorie der matrizen. *Mathematische Annalen*, 64(2):248–263, 1907.
- [31] F. Pu, J. Luo, Y. Zhang, and G. Luo. Argument ranking with categoriser function. In *Knowledge Science, Engineering and Management*, pages 290–301. 2014.
- [32] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.
- [33] Y. Xu and C. Cayrol. The matrix approach for abstract argumentation frameworks. In *Theory and Applications of Formal Argumentation*, pages 243–259. Springer, 2015.
- [34] X. Zhao, A. Strasser, J. N. Cappella, C. Lerman, and M. Fishbein. A measure of perceived argument strength: Reliability and validity. *Communication Methods and Measures*, 5(1):48–75, 2011.

A Dialectical Approach for Argument-Based Judgment Aggregation

Martin CAMINADA and Richard BOOTH

Cardiff University, UK

Abstract. The current paper provides a dialectical interpretation of the argumentation-based judgment aggregation operators of Caminada and Pigozzi. In particular, we define discussion-based proof procedures for the foundational concepts of *down-admissible* and *up-complete*. We then show how these proof procedures can be used as the basis of dialectical proof procedures for the *sceptical*, *credulous* and *super credulous* judgment aggregation operators.

Keywords. judgment aggregation, proof procedures, discussion games

1. Introduction

Given an argumentation framework, there can be more than one reasonable position on which arguments to accept and which arguments to reject [4], and different agents can take different positions. How to aggregate the agents' individual positions to form a group position has been studied by Caminada and Pigozzi [11]. For this, three different operators have been formulated: the *sceptical operator*, the *credulous operator* and the *super credulous operator*. These operators are such that, when each individual position is an admissible labelling, the collective outcome will also be an admissible labelling.¹

Various follow-up research has been done based on the work of Caminada and Pigozzi. Podlaskowski [19], Caminada et al. [12] and Awad et al. [1,3] have examined issues of Pareto optimality and strategy proofness of the three judgment aggregation operators. Awad et al. [1,2] have examined the empirical acceptance of their outcomes, and Booth et al. [6] have recently provided a generalised theory and have shown how the operators of Caminada and Pigozzi fit in.

In the current paper, we examine how the three judgment aggregation operators of Caminada and Pigozzi can be given a dialectical interpretation. This is in line with recent work on argumentation-based discussion games [17,13,14,8,9,15]. However, instead of applying discussion games as proof procedures for argumentation semantics, we apply discussion games as proof procedures for the judgment aggregation operators. That is, we introduce argument games for the sceptical,

¹In some cases, stronger results apply. For instance, when each agent's position is a complete labelling, applying the sceptical operator will yield a complete labelling. We refer to [11] for details.

credulous and super credulous operator, such that the ability to win the game coincides with the argument being accepted by the respective judgment aggregation operator. This is done by defining discussion games for two fundamental concepts used by the judgment aggregation operators: the *down-admissible* and *up-complete* labellings.

The remaining part of this paper is structured as follows. First, in Section 2 we briefly revisit Caminada and Pigozzi's work on argumentation based judgment aggregation. Then, in Section 3 we introduce the down-admissible game, as well as the discussion games for the sceptical and credulous operators based on it. In Section 4 we then introduce the up-complete game, as well as the discussion game for the super credulous operator based on it. We then round off with a discussion of the obtained results in Section 5.

2. Formal Preliminaries

For current purposes, we apply the labelling-based version of argumentation semantics [7,10,4]. In line with [11], we restrict ourselves to finite argumentation frameworks.

Definition 1. An argumentation framework is a pair (Ar, \rightarrow) where Ar is a finite set of arguments² and $\rightarrow \subseteq Ar \times Ar$.

Definition 2. Let (Ar, \rightarrow) be an argumentation framework. A labelling is a total function $\mathcal{Lab} : Ar \rightarrow \{\text{in}, \text{out}, \text{undec}\}$. We write $\text{in}(\mathcal{Lab})$ for $\{A \mid \mathcal{Lab}(A) = \text{in}\}$, $\text{out}(\mathcal{Lab})$ for $\{A \mid \mathcal{Lab}(A) = \text{out}\}$ and $\text{undec}(\mathcal{Lab})$ for $\{A \mid \mathcal{Lab}(A) = \text{undec}\}$. We define a relation \sqsubseteq between labellings s.t. $\mathcal{Lab}_1 \sqsubseteq \mathcal{Lab}_2$ iff $\text{in}(\mathcal{Lab}_1) \subseteq \text{in}(\mathcal{Lab}_2)$ and $\text{out}(\mathcal{Lab}_1) \subseteq \text{out}(\mathcal{Lab}_2)$. We define a function Γ such that $\Gamma(\mathcal{Lab})$ is a labelling with $\text{in}(\Gamma(\mathcal{Lab})) = \{A \mid \mathcal{Lab}(B) = \text{out} \text{ for each } B \rightarrow A\}$ and $\text{out}(\Gamma(\mathcal{Lab})) = \{A \mid \mathcal{Lab}(B) = \text{in} \text{ for some } B \rightarrow A\}$. A labelling \mathcal{Lab} is called admissible iff $\mathcal{Lab} \sqsubseteq \Gamma(\mathcal{Lab})$. A labelling is called complete iff $\mathcal{Lab} = \Gamma(\mathcal{Lab})$.

We will sometimes write a labelling as a triple $(\text{in}(\mathcal{Lab}), \text{out}(\mathcal{Lab}), \text{undec}(\mathcal{Lab}))$. We proceed to define the concepts of down-admissible and up-complete.

Definition 3 ([11]). Let \mathcal{Lab} be a labelling of argumentation framework (Ar, \rightarrow) . The down-admissible labelling of \mathcal{Lab} (written as $\downarrow \mathcal{Lab}$) is the unique biggest (w.r.t. \sqsubseteq) admissible labelling of (Ar, \rightarrow) that is smaller or equal (w.r.t. \sqsubseteq) to \mathcal{Lab} .

Definition 4 ([11]). Let \mathcal{Lab} be an admissible labelling of argumentation framework (Ar, \rightarrow) . The up-complete labelling of \mathcal{Lab} (written as $\uparrow \mathcal{Lab}$) is the unique smallest (w.r.t. \sqsubseteq) complete labelling that is bigger or equal (w.r.t. \sqsubseteq) to \mathcal{Lab} .

²For current purposes, we keep the internal structure of the arguments abstract, although we emphasize that our theory is compatible with instantiated argumentation theories like ASPIC+ [18], ABA [20] and logic-based argumentation [16].

Definition 5 ([11]). Given labellings $\mathcal{L}ab_1, \dots, \mathcal{L}ab_n$ ($n \geq 1$) of argumentation framework (Ar, \rightarrow) , we define $\sqcap(\mathcal{L}ab_1, \dots, \mathcal{L}ab_n)$ as the labelling $(\{A \mid \forall_{i \in \{1 \dots n\}} \mathcal{L}ab_i(A) = \text{in}\}, \{A \mid \forall_{i \in \{1 \dots n\}} \mathcal{L}ab_i(A) = \text{out}\}, \{A \mid \exists_{i \in \{1 \dots n\}} \mathcal{L}ab_i(A) \neq \text{in} \wedge \exists_{i \in \{1 \dots n\}} \mathcal{L}ab_i(A) \neq \text{out}\})$ and $\sqcup(\mathcal{L}ab_1, \dots, \mathcal{L}ab_n)$ as the labelling $(\{A \mid \exists_{i \in \{1 \dots n\}} \mathcal{L}ab_i(A) = \text{in} \wedge \neg \exists_{i \in \{1 \dots n\}} \mathcal{L}ab_i(A) = \text{out}\}, \{A \mid \exists_{i \in \{1 \dots n\}} \mathcal{L}ab_i(A) = \text{out} \wedge \neg \exists_{i \in \{1 \dots n\}} \mathcal{L}ab_i(A) = \text{in}\}, \{A \mid \forall_{i \in \{1 \dots n\}} \mathcal{L}ab_i(A) = \text{undec} \vee (\exists_{i \in \{1 \dots n\}} \mathcal{L}ab_i(A) = \text{in} \wedge \exists_{i \in \{1 \dots n\}} \mathcal{L}ab_i(A) = \text{out})\})$.³

Given the above defined concepts, we proceed to formally state the three judgment aggregation operators of [11].

Definition 6 ([11]). Let $\mathcal{L}ab_1, \dots, \mathcal{L}ab_n$ ($n \geq 1$) be admissible labellings of argumentation framework (Ar, \rightarrow) . We define:

- the sceptical outcome as $\downarrow \sqcap(\mathcal{L}ab_1, \dots, \mathcal{L}ab_n)$
- the credulous outcome as $\downarrow \sqcup(\mathcal{L}ab_1, \dots, \mathcal{L}ab_n)$
- the super credulous outcome as $\Downarrow \sqcup(\mathcal{L}ab_1, \dots, \mathcal{L}ab_n)$

We sometimes refer to $\downarrow \sqcap$ as the sceptical operator, $\downarrow \sqcup$ as the credulous operator and $\Downarrow \sqcup$ as the super credulous operator. We refer to [11,12,19,1,3] for the formal properties of these operators.

3. Dialectical Proof Procedures for the Sceptical and Credulous Operators

Using the formal preliminaries stated above, we now turn our attention to specifying dialectical proof procedures for the three judgment aggregation operators. We start with the sceptical and credulous operators. As these are both based on the down-admissible labelling, we first define a discussion game for the down-admissible, based on the admissible game (for preferred semantics) of [15,9].

Definition 7. Let $\mathcal{L}ab$ be a labelling of argumentation framework (Ar, \rightarrow) . A down-admissible discussion for $A \in Ar$ in $\mathcal{L}ab$ is a sequence of moves $[M_1, \dots, M_m]$ ($m \geq 1$) such that:

- $M_1 = \text{in}(A)$
- each move M_j ($1 \leq j \leq m$) where j is odd (called a proponent move) is of the form $\text{in}(B)$ with $B \in Ar$
- each move M_j ($1 \leq j \leq m$) where j is even (called an opponent move) is of the form $\text{out}(B)$ with $B \in Ar$
- for each opponent move $M_j = \text{out}(B)$ ($2 \leq j \leq m$) there exists a proponent move $M_k = \text{in}(C)$ ($k < j$) such that $B \rightarrow C$
- for each proponent move $M_j = \text{in}(B)$ except the first one ($3 \leq j \leq m$) it holds that M_{j-1} is of the form $\text{out}(C)$ such that $B \rightarrow C$
- there exist no two opponent moves M_j and M_k ($j \neq k$) such that $M_j = M_k$

A down-admissible discussion $[M_1, \dots, M_m]$ is called terminated iff

³ $\sqcap(\mathcal{L}ab_1, \dots, \mathcal{L}ab_n)$ is called the *sceptical initial labelling* in [11] and $\sqcup(\mathcal{L}ab_1, \dots, \mathcal{L}ab_n)$ is called the *credulous initial labelling* in [11].

1. there exists no M_{m+1} such that $[M_1, \dots, M_m, M_{m+1}]$ is a down-admissible discussion, or
2. there exists a proponent move $\text{in}(B)$ and an opponent move $\text{out}(B)$ for the same argument B , or
3. there exists a proponent move $\text{in}(B)$ s.t. $\mathcal{L}ab(B) \neq \text{in}$, or
4. there exists an opponent move $\text{out}(B)$ s.t. $\mathcal{L}ab(B) \neq \text{out}$

and no subsequence $[M_1, \dots, M_l]$ ($l \leq m$) is terminated. A terminated down-admissible discussion is won by the opponent if

1. there exists no M_{m+1} such that $[M_1, \dots, M_m, M_{m+1}]$ is a down-admissible discussion and M_m is an opponent move, or
2. there exists a proponent move $\text{in}(B)$ and an opponent move $\text{out}(B)$ for the same argument B , or
3. there exists a proponent move $\text{in}(B)$ s.t. $\mathcal{L}ab(B) \neq \text{in}$, or
4. there exists an opponent move $\text{out}(B)$ s.t. $\mathcal{L}ab(B) \neq \text{out}$

Otherwise, the terminated down-admissible discussion is won by the proponent.

We observe that the above discussion game is essentially the admissibility game of [15,9], with additional clauses 3 and 4 in both the termination criterion and the winning criterion. These additional clauses essentially state that for the proponent to win, the game has to stay “inside” the initial labelling $\mathcal{L}ab$.

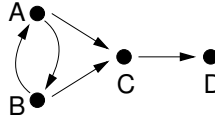


Figure 1. An argumentation framework

As an example of how the down-admissible discussion game works, consider the argumentation framework of Figure 1 and labelling $\mathcal{L}ab = (\{D\}, \{C\}, \{A, B\})$. Here, the discussion $[\text{in}(D), \text{out}(C), \text{in}(A)]$ is terminated and won by the opponent (since $\mathcal{L}ab(A) \neq \text{in}$), as is the discussion $[\text{in}(D), \text{out}(C), \text{in}(B)]$ (since $\mathcal{L}ab(B) \neq \text{in}$). We observe that $D \notin \text{in}(\downarrow \mathcal{L}ab)$ as $\downarrow \mathcal{L}ab$ is the all-undec labelling $(\emptyset, \emptyset, \{A, B, C, D\})$.

We are now ready to formally state soundness and completeness of the down-admissible discussion game.

Theorem 1. *Let $\mathcal{L}ab$ be a labelling of argumentation framework (Ar, \rightarrow) and let $A \in Ar$. A is labelled in by $\downarrow \mathcal{L}ab$ iff the proponent has a winning strategy⁴ in the down-admissible discussion game for A in $\mathcal{L}ab$.*

Proof. “ \Leftarrow ”: Suppose the proponent has a winning strategy for A in $\mathcal{L}ab$. Then, from the definition of a winning strategy, it follows that there exists at least one discussion game for A in $\mathcal{L}ab$ that is won by the proponent. Now consider the labelling $\mathcal{L}ab'$ with $\text{in}(\mathcal{L}ab')$ consisting of all proponent moves and $\text{out}(\mathcal{L}ab')$

⁴We use the term winning strategy in the sense of [9].

consisting of all opponent moves of this discussion.⁵ From the fact that the discussion is won by the proponent (together with winning condition 2) it follows that there is no argument B with both $B \in \text{in}(\mathcal{L}ab')$ and $B \in \text{out}(\mathcal{L}ab')$. This means that $\mathcal{L}ab'$ is a well-defined argument labelling. From the fact that the discussion is won by the proponent, it also follows (termination condition 1 and winning condition 1) that the last move (M_m) is a proponent move. This means that each opponent move in the discussion has been replied to. That is, for each opponent move $\text{out}(B)$ there exists a proponent move $\text{in}(B)$. Hence we obtain that (i) for each $B \in \text{out}(\mathcal{L}ab')$ there exists a $C \in \text{in}(\mathcal{L}ab')$ such that C attacks B . From the fact that the discussion is terminated with the last move (M_m) being a proponent move, it also follows that the opponent cannot make a move M_{m+1} anymore. This means that there is no attacker to any of the proponent's moves that hasn't already been moved. This implies that (ii) for each $B \in \text{in}(\mathcal{L}ab')$ it holds that each attacker C of B has $C \in \text{out}(\mathcal{L}ab')$. From conditions (i) and (ii) it follows that $\mathcal{L}ab'$ is an admissible labelling. From winning conditions 3 and 4 it follows that $\text{in}(\mathcal{L}ab') \subseteq \text{in}(\mathcal{L}ab)$ and $\text{out}(\mathcal{L}ab') \subseteq \text{out}(\mathcal{L}ab)$. That is, $\mathcal{L}ab'$ is an admissible labelling with $\mathcal{L}ab' \sqsubseteq \mathcal{L}ab$. As the down-admissible labelling $\downarrow \mathcal{L}ab$ is the unique biggest (w.r.t. \sqsubseteq) admissible labelling with $\downarrow \mathcal{L}ab \sqsubseteq \mathcal{L}ab$ it follows that $\mathcal{L}ab' \sqsubseteq \downarrow \mathcal{L}ab$. From the fact that $A \in \text{in}(\mathcal{L}ab')$ it directly follows that $A \in \text{in}(\downarrow \mathcal{L}ab)$.

“ \Rightarrow ”: Let A be labelled in by $\downarrow \mathcal{L}ab$. Now consider a discussion that starts with the proponent moving $\text{in}(A)$. As long as each proponent move is labelled in by $\downarrow \mathcal{L}ab$ (as is the case with the first move) the opponent can only move arguments that are labelled out by $\downarrow \mathcal{L}ab$ (this is because $\downarrow \mathcal{L}ab$ is an admissible labelling). Moreover, when each opponent move is labelled out by $\downarrow \mathcal{L}ab$, it is always possible for the proponent to reply with an argument that is labelled in by $\downarrow \mathcal{L}ab$ (this is again because $\downarrow \mathcal{L}ab$ is an admissible labelling). Suppose the proponent follows such a strategy (of choosing only moves that are labelled in by $\downarrow \mathcal{L}ab$). As the opponent cannot repeat his moves, the discussion will terminate in a finite number of steps. Termination cannot be due to termination condition 2 (since the fact that $\downarrow \mathcal{L}ab$ is a well defined labelling implies that there exists no argument B with both $B \in \text{in}(\downarrow \mathcal{L}ab)$ and $B \in \text{out}(\downarrow \mathcal{L}ab)$). Also, termination cannot be due to termination conditions 3 or 4, as the proponent's strategy ensures that (as we have observed) for each proponent move $\text{in}(B)$ it holds that $B \in \text{in}(\downarrow \mathcal{L}ab)$ and for each opponent move $\text{out}(B)$ it holds that $B \in \text{out}(\downarrow \mathcal{L}ab)$. This means that termination must be due to termination condition 1 (meaning no next move is possible). As the proponent's strategy ensures that the proponent can *always* move after an opponent move (as the fact that $\downarrow \mathcal{L}ab$ is an admissible labelling means that for each $B \in \text{out}(\downarrow \mathcal{L}ab)$ there exists a $C \in \text{in}(\downarrow \mathcal{L}ab)$ such that C attacks B) this means that the last move cannot be an opponent move. That is, winning condition 1 cannot be applicable (nor can winning conditions 2, 3 and 4 be applicable). It then directly follows that the proponent wins the discussion. \square

Using the down-admissible game, it becomes fairly straightforward to define a dialectical proof procedure for the sceptical and credulous operators. The eas-

⁵ $\text{undec}(\mathcal{L}ab')$ then consists of all arguments that are neither proponent moves nor opponent moves.

iest way would be simply to start with \mathcal{Lab} being either $\sqcap(\mathcal{Lab}_1, \dots, \mathcal{Lab}_n)$ or $\sqcup(\mathcal{Lab}_1, \dots, \mathcal{Lab}_n)$. Alternatively, it would be possible to define separate dialectical proof procedures for the sceptical and credulous operators, by slightly changing the rules of the down-admissible discussion game. For the sceptical game, we need to change clauses 3 and 4 regarding the termination and winning criterion to:

- 3' there exists a proponent move $\text{in}(B)$ s.t. $\mathcal{Lab}_i(B) \neq \text{in}$ for some $i \in \{1 \dots n\}$
- 4' there exists an opponent move $\text{out}(B)$ s.t. $\mathcal{Lab}_i(B) \neq \text{out}$ for some $i \in \{1 \dots n\}$

As an example of how the sceptical discussion game works, consider the argumentation framework of Figure 1 and labellings $\mathcal{Lab}_1 = (\{A, D\}, \{B, C\}, \emptyset)$ (of agent 1) and $\mathcal{Lab}_2 = (\{B, D\}, \{A, C\}, \emptyset)$ (of agent 2).

- Proponent: "We can all agree that D has to be accepted ($\text{in}(D)$)"
- Opponent: "But then we'd also all have to agree that D 's attacker C has to be rejected ($\text{out}(C)$). Based on what grounds?"
- Proponent: "We can all agree that C has to be rejected because we can all agree that A has to be accepted ($\text{in}(A)$)"
- Agent 2: "Objection! I don't accept A myself."⁶

As the sceptical game is essentially the admissibility game of [15,9] with extra conditions 3' and 4', we can think of the sceptical game as the standard admissibility game with a twist: apart from participants proponent and opponent, there is now also a room full of potential hecklers (the agents whose labellings are being aggregated). If the discussion between the proponent and opponent touches an argument of which one of the agents in the room does not agree on its label, the agent shouts "Objection!" in which case the discussion ends and the proponent loses (regardless of whether it was a proponent or opponent move that was being objected to).

The discussion game for the credulous operator can be defined in a similar way. Again, the easiest way would be to simply start with \mathcal{Lab} being $\sqcup(\mathcal{Lab}_1, \dots, \mathcal{Lab}_n)$. Alternatively, rules 3 and 4 regarding the termination and winning criterion should be changed as follows:

- 3'' there exists a proponent move $\text{in}(B)$ s.t. $\mathcal{Lab}_i(B) = \text{out}$ for some $i \in \{1 \dots n\}$ or $\mathcal{Lab}_i(B) \neq \text{in}$ for each $i \in \{1 \dots n\}$
- 4'' there exists an opponent move $\text{out}(B)$ s.t. $\mathcal{Lab}_i(B) = \text{in}$ for some $i \in \{1 \dots n\}$ or $\mathcal{Lab}_i(B) \neq \text{out}$ for each $i \in \{1 \dots n\}$

As an example of how the credulous discussion game works, consider again the argumentation framework of Figure 1 and labellings $\mathcal{Lab}_1 = (\{A, D\}, \{B, C\}, \emptyset)$

⁶For the sake of the example, we have modelled the objection as a separate move that indicates termination condition 3'.

(of agent 1) and $\mathcal{Lab}_2 = (\{B, D\}, \{A, C\}, \emptyset)$ (of agent 2).

Proponent: “We can all agree that D has to be accepted ($\text{in}(D)$)”

Room: “Aye” (Agent 1) “Aye” (Agent 2)

Opponent: “But then we’d also all have to agree that D ’s attacker C has to be rejected ($\text{out}(C)$). Based on what grounds?”

Room: “Aye” (Agent 1) “Aye” (Agent 2)

Proponent: “We can all agree that C has to be rejected because we can all agree that A has to be accepted ($\text{in}(A)$)”

Room: “Aye” (Agent 1) “Nay” (Agent 2)

As the credulous game is essentially the admissibility game of [15,9] with extra conditions 3'' and 4'', we can think of the credulous game as the standard admissibility game with a twist: after each move of the proponent and opponent, the agents in the room are asked for their opinion. Agents who agree with the label of the argument shout “Aye”. Agents who have the opposite label⁷ shout “Nay”.⁸ If there is at least one agent that shouts “Aye” and no agent that shouts “Nay” then the discussion continues. However, if there is no agent shouting “Aye” or at least one agent that shouts “Nay” then the discussion is terminated and the proponent loses (regardless of whether it was a proponent or opponent move that caused it).

4. A Dialectical Proof Procedure for the Super Credulous Operator

As the super credulous operator is based on the up-complete labelling, we first define an up-complete discussion game, based on the Grounded Discussion Game [8].

The Grounded Discussion Game is a sound and complete dialectical proof procedure to determine whether an argument is in the grounded extension.⁹ It is based on a discussion between two participants (proponent and opponent) who use the following four kind of utterances.

$HTB(A)$ (“ A has to be the case”)

With this move, the proponent claims that A has to be labelled **in**.

$CB(B)$ (“ B can be the case, or at least cannot be ruled out”)

With this move, the opponent claims that B does not have to be labelled **out**.

$CONCEDE(A)$ (“I agree that A has to be the case”)

With this move, the opponent indicates that he now agrees with the proponent (who previously did a $HTB(A)$ move) that A has to be labelled **in**.

⁷with **in** being the opposite of **out**, and **out** being the opposite of **in**

⁸Our naming convention is inspired by the British parliament, where a similar procedure is used before holding a physical vote.

⁹The Grounded Discussion Game has a number of advantages compared to alternative dialectical proof procedures for grounded semantics like the Standard Grounded Game [17] and the Grounded Persuasion Game [13]. We refer to [8] for details.

RETRACT(B) (“I give up that B can be the case”)

With this move, the opponent indicates that he no longer believes that B can be **in** or **undec.** That is, the opponent acknowledges that B has to be labelled **out**.

One of the key ideas of the game is that the proponent has burden of proof. That is, the proponent has to establish the acceptance of the main argument and make sure that the discussion does not go around in circles (meaning that arguments are not mentioned more than once).

Using the four moves of the Grounded Discussion Game, we proceed to define the up-complete discussion game.

Definition 8. *Let (Ar, \rightarrow) be an argumentation framework. An up-complete discussion game is a sequence of discussion moves constructed by applying the following principles.*

BASIS (*HTB*) *If $A \in Ar$ then $[HTB(A)]$ is an up-complete discussion.*

STEP (*HTB*) *If $[M_1, \dots, M_n]$ ($n \geq 1$) is an up-complete discussion without *HTB-CB* repeats,¹⁰ and no *CONCEDE* or *RETRACT* move is applicable, and $M_n = CB(A)$ and B is an attacker of A then $[M_1, \dots, M_n, HTB(B)]$ is also an up-complete discussion.*

STEP (*CB*) *If $[M_1, \dots, M_n]$ ($n \geq 1$) is an up-complete discussion without *HTB-CB* repeats, and no *CONCEDE* or *RETRACT* move is applicable, and M_n is not a *CB* move, and there is a move $M_i = HTB(A)$ ($i \in \{1 \dots n\}$) such that the discussion does not contain *CONCEDE*(A), and for each move $M_j = HTB(A')$ ($j > i$) the discussion contains a move *CONCEDE*(A'), and B is an attacker of A such that the discussion does not contain a move *RETRACT*(B), then $[M_1, \dots, M_n, CB(B)]$ is an up-complete discussion.*

STEP (*CONCEDE*) *If $[M_1, \dots, M_n]$ ($n \geq 1$) is an up-complete discussion without *HTB-CB* repeats, and *CONCEDE*(B) is applicable then $[M_1, \dots, M_n, CONCEDE(B)]$ is an up-complete discussion.*

STEP (*RETRACT*) *If $[M_1, \dots, M_n]$ ($n \geq 1$) is an up-complete discussion without *HTB-CB* repeats, and *RETRACT*(B) is applicable then $[M_1, \dots, M_n, RETRACT(B)]$ is an up-complete discussion.*

A key issue in Definition 8 is when a *CONCEDE* or *RETRACT* move is applicable. In the original Grounded Discussion Game [8], a move *CONCEDE*(B) is applicable iff the discussion contains a move *HTB*(B), the discussion does not already contain a move *CONCEDE*(B) and for every attacker A of B the discussion contains a move *RETRACT*(A). Also, a move *RETRACT*(B) is applicable iff the discussion contains a move *CB*(B), the discussion does not already contain a move *RETRACT*(B), and there is an attacker A of B such that the discussion contains a move *CONCEDE*(A). For the up-complete discussion game, we need to slightly alter this condition as follows.

A move *CONCEDE*(B) is applicable iff

¹⁰We say that there is a *HTB-CB* repeat iff $\exists i, j \in \{1 \dots n\} \exists A \in Ar : (M_i = HTB(A) \vee M_i = CB(A)) \wedge (M_j = HTB(A) \vee M_j = CB(A)) \wedge i \neq j$.

1. the discussion contains a previous move $HTB(B)$, and
2. the discussion does not already contain a move $CONCEDE(B)$, and
3. either
 - a. for every attacker A of B the discussion contains a previous move $RETRACT(A)$, or
 - b. B is labelled in by the initial labelling \mathcal{Lab}

A move $RETRACT(B)$ is applicable iff

1. the discussion contains a previous move $CB(B)$, and
2. the discussion does not already contain a move $RETRACT(B)$, and
3. either
 - a. there exists an attacker A of B such that the discussion contains a previous move $CONCEDE(A)$, or
 - b. B is labelled out by the initial labelling \mathcal{Lab}

The above definition of applicability of $CONCEDE$ and $RETRACT$ is almost the same as in the Grounded Discussion Game [8] (as is the rest of the up-complete game). The only difference is that the condition 3b has been added in regarding the applicability of $CONCEDE$ and the applicability of $RETRACT$.

Just as in the Grounded Discussion Game, the proponent wins the up-complete game iff the opponent concedes the main argument (the argument the discussion started with).

Definition 9. An up-complete discussion $[M_1, \dots, M_n]$ is called terminated iff there exists no move M_{n+1} such that $[M_1, \dots, M_n, M_{n+1}]$ is an up-complete discussion. A terminated up-complete discussion (with A being the main argument) is won by the proponent iff the discussion contains $CONCEDE(A)$, otherwise it is won by the opponent.

As an example of how the up-complete discussion game works, consider the argumentation framework of Figure 1 and labelling $\mathcal{Lab} = (\{A\}, \{B\}, \{C, D\})$. The discussion $[HTB(D), CB(C), HTB(A), CONCEDE(A), RETRACT(C), CONCEDE(D)]$ is terminated and won by the proponent. We observe that $D \in \text{in}(\uparrow\mathcal{Lab})$ as $\uparrow\mathcal{Lab} = (\{A, D\}, \{B, C\}, \emptyset)$.

We are now ready to formally state soundness and completeness of the up-complete discussion game.

Theorem 2. Let \mathcal{Lab} be an admissible labelling (called the initial labelling) of argumentation framework (Ar, \rightarrow) . An argument $A \in Ar$ is labelled in by $\uparrow\mathcal{Lab}$ iff the proponent has a winning strategy for A in the up-complete game.

Proof. We first define the increasing sequence of labellings $L_0 \sqsubseteq \dots \sqsubseteq L_u$ inductively by $L_0 = \mathcal{Lab}$ and $L_{i+1} = \Gamma(L_i)$ for $i \geq 0$, where u is minimal such that $L_u = L_{u+1}$. By results in [5] we know each L_i is admissible and $L_u = \uparrow\mathcal{Lab}$.

“ \Leftarrow ”: Suppose proponent has a winning strategy for A . Then in particular there is a terminated discussion $[M_1, \dots, M_n]$ containing move $CONCEDE(A)$. For each $1 \leq i \leq n$ define labelling N_i by setting $N_i(B) = \mathcal{Lab}(B)$ if $B \in \text{in}(\mathcal{Lab}) \cup \text{out}(\mathcal{Lab})$, $N_i(B) = \text{in}$ if $B \in \text{undec}(\mathcal{Lab})$ and $M_j = CONCEDE(B)$ for

some $j \leq i$, $N_i(B) = \text{out}$ if $B \in \text{undec}(\mathcal{Lab})$ and $M_j = \text{RETRACT}(B)$ for some $j \leq i$, $N_i(B) = \text{undec}$ otherwise. We note that N_i is well-defined due to there being no *HTB-CB* repeats. Then, by induction we can show that for each i we have $N_i \subseteq L_j$ for some j (which depends on i). In particular $N_n \subseteq L_j \subseteq L_u = \uparrow \mathcal{Lab}$ and so, since $N_n(A) = \text{in}$ we have A labelled **in** by $\uparrow \mathcal{Lab}$.

“ \Rightarrow ”: Suppose A is labelled **in** by $\uparrow \mathcal{Lab}$. For any $B \in \text{in}(L_u) \cup \text{out}(L_u)$ let $r(B)$ be minimal such that $L_{r(B)}(B) = L_u(B)$. Note that, for *any* discussion starting with $\text{HTB}(A)$, $L_{r(A)}$ labels the arguments of any *HTB* and *CB* moves with **in**, **out** respectively (this can be proved by induction on the length of the discussion). Then we can define a strategy for the proponent as follows: whenever opponent plays $\text{CB}(B)$ and no *CONCEDE* or *RETRACT* moves are applicable, play $\text{HTB}(C)$ where C is any argument such that $C \rightarrow B$ and $r(C) = r(B) - 1$. (C exists by the admissibility of $L_{r(B)}$ and the minimality of $r(B)$. Note if $r(B) = 0$ then B will be immediately *RETRACTed* after $\text{CB}(B)$ is played.) To show this yields a *winning* strategy we claim that, for any *terminated* discussion following this strategy starting with $\text{HTB}(A)$ and for all arguments B , any move $\text{CB}(B)$ or $\text{HTB}(B)$ will eventually be followed by a $\text{RETRACT}(B)$ and $\text{CONCEDE}(B)$ move respectively. In particular $\text{HTB}(A)$ will eventually be followed by $\text{CONCEDE}(A)$, so the proponent wins. The claim is proved by induction on $r(B)$. If $r(B) = 0$ then the conclusion follows from applicability conditions 3b for *CONCEDE* and *RETRACT*. So suppose $r(B) = i > 0$ and the claim holds for all C such that $r(C) < i$. Suppose $\text{CB}(B)$ is played. If it is not immediately *RETRACTed* then, following the strategy, the next move is $\text{HTB}(C)$ with $C \rightarrow B$ and $r(C) = r(B) - 1$. By induction, C is eventually *CONCEDEd*, at which point B must also be *RETRACTed*. If $\text{HTB}(B)$ is played and is not immediately *CONCEDEd* then eventually all attackers of B must be played as *CB*. However for any such attacker C we have $r(C) < r(B)$ and so, by induction, every attacker must eventually be *RETRACTed*. \square

Given the up-complete game, it becomes possible to combine this with the down-admissible game to provide dialectical proof procedures for the super credulous operator. The idea is to embed the down-admissible game inside of the up-complete game. That is, to determine whether argument A is labelled **in** by the super credulous labelling, we start with running the up-complete game for argument A (first move: $\text{HTB}(A)$). In the up-complete game defined above, the opponent has to move *CONCEDE* when the discussion hits an argument that is labelled **in** by the initial labelling. However, what we are interested in is not so much whether an argument is labelled **in** by the initial labelling, but whether the argument is labelled **in** by the down-admissible of the initial labelling. This can be determined by running the down-admissible game. So whenever the proponent wants to do a *HTB* move for an argument (say B) he thinks is in the down-admissible of the initial labelling, instead of doing an $\text{HTB}(B)$ move, he starts the down-admissible game (first move: $\text{in}(B)$). If the proponent wins the down-admissible game, the entire game counts as an $\text{HTB}(B)$ move with B being

in the down-admissible of the initial labelling. This means that (condition 3b) the opponent has to respond with a *CONCEDE*(B) move.¹¹

If one then substitutes the down-admissible game (inside of the up-complete game) by the credulous game (which after all is the down-admissible game based on the credulous initial labelling $\sqcup(\mathcal{L}ab_1, \dots, \mathcal{L}ab_n)$) one obtains a discussion game for the super-credulous operator.

5. Discussion

In the current paper, we have shown that it is possible to define sound and complete dialectical proof procedures for the down-admissible and up-complete labellings. These proof procedures were obtained by relatively minor changes to the dialectical proof procedures for preferred [9] and grounded [8] semantics. The proof procedure for down-admissible (the down-admissible game) is basically the admissibility game of [9] with the additional constraint that the discussion needs to stay “inside” of the initial labelling (that is, for the proponent to win the game, for each *in*(B) move it has to hold that $B \in \text{in}(\mathcal{L}ab)$, and for each *out*(B) move it has to hold that $B \in \text{out}(\mathcal{L}ab)$). The proof procedure for the up-complete labelling (the up-complete game) is basically the Grounded Discussion Game [8] with immediate *CONCEDE* and *RETRACT* moves whenever the discussion touches arguments in the initial labelling (that is, when uttering a move *HTB*(B) with $B \in \text{in}(\mathcal{L}ab)$ or *CB*(B) with $B \in \text{out}(\mathcal{L}ab)$). In the special case that the initial labelling is the all-undec labelling, the up-complete game coincides precisely with the Grounded Discussion Game, as the grounded labelling is the up-complete of the all-undec labelling.

Based on the proof procedures of the down-admissible and up-complete, we then outlined dialectical proof procedures for the sceptical, credulous and super credulous operator. The proof procedures for the sceptical and credulous operator are based on the down-admissible game with particular initial labellings. The proof procedure of the super credulous operator is based on the up-complete game with the down-admissible game embedded in it.

We have shown that our discussion games can be given an intuitive interpretation. For the sceptical and credulous games, one could envision a panel discussion between a proponent and an opponent, with the audience consisting of the agents whose opinions (labellings) are being aggregated, and who are able to actively interfere (“heckle”) with the panel discussion.¹² By providing such an intuitive interpretation, the game goes beyond the purely computational function of traditional proof procedures. This is in line with what should be arguably one of the main aims of formal argumentation theory: to bridge the gap between computer-based reasoning and human reasoning.

¹¹One could ask whether a similar down-admissible game is necessary when the opponent wants to do a *CB* move for an argument that is labelled *out* by the down-admissible of the initial labelling. The answer is negative, as carrying on for one more step in the up-complete game will yield an (*HTB*) argument that is labelled *in* by the down-admissible of the initial labelling. On this argument we then run the embedded down-admissible game as described above.

¹²For the super credulous game, audience participation is only possible during the credulous subgame.

References

- [1] E. Awad. *Collective Judgement in Contested Domains: The Case of Conflicting Arguments*. PhD thesis, Masdar Institute, 2015.
- [2] E. Awad, J. F. Bonnefon, M. W. A. Caminada, Th. Malone, and I. Rahwan. Experimental assessment of aggregation rules in argumentation-enabled collective intelligence. Technical report, arXiv:1604.00681, 2016.
- [3] E. Awad, M. W. A. Caminada, G. Pigozzi, M. Podlaszewski, and I. Rahwan. Pareto optimality and strategy proofness in group argumentation evaluation. Technical report, arXiv:1604.00693, 2016.
- [4] P. Baroni, M.W.A. Caminada, and M. Giacomin. An introduction to argumentation semantics. *Knowledge Engineering Review*, 26(4):365–410, 2011.
- [5] R. Booth. Judgment aggregation in abstract dialectical frameworks. In *Advances in Knowledge Representation, Logic Programming, and Abstract Argumentation - Essays Dedicated to Gerhard Brewka on the Occasion of His 60th Birthday*, volume 9060 of *Lecture Notes in Computer Science*, pages 296–308. Springer, 2015.
- [6] R. Booth, E. Awad, and I. Rahwan. Interval methods for judgment aggregation in abstract argumentation. In *Proceedings of the 14th International Conference on Principles of Knowledge Representation and Reasoning (KR 2014)*, pages 594–597, 2014.
- [7] M.W.A. Caminada. On the issue of reinstatement in argumentation. In M. Fischer, W. van der Hoek, B. Konev, and A. Lisitsa, editors, *Logics in Artificial Intelligence; 10th European Conference, JELIA 2006*, pages 111–123. Springer, 2006. LNAI 4160.
- [8] M.W.A. Caminada. A discussion game for grounded semantics. In Elizabeth Black, Sanjay Modgil, and Nir Oren, editors, *Theory and Applications of Formal Argumentation (proceedings TAFE 2015)*, pages 59–73. Springer, 2015.
- [9] M.W.A. Caminada, W. Dvořák, and S. Vesic. Preferred semantics as socratic discussion. *Journal of Logic and Computation*, 2014. (in print).
- [10] M.W.A. Caminada and D.M. Gabbay. A logical account of formal argumentation. *Studia Logica*, 93(2-3):109–145, 2009. Special issue: new ideas in argumentation theory.
- [11] M.W.A. Caminada and G. Pigozzi. On judgment aggregation in abstract argumentation. *Autonomous Agents and Multi-Agent Systems*, 22(1):64–102, 2011.
- [12] M.W.A. Caminada, G. Pigozzi, and M. Podlaszewski. Manipulation in group argument evaluation. In Toby Walsh, editor, *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*, pages 121–126, 2011.
- [13] M.W.A. Caminada and M. Podlaszewski. Grounded semantics as persuasion dialogue. In Bart Verheij, Stefan Szeider, and Stefan Woltran, editors, *Computational Models of Argument - Proceedings of COMMA 2012*, pages 478–485, 2012.
- [14] M.W.A. Caminada and M. Podlaszewski. User-computer persuasion dialogue for grounded semantics. In Jos W.H.M. Uiterwijk, Nico Roos, and Mark H.M. Winands, editors, *Proceedings of BNAIC 2012; The 24th Benelux Conference on Artificial Intelligence*, pages 343–344, 2012.
- [15] M.W.A. Caminada and Y. Wu. An argument game of stable semantics. *Logic Journal of IGPL*, 17(1):77–90, 2009.
- [16] N. Gorogiannis and A. Hunter. Instantiating abstract argumentation with classical logic arguments: Postulates and properties. *Artificial Intelligence*, 175(9-10):1479–1497, 2011.
- [17] S. Modgil and M.W.A. Caminada. Proof theories and algorithms for abstract argumentation frameworks. In I. Rahwan and G.R. Simari, editors, *Argumentation in Artificial Intelligence*, pages 105–129. Springer, 2009.
- [18] S. Modgil and H. Prakken. The ASPIC+ framework for structured argumentation: a tutorial. *Argument & Computation*, 5:31–62, 2014. Special Issue: Tutorials on Structured Argumentation.
- [19] M. Podlaszewski. *Poles Apart: Navigating the Space of Opinions in Argumentation*. PhD thesis, Université du Luxembourg, 2015.
- [20] F. Toni. A tutorial on assumption-based argumentation. *Argument & Computation*, 5:89–117, 2014. Special Issue: Tutorials on Structured Argumentation.

Towards a New Framework for Recursive Interactions in Abstract Bipolar Argumentation

Claudette CAYROL^a Andrea COHEN^b M-Christine LAGASQUIE-SCHIEX^{a,1}

^aIRIT, Université de Toulouse, France

^bICIC, CONICET-UNS, Bahía Blanca, Argentina

Abstract. This paper proposes a new framework able to take into account recursive interactions in bipolar abstract argumentation systems. We address issues such as “How an interaction can impact another one?”, or in other words “How can the validity of an interaction be affected if this interaction is attacked or supported by another one?”. Thus, building on numerous examples, a new method for flattening such recursive bipolar abstract argumentation systems (ASAF) using meta-arguments is proposed and compared with the original framework defined in [8].

Keywords. Abstract argumentation, bipolar argumentation, recursive interactions.

1. Introduction

Argumentation has become an essential paradigm especially for reasoning from contradictory information [9,1], and for formalizing the exchange of arguments between agents in, *e.g.*, negotiation [2]. Formal abstract frameworks have greatly eased the modelling and study of argumentation. For instance, a Dung’s argumentation system (AS) [9] consists of a collection of arguments interacting with each other through an *attack* relation, enabling to determine “acceptable” sets of arguments called *extensions*.

In the last decade, extensions of Dung’s AS were proposed for including a positive interaction between arguments, called *support*. The support relation has been first introduced in [10,17]. In [4], the support relation is left general so that the obtained bipolar AS (BAS) keeps a high level of abstraction. However there is no single interpretation of support, and a number of researchers proposed specialized variants of the support relation (deductive support [18], necessary support [13], evidential support [14]). Each specialization was developed quite independently, based on different intuitions and provided with an appropriate formalization. In order to restate those proposals in a common setting, [6] proposed a comparative study using the BAS. Following the same line, recent works have been proposed that enforce the important role of necessary support (see *e.g.*, [15,16,7,12]). Another line of work extending Dung’s AS regards high-order attacks: attacks to the attack relation [11,3] and attacks to attacks and supports [18]. More generally, [8] proposes an *Attack-Support Argumentation Framework* (ASAF) which allows for attack and support to the attack and support relations, at any level.

¹Corresponding Author: lagasq@irit.fr

The authors in [8] encode an ASAF by turning it into a BAS with necessary support, and then into an AS by adding extended attacks. As [7] presents different frameworks for encoding necessary support, it is interesting to enrich them with recursive interactions. So, in this paper we propose to translate the ASAF into a special AS using meta-arguments² (called MAS), and compare this MAS with the AS obtained in [8]. Note that our aim is not to replace the ASAF with the MAS; although recursive interactions can be modelled using the MAS, the ASAF is a more intuitive and visual representation tool.

This paper is organized as follows: BAS (with necessary support) and its axiomatization are presented in Sect. 2. Background on the ASAF is given in Sect. 3. In Sect. 4 we extend the MAS proposed in [7] to model recursive interactions. Then, Sect. 5 compares this MAS with the ASAF. Finally, Sect. 6 concludes and suggests lines of future work.

2. Bipolar abstract argumentation system

The abstract bipolar argumentation system presented in [4,5] extends Dung's AS [9] by adding a positive interaction between arguments: the support relation.

Def. 1 (BAS) A bipolar argumentation system (BAS) is a tuple $\langle \mathbf{A}, \mathbf{R}_{\text{att}}, \mathbf{R}_{\text{sup}} \rangle$, where \mathbf{A} is a finite and non-empty set of arguments, $\mathbf{R}_{\text{att}} \subseteq \mathbf{A} \times \mathbf{A}$ is an attack relation and $\mathbf{R}_{\text{sup}} \subseteq \mathbf{A} \times \mathbf{A}$ is a support relation.

A BAS can be represented by a directed graph with two kinds of edges: $\forall a, b \in \mathbf{A}$, $a\mathbf{R}_{\text{att}}b$ (resp. $a\mathbf{R}_{\text{sup}}b$) is represented by $a \longrightarrow b$ (resp. by $a \Longrightarrow b$). Semantics introduced by Dung for AS can only be used if $\mathbf{R}_{\text{sup}} = \emptyset$. They characterize sets of arguments that satisfy some properties and some form of optimality. For instance:³

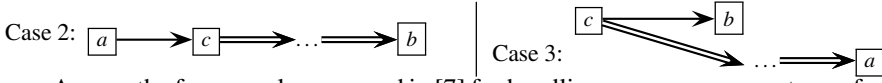
Def. 2 (Preferred extensions of AS) Let $\text{AS} = \langle \mathbf{A}, \mathbf{R}_{\text{att}} \rangle$ and $S \subseteq \mathbf{A}$. S is conflict-free iff $\nexists a, b \in S$, s.t. $a\mathbf{R}_{\text{att}}b$. $a \in \mathbf{A}$ is acceptable wrt S iff $\forall b \in \mathbf{A}$ s.t. $b\mathbf{R}_{\text{att}}a$, $\exists c \in S$ s.t. $c\mathbf{R}_{\text{att}}b$. S is admissible iff it is conflict-free and $\forall b \in S$, b is acceptable wrt S . S is a preferred extension of AS iff it is a maximal (\subseteq) admissible set.

Handling support and attack at an abstract level has the advantage to keep genericity and to give an analytic tool for studying complex attacks and new semantics considering both attack and support relations, among others. However, the drawback is the lack of guidelines for choosing the appropriate definitions and semantics depending on the application. For solving this problem, some variants of the support relation have been proposed recently, including the necessary support. This kind of support was initially proposed in [13] with the following interpretation: If $c\mathbf{R}_{\text{sup}}b$ then the acceptance of c is necessary to get the acceptance of b , or equivalently the acceptance of b implies the acceptance of c . Suppose now that $a\mathbf{R}_{\text{att}}c$. The acceptance of a implies the non-acceptance of c and so the non-acceptance of b . Also, if $c\mathbf{R}_{\text{sup}}a$ and $c\mathbf{R}_{\text{att}}b$, the acceptance of a implies the acceptance of c and the acceptance of c implies the non-acceptance of b . So, the acceptance of a implies the non-acceptance of b . These constraints relating a and b are enforced by adding new complex attacks from a to b :

Def. 3 ([13] Extended attack) Let $\text{BAS} = \langle \mathbf{A}, \mathbf{R}_{\text{att}}, \mathbf{R}_{\text{sup}} \rangle$. There is an extended attack from a to b iff either (1) $a\mathbf{R}_{\text{att}}b$; or (2) $a_1\mathbf{R}_{\text{att}}a_2\mathbf{R}_{\text{sup}}\dots\mathbf{R}_{\text{sup}}a_n$, $n \geq 3$, with $a_1 = a$, $a_n = b$; or (3) $a_1\mathbf{R}_{\text{sup}}\dots\mathbf{R}_{\text{sup}}a_n$, and $a_1\mathbf{R}_{\text{att}}a_p$, $n \geq 2$, with $a_n = a$, $a_p = b$. Graphically:

²A similar idea is presented in [18] for representing defeasible attacks and supports.

³“iff” means “if and only if”, “s.t.” means “such that” and “wrt” means “with respect to”.



Among the frameworks proposed in [7] for handling necessary supports, we focus on the one encoding the following interpretation: If $c \mathbf{R}_{\text{sup}} b$, “the acceptance of c is necessary to get the acceptance of b ” because “ c is the *only* attacker of a particular attacker of b ”:

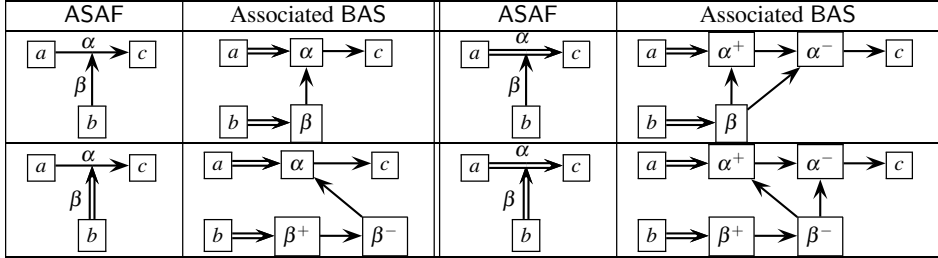
Def. 4 ([7] MAS associated with a BAS) Let $\text{BAS} = \langle \mathbf{A}, \mathbf{R}_{\text{att}}, \mathbf{R}_{\text{sup}} \rangle$ with \mathbf{R}_{sup} being a set of necessary supports. Let $\mathbf{A}_n = \{N_{cb} | (c, b) \in \mathbf{R}_{\text{sup}}\}$ and $\mathbf{R}_n = \{(c, N_{cb}) | (c, b) \in \mathbf{R}_{\text{sup}}\} \cup \{(N_{cb}, b) | (c, b) \in \mathbf{R}_{\text{sup}}\}$. The tuple $\text{MAS} = \langle \mathbf{A} \cup \mathbf{A}_n, \mathbf{R}_{\text{att}} \cup \mathbf{R}_n \rangle$ is the meta-argumentation system associated with BAS (it is a Dung’s AS).

3. Recursive interactions

Recursive interactions were explored in [3] for recursive attacks (the AFRA) and in [8] for recursive supports plus attacks (the ASAF). The idea is to model that the validity of an interaction may depend on other interactions (e.g., because of preferences [11]). Here we focus on the ASAF [8], where arguments interact with other arguments or interactions:

Def. 5 (ASAF) An Attack-Support Argumentation Framework (ASAF) is a tuple $\langle \mathbf{A}, \mathbf{R}_{\text{att}}, \mathbf{R}_{\text{sup}} \rangle$ where \mathbf{A} is a set of arguments, $\mathbf{R}_{\text{att}} \subseteq \mathbf{A} \times (\mathbf{A} \cup \mathbf{R}_{\text{att}} \cup \mathbf{R}_{\text{sup}})$ is an attack relation, and $\mathbf{R}_{\text{sup}} \subseteq \mathbf{A} \times (\mathbf{A} \cup \mathbf{R}_{\text{att}} \cup \mathbf{R}_{\text{sup}})$ is a necessary support relation. Note that \mathbf{R}_{sup} is assumed to be irreflexive and transitive. We assume that $\mathbf{R}_{\text{att}} \cap \mathbf{R}_{\text{sup}} = \emptyset$.

[8] translates an ASAF into an AS in two steps: first, the ASAF is turned into a BAS with necessary support; then, this BAS is turned into an AS by adding extended attacks. For the first step, the following schemas describe the translation of the 4 basic cases:



Given the BAS associated with the ASAF, the second step followed in [8] is to create an AS by adding complex attacks, namely Case 2 - extended attacks (see Def. 3):

Def. 6 (AS associated with BAS and with ASAF) Let $\text{BAS} = \langle \mathbf{A}, \mathbf{R}_{\text{att}}, \mathbf{R}_{\text{sup}} \rangle$ be the BAS associated with ASAF. The pair $\text{AS}' = \langle \mathbf{A}', \mathbf{R}' \rangle$, where $\mathbf{A}' = \mathbf{A}$ and $\mathbf{R}' = \mathbf{R}_{\text{att}} \cup \{(a, b) | \text{there is a sequence } a_1 \mathbf{R}_{\text{att}} a_2 \mathbf{R}_{\text{sup}} \dots \mathbf{R}_{\text{sup}} a_n, n \geq 3, \text{ with } a_1 = a, a_n = b\}$ is the AS associated with BAS and ASAF.

4. Encoding recursive interactions in MAS

In this section, we propose to use the MAS (see Sect. 2) for encoding recursive interactions, addressing the following issues:

- distinguish between labelled and unlabelled interactions, i.e. between interactions that may be involved in a recursion (either as a target, or as targeting another interaction) and the other interactions;
- encode labelled interactions, i.e. the ability to reason about them; and
- encode recursive interactions, i.e. the impact of an interaction on another one.

For this purpose, we need to formalize the notion of labelled interaction. So we propose a slightly modified version of the ASAF, which we call the labelled ASAF.

Def. 7 (Labelled ASAF) A labelled ASAF is a 5-uple $\langle \mathbf{A}, \mathbf{R}_{\text{att}}, \mathbf{R}_{\text{sup}}, \mathcal{V}, \mathcal{L} \rangle$ where \mathbf{A} is a set of arguments, $\mathbf{R}_{\text{att}} \subseteq \mathbf{A} \times (\mathbf{A} \cup \mathbf{R}_{\text{att}} \cup \mathbf{R}_{\text{sup}})$ is an attack relation, $\mathbf{R}_{\text{sup}} \subseteq \mathbf{A} \times (\mathbf{A} \cup \mathbf{R}_{\text{att}} \cup \mathbf{R}_{\text{sup}})$ is a necessary support relation, \mathcal{V} is a set of labels (denoted by greek letters) and \mathcal{L} is a bijection from $\mathbf{R} \subseteq (\mathbf{R}_{\text{att}} \cup \mathbf{R}_{\text{sup}})$ to \mathcal{V} . We still assume that $\mathbf{R}_{\text{att}} \cap \mathbf{R}_{\text{sup}} = \emptyset$.

The above definition enables to distinguish interactions that are not involved in a recursion; they may be considered as always “valid” and will be called “basic” in the following. Since the aim of labels is to enable reasoning about interactions and encode recursive interactions, basic interactions do not require labels.⁴ Moreover, each label corresponds to a unique labelled interaction and vice-versa. The main difference between Def. 7 and Def. 5 is the explicit integration of labels into the ASAF. In the following, a labelled interaction will be confused with its label. In order to define the MAS associated with a labelled ASAF, next we explain the encoding of each component of the ASAF.

Encoding basic attacks/supports. Such interactions correspond to unlabelled interactions and can be directly encoded using the MAS given in [7] (see Def. 4).

Encoding labelled interactions. In order to reason about an interaction that is attacked or supported we must be able to refer to it; hence, it must be labelled and its label will be used as a “meta-argument”. A labelled interaction $\alpha = (a, b)$ encompasses two types of links. One link relates α to b , representing the role of α (either an attack or a support), and will be called the *effect-link*. The other link relates α to its source a , representing the grounding of α , and will be called the *ground-link*. The idea of “grounded” interaction is close to the notion of evidential argumentation in the work of [14,15]. It means that “an interaction makes sense only if its source argument is accepted”.

These two links suggest two kinds of validity for the interaction. We reserve the term *validity* for the effect-link. For instance, in a graph containing only α attacked by b (through an attack $\beta = (b, \alpha)$), α is not valid. Similarly, if α is supported by c (with a support $\gamma = (c, \alpha)$) and c is attacked by d , then α is not valid. Concerning the ground-link, we use the term *grounded*. For instance, $\alpha = (a, b)$ is not grounded if a is attacked and not defended. Note that a support can be valid even though its source is not accepted. So interactions may be valid and not grounded, or grounded and not valid. We call *active* an interaction which is both valid and grounded. Intuitively, if α is only attacked by a non-active interaction (whatever the origin of this non-activation), then α should be valid. If α is supported by an interaction β which is valid but not grounded, then α should not be valid. However, if β is not valid, the validity of α cannot be affected by β . The following table synthetizes the above notions for a labelled interaction $\alpha = (a, b)$.

Type of link	Meaning of the link	Corresponding Notions	
effect-link	describes the role of α wrt b (is affected by interactions on α)	validity	validity + groundness = activation
ground-link	describes the existence of α wrt a (takes into account only the source of α)	groundness	

If an attack $\alpha = (a, b)$ is active, then a and b cannot belong to the same extension. And if a support $\alpha = (a, b)$ is active, then if b is accepted then a must be also accepted.

The ground-link is a necessary support between the meta-argument and the source argument; thus, an interaction $\alpha = (a, b)$ will be “grounded” only if a is accepted. This

⁴Note that if all interactions are unlabelled ($\mathcal{V} = \emptyset$), the labelled ASAF is reduced to a simple BAS.

support is basic since it is not defeasible. The effect-link of an attack (resp. a support) $\alpha = (a, b)$ is a basic attack (resp. support) from α to b . We encode a labelled interaction α with basic interactions since an attack (or a support) to α will be encoded by attacks (supports) to the meta-arguments that are introduced. So, a labelled attack $\alpha = (a, b)$ is encoded by $a \Rightarrow \alpha \rightarrow b$ and a labelled support $\beta = (c, d)$ is encoded by $c \Rightarrow \beta \Rightarrow d$. That is, a labelled interaction is encoded in two steps: first, a meta-argument is introduced with a basic support from its source (ground-link) and a basic attack or support to its target (effect-link); then, the basic supports are encoded in MAS (see Sect. 2). So, α (resp. β) is encoded in the MAS by $a \rightarrow N_{a\alpha} \rightarrow \alpha \rightarrow b$ (resp. $c \rightarrow N_{c\beta} \rightarrow \beta \rightarrow N_{\beta d} \rightarrow d$).

Encoding recursive interactions. Our aim is to encode an attack (resp. a support) on a labelled interaction through attacks (resp. supports) on the meta-arguments associated with it (the labels and the N_{ij} , see the previous paragraph). However, every meta-argument does not play the same role and a deeper analysis is needed in order to identify the meta-arguments that will be affected by the recursive interaction. Let $\alpha = (a, b)$ and $\beta = (c, d)$ be two labelled interactions. We discuss their encoding on two cases, with two sub-cases each, considering the intuitively desirable preferred extension of the MAS denoted by E . All these cases will be synthesized in Def. 8.

Case 1: α is an attack. Encoding α produces the meta-arguments α and $N_{a\alpha}$.

- *Case 1.1: α is attacked by β .* In this case, E should be $\{a, c, \beta, b\}$. So a and c are accepted, β is active, α is not active (it is grounded but not valid) and b can be accepted; this result holds whatever the status of a . Now, if c is attacked by d (with a basic attack) E should be $\{a, d, N_{c\beta}, \alpha\}$, which corresponds to the set $\{a, d, \alpha\}$ ⁵ after removing the meta-argument $N_{c\beta}$. Since c is not accepted β is not grounded nor active, and α can be valid. Also, since a is accepted α is grounded, thus active. So a and b cannot belong to the same extension.
- *Case 1.2: α is supported by β .* E should be $\{a, c, \beta, \alpha\}$. So a and c are accepted, β and α are active and thus, b cannot be accepted. Now, if c is attacked by d (with a basic attack), E should be $\{a, d, N_{c\beta}, N_{\beta\alpha}, b\}$, which corresponds to the set $\{a, d, b\}$ after removing the meta-arguments $N_{c\beta}$ and $N_{\beta\alpha}$. Since c is not accepted, β is not grounded. Furthermore, since β is valid, α is not valid nor active. Thus, a and b can belong to the same extension (whatever the status of a).

Case 2: α is a support. Encoding α produces the meta-arguments α , $N_{a\alpha}$ and N_{ab} .

- *Case 2.1: α is attacked by β .* If β is active, then α is not valid; this is captured by an attack from the meta-argument β to the meta-argument N_{ab} . Thus, α is not active, captured by an attack from β to α in MAS. If β is not active (e.g., if c is attacked) and β is the only interaction that impacts α , then α is valid; hence, b is accepted implies a is accepted. If a is attacked by e (with a basic attack), E should be $\{e, c, N_{a\alpha}, \beta, b\}$, which corresponds to the set $\{e, c, \beta, b\}$. Since c is accepted, β is grounded. Moreover, β is valid and so active. Therefore, α is not valid nor active and b (not being attacked) can be accepted even though a is not accepted. Note that the presence of $N_{a\alpha}$ in E means that α is not grounded. Moreover, if c is attacked by d (with a basic attack), E should be $\{e, d, N_{a\alpha}, N_{c\beta}, N_{ab}\}$, which corresponds to the set $\{e, d\}$. Now, c is not accepted; so β is not grounded nor active, and α is valid. However, since a is not accepted, α is not grounded nor active and b cannot be accepted. Lastly, if we remove the attack on a by e , E should be $\{a, d, \alpha, N_{c\beta}, b\}$, which corresponds to $\{a, d, \alpha, b\}$. c is not accepted, β is

⁵This set could be considered as the extension of the labelled ASAF; however, since this paper reports only a preliminary study, the expected outcomes of the framework following our approach are not yet defined.

not grounded nor active, and α is valid. Also, α is grounded (thus active), so a and b are accepted.

- **Case 2.2:** α is supported by β . If β is valid and not grounded, then α is not valid nor active; this is captured by attacks from $N_{\beta\alpha}$ to $N_{\alpha b}$ and α in MAS. If a is attacked by e and c by d (with basic attacks), E should be $\{e, d, N_{a\alpha}, N_{c\beta}, N_{\beta\alpha}, b\}$, which corresponds to the set $\{e, d, b\}$. Since c is not accepted, β is not grounded nor active. Then, α is not valid (nor active) and b can be accepted even though a is not accepted.

Def. 8 (Attacked or supported attacks/supports in MAS) *The following schemas describe the encoding of an attacked (resp. supported) attack/support in a MAS.*

Labelled ASAF	Associated BAS	Associated MAS
Case 1.1:		
Case 1.2:		
Case 2.1:		
Case 2.2:		

Given a labelled interaction $\alpha = (a, b)$ and an extension E , the following cases can occur:

support α	$\alpha \in E$:	α is active and in that case $N_{a\alpha} \notin E$ and $N_{\alpha b} \notin E$
	$\alpha \notin E$ and $N_{a\alpha} \notin E$:	α is grounded but not active, so it is not valid
	$\alpha \notin E$, $N_{a\alpha} \in E$ and $N_{\alpha b} \in E$:	α is not active, not grounded, but valid
	$\alpha \notin E$, $N_{a\alpha} \in E$ and $N_{\alpha b} \notin E$:	α is not active, not grounded and not valid
attack α	$\alpha \in E$:	α is active and in that case $N_{a\alpha} \notin E$
	$\alpha \notin E$ and $N_{a\alpha} \notin E$:	α is grounded but not active, so it is not valid
	$\alpha \notin E$ and $N_{a\alpha} \in E$:	α is not active nor grounded. The validity of α depends on supporters and attackers of α present in E

5. Comparison with ASAF

Let us compare the ASAF and MAS approaches for encoding labelled and recursive interactions. Both approaches follow two steps. The first step produces a BAS in both cases; however, they differ in the encoding of supports. Moreover, the second step is quite different. Let us first consider the encoding of a labelled attack.

Prop. 1 *Let $\alpha = (a, b)$ be a labelled attack. The translation of α given in Sect. 3 is exactly the same as the one given in Sect. 4: $a \xrightarrow{\alpha} b$ becomes $a \Rightarrow \alpha \rightarrow b$ where α denotes a meta-argument associated with the attack (a, b) .*

Let us now consider a labelled support $\alpha = (a, b)$. The first step of the ASAF approach (Sect. 3) produces $a \Rightarrow \alpha^+ \rightarrow \alpha^- \rightarrow b$, where two meta-arguments are used for representing the support α . In contrast, the first step of the MAS approach (Sect. 4)

produces $a \Rightarrow \alpha \Rightarrow b$, where only one meta-argument is created for representing α . However, when encoding $\alpha \Rightarrow b$, the second step of the MAS approach will produce $\alpha \Rightarrow N_{\alpha b} \Rightarrow b$. So α (resp. $N_{\alpha b}$) in the MAS plays the role of α^+ (resp. α^-) in the associated BAS of the ASAF. Indeed, they mainly differ in the encoding of the ground-link.

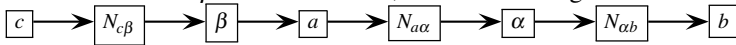
The second step produces different AS. The ASAF approach handles the remaining supports through extended attacks. Thus, since by Def. 6 no extended attacks are added, $a \Rightarrow \alpha^+ \Rightarrow \alpha^- \Rightarrow b$ is turned into: $\boxed{a} \rightarrow \boxed{\alpha^+} \rightarrow \boxed{\alpha^-} \rightarrow \boxed{b}$ and the preferred extension of the AS is $\{a, \alpha^+, b\}$, which corresponds to $\{a, \alpha, b\}$ in the ASAF. In contrast, the MAS approach handles both supports by creating meta-arguments (according to Def. 4). So $a \Rightarrow \alpha \Rightarrow b$ is turned into: $\boxed{a} \rightarrow \boxed{N_{a\alpha}} \rightarrow \boxed{\alpha} \rightarrow \boxed{N_{\alpha b}} \rightarrow \boxed{b}$ and the preferred extension of MAS (and BAS) is $\{a, \alpha, b\}$.

Note that, in the ASAF approach, the resulting AS has no connection between the source a and the meta-arguments associated with α . Differently, the MAS links a and α with a sequence of attacks going through the meta-argument $N_{a\alpha}$. This is because the ASAF approach does not treat all supports in the same way, whereas the MAS approach provides a unified handling through the addition of meta-arguments.

Let us now consider a labelled ASAF represented by $c \xrightarrow{\beta} a \Rightarrow \alpha \Rightarrow b$. On the one hand, following the ASAF approach, the associated BAS is $c \Rightarrow \beta \Rightarrow a \Rightarrow \alpha^+ \Rightarrow \alpha^- \Rightarrow b$. By Def. 6, there is an extended attack from β to α^+ . So the resulting AS is: $\boxed{c} \rightarrow \boxed{\beta} \rightarrow \boxed{a} \rightarrow \boxed{\alpha^+} \rightarrow \boxed{\alpha^-} \rightarrow \boxed{b}$ and the preferred extension of the AS is $\{c, \beta, \alpha^-\}$, which is mapped into $\{c, \beta, \alpha\}$ in the ASAF.

Since the ASAF approach deems all interactions as labelled, the resulting AS can be uselessly complex. To address this issue, the MAS approach offers two alternatives:

- If interactions are always labelled (even though they are not involved in a recursion) the associated BAS is $c \Rightarrow \beta \Rightarrow a \Rightarrow \alpha \Rightarrow b$, and the resulting AS is:



$\{c, \beta, N_{a\alpha}, N_{\alpha b}\}$ is the preferred extension of MAS, corresponding to $\{c, \beta\}$ in BAS.

- If labels are only used for reasoning about interactions involved in a recursion, we can directly apply Def. 4 to obtain a simpler system: $\boxed{c} \rightarrow \boxed{a} \rightarrow \boxed{N_{ab}} \rightarrow \boxed{b}$

Here, the preferred extension is $\{c, N_{ab}\}$, corresponding to the extension $\{c\}$ of the BAS.

A main difference between both approaches regards the presence of interactions in their extensions. Every interaction in an extension of a MAS is active (grounded and valid). In contrast, the meaning ascribed by the ASAF approach differs in the case of attacks and supports. As in the MAS approach, the presence of an attack in the extension obtained by the ASAF approach means that this attack is active. This is because the ASAF approach condenses the validity and groundness of an attack α through the meta-argument α . However, by combining these features, it does not allow to easily identify situations in which α is not grounded but valid (or vice-versa). On the other hand, if a support β belongs to an extension of the ASAF, we can only assure that it is valid. This is because β is represented by meta-arguments $\beta^{+/-}$ in the associated AS, which also capture the groundness (resp. non-groundness) of the support. As a result, the MAS approach is more flexible than the ASAF because of handling the different features of an interaction separately. Also, it ascribes the same meaning to the presence of every interaction in its extensions, in contrast with the ASAF.

6. Conclusion and future works

We have introduced a new framework for handling recursive interactions in bipolar argumentation systems, extending the work of [8]. Our study addresses the following issue: “How can the validity of an interaction be affected if this interaction is attacked or supported by another one?”. Drawing on examples, we identified different kinds of validity of interactions (namely the notions of “grounded interaction”, “valid interaction” and “active interaction”). Then, we proposed a new method for flattening an ASAF using meta-arguments. The comparison with the original approach of [8] highlights the similarities and differences between the frameworks, and confirms the choices given in [8].

Encodings of attacks and supports through meta-arguments can be found in [18] for the purpose of representing second-order attacks. Our work goes further since the MAS enables to represent attack and support to both attack and support relations, at any level.

Our study has been essentially carried out from examples. So it opens several lines for future work: give a formal proof of the intuitions behind the meaning of the meta-arguments; formally define the expected outcomes of our framework; and compare more deeply our proposal with the existing works, particularly in terms of outcomes.

References

- [1] L. Amgoud and C. Cayrol. A reasoning model based on the production of acceptable arguments. *Annals of Mathematics and Artificial Intelligence*, 34:197–216, 2002.
- [2] L. Amgoud, N. Maudet, and S. Parsons. Modelling dialogues using argumentation. In *Proc. of ICMAS*, pages 31–38, 2000.
- [3] P. Baroni, F. Cerutti, M. Giacomin, and G. Guida. AFRA: Argumentation framework with recursive attacks. *Intl. Journal of Approximate Reasoning*, 52:19–37, 2011.
- [4] C. Cayrol and M-C. Lagasque-Schiex. On the acceptability of arguments in bipolar argumentation frameworks. In *Proc. of ECSQARU*, pages 378–389, 2005.
- [5] C. Cayrol and M-C. Lagasque-Schiex. Coalitions of arguments: a tool for handling bipolar argumentation frameworks. *Intl. J. of Intelligent Systems*, 25:83–109, 2010.
- [6] C. Cayrol and M-C. Lagasque-Schiex. Bipolarity in argumentation graphs: towards a better understanding. *IJAR*, 54(7):876–899, 2013.
- [7] C. Cayrol and M-C. Lagasque-Schiex. An axiomatic approach to support in argumentation. In *Proc. of TAFA (LNAI 9524)*, pages 74–91, 2015.
- [8] A. Cohen, S. Gottifredi, A. J. García, and G. R. Simari. An approach to abstract argumentation with recursive attack and support. *J. Applied Logic*, 13(4):509–533, 2015.
- [9] P. M. Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77:321–357, 1995.
- [10] N. Karacapilidis and D. Papadias. Computer supported argumentation and collaborative decision making: the HERMES system. *Information systems*, 26(4):259–277, 2001.
- [11] S. Modgil. Reasoning about preferences in argumentation frameworks. *Artif. Intell.*, 173:901–934, 2009.
- [12] F. Nouioua. AFs with necessities: further semantics and labelling characterization. In *Proc. of SUM*, pages 120–133, 2013.
- [13] F. Nouioua and V. Risch. Argumentation frameworks with necessities. In *Proc. of SUM*, pages 163–176, 2011.
- [14] N. Oren, C. Reed, and M. Luck. Moving between argumentation frameworks. In *Proc. of COMMA*, pages 379–390. IOS Press, 2010.
- [15] S. Polberg and N. Oren. Revisiting support in abstract argumentation systems. In *Proc. of COMMA*, pages 369–376. IOS Press, 2014.
- [16] H. Prakken. On support relations in abstract argumentation as abstraction of inferential relations. In *Proc. of ECAI*, pages 735–740, 2014.
- [17] B. Verheij. Deflog: on the logical interpretation of prima facie justified assumptions. *Journal of Logic in Computation*, 13:319–346, 2003.
- [18] S. Villata, G. Boella, D. M. Gabbay, and L. van der Torre. Modelling defeasible and prioritized support in bipolar argumentation. *AMAI*, 66(1-4):163–197, 2012.

On the Effectiveness of Automated Configuration in Abstract Argumentation Reasoning

Federico CERUTTI ^{a,1}, Mauro VALLATI ^b and Massimiliano GIACOMIN ^c

^a *Cardiff University, UK*

^b *University of Huddersfield, UK*

^c *University of Brescia, Italy*

Abstract. In this paper we investigate the impact of automated configuration techniques on the ArgSemSAT solver—runner-up of the ICCMA 2015—for solving the enumeration of preferred extensions. Moreover, we introduce a fully automated method for varying how argumentation frameworks are represented in the input file, and evaluate how the joint configuration of frameworks and ArgSemSAT parameters can have a remarkable impact on performance. Our findings suggest that automated configuration techniques lead to improved performances in argumentation solvers, an important message for participants to the forthcoming competition.

Keywords. Algorithm Configuration, Argumentation Framework Configuration, Abstract Argumentation

1. Introduction

Dung’s theory of abstract argumentation [7] is a unifying framework able to encompass a large variety of specific formalisms in the areas of nonmonotonic reasoning, logic programming and computational argumentation. It is based on the notion of argumentation framework (AF), that consists of a set of arguments and an *attack* relation between them. Different *argumentation semantics* introduce in a declarative way the criteria to determine which arguments emerge as “justified” from the conflict, by identifying a number of *extensions*, i.e. sets of arguments that can “survive the conflict together” [4].

The first *International Competition on Computational Models of Argumentation* (IC-CMA2015) determined the state-of-the-art of the current implementations for addressing the above problems with respect to the three aforementioned semantics (plus the complete extensions) [14]. In this paper we will focus on ArgSemSAT [6], that scored overall second during ICMMA2015—at one single Borda count point from the winner—despite an implementation bug discovered after the competition.

ArgSemSAT is a rather configurable solver: it allows to select different ways for encoding abstract argumentation problems in SAT, and it is able to exploit external SAT

¹Corresponding Author: Federico Cerutti, Cardiff University, School of Computer Science & Informatics, CF24 3AA, Cardiff, UK; E-mail: CeruttiF@cardiff.ac.uk.

solvers. We manually tuned its parameter before submitting it to ICCMA2015; however, the question naturally arises: *is it possible to improve the chosen configuration?*

We investigated whether automatic configuration systems [12,1,18] can address such a question. In this work we exploit the sequential model-based algorithm configuration method SMAC [11], which represents the state of the art of configuration tools. SMAC uses predictive models of algorithm performance [13] to guide its search for good—according to a chosen metric—configurations.

Surprisingly, we also proved that the way *AFs* are described (for instance, the order in which arguments are listed) can have an effect on the overall performance. This is a remarkable finding that has been proved only (and very recently) in classical planning [16], and here for the second time. This is once again an important element that future organisers of competitions should be aware of and take into serious consideration.

Finally, for the first time—to our knowledge—we are in the position to prove that there is also a significant synergy between solvers' parameter configuration on the one side and knowledge representation (how *AFs* are described) on the other side, leading to increased performance.

Although due to space constraints we report our investigation w.r.t. ArgSemSAT only and the problem of enumeration of preferred extensions, those results can be generalised to other solvers and other semantics and problems.

Let us recall that an argumentation framework [7] consists of a set of arguments and a binary attack relation between them² and that preferred extensions are maximal admissible sets.

Definition 1. An argumentation framework (*AF*) is a pair $\Gamma = \langle \mathcal{A}, \mathcal{R} \rangle$ where \mathcal{A} is a set of arguments and $\mathcal{R} \subseteq \mathcal{A} \times \mathcal{A}$. We say that \mathbf{b} attacks \mathbf{a} , or $\mathbf{b} \rightarrow \mathbf{a}$, iff $\langle \mathbf{b}, \mathbf{a} \rangle \in \mathcal{R}$.

Given an *AF* $\Gamma = \langle \mathcal{A}, \mathcal{R} \rangle$:

- a set $S \subseteq \mathcal{A}$ is a conflict-free set of Γ if $\nexists \mathbf{a}, \mathbf{b} \in S$ s.t. $\mathbf{a} \rightarrow \mathbf{b}$;
- an argument $\mathbf{a} \in \mathcal{A}$ is acceptable with respect to a set $S \subseteq \mathcal{A}$ of Γ if $\forall \mathbf{b} \in \mathcal{A}$ s.t. $\mathbf{b} \rightarrow \mathbf{a}$, $\exists \mathbf{c} \in S$ s.t. $\mathbf{c} \rightarrow \mathbf{b}$;
- a set $S \subseteq \mathcal{A}$ is an admissible set of Γ if S is a conflict-free set of Γ and every element of S is acceptable with respect to S of Γ ;
- a set $S \subseteq \mathcal{A}$ is a preferred extension of Γ iff S is a maximal (w.r.t. set inclusion) admissible set of Γ .

2. Automated Configuration

The description of an abstract argumentation framework can be synthesised by listing all the arguments and all the attacks of the framework. Currently, three main formats for describing frameworks are used: Trivial Graph Format, Aspartix Format and the CNF Format. Here we focus on the most used one, the Aspartix Format [8].

Since this *configuration* of the input file should be performed *online* to lead to improvements of the overall system, we are interested only in information about the *AF* that can be quickly obtained. In particular, we considered the possibility to list arguments ordered according to the following five criteria: (1) the number of attacks received; (2) the

²In this paper we consider only *finite* sets of arguments: see [5] for a discussion on infinite sets of arguments.

number of attacks to other arguments; (3) the presence of self-attacks; (4) the difference between the number of received attacks and the number of attacks to other arguments; and (5) being an argument in a mutual attack. For each of the five mentioned criteria, arguments can be listed following a direct or inverse order.

To order the list of attacks, these five criteria can be applied either to the attacking or to the attacked argument. The choice of the criteria for ordering the list of arguments is independent from the choice of criteria for ordering the list of attacks.

There are different ways for encoding the degrees of freedom in *AF*s descriptions as parameters, mainly because orders are not natively supported by general configuration techniques. Following [16], we generate 10 continuous parameters, which correspond to the aforementioned possible orderings of arguments and attacks in frameworks. An additional categorical selector among 5 alternatives allows to decide how to apply the criteria for ordering the list of attacks, i.e. whether on the first or the second argument, and following same or inverse ordering of arguments.

Each continuous parameter has associated a real value in the interval $[-1.0, +1.0]$ which represents (in absolute value) the *weight* or *precedence* given to an ordering criterion: the criterion corresponding to the parameter with the highest absolute value, is considered first in the ordering. Ties of such ordering are then broken by referring to the criterion associated to the next parameter of high absolute value. Negative values indicate that inverse ordering is used. In the case of two criteria having exactly the same absolute value, they are applied according to their alphabetical ordering. Thus, the configuration space is $\mathcal{C} = [-1.0, +1.0]^{10} \cdot 5$, where 5 are the possible values of the categorical parameter describing the order of the list of attacks.

In order to automatically re-order an argumentation framework according to the specified configuration, we developed a wrapper in Python. On the *AF*s considered in our experimental analysis, composed by hundreds of arguments and few hundreds of thousands of attacks, the re-ordering of the Aspartix format description takes less than 1 CPU-time second.

Joint AF-Solver Configuration As a case study for investigating the synergies of re-ordering a given argumentation framework, and of selecting the most appropriate solver's parameters, we consider ArgSemSAT [6], which is the runner-up of ICCMA 2015. On the one hand, ArgSemSAT exposes a single—critical—parameter which allows to select the encoding for translating the problem of identifying a complete extension into a SAT formula, with remarkable impact on size and structure of the generated CNFs, and on the CPU-time required to enumerate all the preferred extensions. On the other hand, ArgSemSAT allows the use of an external SAT solver, to be used as an NP-oracle. In this work we exploit the Glucose SAT solver [2]: it shows very good performance in recent SAT competitions, and has a large number of parameters that can be tuned and controlled for modifying its behaviour, from decay level of variables and clauses, to the number of restarts. Configuring ArgSemSAT together with Glucose requires to tune 20 parameters (2 categorical and 18 continuous).

In order to maximise the impact of automated configuration on solvers' performance and thus exploiting unforeseen synergies between solver behaviour and specific knowledge descriptions, we use SMAC for configuring at the same time the *AF*s description

and the configuration of ArgSemSAT. The total number of configurable parameters is 31: 3 categorical and 28 continuous.³

SMAC [11] is an *anytime algorithm* (or *interruptible algorithm*) that interleaves the exploration of new configurations with additional runs of the current best configuration to yield both better and more confident results over time. As all anytime algorithms, SMAC improves performance over time, and for finite configuration spaces it is guaranteed to converge to the optimal configuration in the limit of infinite time.

3. Experimental Analysis

Settings. As described in the previous section, in this work we consider ArgSemSAT using Glucose as SAT solver [2] for enumerating preferred extensions. In total, 31 parameters are exposed. Three of them are categorical, while the others are continuous.

We randomly generated 8,000 *AFs*, divided into 4 sets of 2,000 *AFs* each. Three of such sets include only framework based on different graph models: Barabasi-Albert [3], Erdős-Rényi [9] and Watts-Strogatz [17]. The fourth set (“General”) includes mixed-structured *AFs* generated by considering graphs of all the mentioned models.

To identify challenging frameworks *AFs* we followed the protocol suggested in [15] which leads to the selection of *AFs* with a number of arguments between 250 and 650, and number of attacks between (approximately) 400 and 180,000.

Each set of *AFs* has been split into a training set (1,800 *AFs*) and a testing set (200 *AFs*) in order to obtain an unbiased estimate of generalisation performance to previously unseen *AFs* from the same distribution.

Configuration was done using SMAC version 2.10. The performance metric we optimised is the Penalized Average Runtime (PAR), counting runs that crash or do not find a solution as ten times the cutoff time (PAR10).

Experiments were performed on Dual Xeon X5660-2.80GHz with 48GB DDR3 RAM. Each configuration run was limited to a single core, and was given an overall runtime of 5 days and 4 GB of RAM, for ensuring re-usability of results also on less equipped machines. The cutoff time was 500 seconds.

In the following, also the IPC score is used for comparing different configurations performance. For a solver \mathcal{C} and a problem p , $Score(\mathcal{C}, p)$ is 0 if p is unsolved, and $1/(1 + \log_{10}(T_p(\mathcal{C})/T_p^*))$ otherwise (where T_p^* is the minimum amount of time required by any compared system to solve the enumeration problem). The IPC score on a set of instances is given by the sum of the scores achieved on each considered instance.

Results. Table 1 compares the performance of ArgSemSAT using the default configuration, and the specific joint configuration of *AFs* description and ArgSemSAT, obtained by running SMAC. Remarkably, the joint configuration of *AF* description and ArgSemSAT leads to a general performance improvement. In particular, on the Barabasi-Albert and General sets the performance of the configured system are statistically significantly better than the performance achieved by using the default configuration, according to the Wilcoxon test. The significant performance improvement achieved on the General set is of particular interest: it indicates that it is possible to identify a configuration able to

³The interested reader can find the full list of parameters, including default value and valid value range, at <https://helios.hud.ac.uk/scommv/afconf/PARAMS.TXT>.

Table 1. Comparison between the default and tuned configuration, in terms of IPC score, PAR10, and percentage of instances on which a configuration has been the fastest, on the considered *AFs* test sets for enumerating preferred extensions. In bold the best results.

Set	Configuration	IPC Score	PAR10	Fastest
Barabasi-Albert	Default	78.0	1921.0	2.5
	Configured	125.2	1863.1	60.5
Erdős-Rényi	Default	56.8	3426.5	16.5
	Configured	60.4	3329.2	18.0
Watts-Strogatz	Default	116.6	1967.3	28.0
	Configured	118.1	1967.9	23.5
General	Default	110.0	1665.4	11.0
	Configured	143.0	1376.8	62.5

improve the performance across differently-structured graphs. In other words, this is an indication that the default configuration can be improved.

Conversely, the configuration process does not significantly improve the default performance on the Watts-Strogatz set. According to the Wilcoxon test, performance of default and tuned configurations are statistically undistinguishable even though IPC score show slight improvements. This is possibly due to the fact that the default configuration is already showing very good performance. In that scenario, it may be the case that only small portions of the configuration space lead to a significant performance improvement over the default configuration. Given the limited CPU-time made available to the configuration process, SMAC did not identify such portions of the vast configuration space.

Finally, the results on the Erdős-Rényi set deserves a more detailed discussion. On the one hand, the Wilcoxon test indicates that there is not a statistically significant performance improvement. On the other hand, *AFs* from the Erdős-Rényi set are extremely hard for ArgSemSAT, as testified by the PAR10 values. Moreover, those that can be solved are usually solved quickly, i.e. in few CPU-time seconds. This makes the evaluation of configurations' performance, and the exploration of the space of configurations, hard and slow. Despite such issues, SMAC was able to identify a configuration that is able to improve the performance both in terms of runtime (better IPC score) and PAR10. Overall, this is a remarkable result, that shows the ability of automated configuration in improving performance also in unfavourable cases.

To provide a better overview of the impact of different configurations, we ran all the configurations obtained by SMAC from training sets with different graph models on all the considered test sets. Table 2 shows the results of this comparison. Performance of different configurations tend to be similar but for the parameters' configuration derived from Barabasi-Albert training *AFs*. This possibly indicates that there are some parameters' values that can boost performance on differently-structured *AFs*. Remarkably, the configuration identified by training on the General set, is usually able to obtain performance that are close to those of the specific configuration, on each considered test set. Again, this supports the hypothesis that there are some parameters' values that can help improving the general performance. On the contrary, Table 2 also indicates that the configuration derived from Barabasi-Albert training does not generalise well on differently-structured *AFs*. Such behaviour is possibly due to some parameter's value that helps increasing the performance on the Barabasi-Albert test set, but has a detrimental effect on different graph structures. For instance, among considered structures, Barabasi-Albert is

Table 2. Performance of each configuration generated by SMAC (rows) running on the different test sets (columns) for enumerating preferred extensions. IPC Score is evaluated by considering all the four configurations on each single test set. In bold the best results, with respect to each specific test set.

Training sets	Test sets			
	Barabasi-Albert	Erdős-Rényi	Watts-Strogatz	General
Barabasi-Albert	119.2	6.9	34.5	42.8
Erdős-Rényi	92.3	58.6	105.3	125.7
Watts-Strogatz	116.2	52.6	115.6	129.2
General	87.5	57.6	113.5	133.2

Table 3. Most important single parameters (configured value) for SMAC runs on the considered *AF* sets. F-,S- and G- stand for, respectively, Framework, ArgSemSAT and Glucose parameters.

Set	1st	2nd	3rd
Barabasi-Albert	S-ExtEnc (011111)	G-firstReduceDB (1528)	G-cla-decay (0.32)
Erdős-Rényi	F-autoFirst (-1.00)	G-rnd-freq (0.00)	G-K (0.26)
Watts-Strogatz	S-ExtEnc (101010)	G-Grow (0)	G-rnd-freq (0.08)
General	S-ExtEnc (101010)	G-R (2.09)	G-cla-decay (0.99)

the only set of *AF*s with a large number of preferred extensions (up to some thousands) per *AF*.

Discussion. In order to shed some light on the usefulness of algorithm and *AF* tuning, we used fANOVA [10], a recently-released tool for assessing parameter importance after each configuration. fANOVA exploits predictive models of the performance of each configuration for assessing the importance of each parameter, regardless of the value of the others, and the interaction between parameters' values. Table 3 shows the three most important parameters for each configuration. Unsurprisingly, the encoding used by ArgSemSAT for generating the SAT formulae is usually the most important parameter. Its default value (i.e. 101010), is proven to be the best choice for *AF*s belonging to Watts-Strogatz and General sets, but not for *AF*s in the Barabasi-Albert set. After that, the parameters that control the behaviour of Glucose are those with the highest impact on performance, notably: decay value of clauses and size of the DB of learnt clauses are among the aspects with a strongest impact on the performance of ArgSemSAT.

One parameter used for controlling the *AF* description has a significant impact on performance on *AF*s belonging to the Erdős-Rényi set, according to the fANOVA tool. In this case, the order in which arguments are listed is important and, in particular, it is required that self-attacking arguments are provided at the very end.

However, the interaction of parameters controlling the shape of *AF*s with reasoning-related parameters do have a remarkable impact, i.e. the best performance depends on two or more parameters. Parameters used for controlling the order of arguments have strong interactions with the parameter that controls the encoding of ArgSemSAT, as well as with parameters of Glucose controlling the number and type of clauses learnt. Figure 1 (coloured) shows the average PAR10 performance of ArgSemSAT on the Barabasi-Albert set as a function of two interacting parameters. `args_eachOther` is used for listing earlier in the *AF* description arguments that are attacking each other, the other parameter is used for controlling the number of Glucose learnt clauses, according to their

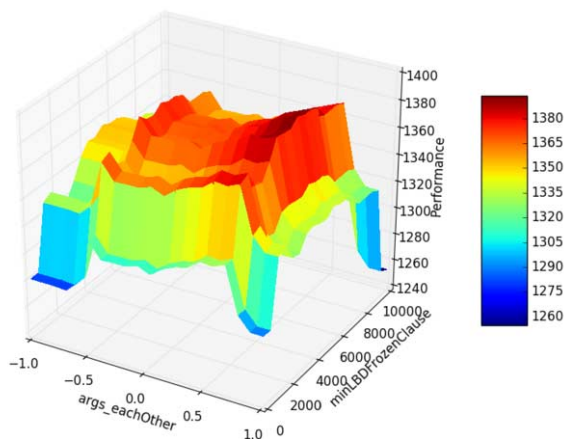


Figure 1. (Coloured) The average PAR10 performance of ArgSemSAT on the Barabasi-Albert set, as a function of the ordering of arguments according to the fact that they attack each other (`args_eachOther`) and of number of clauses stored by Glucose during the search. Lower PAR10 values correspond to better performance.

heuristic LBD evaluation. In order to achieve better performance in terms of PAR10, arguments attacking each other should be listed very late (-1.0 value of the parameter) or very early ($+1.0$ value) and either few or many clauses should be kept (respectively, low and high value of Glucose parameter).

Parameters that control the order in which arguments are listed tend to have a stronger impact on overall performance—either singularly (Table 3) or as a result of their interaction with other parameters (Fig. 1)—than parameters controlling the order in which attacks are listed. At a first sight, this may be seen as counter-intuitive, since the number of attacks in a typical benchmark *AF* is significantly higher than the number of arguments. However, this difference can be due to the data structure used by ArgSemSAT. The set of arguments of the *AF* is stored in a list which is populated according to the order in which the arguments are listed in the input file. Each argument has then an associated data structure with pointers to two other lists of arguments: one for the attacked arguments; and one for the arguments that attack it. Then the list representing the set of arguments of the *AF* is navigated several times when creating CNFs to be evaluated by the SAT solver: these results suggest that not only the encoding of complete labellings in CNF, but also the order of clauses have a remarkable impact on the performance.

4. Conclusion and Future Work

In this paper we proposed an approach for the joint automatic configuration of *AF* descriptions and argumentation solvers. Specifically, we designed a method to automatically order the list of arguments and the list of attacks in argumentation frameworks by tuning 11 parameters, using as a test-case the widely used Aspartix format. We focused our investigation on ArgSemSAT—runner-up of the ICCMA2015—using Glucose as a SAT solver: they export together a further set of 20 parameters.

As described in the previous sections: (i) we demonstrate that joint *AF*-solver configuration has a statistically significant impact on the performance of ArgSemSAT; (ii)

we demonstrate the synergies between *AF*s configuration and SAT solvers behaviour; and (iii) we open new, exciting possibilities in the area of learning for improving performance of abstract argumentation solvers. We believe this work would be particularly beneficial for the participants of the forthcoming competition ICCMA2017.

We see several avenues for future work. We plan to evaluate the proposed joint *AF*-solver configuration approach on different solvers and on different problems and on different semantics. Moreover, we are interested in exploiting the configuration approach for combining different argumentation and SAT solvers into portfolios. Finally, we are considering investigating the presence of *AF* configurations that are able to improve—on average—the performance of all the existing state-of-the-art argumentation solvers: this would provide powerful guidelines for the encoding of frameworks.

Acknowledgement

This work was performed using the computational facilities of the Advanced Research Computing @ Cardiff (ARCCA) Division, Cardiff University.

References

- [1] C. Ansótegui, M. Sellmann, and K. Tierney. A gender-based genetic algorithm for the automatic configuration of algorithms. In *Proc. of CP*, pages 142–157, 2009.
- [2] G. Audemard and L. Simon. Lazy clause exchange policy for parallel sat solvers. In *SAT 2014*, pages 197–205, 2014.
- [3] A. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):11, 1999.
- [4] P. Baroni, M. Caminada, and M. Giacomin. An introduction to argumentation semantics. *Knowledge Engineering Review*, 26(4):365–410, 2011.
- [5] P. Baroni, F. Cerutti, P. E. Dunne, and M. Giacomin. Automata for Infinite Argumentation Structures. *Artif. Intell.*, 203(0):104–150, 2013.
- [6] F. Cerutti, M. Vallati, and M. Giacomin. Argsemsat-1.0: Exploiting sat solvers in abstract argumentation. *System Descriptions of the First International Competition on Computational Models of Argumentation (ICCMA'15)*, page 4, 2015.
- [7] P. M. Dung. On the Acceptability of Arguments and Its Fundamental Role in Nonmonotonic Reasoning, Logic Programming, and n-Person Games. *Artif. Intell.*, 77(2):321–357, 1995.
- [8] U. Egly, S. A. Gaggl, and S. Woltran. Aspartix: Implementing argumentation frameworks using answer-set programming. In *Logic Programming*, pages 734–738, 2008.
- [9] P. Erdős and A. Rényi. On random graphs. I. *Publ. Math. Debrecen*, 6:290–297, 1959.
- [10] F. Hutter, H. Hoos, and K. Leyton-Brown. An efficient approach for assessing hyperparameter importance. In *Proc. of ICML*, pages 754–762, 2014.
- [11] F. Hutter, H. H. Hoos, and K. Leyton-Brown. Sequential model-based optimization for general algorithm configuration. In *Proc. of LION*, pages 507–523, 2011.
- [12] F. Hutter, H. H. Hoos, K. Leyton-Brown, and T. Stützle. Paramils: An automatic algorithm configuration framework. *J. Artif. Intell. Res.*, 36:267–306, 2009.
- [13] F. Hutter, L. Xu, H. H. Hoos, and K. Leyton-Brown. Algorithm runtime prediction: Methods & evaluation. *Artif. Intell.*, 206:79–111, 2014.
- [14] M. Thimm, S. Villata, F. Cerutti, N. Oren, H. Strass, and M. Vallati. Summary report of the first international competition on computational models of argumentation. *AI Magazine*, 2016.
- [15] M. Vallati, L. Chrapa, M. Grzes, T. L. McCluskey, M. Roberts, and S. Sanner. The 2014 international planning competition: Progress and trends. *AI Magazine*, 2015.
- [16] M. Vallati, F. Hutter, L. Chrapa, and T. L. McCluskey. On the effective configuration of planning domain models. In *Proc. of IJCAI*, 2015.
- [17] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, 1998.
- [18] Z. Yuan, T. Stützle, and M. Birattari. Mads/f-race: Mesh adaptive direct search meets f-race. In *Proc. of IEA/AIE*, pages 41–50, 2010.

Where Are We Now? State of the Art and Future Trends of Solvers for Hard Argumentation Problems

Federico CERUTTI ^{a,1}, Mauro VALLATI ^b and Massimiliano GIACOMIN ^c

^aCardiff University, UK

^bUniversity of Huddersfield, UK

^cUniversity of Brescia, Italy

Abstract. We evaluate the state of the art of solvers for hard argumentation problems—the enumeration of preferred and stable extensions—to envisage future trends based on evidence collected as part of an extensive empirical evaluation. In the last international competition on computational models of argumentation a general impression was that reduction-based systems (either SAT-based or ASP-based) are the most efficient.

Our investigation shows that this impression is not true in full generality and suggests the areas where the relatively under-developed non reduction-based systems should focus more to improve their performance. Moreover, it also highlights that the state-of-the-art solvers are very complementary and can be successfully combined in portfolios: our best per-instance portfolio is 51% (resp. 53%) faster than the best single solver for enumerating preferred (resp. stable) extensions.

Keywords. Abstract Argumentation, Solvers for Argumentation Problems, Portfolios methods for Argumentation

1. Introduction

An abstract argumentation framework (AF) consists of a set of arguments and a binary *attack* relation between them. In [9] four semantics were introduced, namely *grounded*, *preferred*, *complete*, and *stable* semantics: each of them lead to a single or to multiple *extensions* (or no extensions in the case of stable semantics) where an *extension* is intuitively a set of arguments which can “survive the conflict together.” We refer the reader to [2] for a detailed analysis. Moreover, for each semantics, several *decision* and *enumeration* problems have been identified. In this paper we focus on the enumeration of preferred and stable extensions because: (i) the solution to the problem of enumerating extensions implies the answer to other problems; (ii) the problems of enumerating preferred and stable extensions are among the hardest in abstract argumentation.

Research around argumentation-based technology is fast growing: for instance, three of the most cited papers (top-25) published on Artificial Intelligence Journal since 2011

¹Corresponding Author: Federico Cerutti, Cardiff University, School of Computer Science & Informatics, CF24 3AA, Cardiff, UK; E-mail: CeruttiF@cardiff.ac.uk.

according to Scopus² are in this field, and the last International Competition on Computational Models of Argumentation (ICCMA-15) received more submissions than the last ASP competition.

The results of ICCMA-15³ [17] suggest that (i) reduction-based systems (either SAT-based or ASP-based) are more efficient than non reduction-based: indeed the best solvers for enumerating stable and preferred extensions are either SAT-based or ASP-based; and (ii) a mixture of approaches can be fruitful: **CoQuiAas**—that scored first among all for each semantics considered in ICCMA-15—uses a variety of approaches.

Here, we test how general such conclusions are with a large empirical investigation focused on enumeration of stable and preferred extensions using the solvers submitted to ICCMA-15. By adopting different metrics, we identified avenues for improvement that we hope will be valuable for solvers' authors and for the argumentation community.

Solvers indeed proved to be very complementary (i.e. a mixture of approaches can be fruitful), and we then exploit portfolio approaches in order to highlight (relative) strengths and weaknesses of solvers. As testified by experiences in other research areas in artificial intelligence, such as planning [19], SAT [21], and ASP [12], portfolios and algorithm selection techniques [14] are very useful tools for understanding the importance of solvers, evaluate the improvements, and effectively combine solvers for increasing overall performance. Existing works [6,5] either focus on algorithm selection for enumerating preferred extensions, with a very small number of solvers and of instances; or on theoretical complementariness of algorithms.

Our findings reshape one of the take-away messages from ICCMA-15, namely that reduction-based systems have higher performance than non reduction-based. This is not always the case, although it is the case that they have better coverage, and ICCMA-15 privileged coverage against speed.

Finally, the analysis of portfolio techniques—and their generalisation capabilities—highlighted that, by combining solvers, it is possible to increase the coverage of 13% (resp. 3%) and the speed of 51% (resp. 53%) against the best single solver for enumerating preferred (resp. stable) extensions.

2. Dung's Argumentation Framework

An argumentation framework [9] consists of a set of arguments and a binary attack relation between them.⁴

Definition 1. An argumentation framework (AF) is a pair $\Gamma = \langle \mathcal{A}, \mathcal{R} \rangle$ where \mathcal{A} is a set of arguments and $\mathcal{R} \subseteq \mathcal{A} \times \mathcal{A}$. We say that **b** attacks **a** iff $\langle \mathbf{b}, \mathbf{a} \rangle \in \mathcal{R}$, also denoted as $\mathbf{b} \rightarrow \mathbf{a}$.

The basic properties of conflict-freeness, acceptability, and admissibility of a set of arguments are fundamental for the definition of argumentation semantics.

Definition 2. Given an AF $\Gamma = \langle \mathcal{A}, \mathcal{R} \rangle$:

²<http://www.journals.elsevier.com/artificial-intelligence/most-cited-articles>, accessed on 10th June 2016.

³<http://argumentationcompetition.org/2015/results.html>

⁴In this paper we consider only *finite* sets of arguments: see [3] for a discussion on infinite sets of arguments.

- a set $S \subseteq \mathcal{A}$ is a conflict-free set of Γ if $\nexists a, b \in S$ s.t. $a \rightarrow b$;
- an argument $a \in \mathcal{A}$ is acceptable with respect to a set $S \subseteq \mathcal{A}$ of Γ if $\forall b \in \mathcal{A}$ s.t. $b \rightarrow a$, $\exists c \in S$ s.t. $c \rightarrow b$;
- a set $S \subseteq \mathcal{A}$ is an admissible set of Γ if S is a conflict-free set of Γ and every element of S is acceptable with respect to S of Γ .

An argumentation semantics σ prescribes for any AF Γ a set of *extensions*, namely a set of sets of arguments satisfying the conditions dictated by σ .

Definition 3. Given an AF $\Gamma = \langle \mathcal{A}, \mathcal{R} \rangle$: a set $S \subseteq \mathcal{A}$ is a:

- preferred extension of Γ iff S is a maximal (w.r.t. set inclusion) admissible set of Γ ;
- stable extension of Γ iff S is a conflict-free set of Γ and $\mathcal{A} \setminus S = \{a \in \mathcal{A} \mid b \rightarrow a \text{ and } b \in S\}$.

3. Generation of Portfolios

In this section we describe the techniques we used for combining solvers into sequential portfolios. Every approach requires as input a set of solvers, a set of training AFs, and measures of performance of solvers on the training set. Solvers' performance are measured in terms of Penalised Average Runtime (PAR) score. This metric trades off coverage and runtime for successfully analysed AFs: runs that do not solve the given problem get ten times the cutoff time (PAR10), other runs get the actual runtime. The PAR10 score of a solver on a set of AFs is the average of the associated scores. Although PAR10 largely emphasises the coverage, it also gives a clear indication on effective performance, thus resulting in an interesting and useful measure. This is also compatible with the ICCMA experience: ties on coverage are automatically solved on the basis of performance.

3.1. Static Portfolios

Static portfolios—as the name suggests—are generated once, according to the performance of the considered solvers on training instances, and never adjusted. Static portfolios are defined by: (i) the selected solvers; (ii) the order in which solvers will be run, and (iii) the runtime allocated to each solver.

We considered two different approaches for configuring static portfolios. First, we generated static portfolios of exactly k components, *Shared- k* . Each component solver has been allocated the same amount of CPU-time, equal to $\text{maxRuntime}/k$ seconds. Solvers are selected and ordered according to overall PAR10 score achieved by the resulting portfolio. We considered values of k between 2 and 5. In fact, $k = 1$ would be equivalent to select the single solver with the best PAR10 score on training instances, which is not relevant for our investigation. For $k > 5$, the CPU-time assigned to each solver tends to be too short hence drastically reducing portfolio performance.

For our second static portfolio approach, named *FDSS*, we adapted the Fast Downward Stone Soup technique [15]. We start from an empty portfolio, and iteratively add

either a new solver component, or extend the allocated CPU-time⁵ of a solver already added to the portfolio, depending on what maximises the increment of the PAR10 score of the portfolio. We continue until the time limit of the portfolio has been reached, or it is not possible to further improve the PAR10 score of the portfolio on the training instances.

3.2. Per-instance Portfolios

Per-instance portfolios rely on instance features for configuring an instance-specific portfolio. For each *AF* a vector of features is computed; each feature is a real number that summarises a potentially important aspect of the considered *AF*. Similar instances should have similar feature vectors, and, on this basis, portfolios are configured using empirical performance models [13].

In this investigation we consider the largest set of features available for *AFs* [6]. Such set includes 50 features, extracted by exploiting the representation of *AFs* both as directed (loss-less) or undirected (lossy) graphs. Features are extracted by considering aspects such as the size of graphs, the presence of connected components, the presence of auto-loops, etc. The features extraction process is usually quick (less than 2 CPU-time seconds on average) and is done by exploiting a wrapper written in Python.

3.2.1. Classification-based approach

The classification-based (hereinafter *Classify*) approach exploits the technique introduced in [6]. It trains a random decision forest classification model to perform algorithm selection. It classifies a given *AF* into a single category which corresponds to the single solver predicted to be the fastest. The difference between solvers' performance is ignored: all the available CPU-time is then allocated to the selected solver.

3.2.2. Regression-based approaches

For regression-based approaches, deciding which solver to execute and its runtime depends on the empirical hardness models learned from the available training data, in particular a M5-Rules [11] model generated for each solver. When executed on a fresh *AF*, the predictive model estimates the CPU-time required by each solver to successfully terminate.

We exploit the regression-based model in two different ways. First, for performing algorithm selection (hereinafter *1-Regression*): given the predicted runtime of each solver, the solver predicted to be the fastest is selected and it has allocated all the available CPU-time. However, such use of the models do not fully exploit the available predicted runtimes. Therefore, we designed a different way for using the regression-based approach, referred to as *M-regression*. As in 1-Regression, we initially select the solver predicted to be the fastest, but we allocate only its predicted CPU-time (increased by 10%). If the selected solver is not able to successfully analyse the given *AF* in the allocated time, it is stopped and no longer available to be selected, and the process iterates by selecting a different solver. The M-regression approach stops when either a solver has successfully analysed the *AF*, or the runtime budget has been exhausted.

With regards to existing well-known portfolio-based solver approaches, it is worthy to remark that SATZilla [21] is a regression-based approach similar to the 1-regression

⁵A granularity of 5 CPU-time seconds is considered.

we introduced. However, since it was developed for competition purposes, SATZilla also exploits pre and backup solvers. These are undoubtedly useful for improving coverage, but not when the main point is to evaluate to which extent solvers composition/selection can improve results, as in our investigation.

4. Experimental Analysis of ICCMA-15 Solvers

We randomly generated 2,000 *AFs* based on four different graph models: Barabasi-Albert [1], Erdős-Rényi [10], Watts-Strogatz [20] and graphs featuring a large number of stable extensions (hereinafter *StableM*).

Erdős-Rényi graphs [10] are generated by randomly selecting attacks between arguments according to a uniform distribution. While Erdős-Rényi was the predominant model used for randomly generated experiments, [4] investigated also other graph structures such as *scale-free* and *small-world* networks. As discussed by Barabasi and Albert [1], a common property of many large networks is that the node connectivities follow a *scale-free* power-law distribution. This is generally the case when: (i) networks expand continuously by the addition of new nodes, and (ii) new nodes attach preferentially to sites that are already well connected. Moreover, Watts and Strogatz [20] show that many biological, technological and social networks are neither completely regular nor completely random, but something in the between. They thus explored simple models of networks that can be tuned through this middle ground: regular networks *rewired* to introduce increasing amounts of disorder. These systems can be highly clustered, like regular lattices, yet have small characteristic path lengths, like random graphs, and they are named *small-world* networks by analogy with the small-world phenomenon. The *AFs* have been generated by using an improved version of AFBenchGen [7]. It is worthy to emphasise that Watts-Strogatz and Barabasi-Albert produce undirected graphs: in this work, differently from [4], each edge of the undirected graph is then associated with a direction following a probability distribution, that can be provided as input to AFBenchGen. Finally, the fourth set has been generated using the code provided in Probo [8] by the organisers of ICCMA-15.⁶

In order to identify challenging frameworks—i.e., neither trivial nor too complex to be successfully analysed in the given CPU-time—*AFs* for each set have been selected using the protocol introduced in the 2014 edition of the International Planning Competition [18]. This protocol lead to the selection of *AFs* with a number of arguments between 250 and 650, and number of attacks between (approximately) 400 and 180,000.

The set of *AFs* has been divided into training and testing sets. For each graph model, we randomly selected 200 *AFs* for training, and the remaining 300 for testing. Therefore, out of the 2,000 *AFs* generated, 800 have been used for training purposes, while the remaining 1,200 have been used for testing and comparing the performance of trained approaches.

We considered all the solvers that took part in the EE-PR and EE-ST tracks of ICCMA-15 [17], respectively 15 and 11 systems. For the sake of clarity and conciseness, we removed from the analysis single solvers that did not successfully analyse at least one *AF* or which were always outperformed by another solver. The interested reader

⁶<http://argumentationcompetition.org/2015/results.html>

Table 1. PAR10 score and coverage (cov.)—percentage of *AFs* successfully analysed—of the considered *basic solvers* for solving the preferred enumeration (upper table) and stable enumeration (lower table) problems on the complete testing set (All) of 1,200 *AFs*, and on testing sets including *AFs* generated by specific graph models. Solvers are ordered according to PAR10 on the All testing set. Ft column indicates the number of times a solver has been the fastest among considered. Best results in bold.

EE-PR											
Solver	All			Barabasi-Albert		Erdős-Rényi		StableM		Watts-Strogatz	
	PAR10	Cov.	Ft	PAR10	Cov.	PAR10	Cov.	PAR10	Cov.	PAR10	Cov.
Cegartix	1350.4	79.1	229	1662.6	74.2	1266.6	81.0	1439.2	77.0	1028.6	84.2
ArgSemSAT	1916.2	69.1	35	3532.3	41.9	433.7	94.2	2530.9	58.7	1171.1	81.5
LabSATsolver	2050.3	66.8	9	3430.7	43.5	261.3	96.5	2869.5	53.0	1657.5	73.9
prefMaxSAT	2057.2	66.8	273	3482.1	42.9	444.0	94.2	3625.2	40.3	697.5	89.4
DIAMOND	2417.0	61.0	1	3447.8	43.2	1366.7	79.0	2831.8	53.7	2026.0	68.0
ASPARTIX-D	2728.6	56.1	4	4101.5	32.6	3067.8	51.6	2068.8	66.7	1630.3	74.3
ASPARTIX-V	2772.2	55.2	21	3646.6	40.3	3292.6	47.1	2340.7	62.0	1772.4	71.9
CoQuiAas	3026.4	50.5	78	3736.1	38.4	2873.4	53.5	2836.4	53.3	2645.1	57.1
ASGL	3477.3	43.2	1	4809.7	20.3	96.1	100.0	4475.4	26.0	4585.5	25.4
Conarg	3696.3	39.3	158	1128.7	81.6	2813.9	55.8	4934.6	18.3	6000.0	0.0
ArgTools	3906.2	35.2	322	3694.4	39.0	45.2	100.0	6000.0	0.0	6000.0	0.0
GRIS	4543.7	24.4	174	254.6	96.1	6000.0	0.0	6000.0	0.0	6000.0	0.0
EE-ST											
Solver	All			Barabasi-Albert		Erdős-Rényi		StableM		Watts-Strogatz	
	PAR10	Cov.	Ft	PAR10	Cov.	PAR10	Cov.	PAR10	Cov.	PAR10	Cov.
ArgTools	440.7	94.5	245	1328.6	78.4	47.4	100.0	144.1	100.0	230.5	100.0
LabSATsolver	641.6	90.0	352	396.2	93.9	22.7	100.0	1497.6	76.0	684.9	90.7
ASPARTIX-D	829.7	87.1	395	412.2	93.5	1194.4	81.6	1187.2	81.0	535.0	93.0
CoQuiAas	1477.2	76.2	372	1453.3	76.5	1485.1	76.5	1879.0	69.3	1106.5	83.3
DIAMOND	1555.4	75.2	42	2527.1	58.7	692.2	89.7	1887.2	69.7	1127.1	83.7
ArgSemSAT	1826.6	70.5	70	4019.0	33.5	408.9	94.5	1970.0	68.0	900.8	87.0
Conarg	1976.4	67.8	292	261.4	96.1	33.6	100.0	3742.1	38.3	4010.0	35.3
ASGL	2647.6	57.3	11	2737.4	56.1	85.2	100.0	3723.8	38.7	4152.8	33.7

is referred to [16] for detailed descriptions of the solvers. Hereinafter, we will refer to such systems as *basic solvers*, regardless of the approach they exploit for solving argumentation-related problems.

Experiments have been run on a cluster with computing nodes equipped with 2.5 Ghz Intel Core 2 Quad Processors, 4 GB of RAM and Linux operating system. A cutoff of 600 seconds was imposed to compute the extensions—either preferred or stable—for each *AF*. For each solver we recorded the overall result: success (if it solved the considered problem), crashed, timed-out or ran out of memory.

In ICCMA, solvers have been evaluated by considering only coverage (in case of ties the overall runtime on solved instances). Here we also evaluate solvers' performance by considering the PAR10 score.

4.1. Hypothesis 1: Reduction-based Solvers Constantly Outperform Others

Table 1 shows the results of this analysis in terms of coverage, PAR10 scores, and number of instances on which a given solver has been the fastest. We considered runtimes below 1 CPU-time second as equally fast.

Each *basic solver* for the EE-PR problem has at least one instance on which it is the fastest. We note that, when considering performance achieved on the whole testing set (All) by solvers, there can be a significant discrepancy between results shown in the coverage and fastest columns. One would expect that the higher the coverage, the larger

the possibility of a solver to be the fastest. Interestingly, we observed that some of the solvers with low coverage tend to be fast on the (few) instances they are able to analyse. For instance, **ArgTools** (a non reduction-based system) achieves low overall coverage, but it is the best solver for handling *AFs* of the Erdős-Rényi set. This contradicts the hypothesis—endorsed by ICCMA-15 results—that reduction-based systems constantly outperform others.

The best *basic solver* for solving the EE-PR problem on the StableM set of *AFs* is **Cegartix**, which is able to solve 77.0% of the instances. This is approximately 10% more than the coverage of the second best solver on such set, **ASPARTIX-D**. The **prefMaxSAT** solver has shown the best performance on the Watts-Strogatz *AFs*. From an (empirical) complexity perspective, we observe that the set with the lowest average coverage is the Barabasi-Albert set of *AFs*. This is possibly due to the very large number (up to few thousands, in some cases) of preferred extensions of such testing frameworks. Conversely, the Erdős-Rényi set is the less complex for the considered *basic solvers* when solving the EE-PR problem. Moreover we can derive that even though there is usually a *basic solver* with best coverage performance on each testing set, such solver is not always the fastest.

As for the EE-ST problem, the results in Table 1 show another interesting scenario. **ArgTools** is able to achieve the best PAR10 and coverage performance on two of the four considered sets, namely StableM and Watts-Strogatz. **LabSATSolver** obtained the best PAR10 score on the Erdős-Rényi set, but four of the considered *basic solvers* successfully analyse each of the 300 *AFs* in such a set. The winner of the EE-ST track of ICCMA-15, **ASPARTIX-D**, has been the fastest solver on 395 of the testing frameworks, but it did never excel in any of the 4 considered subsets. It seems that the *AFs* of the StableM set are (empirically) the most complex to solve for the considered systems.

4.2. Hypothesis 2: Basic Solvers Show Complementary Performance

Table 1 indicate that there is not a *basic solver* that is always the best selection on the vast majority of the testing frameworks. This is evidence that the *basic solvers* are substantially complementary, thus supporting the claim that a mixture of approaches can be fruitful, and justifying the search for improvements via portfolios.

5. Experimental Analysis of Portfolios

First of all, we generated the Virtual Best Solver (VBS) as the (virtual) oracle which always select the best solver (as to PAR10) for the given framework and problem. This provides the upper bound of performance achievable by combining considered solvers.

For the preferred semantics, the solvers included in the Shared-5 portfolio, ordered following their execution order, are: **Cegartix**, **ArgSemSAT**, **prefMaxSAT**, **LabSATSolver** and **DIAMOND**. Smaller static portfolios include subsets of those 5 solvers, not necessarily in that order. FDSS static portfolio includes **ArgSemSAT** and **GRIS**, only.

For the stable semantics, the solvers included in the Shared-5 portfolio, ordered following their execution order, are: **LabSATSolver**, **ArgTools**, **ASPARTIX-D**, **CoQuiAas** and **DIAMOND**. Smaller portfolios include subsets of the listed solvers, not necessarily in that order. The FDSS portfolio includes **LabSATSolver** and **ASPARTIX-D**.

Table 2. Coverage (Cov.) and PAR10 of the systems considered in this study for solving the EE-PR problem (left part) and the EE-ST problem (right part) on the complete set of 1,200 testing *AF*s. VBS indicates the performance of the virtual best solver. Systems are ordered according to PAR10.

EE-PR			EE-ST		
System	Cov.	PAR10	System	Cov.	PAR10
VBS	91.4	562.9	VBS	100.0	39.3
<i>Classify</i>	89.7	665.2	<i>I-Regression</i>	97.4	206.9
<i>I-Regression</i>	88.6	734.7	<i>Classify</i>	97.1	217.5
<i>M-Regression</i>	82.8	1068.3	<i>Shared-2</i>	97.7	262.3
FDSS	80.0	1311.4	<i>M-Regression</i>	94.7	378.4
Cegartix	79.1	1350.4	<i>Shared-3</i>	94.0	420.1
<i>Shared-2</i>	73.2	1678.0	ArgTools	94.5	440.7
<i>Shared-3</i>	69.4	1892.0	LabSATSolver	90.0	641.6
ArgSemSAT	69.1	1916.2	FDSS	89.4	677.4
LabSATSolver	66.8	2050.3	ASPARTIX-D	87.1	829.7
prefMaxSAT	66.8	2057.2	<i>Shared-5</i>	86.3	867.4
<i>Shared-4</i>	65.7	2105.5	<i>Shared-4</i>	86.0	873.8
<i>Shared-5</i>	63.3	2240.3	CoQuiAas	76.2	1477.2
DIAMOND	61.0	2417.0	DIAMOND	75.2	1555.4
ASPARTIX-D	56.1	2728.6	ArgSemSAT	70.5	1826.6
ASPARTIX-V	55.2	2772.2	Conarg	67.8	1976.4
CoQuiAas	50.5	3026.4	ASGL	57.3	2647.6
ASGL	43.2	3477.3			
Conarg	39.3	3696.3			
ArgTools	35.2	3906.2			
GRIS	24.4	4543.7			

We also generated the three per-instance (per-problem) portfolios that exploit predictive models in order to map the features of the given *AF* to a solver selection or combination: *Classify*, *I-Regression*, and *M-Regression*. *Classify* and *I-Regression* select a single solver by relying, respectively, on classification and regression techniques. *M-regression* iteratively selects the next solver to run, and allocates its CPU-time, by considering the predicted runtime of the available solvers for the given framework and problem, increased by 10% in order to mitigate the impact of negligible prediction mistakes.

We trained all the portfolio approaches using our training set of 800 *AF*s, 200 *AF*s from each set. The runtime cutoff once again was 600 CPU-time seconds. Table 2 shows the coverage and PAR10 scores of all portfolios, *basic solvers* and the VBS on the testing frameworks.

5.1. Hypothesis 3: Static Portfolios are more Efficient than Basic Solvers

Results for the static portfolios vary between stable and preferred semantics. When dealing with the EE-PR problem, the FDSS approach is the only technique which is able to outperform the best *basic solver*. *Shared-2* and *Shared-3* achieve performance close to those of the best *basic solver*, while *Shared-4* and *Shared-5* are undistinguishable from average *basic solvers*. FDSS portfolio performs better than *Shared-k* static portfolios because it includes **GRIS**. **ArgSemSAT** has good coverage, and **GRIS** excels on the

Table 3. Number of times each solver has been selected by the Classify (Class.) or M-Regression (M-Reg.) approaches for solving EE-PR (left part) and EE-ST (right part) problems on the testing frameworks. *Basic solvers* are alphabetically ordered. Highest numbers in bold. Empty cells indicate that the corresponding solver is not able to handle the considered problem.

System	EE-PR		EE-ST	
	Class.	M-Reg.	Class.	M-Reg.
ArgSemSAT	0	253	0	212
ArgTools	311	305	138	428
ASGL	6	36	0	35
ASPARTIX-D	2	80	305	409
ASPARTIX-V	1	99		
Cegartix	221	403		
Conarg	157	122	231	337
CoQuiAas	43	44	288	193
DIAMOND	0	65	33	138
GRIS	153	278		
LabSATSolver	13	208	228	548
prefMaxSAT	297	301		

Barabasi-Albert set (Table 1), while *Shared-k* portfolios do not include any solver able to efficiently solve the EE-PR problem on the Barabasi-Albert set.

Conversely, the right part of Table 2 shows that on the EE-ST problem, both *Shared-2* and *Shared-3* are able to achieve better performance than any *basic solver*, and the FDSS portfolio. Shared portfolios performance are boosted by the inclusion of **ArgTools**, which is able to achieve the best performance on three of the considered benchmark set structures, and **CoQuiAas**—that is the second best *basic solver* in terms of number of *AFs* quickly analysed. Moreover, the EE-ST problems are usually quickly solved by the *basic solvers*, therefore 2 or 3 solvers can be easily executed within the 600 CPU-time seconds limit. When more than three solvers are combined by the Shared approach—i.e. the CPU-time allocated to each *basic solver* is less than 200 seconds—performance drops.

5.2. Hypothesis 4: Per-Instance Portfolios are more Efficient than Static Portfolios

When considering per-instance portfolios, Table 2 indicates that they are all able to outperform the best *basic solver* on the considered testing frameworks. This comes as no surprise, since per-instance approaches should be able to select the most promising—ideally, the fastest—algorithm for solving the considered problem on the given *AF*. For both EE-PR and EE-ST problems, the performances of *Classify* and *1-Regression* are very similar, but the *M-Regression* approach performance is always worse. Such results indicate that: (i) the 50 features considered are informative for both EE-PR and EE-ST problems, and allow to effectively select solvers; (ii) classification and regression predictive models have similar performance when used for selecting a single solver to run; and (iii) the regression predictive model tends to underestimate the CPU-time needed by algorithms for solving the considered problem on the given *AF*.

Table 3 shows the number of times each *basic solver* has been executed by either the *Classify* or the *M-Regression* portfolio. *1-Regression* executed solvers are not shown,

Table 4. Coverage (Cov.) and PAR10 of the systems considered in this study on the complete testing set, when trained on a training set not containing *AF*s of that structure (leave-one-set-out scenario). Systems are ordered according to results shown in Table 2. Best results in bold.

EE-PR								
System	Barabasi-Albert		Erdős-Rényi		StableM		Watts-Strogatz	
	Cov.	PAR10	Cov.	PAR10	Cov.	PAR10	Cov.	PAR10
<i>Classify</i>	78.9	1321.4	88.6	745.0	74.4	1574.3	89.5	677.8
<i>l-Regression</i>	76.3	1479.0	63.0	2255.2	76.5	1453.9	83.0	1079.9
<i>M-Regression</i>	70.4	1828.4	67.3	2039.7	77.0	1434.7	79.6	1267.6
<i>FDSS</i>	69.1	1916.2	80.9	1245.5	79.1	1341.9	78.6	1380.0
<i>Shared-2</i>	73.2	1678.0	73.2	1678.0	74.2	1620.4	73.2	1678.0
<i>Shared-3</i>	69.4	1892.0	67.3	2007.9	69.5	1896.7	69.4	1892.0
<i>Shared-4</i>	65.7	2106.2	65.7	2101.1	65.7	2108.1	65.7	2103.9
<i>Shared-5</i>	63.3	2240.9	63.4	2235.8	63.3	2242.9	63.3	2242.9

EE-ST								
System	Barabasi-Albert		Erdős-Rényi		StableM		Watts-Strogatz	
	Cov.	PAR10	Cov.	PAR10	Cov.	PAR10	Cov.	PAR10
<i>l-Regression</i>	88.6	756.9	92.6	508.7	98.6	149.9	81.6	1153.0
<i>Classify</i>	93.0	470.4	92.4	519.6	91.2	575.6	93.4	439.3
<i>Shared-2</i>	97.7	262.3	97.3	285.2	97.7	220.9	97.7	262.3
<i>M-Regression</i>	96.2	297.4	96.4	282.2	95.6	334.9	90.3	636.5
<i>Shared-3</i>	94.0	420.1	94.0	435.5	94.0	420.1	94.0	476.6
<i>FDSS</i>	89.4	677.4	87.1	829.7	89.4	677.4	88.7	714.7
<i>Shared-4</i>	85.9	878.2	86.0	887.5	86.0	873.8	86.8	833.8
<i>Shared-5</i>	86.3	867.4	86.3	870.8	86.3	862.3	84.3	973.4

because they are a subset of the *M-regression* selections. Table 3 shows some remarkable differences in the algorithm selected by the classification and regression approaches, and also those included in the static portfolios. For instance, *Classify* never selects **ArgSemSAT**, while it is largely exploited by *M-regression*, and included in static portfolios generated for solving EE-PR problems. This is because **ArgSemSAT**, and a few other *basic solvers*, has rarely been the fastest: therefore the classification approach—which only focuses on the best solver—ignores its performance. On the contrary, solvers like **ArgTools** (EE-PR) and **ASPARTIX-D** (EE-ST) are usually the fastest, and are often selected by both *Classify* and *M-Regression* approaches.

Finally, by looking at Table 2, it can be noted that the largest performance improvement can be achieved when exploiting portfolio approaches for solving the problem of enumerating preferred extensions of an *AF*: the use of portfolio-based techniques allows to solve up to 10.6% more instances than the best *basic solver*, **Cegartix**. Such margin is reduced to 2.9% when solving the EE-ST problem. This is due to the higher empirical complexity of the EE-PR problem, and to the higher complementarity between *basic solvers* able to handle the EE-PR problem.

5.3. Post-Hoc Analysis: Generalisation of Performance

To assess the ability of our portfolios on testing instances that are dissimilar from instances used for training we generated four different new training sets as follows: starting by the original training set composed by 800 *AFs*, we removed all the frameworks corresponding to one set at a time, and randomly oversampled frameworks from the remaining three sets—in order to have again approximately 800 frameworks for training. We then tested our portfolios on the complete testing set of 1,200 *AFs*, so that performance can be compared with those of *basic solvers* (Table 2). This can be seen as a leave-one-out scenario. The results of such generalisation analysis are shown in Table 4.

Unsurprisingly, static portfolios—particularly *Shared-k*—show the best generalisation performance: their behaviour does not change much with the new training sets. On the other hand, per-instance approaches do not show good generalisation capabilities: their performance varies significantly when the training set is not fully representative of the testing instances. This is true for both EE-PR and EE-ST problems, despite the fact that gaps are smaller in the EE-ST case, although it is true that also the performance of *basic solvers* on EE-ST tends to be closer.

Remarkably, *Classify* (covering up to 89.7%, cf. Table 2) is very sensible to the absence of Barabasi-Albert (−10.8%, cf. Table 4) or StableM (−15.3%, cf. Table 4) frameworks from the training set for EE-PR, while regression-based approaches show scarce generalisation abilities when the Erdős-Rényi frameworks are removed from the training set. On the contrary, *Classify* is very generalisable on the EE-ST set, and the *I-Regression* method is very sensitive when Watts-Strogatz *AFs* are removed. *M-Regression* is more generalisable than *I-Regression* when dealing with the EE-ST problem: this indicates that when testing instances are dissimilar from training ones, the exploitation of more than one solver can be fruitful.

6. Conclusion

We exploit the ICCMA-15 legacy by combining state-of-the-art solvers, able to handle EE-PR and EE-ST problems, using—for the first time in this research area—portfolio-based techniques. In particular, we tested static and per-instance portfolios, exploiting the largest available set of argumentation features [6]. We remark this is the first comprehensive experimental analysis on the performance of different portfolio-based methods, in the argumentation area.

The results of our extensive empirical analysis showed that: (i) the claim that reduction-based solvers always outperform non reduction-based systems—one of the takeaway message from ICCMA-15—is not always the case; (ii) the solvers at the state of the art show a high level of complementarity (specially those able to deal with EE-PR problems), thus they are suitable to be combined in portfolios; (iii) portfolio systems generally outperform *basic solvers*; (iv) if the training instances are representative of testing *AFs*, the existing set of features is informative for selecting most suitable solvers; (v) classification-based portfolios show good generalisation performance; (vi) static portfolios are usually the approaches which are less sensitive to different training sets.

As part of future research, we are interested in further investigating the generalisation capabilities of portfolios performance by considering significantly differently-

structured AFs, including complex frameworks generated by real-world scenarios. We will also extend the portfolio methods considering SATzilla [21] like approaches, or more sophisticated model-based techniques. Finally, we are interested in testing portfolio methods also in other complex argumentation problems.

Acknowledgements

The authors would like to acknowledge the use of the University of Huddersfield Queensgate Grid in carrying out this work.

References

- [1] A. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):11, 1999.
- [2] P. Baroni, M. Caminada, and M. Giacomin. An introduction to argumentation semantics. *Knowl. Eng. Rev.*, 26(4):365–410, 2011.
- [3] P. Baroni, F. Cerutti, P. E. Dunne, and M. Giacomin. Automata for Infinite Argumentation Structures. *Artif. Intell.*, 203(0):104–150, 2013.
- [4] S. Bistarelli, F. Rossi, and F. Santini. Benchmarking Hard Problems in Random Abstract AFs: The Stable Semantics. In *Proc. of COMMA*, pages 153–160, 2014.
- [5] R. Brochenin, T. Linsbichler, M. Maratea, J. P. Wallner, and S. Woltran. *TAFa 2015*, chapter Abstract Solvers for Dung’s Argumentation Frameworks, pages 40–58. 2015.
- [6] F. Cerutti, M. Giacomin, and M. Vallati. Algorithm selection for preferred extensions enumeration. In *Proc. of COMMA*, pages 221–232, 2014.
- [7] F. Cerutti, M. Giacomin, and M. Vallati. Generating challenging benchmark AFs. In *Proc. of COMMA*, pages 457–458, 2014.
- [8] F. Cerutti, N. Oren, H. Strass, M. Thimm, and M. Vallati. A benchmark framework for a computational argumentation competition. In *Proc. of COMMA*, pages 459–460, 2014.
- [9] P. M. Dung. On the Acceptability of Arguments and Its Fundamental Role in Nonmonotonic Reasoning, Logic Programming, and n-Person Games. *Artif. Intell.*, 77(2):321–357, 1995.
- [10] P. Erdős and A. Rényi. On random graphs. I. *Publ. Math-Debrecen*, 6:290–297, 1959.
- [11] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The WEKA data mining software: An update. *SIGKDD Explor.*, 11(1):10–18, 2009.
- [12] H. Hoos, M. Lindauer, and T. Schaub. clasptfolio 2: Advances in algorithm selection for answer set programming. *Theor. Pract. Log. Prog.*, 14(Special Issue 4-5):569–585, 2014.
- [13] F. Hutter, L. Xu, H. H. Hoos, and K. Leyton-Brown. Algorithm runtime prediction: Methods & evaluation. *Artif. Intell.*, 206:79–111, 2014.
- [14] J. R. Rice. The algorithm selection problem. *Adv. Comput.*, 15:65–118, 1976.
- [15] J. Seipp, M. Braun, J. Garimort, and M. Helmert. Learning portfolios of automatically tuned planners. In *Proc. of ICAPS*, pages 369–372, 2012.
- [16] M. Thimm and S. Villata. System descriptions of the first international competition on computational models of argumentation (ICCMa’15). *arXiv preprint arXiv:1510.05373*, 2015.
- [17] M. Thimm, S. Villata, F. Cerutti, N. Oren, H. Strass, and M. Vallati. Summary Report of The First International Competition on Computational Models of Argumentation. *AI Mag.*, 2016.
- [18] M. Vallati, L. Chrapa, M. Grzes, T. L. McCluskey, M. Roberts, and S. Sanner. The 2014 international planning competition: Progress and trends. *AI Mag.*, 2015.
- [19] M. Vallati, L. Chrapa, and D. E. Kitchin. Portfolio-based planning: State of the art, common practice and open challenges. *AI Commun.*, 2015.
- [20] D. J. Watts and S. H. Strogatz. Collective dynamics of ‘small-world’ networks. *Nature*, 393(6684):440–442, 1998.
- [21] L. Xu, F. Hutter, H. H. Hoos, and K. Leyton-Brown. Satzilla: portfolio-based algorithm selection for sat. *J. Artif. Intell. Res.*, pages 565–606, 2008.

Argumentation for Machine Learning: A Survey

Oana COCARASCU ^{a,1}, Francesca TONI ^a

^a*Department of Computing, Imperial College London, UK*

Abstract. Existing approaches using argumentation to aid or improve machine learning differ in the type of machine learning technique they consider, in their use of argumentation and in their choice of argumentation framework and semantics. This paper presents a survey of this relatively young field highlighting, in particular, its achievements to date, the applications it has been used for as well as the benefits brought about by the use of argumentation, with an eye towards its future.

Keywords. Argumentation, Machine Learning

1. Introduction

Machine Learning (ML) [27] amounts to automatically learning from data and improving with experience. Nowadays, its use is becoming more and more important as much of the work on visual processing, language and speech recognition relies on it.

Argumentation (e.g. as overviewed in [37]) has proven successful in several domains, including multi-agent systems [6] and decision support in medicine [16] and engineering [2]. ML and argumentation are brought together in a number of settings, e.g. to support argument mining (e.g. see [26]) as well as to aid ML, in one sense or another. Also the integration of argumentation and applications of ML have been proven to be fruitful (e.g. as in [22]). In this paper we focus on the use of argumentation to aid ML and provide an overview of this relatively young field, with an eye to guide its future developments.

Existing approaches using argumentation for ML differ in the type of ML they consider and the specific method they use. Concretely:

- the Argumentation-Based Machine Learning (ABML) approach of [32] extends the CN2 rule induction algorithm [12] for *supervised learning*;
- the Argument-Based Inductive Logic Programming (ABILP) approach of [4] extends Inductive Logic Programming (ILP) for *supervised learning*;
- the hybrid approach of [21] uses as its starting point the Fuzzy Adaptive Resonance Theory (ART) model [7] for *unsupervised learning*;
- the fully argumentative concept learning method of [1] focuses on the version space learning framework [27] for *supervised learning*;

¹Corresponding Author: Oana Cocarascu, Department of Computing, Imperial College London, United Kingdom; E-mail: oana.cocarascu11@imperial.ac.uk.

- the multi-agent inductive concept learning of [34] and its computational realisation [35] have concept learning [27] for *supervised learning* as their starting point;
- the Argumentation Accelerated Reinforcement Learning (AARL) of [17,18,19] extends SARSA [39] for *reinforcement learning*;
- the Classification enhanced with Argumentation (CleAr) method of [8,9] works with any *supervised learning* technique, and has been experimented with, in particular, Naïve Bayes classifiers [25], Support Vector Machines (SVMs) [13] and Random Forests [5].

Moreover, existing approaches differ in their use of argumentation and in their choice of argumentation framework/method. Finally, different approaches achieve different (desirable) outcomes, ranging from improving performances to rendering the ML process more transparent by improving its explanatory power.

The paper is organized as follows. In Section 2 we give an abstract re-interpretation of ML, in general and for supervised/unsupervised/reinforcement learning, to serve as a basis for a comparison amongst existing ML approaches using argumentation. In Section 3 we overview the different approaches using argumentation for ML, showing in particular how they use argumentation, and which kind thereof, to contribute to particular instantiations of the abstract model of Section 2. In Section 4 we provide a comparative analysis of the different approaches. In Section 5 we conclude, identifying in particular some challenges/open problems for an even more impactful use of argumentation in ML.

2. Machine Learning in the Abstract

In this section we give an abstract re-interpretation of ML, in general and for supervised/unsupervised/reinforcement learning, to serve as a basis for the comparison amongst different existing ML approaches using argumentation. Being tailored to providing an overview of existing approaches to ML *using argumentation*, this abstract interpretation has no pretence of being general or fully covering (e.g. it completely ignores the use of probabilistic information in ML).

In the abstract, a ML method can be characterised in terms of the following notions, that will be instantiated differently for the different ML methodologies (supervised/unsupervised/reinforcement learning) and for the different methods for the methodologies (e.g. CN2 for supervised learning, ART for unsupervised learning, and SARSA for reinforcement learning):

- H is the *hypotheses space*, namely the set of all possible “reasoners” that a ML method may return;
- \mathcal{S} is the *training* input, given to the ML method to trigger the learning process leading to generating a “reasoner” in H ;²
- X is the set of all possible descriptions of inputs for the ML method (the training input) and for the “reasoner” learnt by the ML method (the *unseen* input, e.g. used for testing);
- \mathcal{L} is the set of all possible outputs that “reasoners” computed by a ML method may return, given the inputs.

²Note that the training input excludes any *testing* input, to be deployed after learning has taken place to test the computed “reasoners”.

2.1. Supervised learning

In this setting a “reasoner” in H is a classifier, \mathcal{L} is a set of alternative classifications, X is a set of combinations of features that inputs may exhibit, and each element of the training input includes the correct classification for a given combination of features, whereas the unseen inputs only consist of features:

- X is the *feature space*; for example, a feature may be an attribute/value pair;
- \mathcal{L} is the set of possible *classifications*; for example, if the aim of the supervised learning method is to learn a concept, then $\mathcal{L} = \{0, 1\}$;
- \mathcal{S} is the set of *training instances*; a training instance is of the form (x, l) for $x \in X$ and $l \in \mathcal{L}$; for example, if the aim of the supervised learning method is to learn a concept, then $(\{f_1, f_2\}, 1)$ indicates that the combination of features f_1, f_2 is an example of the concept, and $(\{f_1, f_3\}, 0)$ indicates that the combination of features f_1, f_3 is not;
- a generic member h_s of H can be abstractly seen as a mapping $h_s : X \mapsto \mathcal{L}$; at an abstract level, the goal of a supervised ML method is to determine a classifier h_s such that (i) $h_s(x) = l$ for all (or for as many as possible) $(x, l) \in \mathcal{S}$ and (ii) h_s generalises well by classifying instances not in \mathcal{S} correctly (during testing).

2.2. Unsupervised learning

In this setting, a “reasoner” in H is also a classifier, but the training instances in \mathcal{S} are given in terms of their feature combinations only, as a correct classification for them is not available; the most popular unsupervised ML methods then compute *clusters* an input may belong to and determine the classifier/classification of the input using the clusters: here a cluster is a collection of instances which are “similar”, while being “dissimilar” to instances in other clusters [27]. Thus:

- X is the feature space, as in supervised learning, and $\mathcal{S} \subseteq X$; for example, inputs may be images of different fruits, and features may include pixels in these images;
- \mathcal{L} is obtained from the “learnt” clusters; for example, one cluster may group together apples and another oranges;
- a generic member h_u of H can be seen as a mapping $h_u : X \mapsto \mathcal{L}$; abstractly, the goal of (cluster-based) unsupervised learning is to find a “good” way to assign inputs to clusters, as a basis for classification.

2.3. Reinforcement learning

In this setting, a “reasoner” is a *policy*, that, given inputs in the form of observations of *states*, returns outputs in the form of *actions*. Actions, during learning, are not known to be right or wrong, and thus classifications are not available. Instead, *rewards* are given for states reached by performing actions (these rewards are positive if the states are “desirable” and negative otherwise; negative rewards can be interpreted as punishments). Thus:

- \mathcal{L} is the set of *actions* that can be performed by the learner; for example, if the learner is a robot, actions may include moves in several directions;

- X is the set of all possible states; for example, a state may represent the physical environment a robot is situated in;
- \mathcal{S} is the *reward function*, namely a mapping $\mathcal{S} : X \mapsto \mathbb{R}$; for example, a goal state, that the robot should aim at achieving, could be given a high reward;
- H is the *policy space*, and a generic member h_r of H is a mapping $h_r : X \mapsto \mathcal{L}$; at an abstract level, the goal of a reinforcement learning method is to determine, with as little training as possible, a policy h_r which is optimal, namely it gets the highest possible cumulative reward.

3. Approaches to Argumentation for Machine Learning

Here we show how existing approaches use argumentation, and which form thereof, for ML, in the context of suitable instantiations of the abstract model in Section 2.

3.1. Argumentation for Supervised Learning

ABML [32]. Here arguments are associated with elements of \mathcal{S} and are of the form:

C because *Reasons* or
 C despite *Reasons*

where $C \in \mathcal{L}$ and *Reasons* $\subseteq X$. The first type of argument provides reasons (in terms of combinations of features) for why a certain training instance is classified as is, whereas the second type of argument indicates combinations of features that do not play a role in classifying the training instance the argument is associated with.

For example, let \mathcal{S} represent credit applications and \mathcal{L} represent whether the credit was approved or not. Then, a training instance $(x, l) \in \mathcal{S}$ with

$x = \{\text{PaysRegularly} = \text{no}, \text{Rich} = \text{yes}, \text{HairColor} = \text{blond}\}$
 $l = \text{CreditApproved}$

may be associated with arguments

$$\text{CreditApproved because Rich} = \text{yes} \tag{1}$$

$$\text{CreditApproved despite PaysRegularly} = \text{no} \tag{2}$$

In ABML, arguments of these two forms are used to modify the CN2 rule induction algorithm for supervised ML so as to learn “better” hypotheses, while also reducing the size of H and providing explanations for classifications. Here, hypotheses are rules of the form IF F_1 AND ... AND F_n THEN C , for $n > 0$, $F_i \in X$, $C \in \mathcal{L}$.

For example, from the credit approval example above, CN2 alone may obtain the rule IF *HairColor* = *blond* THEN *CreditApproved*, whereas using argument (1), ABML may obtain the “better” rule IF *HairColor* = *blond* AND *Rich* = *yes* THEN *CreditApproved*.

ABML does not make use of any particular argumentation framework in the literature, but uses ad hoc arguments of the forms given earlier, suitable for the specific ML setting considered. Moreover, ABML does not make use of any argumentation semantics or methodology for assessing the acceptability/strength of arguments (these are taken at face value instead).

ABILP [4]. Here arguments are as in ABML but they are integrated within ILP.

Concept Learning as Argumentation (CLA) [1]. This method reinterprets concept learning in argumentation terms. Here arguments are obtained from \mathcal{S} and H and are of the form $\langle h, x, l \rangle$ for $h \in H \cup \{\emptyset\}$, $x \in X$ and $l \in \mathcal{L}$ such that

if $h = \emptyset$ then $(x, l) \in \mathcal{S}$, and

if $h \neq \emptyset$ then $h(x) = l$,

namely each training instance in \mathcal{S} and each hypothesis in H gives an argument. Moreover, an argument a attacks an argument b by *rebutting* if the two arguments give different classifications for the same features, or by *undercutting* if a is drawn from an example and b is drawn from a hypothesis which disagrees with the example.

This method then uses standard semantics of extensions [14] applied to abstract argumentation frameworks with arguments obtained from \mathcal{S} and H as above, and a relation of *defeat* between arguments such that a defeats b iff a attacks b by rebutting or undercutting and b is not *preferred* to a , where given a preference relation over H , standardly used in concept learning:

- arguments obtained from \mathcal{S} are stronger than arguments obtained from H ;
- arguments obtained from most preferred hypotheses are stronger than arguments obtained from less preferred hypotheses.

For example, consider $X = \{x_1, x_2\}$, $\mathcal{S} = \{(x_1, c_1), (x_1, c_2)\}^3$, $\mathcal{L} = \{c_1, c_2, c_3, c_4\}$ and $H = \{h_1, h_2\}$ with $h_1(x_1) = c_1$, $h_1(x_2) = c_1$, $h_2(x_1) = c_2$, and $h_2(x_2) = c_1$. The corresponding abstract argumentation framework has arguments $a_1 = \langle \emptyset, x_1, c_1 \rangle$, $a_2 = \langle \emptyset, x_1, c_2 \rangle$, $a_3 = \langle h_1, x_1, c_1 \rangle$, $a_4 = \langle h_1, x_2, c_1 \rangle$, $a_5 = \langle h_2, x_1, c_2 \rangle$ and $a_6 = \langle h_2, x_2, c_1 \rangle$. Also, assuming that the two hypotheses are equally preferred, the defeat relation is such that a_1 defeats a_2 , a_1 defeats a_5 , a_1 defeats a_6 , a_2 defeats a_1 , a_2 defeats a_3 and a_2 defeats a_4 . The resulting abstract argumentation framework has an empty grounded extension and two preferred/stable extensions $\mathcal{E}_1 = \{a_1, a_3, a_4\}$ and $\mathcal{E}_2 = \{a_2, a_5, a_6\}$, both classifying x_2 as c_1 .

The grounded extension of the abstract argumentation framework corresponding to a given concept learning setting corresponds to the output of the version space method for concept learning when the latter is applicable, namely when the given \mathcal{S} is not inconsistent. Moreover, if \mathcal{S} is inconsistent (as in our earlier illustration), argumentation can still return an output, e.g. c_1 or c_2 for x_1 .

Argumentation for Multi-Agent Inductive Concept Learning (MAICL) [34,35]. In this approach, $\mathcal{L} = \{0, 1\}$ and \mathcal{S} is assumed to be consistent as well as distributed amongst agents, so that each agent is only aware of some subset of \mathcal{S} . Arguments are hypotheses induced by individual agents from training instances they are aware of. These hypotheses/arguments are rules. For uniformity of presentation, we assume here that these rules/hypotheses/arguments are in the same form as the rules learnt by CN2, presented earlier. Then an argument IF F_1 AND ... AND F_n THEN C attacks an argument IF F'_1 AND ... AND F'_m THEN C' iff $C \neq C'$ and $\{F_1, \dots, F_n\} \supseteq \{F'_1, \dots, F'_m\}$.

For example, let $\mathcal{S} = \{e_1, e_2, e_3\}$ represent a mammal dataset where

$e_1 = (\{\text{hair}, \text{milk}, \text{backbone}\}, 1)$

$e_2 = (\{\text{toothed}, \text{backbone}, \text{twolegged}\}, 1)$

³Note that this set is *inconsistent*, as it classifies differently the same features.

$$e_3 = (\{toothed, backbone\}, 0)$$

and 1 stands for mammal, 0 stands for non-mammal. Consider two agents, ag_1 and ag_2 , aware of $\{e_1, e_2\}$ and $\{e_3\}$, respectively, and let

IF *backbone* THEN 1

IF *backbone* AND *toothed* AND *twolegged* THEN 1

be the rules learnt by ag_1 and

IF *backbone* AND *toothed* THEN 0

be the rule learnt by ag_2 . Then, ag_1 's second rule/argument attacks ag_2 's rule/argument.

In this approach, agents communicate arguments and attacks to construct dialectical trees as defined in [11,38] and determine which arguments are defeated/undefeated. For example, in the earlier illustration, ag_2 's rule/argument is defeated. Here, argumentation helps building hypotheses in a distributed manner when examples are not held centrally. Also, this method is supported by a computational realisation [35].

CleAr [8,9]. In this approach, arguments and relations amongst them are drawn from a given set of templates (an *Argument base*) for a given *testing* instance that has already been classified by means of a "reasoner" (classifier) learnt by any standard supervised learning methods. The relations amongst arguments are of *attack* or *support* and thus the resulting argumentation frameworks, associated with training instances, are *bipolar* [10]. In addition, a *base score* is associated with arguments, as in QuAD frameworks [2,36]. Arguments are either elements of \mathcal{L} or express domain knowledge of the learning task at hand and, in this latter case, are of the form

Premise \Rightarrow *Conclusion*

where *Premise* may represent any information, including, but not limited to, combinations of elements of X , and *Conclusion* is either an element of \mathcal{L} or it represents a statement agreeing or disagreeing with the *Premise* of some other argument.

For example, consider the task of determining sentiment polarity in tweets. Then $\mathcal{L} = \{positive, negative\}$ and X are (syntactic or semantic) features extracted from tweets. Suppose that some existing classifier h assigns positive polarity to the tweet:

'more depressed than you could ever imagine that I wont be going to Vegas.

I hate having to be financially responsible'

The resulting argumentation framework, for this testing instance, may include arguments *positive* and *negative* (the elements of \mathcal{L}) as well as arguments

'hate' occurs in the tweet \Rightarrow *negative*

a negation ('wont') occurs in the tweet \Rightarrow *negative*

and, in addition, that the arguments attack *positive* and/or support *negative*.

In this approach, base scores for the arguments are derived from the output or the performances of h (the given classifier) or are drawn from the given Argument Base.

The dialectical strength of each classification in \mathcal{L} is then computed using a quantitative semantics (e.g. as in [15,2,36]) and the classification with maximal strength is assigned as the final classification for the testing instance. In our earlier illustration, assuming that *positive* and *negative* have a base score of 0.6 and 0.4 respectively and the other two arguments above are supporters of *negative* and have a base score of 0.4, the computed strength may be 0.75 for *negative* and 0.6 for *positive*. Hence, the use of argumentation, in this case, would change the classification to *negative*. In general, in this approach, argumentation contributes a (possibly revised) classification and a justification thereof.

3.2. Argumentation for Unsupervised Learning

Argumentation for ART (A-ART) [21]. In this approach, arguments, attacks and semantics are as in DeLP [20,11,38], but instantiated so as to reason with the output of a fuzzy ART network, when this assigns a training instance to different clusters. In this case, the classification choice for the given instance by the h_u being learnt is, conventionally, that of a randomly chosen cluster. By arguing, instead, this choice can be “reasoned” upon.

As an example, consider a fuzzy ART network which identifies three clusters $c_1^+, c_2^-, c_3^- \in \mathcal{L}$ for an instance e , such that c_1^+ subsumes c_3^- . Suppose also that, from the given DeLP program, DeLP arguments can be constructed with the following informal reading:

- + because e belongs to c_1^+
- because e belongs to c_3^-

with the second argument attacking the first but not vice versa, as c_1^+ subsumes c_3^- . Then, the dialectical analysis of [11,38] gives classification –, drawn from membership of e in c_3^- .

3.3. Argumentation for Reinforcement Learning

AARL [17,18,19]. In this approach, arguments represent recommendations of actions to individual agents in a multi-agent system and are of the form:

Conclusion IF Premise

where *Conclusion* is an action (in \mathcal{L}) to be performed by an agent and *Premise* describes conditions under which the argument is applicable and may, for example, amount to properties of the state (in X) of the environment where the agent is situated. Then an argument attacks another argument iff

- the arguments support the same action but for different agents, or
- the arguments support different actions by the same agent.

For example, in a given state of the environment in which a RoboCup agent is situated, the applicable arguments may be

- agent a_1 should tackle the ball IF a_1 is closest to the ball keeper
- agent a_1 should mark agent a_2 IF a_1 is closest to a_2

with the two arguments attacking one another. At each iteration of learning one such abstract argumentation framework is generated, by instantiating a set of argument templates given up-front, representing domain knowledge.

AARL then uses preferences over arguments and adapts value-based argumentation [3] to choose actions (supported by arguments in some extension, e.g. the grounded extension) and shape rewards, thus modifying the reward function \mathcal{R} . For example, if tackling is more preferred than marking, for our earlier illustration, then the attack from the second to the first argument is deleted, as in value-based argumentation, and tackling gets extra reward at the current iteration of learning.

4. A Comparative Analysis of Argumentation for Machine Learning

In this section we provide a comparative analysis of the different approaches we have overviewed in Section 3. First, we note that existing approaches differ considerably in their choice of argumentation framework/semantics:

- ABML and ABILP use ad hoc arguments and no argumentation framework or semantics;
- CLA and AARL instantiate abstract argumentation, with arguments equipped with preferences, and deploy standard semantics of extensions;
- MAICL uses abstract argumentation, but deploys the dialectical trees of [11,38] as a semantics, rather than extensions;
- A-ART uses the DeLP argumentation framework and again the dialectical trees of [11,38] as a semantics;
- CleAr uses bipolar abstract argumentation extended with base scores or, equivalently, QuAD frameworks, and quantitative semantics.

Moreover, some approaches (i.e. ABML and AARL) use argumentation *during* learning, some (i.e. MAICL, CleAr and A-ART) use argumentation *after* learning, to process the output of standard ML techniques, and some (i.e. CLA) use argumentation *instead of* learning, to re-interpret the learning process. Furthermore, some approaches (i.e. MAICL and AARL) are developed to coordinate agents in multi-agent systems. Finally, different approaches are used for different applications and have different advantages over standard ML techniques, ranging from improving performances to rendering the ML process more transparent by improving its explanatory power or using argumentation to better elicit domain knowledge, of benefit to the learning process, from users. In the remainder of this section we analyse how the approaches overviewed in Section 3 have been applied and evaluated as well as their advantages.

ABML [32]. Compared with standard CN2, ABML has the advantage of *reducing the size of the hypotheses space H* , in that it forces the rules to be learnt to take into account the arguments associated with the examples, and thus allowing fewer rules to be legitimate hypotheses.

ABML was tested on several domains (notably law [31], medicine [40] and zoology [28]), and was shown to *improve classification accuracy* across the board. For example, by including arguments, the accuracy was improved on a zoo dataset from 94.51% to 96.75% [28]. Also, on a dataset related to severe bacterial infections [40], ABML achieved similar accuracy to CN2 and a further ML technique, C4.5 (88%), whilst Naïve Bayes (NB) and Logistic Regression performed worse (with accuracy under 86.5%). Further, using AUC (Area Under the Curve, an alternative measure to standard accuracy), ABML outperformed all other classifiers, the improvement varying between 0.03% and 0.2%. ABML was also tested on chess, improving the initial accuracy of 72% to 95% when learning the concept of bad bishop [29] of 84% to 91% when learning the concept of an attack on the castled king [30].

ABML was shown to be *robust*, in the presence of noise in the examples as well as random arguments. Indeed, ABML performed better in the presence of *noise*, compared to CN2, on a welfare benefit dataset [31]: the class of each example was randomly replaced with a value from \mathcal{L} with probability $p\%$ (for $p \in \{0, 2, 5, 10, 20, 40\}$) with dis-

tribution (0.5, 0.5), and the average accuracy of ABML was better than CN2 by 0.3% at 0% noise, by 3.3% at 20% noise and by 1.7% at 40% noise. Moreover, ABML was tested in the presence of *random* arguments, and shown to still outperform or perform similarly to the original CN2 [32]: here, random arguments were given for k randomly selected examples ($k \in \{2, 5, 10, 20\}$), each example could have up to five random arguments and each argument could have up to five random reasons. Thus, ABML is robust in that it is not negatively affected by “bad” domain knowledge.

ABML has been shown to support *knowledge elicitation* well [23,24,41] by identifying *critical examples* (namely instances that the learnt hypotheses, using ABML, do not classify well) and eliciting arguments for them and retraining, using ABML again. On a medical dataset, this knowledge elicitation-enriched ABML increased the performance from 60% to 80% for CN2 [23] and, on a larger medical dataset, from 52% to 82% [24]. Knowledge elicitation was also employed during an interactive learning session using python code [41] to distinguish between classifications in $\mathcal{L} = \{\textit{basic}, \textit{advanced}\}$ programming style achieving 87.1% accuracy when using ABML compared to 86.7% manual student classification.

ABILP [4]. This approach is in the same spirit as ABML. The advantages of this approach are potentially the same as for ABML.

CLA [1]. The advantages of this approach are theoretical, rather than of an experimental nature. Indeed, CLA can handle inconsistent sets \mathcal{S} of training instances, whereas standard concept learning cannot. Thus, the method is *robust*. In addition, by using argumentation, CLA supports in principle the generation of *explanations* for classifications.

MAICL [34,35]. At a theoretical level, MAICL allows agents to agree classifications even when they hold *partial information*, in the form of subsets of the set \mathcal{S} of training instances. A-MAIL [35], an implementation of a generalisation of MAICL, not restricted to $\mathcal{L} = \{0, 1\}$, uses four datasets [33] to test experimentally whether this method can work in practice and, in particular, whether the method can cope with a large number of agents and several forms of data distribution. The experiments showed, in particular, that the use of A-MAIL can lead to a *recall increase*, which is higher for five agents, each having the same portion of \mathcal{S} , than with two agents. In the case of more agents (10 or 20), more examples need to be exchanged by communication, as expected, but recall increases can still be observed (e.g., with 20 agents, from 0.35% to 0.88%). In the case of unbalanced distributions of training instances between two agents when ag_1 receives only $p\%$ of \mathcal{S} ($p \in \{50, 30, 10, 0\}$), using A-MAIL results in an improvement in recall for ag_1 at the cost of arguments exchanged as ag_1 has more information to obtain from ag_2 . Overall, the experiments show that A-MAIL can improve performances at a relatively reasonable cost in terms of number of messages being exchanged.

CleAr [8,9]. CleAr has been applied to two problems within the computational linguistic setting: cross-domain sentiment polarity classification [8,9], with $\mathcal{L} = \{\textit{Positive}, \textit{Negative}\}$, and relation-based argument mining to determine relations between pieces of text [9], with $\mathcal{L} = \{\textit{Attack}, \textit{Support}, \textit{Neither}\}$. In these two settings, CleAr has been instantiated with three types of supervised ML methods (i.e. NB, Support Vector Machines (SVM) and Random Forests (RF)) with suitably defined Argument Bases. Deploying CleAr with these Argument Bases gives an *increase in accuracy* of up to 14% for Sentiment Polarity Classification, from 50% to 64%, and *performance im-*

provements varying between 0.006% and 0.022% on various datasets for relation-based argument mining, with respect to using the standard ML methods alone.

A-ART [21]. The advantages of this approach, as presented in [21], are theoretical, rather than of an experimental nature. Here, argumentation is used to *resolve inconsistency* amongst classifications of clusters to which an instance is assigned as well as to *explain* the final classification dialectically.

AARL [17,18,19]. AARL has been deployed in RoboCup, and in particular for Keep-Away and TakeAway games, as well as other standard RL benchmarks. Experimentally, AARL, combined with a distance-oriented reward system, performs better overall when compared with SARSA or hand-coded strategies in terms of *stability*, *average convergence time* and *average optimal performance*. Moreover, this method is *robust* to errors in arguments.

Method	ML method	AF	Semantics	D/A ML	Multi agent	Advantages	Apps.
ABML	CN2	✗	✗	D		experimental (accuracy, robustness); elicitation	law; medicine; zoology; chess; coding
ABILP	ILP	✗	✗	D			
CLA	concept learning	AA with prefs.	extensions	✗		theoretical (inconsistency tolerance); explanation	
MAICL	concept learning	AA	dialectical trees	A	✓	experimental (recall); partial info	
CleAr	Random Forests; NB; SVM	Bipolar AA/ QuAD	quantitative	A		experimental (accuracy)	Sentiment Analysis; Argument Mining
A-ART	Fuzzy ART	DeLP	dialectical trees	A		explanation; inconsistency resolution	
AARL	SARSA	Value-based AA	extensions	D	✓	experimental (stability; convergence time; optimal performance)	RoboCup; Wumpus

Table 1. Overview of approaches using argumentation to aid ML (D=During, A=After, Apps. = Applications).

5. Conclusion

We have surveyed existing approaches using argumentation to aid ML, focusing on the type of ML method they augment, the form of arguments and argumentation frameworks and semantics they deploy, as well as their advantages, ranging from improving performances to rendering the ML process more transparent by improving its explanatory power. Table 1 summarises our analysis.

The existing approaches show promise for further future developments and substantial potential impact in ML, to improve performances and allow the incorporation of domain knowledge by users as well as user-friendly explanations and transparency of the output of ML.

References

- [1] Amgoud, L., Serrurier, M.: Agents that argue and explain classifications. *Autonomous Agents and Multi-Agent Systems* 16(2), 187–209 (2007)
- [2] Baroni, P., Romano, M., Toni, F., Aurisicchio, M., Bertanza, G.: Automatic evaluation of design alternatives with quantitative argumentation. *Argument and Computation* (2015)
- [3] Bench-Capon, T.J.M., Atkinson, K.: Abstract argumentation and values. In: Rahwan, L., Simari, G. (eds.) *Argumentation in Artificial Intelligence*. Springer (2009)
- [4] Bratko, I., Žabkar, J., Možina, M.: Argument-based machine learning. In: Simari, G., Rahwan, I. (eds.) *Argumentation in Artificial Intelligence*, pp. 463–482. Springer, 1st edn. (2009)
- [5] Breiman, L.: Random forests. *Machine Learning* 45(1), 5–32 (Oct 2001)
- [6] Bromuri, S., Urovi, V., Morge, M., Stathis, K., Toni, F.: A multi-agent system for service discovery, selection and negotiation. In: *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems - Volume 2*. pp. 1395–1396. AAMAS '09 (2009)
- [7] Carpenter, G.A., Grossberg, S., Rosen, D.B.: Fuzzy art: Fast stable learning and categorization of analog patterns by an adaptive resonance system. *Neural Networks* 4(6), 759–771 (Nov 1991)
- [8] Carstens, L., Toni, F.: Improving out-of-domain sentiment polarity classification using argumentation. In: *IEEE International Conference on Data Mining Workshop, ICDMW*. pp. 1294–1301 (2015)
- [9] Carstens, L., Toni, F.: Using Argumentation to improve classification in Natural Language problems. Ph.D. thesis, Imperial College London (2016)
- [10] Cayrol, C., Lagasque-Schiex, M.C.: Symbolic and Quantitative Approaches to Reasoning with Uncertainty: 8th European Conference, ECSQARU 2005, chap. On the Acceptability of Arguments in Bipolar Argumentation Frameworks, pp. 378–389. Springer (2005)
- [11] Chesñevar, C.I., Simari, G.R.: A lattice-based approach to computing warranted beliefs in skeptical argumentation frameworks. In: *Proceedings of the 20th International Joint Conference on Artificial Intelligence*. pp. 280–285. IJCAI'07 (2007)
- [12] Clark, P., Niblett, T.: The CN2 induction algorithm. *Machine Learning* 3(4), 261–283 (1989)
- [13] Cortes, C., Vapnik, V.: Support-vector networks. *Machine Learning* 20(3), 273–297 (Sep 1995)
- [14] Dung, P.M.: On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence* 77(2), 321 – 357 (1995)
- [15] Evripidou, V., Toni, F.: Argumentation and voting for an intelligent user empowering business directory on the web. In: *Web Reasoning and Rule Systems - 6th International Conference, RR*. pp. 209–212 (2012)
- [16] Fox, J., Glasspool, D., Grecu, D., Modgil, S., South, M., Patkar, V.: Argumentation-based inference and decision making—a medical perspective. *IEEE Intelligent Systems* 22(6), 34–41 (2007)
- [17] Gao, Y., Toni, F.: Argumentation accelerated reinforcement learning for robocup keepaway-takeaway. In: *Theory and Applications of Formal Argumentation - Second International Workshop, TAFE*. vol. 8306, pp. 79–94 (2013)
- [18] Gao, Y., Toni, F.: Argumentation accelerated reinforcement learning for cooperative multi-agent systems. In: *ECAI 2014 - 21st European Conference on Artificial Intelligence*. pp. 333–338 (2014)

- [19] Gao, Y., Toni, F.: Argumentation accelerated reinforcement learning. Ph.D. thesis, Imperial College London (2015)
- [20] García, A.J., Simari, G.R.: Defeasible logic programming: An argumentative approach. *Theory and Practice of Logic Programming* 4(1-2), 95–138 (2004)
- [21] Gómez, S.A., Chesñevar, C.I.: A hybrid approach to pattern classification using neural networks and defeasible argumentation. In: *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference*, Miami Beach, Florida, USA. pp. 393–398 (2004)
- [22] Grosse, K., González, M.P., Chesñevar, C.I., Maguitman, A.G.: Integrating argumentation and sentiment analysis for mining opinions from twitter. *AI Communications* 28(3), 387–401 (2015)
- [23] Groznik, V., Guid, M., Sadikov, A., Možina, M., Georgiev, D., Kragelj, V., Ribaric, S., Pirtosek, Z., Bratko, I.: Elicitation of neurological knowledge with ABML. In: *Artificial Intelligence in Medicine - 13th Conference on Artificial Intelligence in Medicine, AIME*. vol. 6747, pp. 14–23 (2011)
- [24] Guid, M., Možina, M., Groznik, V., Georgiev, D., Sadikov, A., Pirtosek, Z., Bratko, I.: ABML knowledge refinement loop: A case study. In: *Foundations of Intelligent Systems - 20th International Symposium, ISMIS*. vol. 7661, pp. 41–50 (2012)
- [25] John, G.H., Langley, P.: Estimating continuous distributions in bayesian classifiers. In: *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*. pp. 338–345. UAI'95 (1995)
- [26] Lippi, M., Torroni, P.: Argumentation mining: State of the art and emerging trends. *ACM Transactions on Internet Technology* 16(2), 10:1–10:25 (Mar 2016)
- [27] Mitchell, T.M.: *Machine Learning*. McGraw-Hill, Inc., 1 edn. (1997)
- [28] Možina, M., Giuliano, C., Bratko, I.: Argument based machine learning from examples and text. In: *First Asian Conference on Intelligent Information and Database Systems, ACIIDS*. pp. 18–23 (2009)
- [29] Možina, M., Guid, M., Krivec, J., Sadikov, A., Bratko, I.: Fighting knowledge acquisition bottleneck with argument based machine learning. In: *Proceedings of the 2008 Conference on ECAI 2008: 18th European Conference on Artificial Intelligence*. pp. 234–238 (2008)
- [30] Možina, M., Guid, M., Krivec, J., Sadikov, A., Bratko, I.: Learning to explain with ABML. In: *Explanation-aware Computing, Papers from the 2010 ECAI Workshop*. pp. 37–48 (2010)
- [31] Možina, M., Zabkar, J., Bench-Capon, T.J.M., Bratko, I.: Argument based machine learning applied to law. *Artificial Intelligence and Law* 13(1), 53–73 (2005)
- [32] Možina, M., Zabkar, J., Bratko, I.: Argument based machine learning. *Artificial Intelligence* 171(10-15), 922–937 (2007)
- [33] Murphy, P., Aha, D.: *UCI Repository of machine learning databases*. Tech. rep., University of California, Department of Information and Computer Science, Irvine, CA, US. (1994)
- [34] Ontañón, S., Dellunde, P., Godo, L., Plaza, E.: A defeasible reasoning model of inductive concept learning from examples and communication. *Artificial Intelligence* 193, 129–148 (2012)
- [35] Ontañón, S., Plaza, E.: Coordinated inductive learning using argumentation-based communication. *Autonomous Agents and Multi-Agent Systems* 29(2), 266–304 (2014)
- [36] Rago, A., Toni, F., Aurisicchio, M., Baroni, P.: Discontinuity-free decision support with quantitative argumentation debates. In: *Principles of Knowledge Representation and Reasoning: Proceedings of the Fifteenth International Conference*. pp. 63–73 (2016)
- [37] Rahwan, I., Simari, G.R.: *Argumentation in Artificial Intelligence*. Springer, 1st edn. (2009)
- [38] Rotstein, N.D., Moguillansky, M.O., Simari, G.R.: Dialectical abstract argumentation: A characterization of the marking criterion. In: *Proceedings of the 21st International Joint Conference on Artificial Intelligence*. pp. 898–903. IJCAI'09 (2009)
- [39] Rummery, G.A., Niranjan, M.: *On-line Q-learning using connectionist systems*. Tech. rep., Cambridge University Engineering Department (1994)
- [40] Zabkar, J., Možina, M., Videcnik, J., Bratko, I.: Argument based machine learning in a medical domain. In: *Computational Models of Argument: Proceedings of COMMA. Frontiers in Artificial Intelligence and Applications*, vol. 144, pp. 59–70 (2006)
- [41] Zapašek, M., Možina, M., Bratko, I., Rugelj, J., Guid, M.: *Intelligent Tutoring Systems: 12th International Conference*, chap. Designing an Interactive Teaching Tool with ABML Knowledge Refinement Loop, pp. 575–582 (2014)

On the Acceptability Semantics of Argumentation Frameworks with Recursive Attack and Support

Andrea COHEN¹ Sebastian GOTTIFREDI Alejandro J. GARCÍA
Guillermo R. SIMARI

ICIC, CONICET-UNS, Bahía Blanca, Argentina

Abstract. The Attack-Support Argumentation Framework (ASAF) is an abstract argumentation framework that provides a unified setting for representing attack and support for arguments, as well as attack and support for the attack and support relations at any level. Currently, the extensions of the ASAF are obtained by translating it into a Dung's Argumentation Framework (AF). In this work we provide the ASAF with the ability of determining its extensions without requiring such a translation. We follow an extension-based approach for characterizing the acceptability semantics directly on the ASAF, considering the complete, preferred, stable and grounded semantics. Finally, we show that the proposed characterization satisfies different results from Dung's argumentation theory.

Keywords. abstract argumentation, bipolar argumentation, recursive interactions, acceptability semantics.

1. Introduction

The study of Abstract Argumentation Frameworks (AFs) has proved to be of great interest within the argumentation community since they allow to explore different properties on arguments and their relationships, as well as providing various characterizations for their acceptability status [8,15]. Whereas Dung's AFs only account for the existence of an attack relation between arguments, in the last decade, several proposals have been developed in order to enrich such AFs with a positive interaction between arguments: a support relation. A first line of work on such AFs, called Bipolar Argumentation Frameworks (BAFs) in [3], introduced a general support relation between arguments and proposed a series of complex attacks [4] enforcing acceptability constraints derived from the coexistence of attacks and supports. Later, alternative interpretations for the notion of support were proposed, the most well-known being evidential support [12], deductive support [16] and necessary support [11].

Starting from [4] and [6], where different interpretations of support are compared and discussed, the interest in studying AFs with support relations has greatly increased. Furthermore, recent works have focused on a deeper study of the necessary support re-

¹Corresponding author: ac@cs.uns.edu.ar

lation (see [10,14,13,5]). For instance, in [14] the author gives an instantiation of necessary support in ASPIC+ using sub-arguments; and in [5] an axiomatization of necessary support is proposed through different frameworks.

Another line of work extending AFs that has gained attention amongst the researchers regards the consideration of high-order interactions. Motivated by [9], where second-order attacks are used for representing preferences between arguments, in [1] the authors proposed an AF with recursive attacks (AFRA). Moreover, in [16], the authors allow the attack and support relations of an AF to be attacked in order to model their defeasible nature. Further research on this area combined the above results by characterizing the *Attack-Support Argumentation Framework* (ASAF) [7], an AF that allows for attacks and supports between arguments, as well as attacks and supports from an argument to the attack and support relations, at any level.

A key feature of any argumentation system consists in determining the conditions under which the arguments are accepted, after accounting for their interactions [8,2]. A criticism on [7] is that such conditions are not specified directly on the ASAF; instead, the collectively acceptable sets of arguments are obtained by translating the ASAF into a Dung's AF. In this work we will provide the means for characterizing the acceptability semantics of the ASAF, hence addressing the above mentioned criticism. Since attacks and supports in an ASAF may be affected by other interactions, we will have to account for the conditions under which these attacks and supports are considered as accepted. Moreover, we will show that the characterization of the semantics proposed here satisfies properties given in [8] for Dung's argumentation theory.

The rest of this paper is organized as follows. Section 2 introduces some background notions, including definitions from Dung's theory [8] and the definition of the ASAF [7]. Then, Section 3 identifies conflicts between the elements of the ASAF, leading to the characterization of different kinds of defeat. Given those defeats, Section 4 starts by adapting Dung's basic semantic notions to then characterize the acceptability semantics of the ASAF. Finally, Section 5 discusses related work, presents some conclusions and comments on future lines of research.

2. Background

In this section we include the background required for characterizing the acceptability semantics of the ASAF. We first present some basic notions related to Dung's AFs [8] and then, the definition of the ASAF provided in [7].

The Abstract Argumentation Framework defined in [8] consists of a set of arguments and a set of conflicts between them:

Definition 1 (AF). *An abstract argumentation framework (AF) is a pair $\langle \mathbb{A}, \mathbb{R} \rangle$, where \mathbb{A} is finite and non-empty set of arguments and $\mathbb{R} \subseteq \mathbb{A} \times \mathbb{A}$ is an attack relation.*

Given an AF, [8] defines a series of semantic notions, leading to the characterization of collectively acceptable sets of arguments.

Definition 2 (Conflict-freeness, acceptability, admissibility). *Let $AF = \langle \mathbb{A}, \mathbb{R} \rangle$ and $\mathbf{S} \subseteq \mathbb{A}$. \mathbf{S} is conflict-free if $\nexists \mathcal{A}, \mathcal{B} \in \mathbf{S}$ s.t. $(\mathcal{A}, \mathcal{B}) \in \mathbb{R}$. $\mathcal{A} \in \mathbb{A}$ is acceptable w.r.t. \mathbf{S} if $\forall \mathcal{B} \in \mathbb{A}$ s.t. $(\mathcal{B}, \mathcal{A}) \in \mathbb{R}$, $\exists \mathcal{C} \in \mathbf{S}$ s.t. $(\mathcal{C}, \mathcal{B}) \in \mathbb{R}$. \mathbf{S} is admissible if it is conflict-free and $\forall \mathcal{A} \in \mathbf{S}$: \mathcal{A} is acceptable w.r.t. \mathbf{S} .*

Then, by adding restrictions to the notion of admissibility, the complete, preferred, stable and grounded extensions of an AF are defined as follows:

Definition 3 (AF Extensions). *Let $AF = \langle \mathbb{A}, \mathbb{R} \rangle$ and $\mathbf{S} \subseteq \mathbb{A}$. \mathbf{S} is a complete extension of AF iff it is admissible and $\forall \mathcal{A} \in \mathbb{A}$: if \mathcal{A} is acceptable w.r.t. \mathbf{S} , then $\mathcal{A} \in \mathbf{S}$. \mathbf{S} is a preferred extension of AF iff it is a maximal (w.r.t. \subseteq) admissible set of AF . \mathbf{S} is a stable extension of AF iff it is conflict-free and $\forall \mathcal{A} \in \mathbb{A} \setminus \mathbf{S}$: $\exists \mathcal{B} \in \mathbf{S}$ s.t. $(\mathcal{B}, \mathcal{A}) \in \mathbb{R}$. \mathbf{S} is the grounded extension of AF iff it is the smallest (w.r.t. \subseteq) complete extension of AF .*

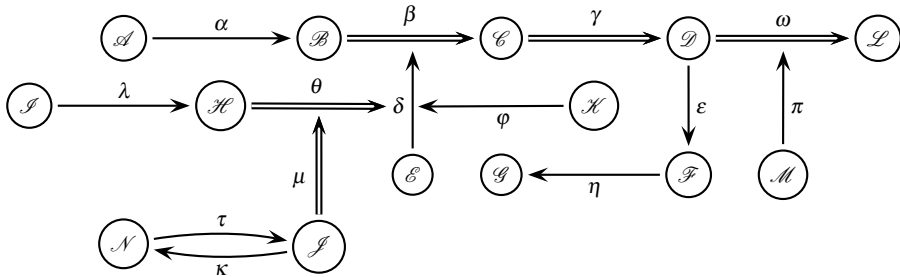
Next, we include the definition of the ASAF given in [7], corresponding to an AF with recursive attack and support relations.

Definition 4 (ASAF). *An Attack-Support Argumentation Framework (ASAF) is a tuple $\langle \mathbb{A}, \mathbb{R}, \mathbb{S} \rangle$ where \mathbb{A} is a set of arguments, $\mathbb{R} \subseteq \mathbb{A} \times (\mathbb{A} \cup \mathbb{R} \cup \mathbb{S})$ is an attack relation and $\mathbb{S} \subseteq \mathbb{A} \times (\mathbb{A} \cup \mathbb{R} \cup \mathbb{S})$ is a support relation. We assume that \mathbb{S} is acyclic and $\mathbb{R} \cap \mathbb{S} = \emptyset$.*

The support relation of the ASAF follows the necessity interpretation of the Argumentation Framework with Necessities (AFN) [11], where if \mathcal{A} supports \mathcal{B} then it means that the acceptance of \mathcal{A} is necessary to get the acceptance of \mathcal{B} ; in other words, the acceptance of \mathcal{B} implies the acceptance of \mathcal{A} or, equivalently, the non-acceptance of \mathcal{A} implies the non-acceptance of \mathcal{B} . As a result, the attack relation of the ASAF not only extends the attack relation of the AFRA [1] by allowing for attacks to the support relation, but it also extends the attack and support relations of the AFN [11] by allowing for recursive attacks and supports, as well as attacks to the support relation and vice-versa.

Given an attack or a support $\alpha = (\mathcal{A}, X) \in (\mathbb{R} \cup \mathbb{S})$, \mathcal{A} is called the source of α and X is called the target of α , and they can be referred to as $\text{src}(\alpha)$ and $\text{trg}(\alpha)$, respectively. Moreover, an ASAF can be graphically represented using a graph-like notation: an argument $\mathcal{A} \in \mathbb{A}$ will be denoted as a node in the graph, an attack $\alpha = (\mathcal{A}, X) \in \mathbb{R}$ will be denoted as $\mathcal{A} \xrightarrow{\alpha} X$, and a support $\beta = (\mathcal{B}, Y) \in \mathbb{S}$ will be denoted as $\mathcal{B} \xrightarrow{\beta} Y$. To simplify the notation, the attack from an argument \mathcal{C} to an attack or a support $\alpha = (\mathcal{A}, X)$ will be referred to as (\mathcal{C}, α) . Similarly, the support from an argument \mathcal{D} to an attack or a support $\beta = (\mathcal{B}, Y)$ will be referred to as (\mathcal{D}, β) . Since, as mentioned before, the attack and support relations of an ASAF are assumed to be disjoint, a pair $\gamma = (\mathcal{E}, Z)$ in the attack relation or the support relation will be unequivocally identified by γ . Thus, when referring to γ , it will be possible to identify the attack or support it represents. To illustrate this, let us consider the following example.

Example 1. *Let us consider the ASAF Δ_1 with the following graphical representation:*



We have the first-level attacks $\alpha = (\mathcal{A}, \mathcal{B})$, $\varepsilon = (\mathcal{D}, \mathcal{F})$, $\eta = (\mathcal{F}, \mathcal{G})$, $\lambda = (\mathcal{J}, \mathcal{H})$, $\tau = (\mathcal{N}, \mathcal{J})$ and $\kappa = (\mathcal{J}, \mathcal{N})$. The first-level supports are $\beta = (\mathcal{B}, \mathcal{C})$, $\gamma = (\mathcal{C}, \mathcal{D})$ and $\omega = (\mathcal{D}, \mathcal{L})$. The second-level interactions are the attacks $\delta = (\mathcal{E}, \beta)$ and $\pi = (\mathcal{M}, \omega)$. Then, we have the third-level attack and support on δ : respectively, $\varphi = (\mathcal{K}, \delta)$ and $\theta = (\mathcal{K}, \delta)$. Finally, the only fourth-level interaction is the support $\mu = (\mathcal{J}, \theta)$.

3. Defeats in the ASAF

Before characterizing the acceptability semantics of the ASAF we need to clearly identify all the conflicts between its elements, in addition to those already expressed in the attack relation. The set of all conflicts between the elements of the ASAF will be called the set of *defeats*, in order to distinguish them from the original attacks. In particular, similarly to [1], we consider a notion of defeat which regards attacks, rather than their source arguments, as the subjects able to defeat arguments, attacks or supports.

In the following we will distinguish between two types of defeats: those that can be inferred directly by looking at the attack relation of the ASAF, and those that are conditioned by the existence of supports. The former will be referred to as *unconditional defeats*, and are defined in Section 3.1, whereas the latter are the *conditional defeats*, defined in Section 3.2.

3.1. Unconditional Defeats

The first case of unconditional defeats corresponds to conflicts already captured by the attack relation of the ASAF, which we call *direct defeats*.

Definition 5 (Direct Defeat). *Let $\Delta = \langle \mathbb{A}, \mathbb{R}, \mathbb{S} \rangle$ be an ASAF, $\alpha \in \mathbb{R}$ and $X \in (\mathbb{A} \cup \mathbb{R} \cup \mathbb{S})$. We say that α directly defeats X , noted α d-def X , iff $\text{trg}(\alpha) = X$.*

The other kind of defeat that may be inferred directly from the attack relation of the ASAF is the *indirect defeat*, which captures the intuition that attacks are strictly related to their source, as in the AFRA [1].

Definition 6 (Indirect Defeat). *Let $\Delta = \langle \mathbb{A}, \mathbb{R}, \mathbb{S} \rangle$ be an ASAF and $\alpha, \beta \in \mathbb{R}$. We say that α indirectly defeats β , noted α i-def β , iff α d-defsrc(β).*

These two kinds of unconditional defeat are grouped together in the following definition and illustrated below.

Definition 7 (Unconditional Defeat). *Let $\Delta = \langle \mathbb{A}, \mathbb{R}, \mathbb{S} \rangle$ be an ASAF, $\alpha \in \mathbb{R}$ and $X \in (\mathbb{A} \cup \mathbb{R} \cup \mathbb{S})$. We say that α unconditionally defeats X , noted α u-def X , iff α d-def X or α i-def X .*

Example 2. *Given the ASAF Δ_1 from Ex. 1, the following unconditional defeats occur. The direct defeats: α d-def \mathcal{B} , ε d-def \mathcal{F} , η d-def \mathcal{G} , λ d-def \mathcal{H} , τ d-def \mathcal{J} , κ d-def \mathcal{N} , δ d-def β , φ d-def δ , π d-def ω ; and the indirect defeats: ε i-def η , τ i-def κ , κ i-def τ .*

3.2. Conditional Defeats

As mentioned before, the coexistence of attacks and supports may lead to having additional conflicts between the elements of the ASAF. These conflicts will be identified as *conditional defeats* since, unlike the defeats defined in Section 3.1, their existence depends on the consideration of the support relation of the ASAF. Following the necessary interpretation of support, such conflicts are handled in [11] by characterizing the notion of *extended attack*, which reinforces the acceptability constraints presented in Section 2: given an attack $\mathcal{A} \rightarrow \mathcal{B}$ and a sequence of necessary supports $\mathcal{B} \Rightarrow \dots \Rightarrow \mathcal{C}$, there is an extended attack from \mathcal{A} to \mathcal{C} .

The intuitions presented above are captured in the ASAF by defining the notion of *extended defeat*. In particular, we will distinguish the *support sequence* involved in this kind of defeat, and the corresponding supports will be referred to as the *support set*.

Definition 8 (Support Sequence and Support Set). *Let $\Delta = \langle \mathbb{A}, \mathbb{R}, \mathbb{S} \rangle$ be an ASAF and $X \in (\mathbb{A} \cup \mathbb{R} \cup \mathbb{S})$. We say that $\Sigma = [\mathcal{A}_1, \dots, \mathcal{A}_n]$ is a support sequence for X ($n \geq 2$) iff $\mathcal{A}_n = X$ and for every \mathcal{A}_i ($1 \leq i \leq n-1$) it holds that $(\mathcal{A}_i, \mathcal{A}_{i+1}) \in \mathbb{S}$. We define the support set of Σ as $\mathbf{S} = \bigcup_{i=1}^{n-1} S_i$, with $S_i = (\mathcal{A}_i, \mathcal{A}_{i+1})$.*

Definition 9 (Extended Defeat). *Let $\Delta = \langle \mathbb{A}, \mathbb{R}, \mathbb{S} \rangle$ be an ASAF, $\alpha \in \mathbb{R}$, $X \in (\mathbb{A} \cup \mathbb{R} \cup \mathbb{S})$ and $\mathbf{S} \subseteq \mathbb{S}$. We say that α extendedly defeats X given \mathbf{S} , noted α e-def X given \mathbf{S} , if exists a support sequence $\Sigma = [\mathcal{A}_1, \dots, X]$ for X s.t. $\text{trg}(\alpha) = \mathcal{A}_1$ and \mathbf{S} is the support set of Σ .*

Extended defeats in the ASAF are illustrated by the following example.

Example 3. *Let Δ_1 be the ASAF from Ex. 1. Then, we have the following extended defeats: α e-def \mathcal{C} given $\{\beta\}$, α e-def \mathcal{D} given $\{\beta, \gamma\}$, α e-def \mathcal{L} given $\{\beta, \gamma, \omega\}$, λ e-def δ given $\{\theta\}$, and τ e-def θ given $\{\mu\}$.*

It can be noted that Def. 9 explicitly identifies the support sequence originating the extended defeat. Therefore, as shown by the following Proposition, adding a support link to a support sequence results in a new extended defeat.

Proposition 1. *Let $\Delta = \langle \mathbb{A}, \mathbb{R}, \mathbb{S} \rangle$ be an ASAF, $\alpha \in \mathbb{R}$, $\beta \in \mathbb{S}$ and $\mathbf{S} \subseteq \mathbb{S}$. If α e-def $\text{src}(\beta)$ given \mathbf{S} , then α e-def $\text{trg}(\beta)$ given $\mathbf{S} \cup \{\beta\}$.*

Proof. If α e-def $\text{src}(\beta)$ given \mathbf{S} , then, by Def. 9, there exists a support sequence $\Sigma = [\mathcal{A}_1, \dots, \text{src}(\beta)]$ for $\text{src}(\beta)$ s.t. \mathbf{S} is the support set of Σ . Since by hyp. $\beta = (\text{src}(\beta), \text{trg}(\beta)) \in \mathbb{S}$, by Def. 8, $\Sigma' = [\mathcal{A}_1, \dots, \text{src}(\beta), \text{trg}(\beta)]$ is a support sequence for $\text{trg}(\beta)$ and $\mathbf{S} \cup \{\beta\}$ is the support set of Σ' . Thus, by Def. 9, α e-def $\text{trg}(\beta)$ given $\mathbf{S} \cup \{\beta\}$. \square

Given that the ASAF combines intuitions and results from the AFRA [1] and the AFN [11], it is reasonable to combine the intuitions behind the notions of indirect defeat and extended defeat to identify additional conflicts between the elements of the ASAF. In other words, similarly to the indirect defeat, we define the notion of *extended-indirect defeat* where an extended defeat on an argument is propagated to the attacks it originates. This kind of defeat is also conditional since it relies on the existence of an extended defeat, hence on the existence of supports.

Definition 10 (Extended-Indirect Defeat). *Let $\Delta = \langle \mathbb{A}, \mathbb{R}, \mathbb{S} \rangle$ be an ASAF, $\alpha, \beta \in \mathbb{R}$ and $\mathbf{S} \subseteq \mathbb{S}$. We say that α extended-indirectly defeats β given \mathbf{S} , noted α ei-def β given \mathbf{S} , iff α e-def src(β) given \mathbf{S} .*

This is illustrated by the following example.

Example 4. *Given the ASAF Δ_1 from Ex. 1, the only extended-indirect defeat is α ei-def ε given $\{\beta, \gamma\}$. This is because, as shown in Ex. 3, α e-def \mathcal{D} given $\{\beta, \gamma\}$ and, as it can be observed in Ex. 1, $\mathcal{D} = \text{src}(\varepsilon)$.*

Then, similarly to the case of unconditional defeats, the extended and extended-indirect defeats are grouped together in the following definition.

Definition 11 (Conditional Defeat). *Let $\Delta = \langle \mathbb{A}, \mathbb{R}, \mathbb{S} \rangle$ be an ASAF, $\alpha \in \mathbb{R}$, $X \in (\mathbb{A} \cup \mathbb{R} \cup \mathbb{S})$ and $\mathbf{S} \subseteq \mathbb{S}$. We say that α conditionally defeats X given \mathbf{S} , noted α c-def X given \mathbf{S} , iff α e-def X given \mathbf{S} or α ei-def X given \mathbf{S} .*

4. Acceptability Semantics of the ASAF

Having identified the situations in which defeats between the elements of the ASAF occur, in this section we will characterize the acceptability semantics of the ASAF following an extension-based approach. In particular, as stated in [7], the extensions of the ASAF may not only include arguments, but also attacks and supports. This is to reflect the fact that attacks and supports may be affected by other interactions and thus, the presence of an attack or a support in an extension of the ASAF will imply that it is “active”.

Following the methodology of [8], in Section 4.1 we will first define some basic semantic notions for the ASAF. In particular, we will show that the notion of acceptability complies with the constraints imposed by the attack and support relations of the ASAF. Moreover, we will show that results from [8] regarding the notions of acceptability and admissibility also hold for the ASAF. Then, in Section 4.2, we will define the acceptability semantics of the ASAF by characterizing its complete, preferred, stable and grounded extensions. Furthermore, we will show that the ASAF satisfies the relationships between the complete, preferred, stable and grounded extensions given in [8].

4.1. Semantic Notions

Analogously to [8], the notion of conflict-freeness establishes the minimum requirements a set of elements of the ASAF should satisfy in order to be collectively accepted.

Definition 12 (Conflict-Freeness). *Let $\Delta = \langle \mathbb{A}, \mathbb{R}, \mathbb{S} \rangle$ be an ASAF and $\mathbf{S} \subseteq (\mathbb{A} \cup \mathbb{R} \cup \mathbb{S})$. We say that \mathbf{S} is conflict-free iff:*

- $\nexists \alpha, X \in \mathbf{S}$ s.t. α u-def X ; and
- $\nexists \beta, Y \in \mathbf{S}$, $\nexists \mathbf{S}' \subseteq \mathbf{S}$ s.t. β c-def Y given \mathbf{S}' .

Example 5. *Let Δ_1 be the ASAF from Ex. 1. Some conflict-free sets of Δ_1 are: \emptyset , $\{\mathcal{M}, \omega\}$, $\{\mathcal{N}, \mathcal{J}\}$, $\{\lambda, \delta\}$, $\{\mu, \mathcal{E}, \delta\}$, $\{\alpha, \beta, \varepsilon\}$, $\{\mathcal{A}, \mathcal{B}, \mathcal{C}, \mathcal{D}, \mathcal{E}, \mathcal{F}, \mathcal{G}, \mathcal{H}, \mathcal{I}, \mathcal{J}, \mathcal{K}, \mathcal{L}, \mathcal{M}, \mathcal{N}, \beta, \gamma, \omega, \theta, \mu\}$ and $\{\mathcal{A}, \alpha, \gamma, \mathcal{M}, \pi, \mathcal{L}, \mathcal{I}, \lambda, \mathcal{K}, \varphi, \beta, \mathcal{F}, \eta, \mathcal{E}, \mu\}$. In contrast, the sets $\{\alpha, \mathcal{B}\}$, $\{\lambda, \theta, \delta\}$, $\{\pi, \omega\}$ and $\{\tau, \kappa\}$, among others, are not conflict-free.*

As expressed in Def. 12, if a set \mathbf{S} includes all the elements required for the existence of a defeat in the ASAF, then \mathbf{S} will not be conflict-free. This implies that, in particular, any set of elements of an ASAF which does not include an attack will be conflict-free. This is the case of the set $\{\mathcal{A}, \mathcal{B}, \mathcal{C}, \mathcal{D}, \mathcal{E}, \mathcal{F}, \mathcal{G}, \mathcal{H}, \mathcal{I}, \mathcal{J}, \mathcal{K}, \mathcal{L}, \mathcal{M}, \mathcal{N}, \beta, \gamma, \omega, \theta, \mu\}$ illustrated in Ex. 5, which includes every argument and support of the ASAF Δ_1 but none of its attacks. Moreover, when considering conditional defeats, all the elements required for the existence of a defeat must be included in a non-conflict-free set. Hence, if one of the supports in the corresponding support sequence is missing, the resulting set is conflict-free. This situation is illustrated by the conflict-free sets $\{\lambda, \delta\}$ and $\{\alpha, \beta, \varepsilon\}$ in Ex. 5.

Then, we define the notion of acceptability in the context of an ASAF, which characterizes the defense by a set of arguments, attacks and supports against the occurrence of defeats on its elements. Hence, since the ASAF allows for unconditional and conditional defeats, we need to consider all the defeats that may occur, as well as the different ways for providing defense against them.

Definition 13 (Acceptability). *Let $\Delta = \langle \mathbb{A}, \mathbb{R}, \mathbb{S} \rangle$ be an ASAF, $X \in (\mathbb{A} \cup \mathbb{R} \cup \mathbb{S})$ and $\mathbf{S} \subseteq (\mathbb{A} \cup \mathbb{R} \cup \mathbb{S})$. We say that X is acceptable w.r.t. \mathbf{S} iff it holds that:*

1. $\forall \alpha \in \mathbb{R}$ s.t. α u-def X , either:
 - (a) $\exists \beta \in \mathbf{S}$ s.t. β u-def α ; or
 - (b) $\exists \beta \in \mathbf{S}, \exists \mathbf{S}' \subseteq \mathbf{S}$ s.t. β c-def α given \mathbf{S}' .
2. $\forall \alpha \in \mathbb{R}, \forall \mathbf{T} \subseteq \mathbb{S}$ s.t. α c-def X given \mathbf{T} , either:
 - (a) $\exists \beta \in \mathbf{S}$ s.t. β u-def α ;
 - (b) $\exists \beta \in \mathbf{S}, \exists \gamma \in \mathbf{T}$ s.t. β u-def γ ;
 - (c) $\exists \beta \in \mathbf{S}, \exists \mathbf{S}' \subseteq \mathbf{S}$ s.t. β c-def α given \mathbf{S}' ; or
 - (d) $\exists \beta \in \mathbf{S}, \exists \mathbf{S}' \subseteq \mathbf{S}, \exists \gamma \in \mathbf{T}$ s.t. β c-def γ given \mathbf{S}' .

As the preceding definition shows, defense against an unconditional defeat may only be achieved by defeating the corresponding attack. On the other hand, a conditional defeat may be repelled by defeating the corresponding attack or one of the supports required by the conditional defeat. In either case, defense can be provided by both unconditional and conditional defeats. Moreover, it should be noted that, although Def. 13 accounts for a set \mathbf{S} of arguments, attacks and supports of an ASAF, the only elements contributing to the defense are the attacks and supports. This is because attacks and supports are ones leading to the existence of defeats (see Defs. 7 and 11). In other words, similarly to the AFRA, defense through an unconditional defeat can only be provided by an attack. In contrast, defense by a conditional defeat is given by an attack and a set of supports. These intuitions are illustrated in the following example.

Example 6. *For instance, given the ASAF Δ_1 from Ex. 1, \mathcal{A} and φ are acceptable w.r.t. \emptyset , β is acceptable w.r.t. $\{\varphi\}$, \mathcal{N} is acceptable w.r.t. $\{\tau\}$, \mathcal{D} is acceptable w.r.t. $\{\delta\}$, θ is acceptable w.r.t. $\{\kappa\}$, and \mathcal{F} and η are acceptable w.r.t. $\{\alpha, \beta, \gamma\}$. In contrast, for example, \mathcal{B} is not acceptable w.r.t. \emptyset and δ is not acceptable w.r.t. $\{\kappa\}$.*

The following proposition shows that, like in the AFRA, the acceptability of an attack implies the acceptability of its source.

Proposition 2. Let $\Delta = \langle \mathbb{A}, \mathbb{R}, \mathbb{S} \rangle$ be an ASAF, $\alpha \in \mathbb{R}$ and $\mathbf{S} \subseteq (\mathbb{A} \cup \mathbb{R} \cup \mathbb{S})$. If α is acceptable w.r.t. \mathbf{S} , then $\text{src}(\alpha)$ is acceptable w.r.t. \mathbf{S} .

Proof. Suppose by contradiction that α is acceptable w.r.t. \mathbf{S} and $A = \text{src}(\alpha)$ is not acceptable w.r.t. \mathbf{S} . Then, either (a) $\exists \beta \in \mathbb{R}$ s.t. β u-def A , and $\nexists \gamma \in \mathbf{S}$, $\nexists \mathbf{S}' \subseteq \mathbf{S}$ s.t. γ u-def β or γ c-def β given \mathbf{S}' ; or (b) $\exists \beta \in \mathbb{R}$, $\exists \mathbf{T} \subseteq \mathbb{S}$ s.t. β c-def A given \mathbf{T} , and $\nexists \gamma \in \mathbf{S}$, $\nexists \mathbf{S}' \subseteq \mathbf{S}$, $\nexists \delta \in \mathbf{T}$ s.t. γ u-def β , γ c-def β given \mathbf{S}' , γ u-def δ or γ c-def δ given \mathbf{S}' .

- (a) By Def. 4, it holds that $A = \text{src}(\alpha) \in \mathbb{A}$. Then, if β u-def A , by Defs. 7 and 5, it must be the case that β d-def A . Therefore, by Def. 6, β i-def α .
- (b) By Def. 4, it holds that $A = \text{src}(\alpha) \in \mathbb{A}$. Then, if β c-def A given \mathbf{T} , by Defs. 11 and 9, β e-def A given \mathbf{T} . Therefore, by Def. 10, β ei-def α given \mathbf{T} .

Then, by Def. 13, α would not be acceptable w.r.t. \mathbf{S} , contradicting the hypothesis. \square

The following proposition shows that the notion of acceptability complies with the constraints imposed by the necessary interpretation of support adopted by the ASAF.

Proposition 3. Let $\Delta = \langle \mathbb{A}, \mathbb{R}, \mathbb{S} \rangle$ be an ASAF, $\mathbf{S} \subseteq (\mathbb{A} \cup \mathbb{R} \cup \mathbb{S})$ a conflict-free set and $\alpha \in \mathbb{S}$ acceptable w.r.t. \mathbf{S} . If $\text{trg}(\alpha)$ is acceptable w.r.t. \mathbf{S} , then $\text{src}(\alpha)$ is acceptable w.r.t. \mathbf{S} ; equivalently, if $\text{src}(\alpha)$ is not acceptable w.r.t. \mathbf{S} , then $\text{trg}(\alpha)$ is not acceptable w.r.t. \mathbf{S} .

Proof. If $A = \text{src}(\alpha)$ is not acceptable w.r.t. \mathbf{S} , then it holds that either (a) $\exists \beta \in \mathbb{R}$ s.t. β u-def A , and $\nexists \gamma \in \mathbf{S}$, $\nexists \mathbf{S}' \subseteq \mathbf{S}$ s.t. γ u-def β or γ c-def β given \mathbf{S}' ; or (b) $\exists \beta \in \mathbb{R}$, $\exists \mathbf{T} \subseteq \mathbb{S}$ s.t. β c-def A given \mathbf{T} , and $\nexists \gamma \in \mathbf{S}$, $\nexists \mathbf{S}' \subseteq \mathbf{S}$, $\nexists \delta \in \mathbf{T}$ s.t. γ u-def β , γ c-def β given \mathbf{S}' , γ u-def δ or γ c-def δ given \mathbf{S}' .

- (a) By Def. 4, it holds that $A = \text{src}(\alpha) \in \mathbb{A}$. Then, if β u-def A , by Defs. 7 and 5, it must be the case that β d-def A . Therefore, by Def. 9, β e-def $\text{trg}(\alpha)$ given $\{\alpha\}$.
- (b) By Def. 4, it holds that $A = \text{src}(\alpha) \in \mathbb{A}$. Then, if β c-def A given \mathbf{T} , by Defs. 11 and 9, it must be the case that β e-def A given \mathbf{T} . Therefore, by Prop. 1, β e-def $\text{trg}(\alpha)$ given $\mathbf{T} \cup \{\alpha\}$.

Since by hyp. α is acceptable w.r.t. \mathbf{S} and \mathbf{S} is conflict-free, $\nexists \lambda \in \mathbf{S}$, $\nexists \mathbf{S}'' \subseteq \mathbf{S}$ s.t. λ u-def α or λ c-def α given \mathbf{S}'' . As a result, by Def. 13, $\text{trg}(\alpha)$ is not acceptable w.r.t. \mathbf{S} . \square

The following proposition shows that the notion of acceptability is monotonic with respect to set inclusion.

Proposition 4. Let $\Delta = \langle \mathbb{A}, \mathbb{R}, \mathbb{S} \rangle$ be an ASAF, $X \in (\mathbb{A} \cup \mathbb{R} \cup \mathbb{S})$ and $\mathbf{S} \subseteq (\mathbb{A} \cup \mathbb{R} \cup \mathbb{S})$. If X is acceptable w.r.t. \mathbf{S} , then $\forall \mathbf{S}' \subseteq (\mathbb{A} \cup \mathbb{R} \cup \mathbb{S})$ s.t. $\mathbf{S} \subseteq \mathbf{S}'$: X is acceptable w.r.t. \mathbf{S}' .

Proof. Suppose by contradiction that X is acceptable w.r.t. \mathbf{S} and $\exists \mathbf{S}' \subseteq (\mathbb{A} \cup \mathbb{R} \cup \mathbb{S})$ s.t. $\mathbf{S} \subseteq \mathbf{S}'$ and X is not acceptable w.r.t. \mathbf{S}' . Then, it holds that either (a) $\exists \alpha \in \mathbb{R}$ s.t. α u-def X and $\nexists \beta \in \mathbf{S}'$, $\nexists \mathbf{S}'' \subseteq \mathbf{S}'$ s.t. β u-def α or β c-def α given \mathbf{S}'' ; or (b) $\exists \alpha \in \mathbb{R}$, $\exists \mathbf{T} \subseteq \mathbb{S}$ s.t. α c-def X given \mathbf{T} and $\nexists \beta \in \mathbf{S}'$, $\nexists \mathbf{S}'' \subseteq \mathbf{S}'$, $\nexists \gamma \in \mathbf{T}$ s.t. β u-def α , β c-def α given \mathbf{S}'' , β u-def γ or β c-def γ given \mathbf{S}'' . Thus, since $\mathbf{S} \subseteq \mathbf{S}'$, by Def. 13, X would not be acceptable w.r.t. \mathbf{S} , contradicting the hypothesis. \square

Next, like in [8], admissible sets of the ASAF are defined by combining the notions of conflict-freeness and acceptability.

Definition 14 (Admissibility). Let $\Delta = \langle \mathbb{A}, \mathbb{R}, \mathbb{S} \rangle$ be an ASAF and $\mathbf{S} \subseteq (\mathbb{A} \cup \mathbb{R} \cup \mathbb{S})$. We say that \mathbf{S} is admissible iff it is conflict-free and $\forall X \in \mathbf{S}: X$ is acceptable w.r.t. \mathbf{S} .

Example 7. Let Δ_1 be the ASAF from Ex. 1. Some admissible sets of Δ_1 are \emptyset , $\{\alpha, \beta, \gamma, \varphi, \mathcal{F}, \mathcal{M}\}$ and $\{\mathcal{A}, \alpha, \gamma, \mathcal{M}, \pi, \mathcal{L}, \mathcal{J}, \lambda, \mathcal{H}, \varphi, \beta, \mathcal{F}, \eta, \mathcal{E}, \mu, \tau, \mathcal{N}\}$. In contrast, for instance, the sets $\{\beta, \theta, \lambda, \mathcal{J}, \kappa\}$ and $\{\varepsilon, \mathcal{G}\}$ are not admissible; the former because β is not defended against the direct defeat by δ , whereas the latter because ε is not defended against the extended-indirect defeat by α .

The following proposition shows that the notions of acceptability and admissibility allow for Dung's fundamental lemma to hold in the context of an ASAF.

Lemma 1. Let $\Delta = \langle \mathbb{A}, \mathbb{R}, \mathbb{S} \rangle$ be an ASAF, $\mathbf{S} \subseteq (\mathbb{A} \cup \mathbb{R} \cup \mathbb{S})$ an admissible set of Δ , and $X, Y \in (\mathbb{A} \cup \mathbb{R} \cup \mathbb{S})$ s.t. X and Y are acceptable w.r.t. \mathbf{S} . Then, it holds that (1) $\mathbf{S}' = \mathbf{S} \cup \{X\}$ is admissible, and (2) Y is acceptable w.r.t. \mathbf{S}' .

Proof.

1. To prove that \mathbf{S}' is admissible we have to prove that X is acceptable w.r.t. \mathbf{S}' and \mathbf{S}' is conflict-free. Since $\mathbf{S} \subseteq \mathbf{S}'$ and, by hypothesis, X is acceptable w.r.t. \mathbf{S} , by Prop. 4, X is acceptable w.r.t. \mathbf{S}' . Now, suppose by contradiction that \mathbf{S}' is not conflict-free. Then, since by hypothesis \mathbf{S} is admissible, it must be the case that $\exists W, Z \in \mathbf{S}, \exists \mathbf{T} \subseteq \mathbf{S}$ s.t. either (a) X u-def W ; (b) W u-def X ; (c) X c-def W given \mathbf{T} ; (d) W c-def X given \mathbf{T} ; or (e) W c-def Z given $\mathbf{T} \cup \{X\}$.

(a) If X u-def W , since by hypothesis \mathbf{S} is admissible, it must be the case that $\exists \alpha \in \mathbf{S}, \exists \mathbf{S}_1 \subseteq \mathbf{S}$ s.t. α u-def X or α c-def X given \mathbf{S}_1 . Furthermore, since by hypothesis X is acceptable w.r.t. \mathbf{S} , it must be the case that $\exists \beta \in \mathbf{S}, \exists \mathbf{S}_2 \subseteq \mathbf{S}, \exists \gamma \in \mathbf{S}_1$ s.t. β u-def α , β c-def α given \mathbf{S}_2 , β u-def γ , or β c-def γ given \mathbf{S}_2 . As a result, the set \mathbf{S} would not be conflict-free, contradicting the hypothesis that \mathbf{S} is admissible.

(b) If W u-def X , since by hypothesis X is acceptable w.r.t. \mathbf{S} , then $\exists \alpha \in \mathbf{S}, \exists \mathbf{S}_1 \subseteq \mathbf{S}$ s.t. α u-def W or α c-def W given \mathbf{S}_1 . As a result, in each case, the set \mathbf{S} would not be conflict-free, contradicting the hypothesis that \mathbf{S} is admissible.

(c) If X c-def W given \mathbf{T} , since by hypothesis \mathbf{S} is admissible, it must be the case that $\exists \alpha \in \mathbf{S}, \exists \mathbf{S}_1 \subseteq \mathbf{S}, \exists \gamma \in \mathbf{T}$ s.t. either (i) α u-def X , (ii) α c-def X given \mathbf{S}_1 , (iii) α u-def γ or (iv) α c-def γ given \mathbf{S}_1 . Cases (c.i) and (c.ii) are analogous to case (b) and thus, \mathbf{S} would not be conflict-free, contradicting the hypothesis that \mathbf{S} is admissible. In cases (c.iii) and (c.iv), since $\alpha \in \mathbf{S}, \gamma \in \mathbf{T} \subseteq \mathbf{S}$ and $\mathbf{S}_1 \subseteq \mathbf{S}$, the set \mathbf{S} would not be conflict-free, contradicting the hypothesis that \mathbf{S} is admissible.

(d) This case is analogous to case (b) and thus, \mathbf{S} would not be conflict-free, contradicting the hypothesis that \mathbf{S} is admissible.

(e) If W c-def Z given $\mathbf{T} \cup \{X\}$, since by hypothesis \mathbf{S} is admissible, then $\exists \alpha \in \mathbf{S}, \exists \mathbf{S}_1 \subseteq \mathbf{S}, \exists \gamma \in \mathbf{T}$ s.t. either (i) α u-def W , (ii) α c-def W given \mathbf{S}_1 , (iii) α u-def γ , (iv) α c-def γ given \mathbf{S}_1 , (v) α u-def X or (vi) α c-def X given \mathbf{S}_1 . Thus, in cases (e.i)-(e.iv), the set \mathbf{S} would not be conflict-free, contradicting the hypothesis that \mathbf{S} is admissible. In cases (e.v) and (e.vi), similarly to case (a), since by hypothesis X is acceptable w.r.t. \mathbf{S} , it would be the case that $\exists \beta \in \mathbf{S}, \exists \mathbf{S}_2 \subseteq \mathbf{S}, \exists \lambda \in \mathbf{S}_1$ s.t. β u-def α , β c-def α given \mathbf{S}_2 , β u-def λ or β c-def λ given \mathbf{S}_2 ; in all cases, the set \mathbf{S} would not be conflict-free, contradicting the hypothesis that \mathbf{S} is admissible.

2. Since $\mathbf{S} \subseteq \mathbf{S}'$ and, by hyp., Y is acceptable w.r.t. \mathbf{S} , by Prop. 4, Y is acceptable w.r.t. \mathbf{S}' . \square

4.2. Extensional Semantics of the ASAF

Starting from the semantic notions defined in Section 4.1, we characterize the complete, preferred, stable and grounded extensions of the ASAF as follows.

Definition 15 (ASAF Extensions). *Let $\Delta = \langle \mathbb{A}, \mathbb{R}, \$ \rangle$ be an ASAF and $\mathbf{S} \subseteq (\mathbb{A} \cup \mathbb{R} \cup \$)$.*

- *\mathbf{S} is a complete extension of Δ iff it is admissible and $\forall X \in (\mathbb{A} \cup \mathbb{R} \cup \$)$: if X is acceptable w.r.t. \mathbf{S} , then $X \in \mathbf{S}$.*
- *\mathbf{S} is a preferred extension of Δ iff it is a maximal (w.r.t. \subseteq) admissible set of Δ .*
- *\mathbf{S} is a stable extension of Δ iff it is conflict-free and $\forall X \in (\mathbb{A} \cup \mathbb{R} \cup \$) \setminus \mathbf{S}$: $\exists \alpha \in \mathbf{S}$, $\exists \mathbf{S}' \subseteq \mathbf{S}$ s.t. α u-def X or α c-def X given \mathbf{S}' .*
- *\mathbf{S} is the grounded extension of Δ iff it is the smallest (w.r.t. \subseteq) complete extension of Δ .*

Example 8. *Let us consider the ASAF Δ_1 from Ex. 1 and the grounded and preferred semantics. The grounded extension of Δ_1 is $\mathbf{G} = \{\mathcal{A}, \alpha, \gamma, \mathcal{M}, \pi, \mathcal{L}, \mathcal{I}, \lambda, \mathcal{K}, \varphi, \beta, \mathcal{F}, \eta, \mathcal{E}, \mu\}$. Note that although $\text{src}(\mu)$ is involved in an attack cycle that is not resolved when considering the grounded semantics, the support μ is active and thus, $\mu \in \mathbf{G}$. Then, when considering the preferred semantics, there are two alternatives for resolving the attack cycle involving $\text{src}(\mu)$, leading to the existence of two preferred extensions of Δ_1 : $\mathbf{P}_1 = \mathbf{G} \cup \{\tau, \mathcal{N}\}$ and $\mathbf{P}_2 = \mathbf{G} \cup \{\kappa, \mathcal{J}, \theta\}$. In particular, even though $\{\tau, \mu\} \subseteq \mathbf{P}_1$ defends δ against the extended defeat by λ given $\{\theta\}$, \mathbf{P}_1 does not defend δ against the direct defeat by φ ; therefore, $\delta \notin \mathbf{P}_1$.*

Next, we will show that the ASAF semantics from Def. 15 fulfill the relationships between the corresponding semantics proposed in [8]. The following lemma illustrates the relationship between the preferred and complete extensions of an ASAF.

Lemma 2. *Let $\Delta = \langle \mathbb{A}, \mathbb{R}, \$ \rangle$ be an ASAF. Every preferred extension of Δ is also a complete extension of Δ , but not vice-versa.*

Proof. Suppose that $\exists \mathbf{S} \subseteq (\mathbb{A} \cup \mathbb{R} \cup \$)$ s.t. \mathbf{S} is a preferred extension of Δ but not a complete extension of Δ . Then, by Def. 15, it would be the case that $\exists X \in (\mathbb{A} \cup \mathbb{R} \cup \$)$ s.t. X is acceptable w.r.t. \mathbf{S} and $X \notin \mathbf{S}$. By Lemma 1, $\mathbf{S} \cup \{X\}$ is admissible. Therefore, \mathbf{S} would not be a maximal admissible set, contradicting the assumption that \mathbf{S} is a preferred extension of Δ . To show that the reverse does not hold let us consider the ASAF $\Delta = \langle \mathbb{A}, \mathbb{R}, \emptyset \rangle$, with $\mathbb{A} = \{\mathcal{A}, \mathcal{B}\}$ and $\mathbb{R} = \{(\mathcal{A}, \mathcal{B}), (\mathcal{B}, \mathcal{A})\}$. By Def. 15, \emptyset is a complete extension of Δ , whereas the only preferred extensions of Δ are $\{\mathcal{A}\}$ and $\{\mathcal{B}\}$. \square

Similarly, the following lemma relates the stable and preferred extensions of an ASAF.

Lemma 3. *Let $\Delta = \langle \mathbb{A}, \mathbb{R}, \$ \rangle$ be an ASAF. Every stable extension of Δ is also a preferred extension of Δ , but not vice-versa.*

Proof. It is clear that every stable extension of Δ is a maximal (w.r.t. \subseteq) admissible set of Δ , hence a preferred extension of Δ . To show that the reverse does not hold, let us consider the ASAF $\Delta = \langle \mathbb{A}, \mathbb{R}, \emptyset \rangle$, with $\mathbb{A} = \{\mathcal{A}\}$ and $\mathbb{R} = \{(\mathcal{A}, \mathcal{A})\}$. By Def. 15, \emptyset is a preferred extension of Δ but not a stable extension of Δ . \square

Finally, by Def. 15, the grounded extension of an ASAF is also its complete extension.

5. Related Work and Conclusions

In this work we have proposed an approach for characterizing the acceptability semantics of the ASAF introduced in [7]. On the one hand, similarly to [7], we adopted an extension-based approach. On the other hand, differently from [7], we did not make use of a translation into a Dung's AF for obtaining the extensions of the ASAF; instead, we characterized the acceptability semantics directly on the ASAF. This constitutes the main contribution of the paper.

In order to do this, we first identified the different defeats that may occur between the elements of the ASAF. We distinguished between those defeats that can be inferred directly from the attack relation and those that require the consideration of the support relation (respectively, the unconditional and conditional defeats). Therefore, when defining the notion of acceptability, it was necessary to account for all the ways in which defense against a defeat can be provided: either by defeating the corresponding attack or, in the case of conditional defeats, by defeating one of the involved supports. Finally, using the basic notions defined in Section 4.1, a characterization of the acceptability semantics of the ASAF was given in Section 4.2.

Another difference between our approach for obtaining the extensions of the ASAF and the one proposed in [7] regards the presence of supports in the corresponding extensions. For instance, let us consider the ASAF Δ_1 from Ex. 1, whose grounded and preferred extensions were illustrated in Ex. 8. As explained before, even though the source of μ is involved in an attack cycle that is not resolved by the grounded semantics, the support μ is still active. This intuition is captured by our characterization of the ASAF semantics since μ belongs to the grounded extension \mathbf{G} of Δ_1 . In contrast, if we consider the same scenario following the approach given in [7], the ASAF would be translated into an AF such that no support-argument related with μ is in the grounded extension, thus failing to capture the intuition that μ is active.

Our work relates to [1] since the characterization of the ASAF semantics follows the methodology adopted by the AFRA. In particular, when considering an ASAF with an empty support relation (*i. e.*, an AFRA) and the complete, preferred, stable or grounded semantics, the results obtained following the approach by [1] and ours coincide. This is because given such an ASAF only direct and indirect defeats will occur, and the definition of those defeats in Section 3 follows the intuitions of [1]. Moreover, like in the AFRA, when considering defense against defeats in such an ASAF we will only have to account for direct and indirect defeats. Therefore, the formalization of the ASAF can be seen as an extension of the AFRA.

An ASAF where only attacks and supports between arguments may occur can be considered as an AFN [11]. Differently from [11], where arguments attack (here, defeat) other arguments, in our approach only attacks are able to defeat other elements of the ASAF. Nevertheless, the definition of acceptability in both approaches follows the same intuitions. Whereas in the AFN defense against a defeat from an argument \mathcal{A} is provided by defeating \mathcal{A} , in the ASAF this is achieved by defeating the attacks \mathcal{A} originates (through indirect defeats). Hence, following the approach of [11], the extensions of such an ASAF will coincide with the ones obtained through our approach after filtering out the attacks and supports.

In [16] the authors present a formalism that, similarly to ours, extends Dung's framework by adding a support relation and a second-order attack relation that can target at-

tacks and supports. However, in contrast with our approach, their second-order attack relation only allows for attacks to first-order supports and attacks. That is, the interaction is fixed, not being able to combine and nest the attack and support relations at any level. In addition, their second-order attack relation and the attack relation of the ASAF differ in that the former can be originated from first-order attacks, whereas the latter originates only from arguments. Another difference between their approach and ours regards the treatment of support. In contrast with our support relation, in [16] only supports between arguments are allowed. Also, they adopt a deductive interpretation of support which, as shown in [4], corresponds to a dual interpretation of our necessary support.

In this work we defined the complete, preferred, stable and grounded semantics of the ASAF, which correspond to the four classical semantics given in [8]. In particular, Lemmas 2 and 3 show that our characterization of these semantics satisfies the relationships proposed in [8]. Notwithstanding this, the results shown in this work could be extended to other semantics such as semi-stable or ideal; we aim to address this as future work. We also plan to formalize the relationship between the ASAF and the AFRA, as well as the relationship between the ASAF and the AFN. Finally, we intend to study the relationship between the outcome obtained by following the approach of [7] and the outcome obtained by following the approach for determining the extensions of the ASAF proposed in this paper. Moreover, we aim at exploring the computational cost of the acceptability calculus in both approaches, and contrast the results.

References

- [1] P. Baroni, F. Cerutti, M. Giacomin, and G. Guida. AFRA: Argumentation Framework with Recursive Attacks. *Int. J. Approx. Reasoning*, 52(1):19–37, 2011.
- [2] M. W. A. Caminada and L. Amgoud. On the evaluation of argumentation formalisms. *Artif. Intell.*, 171(5–6):286–310, 2007.
- [3] C. Cayrol and M-C. Lagasque-Schiex. On the acceptability of arguments in bipolar argumentation frameworks. In *Proc. of ECSQARU*, pages 378–389. LNAI 3571, Springer, 2005.
- [4] C. Cayrol and M-C. Lagasque-Schiex. Bipolarity in argumentation graphs: Towards a better understanding. *Int. J. Approx. Reasoning*, 54(7):876–899, 2013.
- [5] C. Cayrol and M-C. Lagasque-Schiex. An axiomatic approach to support in argumentation. In *Proc. of TAFA*, pages 74–91. LNAI 9524, Springer, 2015.
- [6] A. Cohen, S. Gottifredi, A. J. García, and G. R. Simari. A survey of different approaches to support in argumentation systems. *Knowledge Eng. Review*, 29:513–550, 2014.
- [7] A. Cohen, S. Gottifredi, A. J. García, and G. R. Simari. An approach to abstract argumentation with recursive attack and support. *J. Applied Logic*, 13(4):509–533, 2015.
- [8] P. M. Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artif. Intell.*, 77(2):321–358, 1995.
- [9] S. Modgil. Reasoning about preferences in argumentation frameworks. *Artif. Intell.*, 173:901–934, 2009.
- [10] F. Nouioua. AFs with necessities: further semantics and labelling characterization. In *Proc. of SUM*, pages 120–133. LNAI 8078, Springer, 2013.
- [11] F. Nouioua and V. Risch. Argumentation frameworks with necessities. In *Proc. of SUM*, pages 163–176. LNAI 6717, Springer, 2011.
- [12] N. Oren and T. J. Norman. Semantics for evidence-based argumentation. In *Proc. of COMMA*, pages 276–284. FAIA 172, IOS Press, 2008.
- [13] S. Polberg and N. Oren. Revisiting support in abstract argumentation systems. In *Proc. of COMMA*, pages 369–376. FAIA 266, IOS Press, 2014.
- [14] H. Prakken. On support relations in abstract argumentation as abstraction of inferential relations. In *Proc. of ECAI*, pages 735–740. FAIA 263, IOS Press, 2014.
- [15] I. Rahwan and G. R. Simari. *Argumentation in Artificial Intelligence*. Springer, 2009.
- [16] S. Villata, G. Boella, D. M. Gabbay, and L. W. N. van der Torre. Modelling defeasible and prioritized support in bipolar argumentation. *Ann. Math. Artif. Intell.*, 66(1-4):163–197, 2012.

Explanation for Case-Based Reasoning via Abstract Argumentation

Kristijonas ČYRAS^{a,1}, Ken SATOH^b and Francesca TONI^a

^aImperial College London, UK

^bNational Institute of Informatics, Tokyo, Japan

Abstract. Case-based reasoning (CBR) is extensively used in AI in support of several applications, to assess a new situation (or case) by recollecting past situations (or cases) and employing the ones most similar to the new situation to give the assessment. In this paper we study properties of a recently proposed method for CBR, based on instantiated Abstract Argumentation and referred to as AA-CBR, for problems where cases are represented by abstract factors and (positive or negative) outcomes, and an outcome for a new case, represented by abstract factors, needs to be established. In addition, we study properties of explanations in AA-CBR and define a new notion of lean explanations that utilize solely relevant cases. Both forms of explanations can be seen as dialogical processes between a proponent and an opponent, with the burden of proof falling on the proponent.

Keywords. Case-Based Reasoning, Abstract Argumentation, Explanation

1. Introduction

Case-based reasoning (CBR), as overviewed in [28], is extensively used in various applications of AI (see e.g. [23,28]). At a high-level, in CBR a reasoner in need to assess a new situation, or *new case*, recollects past situations, or *past cases*, and employs the ones most similar to the new situation to give the assessment. Several approaches to CBR use (forms of) argumentation, e.g. [1,27] and, more recently, the AA-CBR approach of [11].

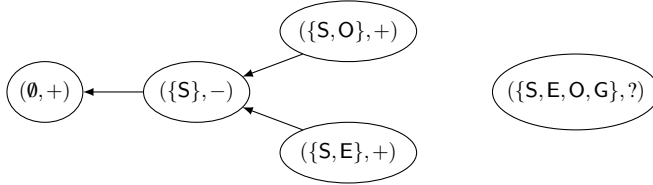
AA-CBR instantiates Abstract Argumentation (AA) [12] to resolve conflicts amongst most similar past cases with diverging outcomes. It provides: 1) a method for computing outcomes for new cases, given past cases and a *default* outcome; and 2) *explanations* for computed outcomes, as dialogical exchanges between a proponent, in favour of the default outcome for the new case, and an opponent, against the default outcome.

As common in the literature (see e.g. [5,25,28]), in AA-CBR past cases are represented as sets of *factors* (also known as features or attribute-value pairs, cf. [30]) together with an outcome, which may be positive (+) or negative (−). AA-CBR then relies upon the *grounded extension* [12] of an AA framework with, as arguments, a default case (with an empty set of factors and the default outcome), past cases (with their outcomes) and a new case (with unknown outcome). A past case attacks another past case or the default

¹Corresponding Author: Kristijonas Čyras, Department of Computing, Imperial College London, United Kingdom; E-mail: k.cyras13@imperial.ac.uk.

case if they have a different outcomes, the former is more specific than the latter and at least as concise as any other similarly more specific, conflicting past case. The following example, used or adapted throughout the paper, illustrates AA-CBR.

Example 1. Suppose Bob wishes to rent his spare room to get between £800 and £900 per month, and decides to use an online AA-CBR system to determine whether this amount is reasonable and why, based on similar lodgings being rented. Let N , the new case, represent the set of features of Bob's room, e.g. $N = \{S, E, O, G\}$ (the room is Small, with an En-suite bathroom in an Open-plan flat with a Gym in the building). Here the default outcome is $+$, indicating Bob's bias for the price range £800–£900. The past cases are either of the form $(X, +)$, for lodgings in the desired price range, or $(Y, -)$, for lodgings in different (lower or higher) price ranges, with X, Y the feature sets of these lodgings. For example, suppose the past cases are $(\{S\}, -)$ (Small rooms go for lower prices), $(\{S, E\}, +)$ (En-suite compensates for Small room), $(\{S, O\}, +)$ (Open-plan flat compensates for Small). Then, the corresponding (instantiated) AA framework [11] is depicted below (with attacks represented by arrows, $(\emptyset, +)$ the argument for the default case, and $(\{S, E, O, G\}, ?)$ the argument for the new case):



$\mathbb{G} = \{(\{S, E, O, G\}, ?), (\{S, E\}, +), (\{S, O\}, +), (\emptyset, +)\}$ is the grounded extension of this AA framework. Since $(\emptyset, +) \in \mathbb{G}$, the outcome for the new case determined by AA-CBR is $+$, with two possible explanations \mathcal{T}_P and \mathcal{T}'_P depicted below (with P standing for proponent and O standing for opponent):



Thus, for example, \mathcal{T}_P explains the recommendation $+$ dialectically as follows: the default outcome $+$ needs to be defended against the objection posed by past case $(\{S\}, -)$, and this can be achieved by using past case $(\{S, E\}, +)$, that cannot be objected against.

In this paper we propose a novel form of explanations, called *lean explanations*, and study properties of both forms of explanations in AA-CBR. Explanations can naturally be seen as dialogical exchanges between a proponent and an opponent, the former having the burden of proof for explaining as well as establishing the outcome of the new case.

The paper is organized as follows. We first recall, in Section 2, the necessary background. In Section 3 we prove some properties of AA-CBR, in the context of some related work, and in Section 4 we investigate properties of explanations in AA-CBR. We then allot Section 5 to relate AA-CBR with proof standards and burden of proof. We conclude with a discussion on related and future work in Section 6.

2. Background

AA-CBR [11] assumes a fixed but otherwise arbitrary (possibly infinite) set \mathbb{F} of *factors*, and a set $\{+, -\}$ of *outcomes*, one of which is singled out as the *default outcome* d . The *complement* of d is indicated as \bar{d} , and is: $+$ if $d = -$; and $-$ if $d = +$. A *case* is a pair (X, o) with $X \subseteq \mathbb{F}$ and $o \in \{+, -\}$; a *case base* is a finite set $CB \subseteq \wp(\mathbb{F}) \times \{+, -\}$ such that for $(X, o_X), (Y, o_Y) \in CB$, if $X = Y$, then $o_X = o_Y$; a *new case* is a set $N \subseteq \mathbb{F}$.

AA-CBR maps the problem of determining the outcome for a new case into a membership problem within the grounded extension of an AA framework [12] obtained from the case base CB , the new case N and the default outcome d . In general, following [12], an *AA framework* is a pair $(Args, \rightsquigarrow)$, where $Args$ is a set (of *arguments*) and \rightsquigarrow is a binary relation on $Args$ (where, for $\mathbf{a}, \mathbf{b} \in Args$, if $\mathbf{a} \rightsquigarrow \mathbf{b}$, then we say that \mathbf{a} *attacks* \mathbf{b}). For a set of arguments $E \subseteq Args$ and an argument $\mathbf{a} \in Args$, E *defends* \mathbf{a} if for all $\mathbf{b} \rightsquigarrow \mathbf{a}$ there exists $\mathbf{c} \in E$ such that $\mathbf{c} \rightsquigarrow \mathbf{b}$. Then, the *grounded extension* of $(Args, \rightsquigarrow)$ can be constructed as $\mathbb{G} = \bigcup_{i \geq 0} G_i$, where G_0 is the set of all unattacked arguments, and $\forall i \geq 0$, G_{i+1} is the set of arguments that G_i defends. For any $(Args, \rightsquigarrow)$, the grounded extension \mathbb{G} always exists and is unique, and, if $(Args, \rightsquigarrow)$ is well-founded [12], extensions under other semantics are equal to \mathbb{G} . AA-CBR uses the following instance of AA [11]:

Definition 2. The *AA framework corresponding to a case base CB , a default outcome $d \in \{+, -\}$ and a new case N* is $(Args, \rightsquigarrow)$ satisfying the following conditions:

- $Args = CB \cup \{(\emptyset, d)\} \cup \{(N, ?)\}$;
- for $(X, o_X), (Y, o_Y) \in CB \cup \{(\emptyset, d)\}$, it holds that $(X, o_X) \rightsquigarrow (Y, o_Y)$ iff
 - * $o_X \neq o_Y$, and (different outcomes)
 - * $Y \subsetneq X$, and (specificity)
 - * $\nexists (Z, o_Z) \in CB$ with $Y \subsetneq Z \subsetneq X$; (concision)
- for $(Y, o_Y) \in CB$, $(N, ?) \rightsquigarrow (Y, o_Y)$ holds iff $Y \not\subseteq N$.

$(N, ?)$ is referred to as the *new case argument* and (\emptyset, d) as the *default case*.

In what follows, $(Args, \rightsquigarrow)$ is the AA framework corresponding to a given, generic CB , d and N , and \mathbb{G} is its grounded extension. Note the following: $(Args, \rightsquigarrow)$ is finite (as case bases are); $\mathbb{G} \neq \emptyset$ (as $(N, ?)$ is unattacked); $(Args, \rightsquigarrow)$ is well-founded (due to the specificity requirement in Definition 2), so that \mathbb{G} is a unique extension under other semantics. AA-CBR decides the outcome for the new case as follows [11]:

Definition 3. The *AA outcome* of the new case N is:

- the default outcome d , if $(\emptyset, d) \in \mathbb{G}$;
- \bar{d} , otherwise, if $(\emptyset, d) \notin \mathbb{G}$.

In AA-CBR, explanations for AA outcomes are defined in terms of *dispute trees* [11,13,14], where a *dispute tree* for $\mathbf{a} \in Args$ is a tree \mathcal{T} such that:

1. every node of \mathcal{T} is of the form $[L:\mathbf{x}]$, with $L \in \{P, 0\}$, $\mathbf{x} \in Args$: the node is *labelled* by argument \mathbf{x} and assigned the *status* of either *proponent* (P) or *opponent* (0);
2. the root of \mathcal{T} is a P node labelled by \mathbf{a} ;
3. for every P node n , labelled by some $\mathbf{b} \in Args$, and for every $\mathbf{c} \in Args$ such that $\mathbf{c} \rightsquigarrow \mathbf{b}$, there exists a child of n , which is an 0 node labelled by \mathbf{c} ;
4. for every 0 node n , labelled by some $\mathbf{b} \in Args$, there exists at most one child of n which is a P node labelled by some $\mathbf{c} \in Args$ such that $\mathbf{c} \rightsquigarrow \mathbf{b}$;
5. there are no other nodes in \mathcal{T} except those given by 1–4.

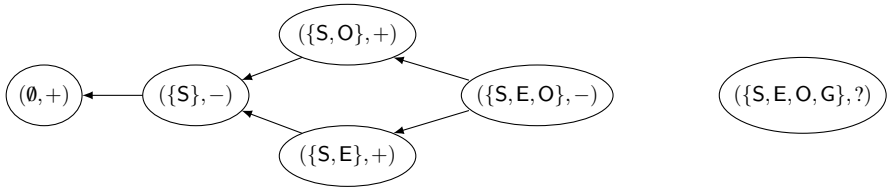
The *defence set* of \mathcal{T} , denoted by $\mathcal{D}(\mathcal{T})$, is the set of all arguments labelling P nodes in \mathcal{T} . A dispute tree \mathcal{T} is an *admissible dispute tree* iff (i) every 0 node in \mathcal{T} has a child, and (ii) no argument in \mathcal{T} labels both P and 0 nodes. A dispute tree \mathcal{T} is a *maximal dispute tree* [11] iff for all opponent nodes $[0:x]$ which are leaves in \mathcal{T} there is no $y \in \text{Args}$ such that $y \rightsquigarrow x$. Note that an admissible dispute tree \mathcal{T} for some $a \in \text{Args}$ is also a maximal dispute tree for $a \in \text{Args}$ [11, Lemma 4]. Indeed, in an admissible dispute tree, each 0 node has exactly one child (a P node); thus, no 0 node is a leaf, and so the dispute tree is maximal.

Explanations in AA-CBR are defined as follows [11]:

Definition 4. If the AA outcome of N is d , then an *explanation for why the AA outcome of N is d* is any admissible dispute tree for (\emptyset, d) . If the AA outcome of N is \bar{d} , then an *explanation for why the AA outcome of N is \bar{d}* is any maximal dispute tree for (\emptyset, d) .

Example 1 illustrates the notion of explanation for why the outcome is d . The following example illustrates the notion of explanation for why the outcome is \bar{d} .

Example 5 (Example 1 ctd.). Suppose there is an additional case $(\{S, E, O\}, -)$ in CB . Then the corresponding AA framework is depicted below:



Here, $\mathbb{G} = \{(\{S, E, O, G\}, ?), (\{S, E, O\}, -), (\{S\}, -)\}$, so the AA outcome of $\{S, E, O, G\}$ is $-$, for which the dispute trees (in linear notation) $\mathcal{T}_0 : [P : (\emptyset, +)] \text{ --- } [O : (\{S\}, -)] \text{ --- } [P : (\{S, E\}, +)] \text{ --- } [O : (\{S, E, O\}, -)]$ and $\mathcal{T}'_0 : [P : (\emptyset, +)] \text{ --- } [O : (\{S\}, -)] \text{ --- } [P : (\{S, O\}, +)] \text{ --- } [O : (\{S, E, O\}, -)]$ are explanations.

3. Properties of AA outcomes

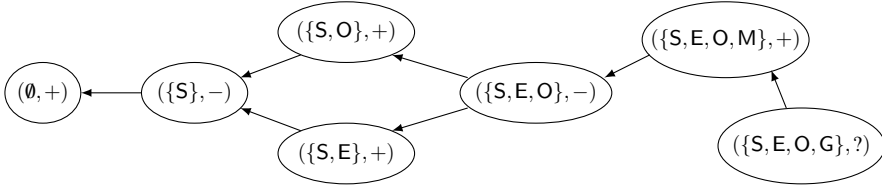
In this section, we prove several properties of AA outcomes, focusing on aspects that have been considered in some related work on CBR. Where indicated, these properties were stated in [11] already, but their proofs omitted there.

In the context of, particularly, legal CBR, as well as CBR in general, two properties are identified as important [3], namely that cases employed in determining the outcome of a new case N should be *most on point* and *untrumped*, where (X, o) is:

most on point iff no other case with the same outcome shares a more inclusive set of factors with the new case, i.e. $X \cap N$ is \subseteq -maximal for $X' \cap N$ with $(X', o) \in CB$;
untrumped iff no counterexample is more on point, i.e. there is no $(Y, o_Y) \in CB$ satisfying both $o_Y \neq o$ and $X \cap N \subsetneq Y \cap N$.

These two constraints together can be summarized into a single condition of $X \cap N$ being \subseteq -maximal among all $Y \cap N$ with $(Y, o') \in CB$. These properties—of being most on point and untrumped—allow for ‘deviating’ factors (i.e. factors not present in N) amongst past cases to be used to determine the outcome of the new case, whereas AA outcome does not. For an illustration, consider the following modification of our running example.

Example 6. Suppose there is an additional past case $(\{S, E, O, M\}, +)$ in Example 5: a Motorway next to the building is disadvantageous and the price of a Small En-suite room in an Open plan flat next to a Motorway falls into Bob's price range:



Here, both $(\{S, E, O\}, -)$ and $(\{S, E, O, M\}, +)$ are most on point and untrumped, yet $(\{S, E, O, M\}, +)$ is attacked by the new case and so effectively discarded from influencing the AA outcome, which is $-$, as in Example 5.

The notion of AA outcome fulfils a variant of the properties of being most on point and untrumped, that disregards ‘deviating’ factors and focuses instead on *nearest* cases [11], where (X, o_X) is

nearest to N iff $X \subseteq N$, and $\nexists (Y, o_Y) \in CB$ with $Y \subseteq N$ and $X \subsetneq Y$.

In other words, (X, o_X) is nearest to N iff $X \subseteq N$ is \subseteq -maximal in the case base. In Example 6, $(\{S, E, O\}, -)$ is nearest, but $(\{S, E, O, M\}, +)$ is not, because $M \notin N$.

Like elsewhere in the literature, e.g. [22,23,28], in AA-CBR nearest cases are very important. In particular, when CB contains a single nearest case (X, o) to N , the AA outcome of N is fully determined by (X, o) , independently of what d is, as follows:

Proposition 1 ([11, Proposition 2]). *If there is a unique nearest case (X, o) to N , then, for any $d \in \{+, -\}$, the AA outcome of N is o .*

Proof. Let $(X, o) \in CB$ be the unique nearest case to N . Consider a chain of attacks $(Y, o_Y) \rightsquigarrow \dots \rightsquigarrow (\emptyset, d)$, with $n \geq 1$ arguments and (Y, o_Y) unattacked in $(Args, \rightsquigarrow)$. First, we know that $(Y, o_Y) \in \mathbb{G}$. Assuming $o_Y \neq o$, we find $Y \subsetneq X$ (as (X, o) is unique nearest to N), whence $(X, o) \rightsquigarrow (Y, o_Y)$ gives a contradiction. So $o_Y = o$. Thus, if $o = d$, then n is odd, and so \mathbb{G} defends (\emptyset, d) , so that $(\emptyset, d) \in \mathbb{G}$. Else, if $o = \bar{d}$, then n is even, so that \mathbb{G} attacks (\emptyset, d) , and so $(\emptyset, d) \notin \mathbb{G}$. In any case, the AA outcome of N is o . \square

In Example 6, the AA outcome of $\{S, E, O, G\}$ is $-$, the outcome of the unique nearest case $(\{S, E, O\}, -)$ (and the complement of the default outcome $+$). If instead the default outcome was $-$, the structure of the AA framework would change (in particular, the attack relation would be different), but the AA outcome would remain unchanged.

In AA-CBR, nearest cases are important as they belong to the grounded extension:

Proposition 2 ([11, Lemma 1]). *\mathbb{G} contains all the nearest past cases to N .*

Proof. Let $(X, o_X) \in CB$ be nearest to N . Then $X \subseteq N$, so $(N, ?) \not\rightsquigarrow (X, o_X)$. Now assume that $(Y, o_Y) \rightsquigarrow (X, o_X)$, for some $(Y, o_Y) \in CB$. Then $Y \not\subseteq N$, whence $(N, ?) \rightsquigarrow (Y, o_Y)$. Since the new case argument $(N, ?)$ is unattacked in $(Args, \rightsquigarrow)$, we have $(N, ?) \in \mathbb{G}$. As (Y, o_Y) was arbitrary, we know that \mathbb{G} defends (X, o_X) , so that $(X, o_X) \in \mathbb{G}$. \square

This result shows that AA-CBR takes into account all the most similar past cases when determining the outcome of a new case. This is in contrast with some forms of the

conventional k -nearest neighbour approaches to CBR, where some of the nearest cases may be ignored in order to decide the new case [28] (see also Section 6 for a discussion).

Note that \mathbb{G} contains not only the nearest cases (as well as the new case N), but also some other past cases: in Examples 5, 6, \mathbb{G} includes $(\{S\}, -)$, which is not nearest to N , but still ‘relevant’ to the AA outcome. Overall, past cases, as arguments, can be classified into those deemed *relevant* and *irrelevant* for deciding the new case, as follows:

Definition 7. An argument $(X, o) \in \text{Args} \setminus \{(\emptyset, d)\}$ is said to be:

- **relevant** if $X \cap N \neq \emptyset$;
- **irrelevant** otherwise, if $X \cap N = \emptyset$.

By convention, the default case (\emptyset, d) is also deemed relevant.

Since arguments in AA-CBR are cases, with an abuse of notation we sometimes talk about cases being relevant and irrelevant.

In Example 6, all cases (including the default case) are relevant. If there was, say, a case $(\{H\}, -)$ in CB , it would be irrelevant, as $\{H\} \cap \{S, E, O, G\} = \emptyset$.

The relevance criteria defined above will play a role in characterizing explanations of AA outcomes, which we will investigate in the next section.

4. Explanations of AA Outcomes

The notion of AA outcome allows to determine algorithmically whether a new case N should be assigned the default outcome (d) or not (\bar{d}), by determining whether the default case (\emptyset, d) belongs or not (respectively) to the grounded extension \mathbb{G} of the AA framework $(\text{Args}, \rightsquigarrow)$ corresponding to the given case base CB , d and N . Explanations of AA outcomes (Section 2) exploit the argumentative re-interpretation afforded by AA-CBR utilizing not only the nearest cases, but also dialectical exchanges of relevant arguments (cf. e.g. [26]). In this section, we prove several properties of explanations in AA-CBR. Where indicated, these properties were stated in [11], but their proofs omitted there.

The following result will help us to characterize explanations of AA outcomes.

Theorem 3. $(\emptyset, d) \in \mathbb{G}$ iff there exists an admissible dispute tree \mathcal{T} for (\emptyset, d) .

Proof. By Theorem 3.2 in [14], there is an admissible dispute tree \mathcal{T} for (\emptyset, d) iff (\emptyset, d) is in some admissible extension. Every admissible extension is contained in some preferred extension [12] and, as $(\text{Args}, \rightsquigarrow)$ is well founded, \mathbb{G} is the only preferred extension. Thus, (\emptyset, d) is in some admissible extension iff $(\emptyset, d) \in \mathbb{G}$, and so the claim follows. \square

Thus, an explanation for the default outcome always exists:

Proposition 4 ([11, Proposition 3]). *If the AA outcome of the new case N is the default outcome d , then there is an explanation \mathcal{T} for why the AA outcome of N is d , which is moreover such that the defence set $\mathcal{D}(\mathcal{T})$ is admissible and $\mathcal{D}(\mathcal{T}) \subseteq \mathbb{G}$.*

Proof. Existence of explanations follows from Theorem 3. Further, Theorem 3.2, part (ii), in [14], says that if $\mathbf{a} \in \text{Args}$ belongs to an admissible set $A \subseteq \text{Args}$ of arguments, then there exists an admissible dispute tree \mathcal{T} for \mathbf{a} such that $\mathcal{D}(\mathcal{T}) \subseteq A$ and $\mathcal{D}(\mathcal{T})$ is admissible. Since \mathbb{G} is admissible and $(\emptyset, d) \in \mathbb{G}$, there is an admissible dispute tree \mathcal{T} for (\emptyset, d) with $\mathcal{D}(\mathcal{T})$ admissible and $\mathcal{D}(\mathcal{T}) \subseteq \mathbb{G}$. \square

An analogous result holds regarding explanations for the non-default outcome:

Proposition 5 ([11, Proposition 5]). *If the AA outcome of the new case N is \bar{d} , then there is an explanation \mathcal{T} for why the AA outcome of N is \bar{d} , and moreover $\mathcal{D}(\mathcal{T}) \not\subseteq \mathbb{G}$.*

Proof. Theorem 3.1 in [14] states that a dispute tree \mathcal{T} such that every 0 node in \mathcal{T} has a child, is necessarily admissible if it is *finite*. Since dispute trees in our setting are guaranteed to be finite, any dispute tree with all leaves labelled P would be admissible, yielding $(\emptyset, d) \in \mathbb{G}$ (by Theorem 3), contradicting the AA outcome of N being \bar{d} . Thus, some dispute tree \mathcal{T} for (\emptyset, d) will have all 0 unattacked in $(\text{Args}, \rightsquigarrow)$, and so be maximal, as required. Further, if $\mathcal{D}(\mathcal{T}) \subseteq \mathbb{G}$ then, by definition of dispute trees and grounded extensions, $(\emptyset, d) \in \mathbb{G}$, which is again a contradiction. Hence, $\mathcal{D}(\mathcal{T}) \not\subseteq \mathbb{G}$. \square

In Example 1, dispute trees $\mathcal{T}_P : [P : (\emptyset, +)] \text{ --- } [0 : (\{S\}, -)] \text{ --- } [P : (\{S, E\}, +)]$ and $\mathcal{T}'_P : [P : (\emptyset, +)] \text{ --- } [0 : (\{S\}, -)] \text{ --- } [P : (\{S, O\}, +)]$ are explanations for why the AA outcome of $\{S, E, O, G\}$ is $+$, with $\mathcal{D}(\mathcal{T}_P) = \{(\{S, E\}, +), (\emptyset, +)\} \subseteq \mathbb{G}$ and $\mathcal{D}(\mathcal{T}'_P) = \{(\{S, O\}, +), (\emptyset, +)\} \subseteq \mathbb{G}$, both admissible. Each explanation serves Bob to legitimize why he is justified in asking the price he has in mind. Similarly, in Example 5 (where CB from Example 1 is augmented with $(\{S, E, O\}, -)$), the trees $\mathcal{T}_0 : [P : (\emptyset, +)] \text{ --- } [0 : (\{S\}, -)] \text{ --- } [P : (\{S, E\}, +)] \text{ --- } [0 : (\{S, E, O\}, -)]$ and $\mathcal{T}'_0 : [P : (\emptyset, +)] \text{ --- } [0 : (\{S\}, -)] \text{ --- } [P : (\{S, O\}, +)] \text{ --- } [0 : (\{S, E, O\}, -)]$ are explanations for why the AA outcome of $\{S, E, O, G\}$ is $-$, with the same defence sets as \mathcal{T}_P and \mathcal{T}'_P (respectively), yet no longer contained in \mathbb{G} . Each explanation indicates that Bob should reconsider his price tag.

The next result says that every case that should be considered in explaining as well as determining the AA outcome is indeed considered.

Proposition 6. *For every nearest case (X, o) , there is an explanation \mathcal{T} (for why the AA outcome of N is either d or \bar{d}) s.t. for some (X', o) with $X' \subseteq X$ we find $(X', o) \in \mathcal{D}(\mathcal{T})$.*

Proof. If a nearest case (X, o) does not itself appear in any explanation, then some (X', o) with $X' \subseteq X$ must appear in some explanation \mathcal{T} . In any event, if either $o = d$ or $o = \bar{d}$, we find $(X', o) \in \mathcal{D}(\mathcal{T})$ by construction of \mathbb{G} and \mathcal{T} . \square

In Example 5, the unique nearest case $(\{S, E, O\}, -)$ to $N = \{S, E, O, G\}$ labels a node in both explanations \mathcal{T}_0 and \mathcal{T}'_0 (as above) for why the AA outcome of N is $-$. Observe that $(\{S, E, O\}, -)$ is unattacked. This need not always happen: in Example 6, $(\{S, E, O\}, -)$ is still a unique nearest case to N , but this time attacked by $(\{S, E, O, M\}, +)$, which is in turn attacked by $(N, ?)$. In any event, $(\{S, E, O\}, -)$ labels a node in both possible explanations, namely $\mathcal{T}_1 : [P : (\emptyset, +)] \text{ --- } [0 : (\{S\}, -)] \text{ --- } [P : (\{S, E\}, +)] \text{ --- } [0 : (\{S, E, O\}, -)] \text{ --- } [P : (\{S, E, O, M\}, +)] \text{ --- } [0 : (\{S, E, O, G\}, ?)]$ and $\mathcal{T}_2 : [P : (\emptyset, +)] \text{ --- } [0 : (\{S\}, -)] \text{ --- } [P : (\{S, O\}, +)] \text{ --- } [0 : (\{S, E, O\}, -)] \text{ --- } [P : (\{S, E, O, M\}, +)] \text{ --- } [0 : (\{S, E, O, G\}, ?)]$, for why the AA outcome of N is $-$.

A nearest case need not itself appear in any explanation; instead, some of its ‘proper subsets’ will. For instance, suppose that in Example 5, instead of $(\{S, E, O\}, -)$, we have $(\{S, E, O\}, +)$, which is then a unique nearest case to N . The AA outcome of N is then $+$, for which $\mathcal{T}_P : [P : (\emptyset, +)] \text{ --- } [0 : (\{S\}, -)] \text{ --- } [P : (\{S, E\}, +)]$ and $\mathcal{T}'_P : [P : (\emptyset, +)] \text{ --- } [0 : (\{S\}, -)] \text{ --- } [P : (\{S, O\}, +)]$ are explanations, with $\mathcal{D}(\mathcal{T}_P) = \{(\{S, E\}, +), (\emptyset, +)\}$ and $\mathcal{D}(\mathcal{T}'_P) = \{(\{S, O\}, +), (\emptyset, +)\}$. Thus, $(\{S, E, O\}, +)$ does not label any node in either \mathcal{T}_P or \mathcal{T}'_P , but $(\{S, E\}, +)$ and $(\{S, O\}, +)$ do.

In general, every argument in an explanation has a reason to appear there:

Proposition 7. *Every argument labelling a node in an explanation \mathcal{T} (for why the AA outcome of N is either d or \bar{d}), is either relevant or attacked by $(N, ?)$.*

Proof. By definition of relevance, if $(X, o) \neq (\emptyset, d)$ labels a node in \mathcal{T} and is irrelevant, then $X \cap N = \emptyset$, so that $(N, ?) \rightsquigarrow (X, o)$, by definition of attack. \square

In Example 5, every argument labelling a node in any of the explanations \mathcal{T}_0 and \mathcal{T}'_0 is relevant. To see that irrelevant arguments can also appear in explanations, consider a single past case $(\{A\}, +)$, default outcome $-$ and a new case $\{B\}$. In the corresponding $(Args, \rightsquigarrow)$, we find $(\{B\}, ?) \rightsquigarrow (\{A\}, +) \rightsquigarrow (\emptyset, -)$ and $\mathbb{G} = \{(\{B\}, ?), (\emptyset, -)\}$, so that the AA outcome of $\{B\}$ is $-$, for which $\mathcal{T} : [P : (\emptyset, -)] - [O : (\{A\}, +)] - [P : (\{B\}, ?)]$ is an explanation. Here, $(\{A\}, +)$ is irrelevant. Observe further that there could be many more similar irrelevant cases $(\{A_1\}, +), (\{A_2\}, +), \dots$, whence there would be as many explanations, all of them containing an irrelevant case.

To avoid overpopulation of explanations with irrelevant arguments, we next propose a leaner version of explanations that contain only relevant arguments.

Definition 8. Let a **relevant dispute tree** be a dispute tree in construction of which only relevant arguments can label nodes. A **lean explanation** for why the AA outcome of N is d (resp., \bar{d}) is an admissible (resp., maximal) relevant dispute tree for (\emptyset, d) .

As for (standard) explanations, the *defence set* of a lean explanation \mathcal{T}^L , denoted by $\mathcal{D}(\mathcal{T}^L)$, is the set of all arguments labelling P nodes in \mathcal{T}^L .

The explanations discussed in Examples 1, 5 and 6 are lean, whereas the explanations discussed in the example before Definition 8 are not: the only lean explanation there is simply $\mathcal{T}^L : [P : (\emptyset, -)]$.

Note that a lean explanation for why the AA outcome of N is d (resp., \bar{d}) is a maximal subtree of an explanation for why the AA outcome of N is d (resp., \bar{d}) such that no parent node is labelled by an irrelevant argument. Plainly, lean explanations can be obtained from (standard) explanations by removing irrelevant nodes, as well as their children.

From Proposition 7 and Definition 8 it trivially follows that nothing is irrelevant in lean explanations:

Corollary 8. *Every argument labelling a node in a lean explanation \mathcal{T}^L (for why the AA outcome of N is either d or \bar{d}) is relevant.*

Simultaneously, lean explanations keep desirable properties in the following sense.

Corollary 9. *If the AA outcome of N is d , then there is a lean explanation \mathcal{T}^L for why the AA outcome of N is d , such that $\mathcal{D}(\mathcal{T}^L) \cup \{(N, ?)\}$ is admissible and $\mathcal{D}(\mathcal{T}^L) \subseteq \mathbb{G}$.*

If the AA outcome of N is \bar{d} , then there is a lean explanation \mathcal{T}^L for why the AA outcome of N is \bar{d} , such that $\mathcal{D}(\mathcal{T}^L) \not\subseteq \mathbb{G}$.

Proof. Follows from Propositions 4 and 5, in the first instance noticing that $(N, ?)$ need not have a relevant parent in an explanation (and hence $(N, ?) \notin \mathcal{D}(\mathcal{T}^L)$). \square

Utilizing the following definition, we see that lean explanations also impose a certain structure to the otherwise unstructured collection of relevant cases.

Definition 9. Let $(X, o_X), (Y, o_Y) \in Args$ be relevant. We say that (X, o_X) is **more relevant** than (Y, o_Y) if either $Y \subsetneq X$ or $(X, o_X) = (N, ?)$.

Proposition 10. Every argument labelling a node in a lean explanation (for why the AA outcome of N is either d or \bar{d}) is more relevant than the argument labelling its parent.

Proof. Follows from Definitions 2 (attack), 8 (lean explanations) and 9. \square

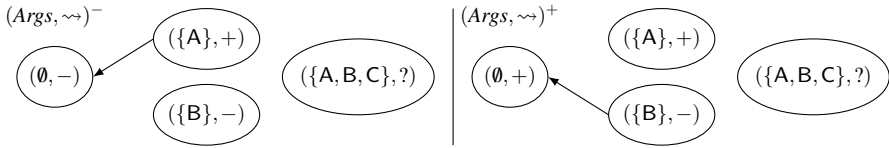
Apart from structuring past cases and providing dialogical justifications for why a particular outcome is assigned to a new case, explanations also yield hints on modifying the situation so as to achieve the desired outcome. For example, imagine that Alice wants to let a Small room with En-suite ($N = \{S, E\}$) for over £900, and past cases are $(\{E\}, -)$ (En-suite falls short) and $(\{S, E, H\}, +)$ (High-speed internet brings the rent over £900). The AA outcome of N is $-$ (as $(\{E\}, -)$ is a unique nearest case to N), with an explanation $[P: (\emptyset, +)] - [O: (\{E\}, -)] - [P: (\{S, E, H\}, +)] - [O: (\{S, E\}, ?)]$, from which Alice sees that installing High-speed internet would allow her to ask the price she wants.

This last illustration, together with the previously considered examples, hint at a feature of AA-CBR, namely that AA outcomes and (lean) explanations exhibit a certain asymmetry between the proponent and the opponent. This is in line with the asymmetry observed in the context of CBR, e.g. in [26].

5. Proof Standards for AA Outcome

In this section we show that the asymmetry described above is a manifestation of the *burden of proof* [18] falling onto the proponent, by introducing, for AA-CBR frameworks, a variant of a well-known *proof standard*. Consider the following example.

Example 10. Let $N = \{A, B, C\}$ and $CB = \{(\{A\}, +), (\{B\}, -)\}$. Consider the two default outcomes $-$ and $+$ in turn. Below are depicted the AA frameworks $(Args, \rightsquigarrow)^d$ corresponding to CB , $d \in \{-, +\}$ and N :



In $(Args, \rightsquigarrow)^-$, the AA outcome of N is $+$ (non-default), with a unique (lean) explanation $\mathcal{T}^- : [P: (\emptyset, -)] - [O: (\{A\}, +)]$. Likewise, In $(Args, \rightsquigarrow)^+$, the AA outcome is $-$, with a unique (lean) explanation $\mathcal{T}^+ : [P: (\emptyset, +)] - [O: (\{B\}, -)]$. Thus, no matter what the default outcome is, the burden of establishing as well as explaining the AA outcome is on the proponent's side.

In what follows, we formalize this feature of AA-CBR in terms of proof standards.

In our setting, following [20], a proof standard can be seen as a function taking a statement and an AA framework and returning an element of $\{\text{TRUE}, \text{FALSE}\}$. Then, a statement s is satisfied by a proof standard STD in $(Args, \rightsquigarrow)$ iff $\text{STD}(s, (Args, \rightsquigarrow)) = \text{TRUE}$. In the context of AA-CBR, the following statement is of interest:

s_d : “given a case base CB and a default outcome d , the outcome of the new case N is d ”.

We identify a proof standard that meets this statement:

Definition 11. The **Scintilla of Evidence** proof standard SE is defined as follows:

$$SE(s_d, (Args, \rightsquigarrow)) = \text{TRUE} \text{ iff there exists an admissible dispute tree } \mathcal{T} \text{ for } (\emptyset, d).$$

Intuitively, the SE proof standard amounts to the proponent P having a good line of defence (a tree of arguments and attacks) in a dialectical exchange of arguments for and against the default outcome. [7,18,20], to name a few, give proof standards with the same name. Our variant is in the same spirit in that a statement by a proponent P meets the standard “if it is supported by at least one *defensible* P argument”, where, in our variant, we interpret support as a dispute tree, and defensible as admissible.

Directly from Theorem 3, we get that AA outcome meets the SE proof standard, in that accepting the default case equates with the satisfaction of the SE proof standard:

Theorem 11. $(\emptyset, d) \in \mathbb{G}$ iff $SE(s_d, (Args, \rightsquigarrow)) = \text{TRUE}$.

This result also indicates that no matter what the default outcome d is, the burden of proof to establish that the AA outcome of the new case N is d falls upon the proponent, in that the proponent needs to construct an admissible dispute tree that dialectically justifies the outcome. This is witnessed in examples we considered, particularly Example 10.

6. Related and Future Work

Argumentation has perhaps been most prominently applied to legal CBR. For example, [6] use AA frameworks to reason with particular types of animal cases by representing legal natural language arguments involved in cases as formal arguments, at the same time taking into account preference information over values promoted by arguments. [1] show how to represent the well known legal CBR systems HYPO [4], CATO [2] and IBP [8] in Abstract Dialectical Frameworks [7]. Another strand of research concerns argumentation schemes—patterns to create and/or classify arguments in order to decide how precedent cases determine the new case (e.g. [20,21,27]). There, proof standards can be employed to evaluate arguments based on argumentation schemes, as in e.g. [20].

In contrast, we are not focused on legal CBR, but rather on general CBR, as overviewed in [28]. In that setting, our work stands out in its aim to provide explanations as to why a particular outcome was obtained in solving CBR problems. To this end we exploit the dialectical aspect that AA supplies by way of dispute trees. Explaining AA outcome can be seen through a dialogical exchange of arguments (namely, past cases) between a proponent in favour of the default outcome, and an opponent against the default outcome. Explanations in AA-CBR relate to the notion of burden of proof from legal CBR, see e.g. [18,26]. However, legal CBR exhibits characteristics not necessarily applicable to CBR in general, or at least not to the type of problems we consider. For instance, in legal CBR, there is usually more granularity to factors [22]; certain hypothetical reasoning and/or background knowledge is involved [3]. Whether and how our approach can be applied to, for instance, legal CBR, is a line of future work: it would be interesting to look at other proof standards as well as burdens of production and persuasion [18]; relating AA-CBR to argumentation based on the discovery of association

rules, as in e.g. [31], and to argument and theory construction from legal cases, as in e.g. [10], would be interesting too.

In terms of general CBR, determining *why* an outcome is computed is deemed crucial, but is inherently hard to define formally [30]. A common form of explanation in CBR amounts to displaying the most similar past cases. In particular, *transparency*, in not trying to “hide conflicting evidence” [30, p. 134], is identified as desirable. This is fulfilled in AA-CBR, as the grounded extension contains all past cases nearest to the new case, be they of agreeing or diverging outcomes (cf. Proposition 2). However, merely displaying the nearest cases (especially with contrasting outcomes) is not always sufficient to explain the proposed outcome. To address this issue, *k*-nearest neighbour approaches produce only the most similar among the nearest neighbours. But then, “the transparency goal is no longer fulfilled [...] if $k > 1$ ” [30, p. 136]. In contrast, (lean) explanations in AA-CBR amount to (relevant) dispute trees, where not only the nearest cases, but also cases relevant to the AA outcome play a role.

In our setting, relevance of past cases is defined via their commonalities with the new case, in terms of factors shared. By contrast, [24] proposed *supporting/opposition* criteria based on counting the ratio (or probability) of how often a factor appears in a case with the outcome d/\bar{d} . We provide explanations without quantifying the appearance of factors, but we plan to investigate such a possibility in the future.

Several works define methods for determining explanations for the (non-)acceptability of arguments in argumentation, see e.g. [16,17,19,29]. These works use trees as the underlying mechanism for computing explanations, but not in a CBR setting. Study of formal relationships with these works is left for future work. Other work in argumentation, e.g. [9], investigates the usefulness of explanation in argumentation with users. Similar explorations for our approach are also left for the future.

Last but not least, computational complexity is an important aspect of explanations in CBR. The construction of the grounded extension \mathbb{G} of a given $(Args, \rightsquigarrow)$ is P-complete [15], so we conjecture that extracting explanations from \mathbb{G} results in a *low construction overhead* [30], as follows: if f is the (fixed) number $|\mathbb{F}|$ of factors, letting n to be the number $|Args|$ of arguments, to construct a (maximal or admissible) dispute tree for (\emptyset, d) we need to traverse the constructed graph of \mathbb{G} from (\emptyset, d) in depth at most f , in every layer exploring at most n^f arguments, so the process is polynomial in n with $O(n^{f^2})$. Precise analysis of this conjecture, as well as the complexity of construction of AA frameworks corresponding to case bases, is left for future work.

References

- [1] Al-Abdulkarim, L., Atkinson, K., Bench-Capon, T.J.M.: Abstract Dialectical Frameworks for Legal Reasoning. In: Hoekstra, R. (ed.) *Leg. Knowl. Inf. Syst.*, pp. 61–70. IOS Press (2014)
- [2] Alevén, V.: Teaching Case-based Argumentation Through a Model and Examples. Ph.D. thesis, University of Pittsburgh (1997)
- [3] Alevén, V.: Using Background Knowledge in Case-Based Legal Reasoning: A Computational Model and an Intelligent Learning Environment. *Artif. Intell.* 150, 183–237 (2003)
- [4] Ashley, K.D.: *Modeling Legal Argument - Reasoning With Cases and Hypotheticals*. MIT Press (1990)
- [5] Athakravi, D., Satoh, K., Law, M., Broda, K., Russo, A.: Automated Inference of Rules with Exception from Past Legal Cases Using ASP. In: Calimeri, F., Ianni, G., Truszczyński, M. (eds.) *LPNMR. Lecture Notes in Computer Science*, vol. 9345, pp. 83–96. Springer, Lexington (2015)

- [6] Bench-Capon, T.J.M., Modgil, S.: Case Law in Extended Argumentation Frameworks. In: ICAIL. pp. 118–127. ACM, Barcelona (2009)
- [7] Brewka, G., Woltran, S.: Abstract Dialectical Frameworks. In: Lin, F., Sattler, U., Truszczyński, M. (eds.) KR. AAAI Press, Toronto (2010)
- [8] Brüninghaus, S., Ashley, K.D.: Predicting Outcomes of Case-Based Legal Arguments. In: Zelezniuk, J., Sartor, G. (eds.) ICAIL. pp. 233–242. ACM, Edinburgh (2003)
- [9] Cerutti, F., Tintarev, N., Oren, N.: Formal Arguments, Preferences, and Natural Language Interfaces to Humans: an Empirical Evaluation. In: Schaub, T., Friedrich, G., O’Sullivan, B. (eds.) ECAI. Frontiers in Artificial Intelligence and Applications, vol. 263, pp. 207–212. IOS Press, Prague (2014)
- [10] Chorley, A., Bench-Capon, T.J.M.: AGATHA: Using Heuristic Search to Automate the Construction of Case Law Theories. *Artif. Intell. Law* 13(1), 9–51 (2005)
- [11] Čyras, K., Satoh, K., Toni, F.: Abstract Argumentation for Case-Based Reasoning. In: Baral, C., Delgrande, J.P., Wolter, F. (eds.) KR. pp. 549–552. AAAI Press, Cape Town (2016)
- [12] Dung, P.M.: On the Acceptability of Arguments and its Fundamental Role in Nonmonotonic Reasoning, Logic Programming and n-person Games. *Artif. Intell.* 77, 321–357 (1995)
- [13] Dung, P.M., Kowalski, R., Toni, F.: Dialectic Proof Procedures for Assumption-Based, Admissible Argumentation. *Artif. Intell.* 170(2), 114–159 (2006)
- [14] Dung, P.M., Mancarella, P., Toni, F.: Computing Ideal Sceptical Argumentation. *Artif. Intell.* 171(10–15), 642–674 (2007)
- [15] Dunne, P.E., Hunter, A., McBurney, P., Parsons, S., Wooldridge, M.: Weighted Argument Systems: Basic Definitions, Algorithms, and Complexity Results. *Artif. Intell.* 175(2), 457–486 (2011)
- [16] Fan, X., Toni, F.: On Computing Explanations in Abstract Argumentation. In: Bonet, B., Koenig, S. (eds.) AAAI. pp. 1496–1502. AAAI Press, Austin, Texas (2015)
- [17] Fan, X., Toni, F.: On Explanations for Non-Acceptable Arguments. In: Black, E., Modgil, S., Oren, N. (eds.) TAFA. Lecture Notes in Computer Science, vol. 9524, pp. 112–127. Springer, Buenos Aires (2015)
- [18] Farley, A.M., Freeman, K.: Burden of Proof in Legal Argumentation. In: McCarty, T. (ed.) ICAIL. pp. 156–164. ACM, College Park (1995)
- [19] García, A.J., Chesñevar, C., Rotstein, N., Simari, G.R.: Formalizing Dialectical Explanation Support for Argument-Based Reasoning in Knowledge-Based Systems. *Expert Syst. Appl.* 40, 3233–3247 (2013)
- [20] Gordon, T., Prakken, H., Walton, D.: The Carneades Model of Argument and Burden of Proof. *Artif. Intell.* 171(10–15), 875–896 (2007)
- [21] Gordon, T., Walton, D.: Legal Reasoning with Argumentation Schemes. In: ICAIL. pp. 137–146. ACM, Barcelona (2009)
- [22] Horty, J., Bench-Capon, T.J.M.: A Factor-Based Definition of Precedential Constraint. *Artif. Intell. Law* 20(2), 181–214 (2012)
- [23] de Mántaras, R.L.: Case-Based Reasoning. In: Paliouras, G., Karkaletsis, V., Spyropoulos, C.D. (eds.) Mach. Learn. Its Appl. Adv. Lect., Lecture Notes in Computer Science, vol. 2049, pp. 127–145. Springer (2001)
- [24] McSherry, D.: Explaining the Pros and Cons of Conclusions in CBR. In: Funk, P., González-Calero, P.A. (eds.) ECCBR. Lecture Notes in Computer Science, vol. 3155, pp. 317–330. Springer, Madrid (2004)
- [25] Ontañón, S., Plaza, E.: An Argumentation-Based Framework for Deliberation in Multi-Agent Systems. In: Rahwan, I., Parsons, S., Reed, C. (eds.) ArgMAS. pp. 178–196. Springer, Honolulu (2007)
- [26] Prakken, H., Sartor, G.: Modelling Reasoning with Precedents in a Formal Dialogue Game. *Artif. Intell. Law* 6(2–4), 231–287 (1998)
- [27] Prakken, H., Wyner, A., Atkinson, K., Bench-Capon, T.J.M.: A Formalization of Argumentation Schemes for Legal Case-Based Reasoning in ASPIC+. *J. Log. Comput.* 25(5), 1141–1166 (2015)
- [28] Richter, M., Weber, R.: Case-Based Reasoning. Springer (2013)
- [29] Schulz, C., Toni, F.: Justifying Answer Sets Using Argumentation. *Theory Pract. Log. Program.* 16(1), 59–110 (2016)
- [30] Sørmo, F., Cassens, J., Aamodt, A.: Explanation in Case-Based Reasoning—Perspectives and Goals. *Artif. Intell. Rev.* 24(2), 109–143 (2005)
- [31] Wardeh, M., Bench-Capon, T.J.M., Coenen, F.: PADUA: A Protocol for Argumentation Dialogue Using Association Rules. *Artif. Intell. Law* 17(3), 183–215 (2009)

Quantifying the Difference Between Argumentation Semantics

Sylvie DOUTRE^a

^a *IRIT, University of Toulouse 1, France, doutre@irit.fr*

Jean-Guy MAILLY^b

^b *Institute of Information Systems, TU Wien, Austria,
jmailly@dbai.tuwien.ac.at*

Abstract. Properties of argumentation semantics have been widely studied in the last decades. However, there has been no investigation on the question of difference measures between semantics. Such measures turn helpful when the semantics associated to an argumentation framework may have to be changed, in a way that ensures that the new semantics is not too dissimilar from the old one. Three main notions of difference measures between semantics are defined in this paper. Some of these measures are shown to be distances or semi-distances.

Keywords. Abstract argumentation, extension-based semantics

1. Introduction

Abstract argumentation frameworks (AFs) are classically associated with a semantics which allows to evaluate arguments' statuses, determining sets of jointly acceptable arguments called extensions [7,1]. In [2], a method to modify an AF in order to satisfy a constraint (a given set of arguments should be an extension, or at least included in an extension) is defined; this process is called extension enforcement. The authors distinguish between conservative enforcement when the semantics does not change (only the AF changes) and liberal enforcement when the semantics changes. But they do not explain why the semantics should change, nor which semantics should be the new one.

Apart from this use of a semantic change for an extension enforcement purpose, a change of the semantics may be necessary for other reasons, for instance, for computational purposes: if a given semantics was appropriate at some point in a certain context for some AF, one may imagine that changes over time on the structure of the AF (number of arguments, of attacks) may make this semantics too "costly" to compute, and then not appropriate anymore. It may be interesting to pick up another semantics to apply to the AF, possibly not too dissimilar to the former one.

In another revision context, [5] defines revision operators for AFs which proceed in two steps. First, revised extensions are computed, then a set of AFs is

associated with these revised extensions. Indeed, it is not possible in general to associate a single AF with an arbitrary set of extensions, under a chosen semantics. Other revision approaches for argumentation may also result in a set of AFs [6]. Modifying the semantics in the revision process may permit to obtain a single AF in some situations, or at least to minimize the number of AFs in the result.

Whatever be the context where a semantic change is necessary, we think that such a semantic change should not be performed any old how, and should respect some kind of minimality, exactly as belief change operations usually require minimal change (see e.g. [9] for belief revision in a propositional setting). Defining *difference measures* between semantics, to quantify how much a semantics is dissimilar to another one, allows to define different minimality criteria. Such criteria can be used to select the new semantics among several options when a semantic change occurs.

Main contribution We propose in this paper three sensible ways to quantify the difference between two semantics:

- depending on the properties which characterize the semantics;
- depending on the relations between semantics;
- depending on the acceptance statuses of arguments the semantics lead to.

The first ones (property-based and relation-based) are said to be *absolute* measures, since they only depend on the considered semantics; they apply to any graph. The last one (acceptance-based) is said to be *relative*: the definition of the measure depends on a particular AF. We study the properties of our measures, in particular we show that some of them are distances or semi-distances.

2. Background Notions

An argumentation framework (AF) [7] is a directed graph $\langle A, R \rangle$ where the nodes in A represent abstract entities called *arguments* and the edges in R represent *attacks* between arguments. $(a_i, a_j) \in R$ means that a_i attacks a_j ; a_i is called an *attacker* of a_j . We say that an argument a_i (resp. a set of arguments S) defends the argument a_j against its attacker a_k if a_i (resp. any argument in S) attacks a_k . The *range* of a set of arguments S w.r.t. R , denoted S_R^+ , is the subset of A which contains S and the arguments attacked by S ; formally $S_R^+ = S \cup \{a_j \mid \exists a_i \in S \text{ s.t. } (a_i, a_j) \in R\}$. Different semantics allow to determine which sets of arguments can be collectively accepted [7,1].

Definition 1. Let $F = \langle A, R \rangle$ be an AF. A set of arguments $S \subseteq A$ is

- conflict-free w.r.t. F if $\nexists a_i, a_j \in S$ s.t. $(a_i, a_j) \in R$;
- admissible w.r.t. F if S is conflict-free and S defends each of its arguments against all of their attackers;
- a naive extension of F if S is a maximal conflict-free set (w.r.t. \subseteq);
- a complete extension of F if S is admissible and S contains all the arguments that it defends;
- a preferred extension of F if S is a maximal complete extension (w.r.t. \subseteq);

- a stable extension of F if S is conflict-free and S attacks each argument in $A \setminus S$;
- a grounded extension of F if S is a minimal complete extension (w.r.t. \subseteq);
- a stage extension of F if S is conflict-free and there is no conflict-free T such that $S_R^+ \subset T_R^+$;
- a semi-stable extension of F if S is admissible and there is no admissible T such that $S_R^+ \subset T_R^+$.

These semantics are denoted, respectively, *cf*, *adm*, *na*, *co*, *pr*, *st*, *gr*, *stg*, *sem*. For each σ of them, $Ext_\sigma(F)$ denotes the set of σ -extensions of F .

Example 1. Let us consider the argumentation framework F_1 given at Figure 1, and let us illustrate some of the semantics. $Ext_{adm} = \{\emptyset, \{a_4, a_6\}, \{a_1, a_4, a_6\}, \{a_1, a_3\}, \{a_1, a_4\}, \{a_1\}, \{a_4\}\}$, $Ext_{st}(F) = \{\{a_1, a_4, a_6\}\}$, $Ext_{pr}(F) = \{a_1, a_3\}, \{a_1, a_4, a_6\}\}$, $Ext_{co}(F) = \{\{a_1, a_4, a_6\}, \{a_1, a_3\}, \{a_1\}\}$, $Ext_{gr}(F) = \{\{a_1\}\}$.

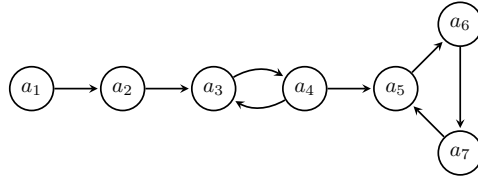


Figure 1. The AF F_1

In order to compare, in the following section, the semantics, and propose measures of their differences, let us introduce a useful notation: given two sets X, Y , $X \Delta Y$ is the symmetric difference between X and Y . Let us recall also the definition of a distance and of a semi-distance.

Definition 2. Given a set E , a mapping d from $E \times E$ to \mathbb{R}^+ satisfies:

coincidence if, $\forall x, y \in E, d(x, y) = 0$ iff $x = y$;

symmetry if $\forall x, y \in E, d(x, y) = d(y, x)$;

triangular inequality if $\forall x, y, z \in E, d(x, y) + d(y, z) \geq d(x, z)$.

Such a mapping d is then:

- a semi-distance if it satisfies coincidence and symmetry;
- a distance if it satisfies coincidence, symmetry and triangular inequality.

3. Property-based Difference Measures

We propose a first way to measure how much two semantics are different. This way relies upon the idea of splitting a semantics into a set of properties, or principles (following the idea of [3]), which characterize it. A weight can then be given to each property, these weights corresponding to the importance of the property in the context where the semantics have to be compared. Then, measuring the difference between two semantics is equivalent to adding the weight of the properties which appear in the characterization of exactly one of the semantics.

Definition 3. A set of properties \mathcal{P} characterizes a semantics σ if for each AF F ,

1. each σ -extension of F satisfies each property from \mathcal{P} ,
2. each set of arguments which satisfies \mathcal{P} is a σ -extension of F ,
3. \mathcal{P} is a minimal set (w.r.t \subseteq) among those which satisfy 1. and 2.

$\text{Prp}(\sigma)$ denotes the set of properties that characterizes a semantics σ .

Beyond the use of characterizations to define difference measures, let us point out the fact that they can have a computational interest. For instance, verifying if a set of arguments is a σ -extension can be done by checking if it satisfies all the properties in $\text{Prp}(\sigma)$. In this case, the computation can stop as soon as one of the properties is not satisfied.

Let us point out interesting properties, and establish which ones characterize each semantics. We distinguish between absolute properties (which concern only a set of arguments itself, Definition 4) and relative properties (which concern a set of arguments with respect to other sets of arguments, Definition 5).

Definition 4. Given an AF $F = \langle A, R \rangle$, a set of arguments S satisfies

- *conflict-freeness* if S is conflict-free;
- *acceptability* (*accpt.*) if S defends itself against each attacker;
- *reinstatement* (*reins.*) if S contains all the arguments that it defends;
- *complement attack* (*comp. att.*) if each argument in $A \setminus S$ is attacked by S .

Definition 5. Given an AF $F = \langle A, R \rangle$ and a set of properties \mathcal{P} , a set of arguments S satisfies

- \mathcal{P} -max if S is \subseteq -maximal among the sets of arguments which satisfy \mathcal{P} ;
- \mathcal{P} -min if S is \subseteq -minimal among the sets of arguments which satisfy \mathcal{P} ;
- \mathcal{P} -R-max if S has a \subseteq -maximal range among the sets of arguments which satisfy \mathcal{P} .

Now, we establish a characterization of the different semantics, that follows from the previous definitions.

Proposition 1. The extension-based semantics considered in this paper can be characterized as follows:

$$\begin{array}{ll}
 \text{Prp}(cf) = \{\text{conflict-freeness}\} & \text{Prp}(sem) = \text{Prp}(adm)\text{-R-max} \\
 \text{Prp}(adm) = \text{Prp}(cf) \cup \{\text{accpt}\} & \text{Prp}(stg) = \text{Prp}(cf)\text{-R-max} \\
 \text{Prp}(na) = \text{Prp}(cf)\text{-max} & \text{Prp}(st) = \text{Prp}(cf) \cup \{\text{comp. att.}\} \\
 \text{Prp}(co) = \text{Prp}(adm) \cup \{\text{reins.}\} & \text{Prp}(gr) = \text{Prp}(co)\text{-min} \\
 \text{Prp}(pr) = \text{Prp}(adm)\text{-max} &
 \end{array}$$

Let us notice that we may consider other properties, and give alternative characterizations of the semantics. Even if the value of the difference between two semantics (obviously) depends of the chosen characterizations, the general definition of property-based difference measures is the same whatever the characterizations.

Our intuition which leads to define the characterization as the minimal set of properties is related to computational issues. Indeed, computing some reasoning

tasks related to the semantics thanks to the semantics characterization can be done more efficiently with this definition. For instance, to determine whether a set of arguments is a stable extension of a given AF, checking the satisfaction of conflict-freeness and complement attack proves enough. We may add $\text{Prp}(adm)\text{-max}$ in the characterization of the stable semantics, but computing the result of our problem would then be harder.

A weight can be associated to each property, depending on the importance of the property in a certain context.

Definition 6. Let \mathcal{P} be a set of properties. Let w be a function which maps each property $p \in \mathcal{P}$ to a strictly positive real number $w(p)$. Given σ_1, σ_2 two semantics such that $\text{Prp}(\sigma_1) \subseteq \mathcal{P}$ and $\text{Prp}(\sigma_2) \subseteq \mathcal{P}$, the property-based difference measure δ_{prop}^w between σ_1 and σ_2 is defined as $\delta_{prop}^w(\sigma_1, \sigma_2) = \sum_{p_i \in \text{Prp}(\sigma_1) \Delta \text{Prp}(\sigma_2)} w(p_i)$.

The specific property-based difference measure when all the properties have the same importance is defined as follows.

Definition 7. Given two semantics σ_1, σ_2 , the property-based difference measure δ_{prop} is defined by $\delta_{prop}(\sigma_1, \sigma_2) = |\text{Prp}(\sigma_1) \Delta \text{Prp}(\sigma_2)|$.

Example 2. Let us suppose that the initial semantics is the admissible one. When we consider δ_{prop} , $\delta_{prop}(adm, co) = 1$ and $\delta_{prop}(adm, st) = 2$; in other words, the complete semantics is “better” than the stable semantics, because closer to the admissible semantics. However, with a weighted measure δ_{prop}^w such that $w(reins.) = 2$ and the weight of the other properties is 1, the complete and the stable semantics turn “equivalent” since $\delta_{prop}(adm, co) = \delta_{prop}(adm, st) = 2$.

Proposition 2. Given a set of semantics \mathcal{S} , the property-based measures defined on \mathcal{S} are distances.

4. Relation-based Difference Measures

The second absolute method to measure the difference between semantics that we propose, is based on the fact that most of the usual semantics are related according to some notions. For instance, it is well-known that each preferred extension of an AF is also a complete extension of it, and the grounded extension is also complete, but in general it is not a preferred extension. The preferred semantics may thus be seen as closer to the complete semantics, than to the grounded semantics. We formalize this idea with the notion of semantics relation graph.

Definition 8. Let $\mathcal{S} = \{\sigma_1, \dots, \sigma_n\}$ a set of semantics. A semantics relation graph on \mathcal{S} is defined by $\text{Rel}(\mathcal{S}) = \langle \mathcal{S}, D \rangle$ with $D \subseteq \mathcal{S} \times \mathcal{S}$.

This abstract notion of relation graph, where the nodes are semantics, can be instantiated with the inclusion relation between the extensions of an AF.

Definition 9. Let $\mathcal{S} = \{\sigma_1, \dots, \sigma_n\}$ a set of semantics. The extension inclusion graph of \mathcal{S} is defined by $\text{Inc}(\mathcal{S}) = \langle \mathcal{S}, D \rangle$ with $D \subseteq \mathcal{S} \times \mathcal{S}$ such that $(\sigma_i, \sigma_j) \in D$ if and only if:

- for each AF F , $Ext_{\sigma_i}(F) \subseteq Ext_{\sigma_j}(F)$;
- there is no $\sigma_k \in \mathcal{S}$ ($k \neq i, k \neq j$) such that for each AF F , $Ext_{\sigma_i}(F) \subseteq Ext_{\sigma_k}(F)$ and $Ext_{\sigma_k}(F) \subseteq Ext_{\sigma_j}(F)$.

This idea is discussed in [1], but that paper does not formalize the notion of relation between semantics as we do here.

Example 3. For instance, when $\mathcal{S} = \{co, pr, st, gr, stg, sem, adm, cf, na\}$, $Inc(\mathcal{S})$ is the graph given at Figure 2.

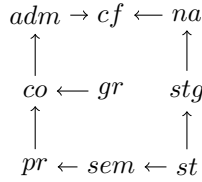


Figure 2. Extension Inclusion Graph $Inc(\mathcal{S})$

Now, we define a family of difference measures between semantics which is based on the semantics relation graphs.

Definition 10. Given \mathcal{S} a set of semantics, a \mathcal{S} -relation difference measure is the mapping from two semantics $\sigma_1, \sigma_2 \in \mathcal{S}$ to the non-negative integer $\delta_{Rel, \mathcal{S}}(\sigma_1, \sigma_2)$ which is the length of the shortest non-oriented path between σ_1 and σ_2 in $Rel(\mathcal{S})$. In particular, the \mathcal{S} -inclusion measure is the length of the shortest non-oriented path between σ_1 and σ_2 in $Inc(\mathcal{S})$, denoted by $\delta_{Inc, \mathcal{S}}(\sigma_1, \sigma_2)$.

Example 4. Given two semantics σ_1 and σ_2 which are neighbours in the graph given at Figure 2, the difference measure $\delta_{Inc, \mathcal{S}}(\sigma_1, \sigma_2)$ is obviously 1. Otherwise, if several paths allow to reach σ_2 from σ_1 , then the difference is the length of the minimal one. For instance, $\delta_{Inc, \mathcal{S}}(st, cf) = 3$ since the minimal path is $st \rightarrow stg \rightarrow na \rightarrow cf$, but other paths exist (for instance, $st \rightarrow sem \rightarrow pr \rightarrow co \rightarrow adm \rightarrow cf$).

Proposition 3. The \mathcal{S} -inclusion difference measure is a distance.

For the possible instantiations of the relation graph that have been proposed, we can also define a relative version. In this case, the edges in the graph depend on the inclusion relations for a given AF, while our first proposal considers the inclusion relations which are true for any AF. This AF-based relation graph can lead to an interesting new measure.

We may instantiate the relation graph with another relation between semantics such as, for instance, the graph resulting from the intertranslatability relationship of semantics [8].

5. Acceptance-based Difference Measures

We have previously defined two approaches to quantify the difference between semantics which are absolute, which means that the difference between two semantics is always the same, whatever the situation and the AF. It may be interesting for some applications to take into account the AF of the agent to measure the difference between the semantics. We propose here such a family of measures. Now, the difference between two semantics σ_1 and σ_2 depends on the acceptance status of arguments in a given AF, w.r.t. the different semantics into consideration.

Our first acceptance-based measure quantifies the difference between the σ_1 -extensions and the σ_2 -extension of the AF.

Definition 11. Let F be an AF, and d be a distance between sets of arguments. The F - d -extension-based difference measure δ_F^d is defined by $\delta_F^d(\sigma_1, \sigma_2) = \sum_{\epsilon \in \text{Ext}_{\sigma_1}(F)} \min_{\epsilon' \in \text{Ext}_{\sigma_2}(F)} d(\epsilon, \epsilon')$.

In general, the F - d -extension-based difference measures are not distances, they do not satisfy coincidence, symmetry.

Example 5. For instance, we consider the Hamming distance between sets of arguments, defined as $d_H(s_1, s_2) = |s_1 \Delta s_2|$. Now, we define the F_1 - d_H -extension-based difference measure $\delta_{F_1}^{d_H}$ from d_H and the AF F_1 given at Figure 1. Its set of stable extensions is $\text{Ext}_{st}(F_1) = \{\{a_1, a_4, a_6\}\}$.

When measuring the difference between the stable semantics and the grounded semantics, we obtain $\delta_{F_1}^{d_H}(st, gr) = 2$ since $\text{Ext}_{gr}(F_1) = \{\{a_1\}\}$. $\delta_{F_1}^{d_H}(st, pr) = 0$ since $\text{Ext}_{pr}(F_1) = \{\{a_1, a_3\}, \{a_1, a_4, a_6\}\}$; on the opposite, $\delta_{F_1}^{d_H}(pr, st) = 3$.

From this measure, a new one, which satisfies symmetry, can be defined.

Definition 12. Let F be an AF, and d be a distance between sets of arguments. The symmetric F - d -extension-based difference measure $\delta_{F, sym}^d$ is defined by $\delta_{F, sym}^d(\sigma_1, \sigma_2) = \max(\delta_F^d(\sigma_1, \sigma_2), \delta_F^d(\sigma_2, \sigma_1))$.

This measure satisfies the semi-distance properties under some conditions.

Proposition 4. For a given F and a given set of semantics $\mathcal{S} = \{\sigma_1, \dots, \sigma_n\}$, if for all $\sigma_i, \sigma_j \in \mathcal{S}$ such that $\sigma_i \neq \sigma_j$, $\text{Ext}_{\sigma_i}(F) \neq \text{Ext}_{\sigma_j}(F)$, then the symmetric extension-based measure $\delta_{F, sym}^{d_H}$ is a semi-distance.

We can also define similar measures based on the set of credulously (resp. skeptically) accepted arguments, instead of the whole set of extensions.

6. Conclusion

In this paper, we have defined several ways to quantify the difference between extension-based semantics. Some of them are absolute (they only depend on the semantics), while the other ones are relative (they depend on the considered AF). Let us mention the fact that there is no general relation between these

difference measures; for instance it may occur that $\delta_1(\sigma_1, \sigma_2) > \delta_1(\sigma_1, \sigma_3)$ while $\delta_2(\sigma_1, \sigma_2) < \delta_2(\sigma_1, \sigma_3)$ (e.g. $\delta_{F_1, sym}^{d_H}(st, gr) < \delta_{F_1, sym}^{d_H}(st, pr)$ while $\delta_{Inc, S}(st, gr) > \delta_{Inc, S}(st, pr)$ for S as in Example 3). When a semantic change occurs, this permits the agent to use some very different notions of minimality to select the new semantics, depending on which difference measures make sense in the context of her application. In addition, the combination of these “basic” measures permits to express even more notions of minimality.

Let us notice that only the relation-based and property-based measures are distances, other methods failing in general to satisfy the distance properties, which seem to be desirable to quantify the difference between objects. Further study could lead to identify the necessary conditions that a set of semantics must satisfy to ensure that these are distances.

We consider several tracks for future work. We have noticed that we can order semantics, with respect to an initial semantics σ and a measure δ : $\sigma_1 \leq_{\sigma, \delta} \sigma_2$ if and only if $\delta(\sigma, \sigma_1) \leq \delta(\sigma, \sigma_2)$. In this case, we can investigate the relation of the orderings defined by different measures. For instance, if some pairs (σ, δ_1) and (σ, δ_2) lead to the same ordering, then we can choose to use the measure which is the least expensive one to compute among δ_1 and δ_2 .

We also plan to define a similar notion of difference measures for labelling-based semantics [1], and for ranking-based semantics [4]. In this last context, we need to determine whether some relevant properties characterize the ranking which is used to evaluate arguments, or to determine meaningful notions of difference between the rankings.

Finally, we will investigate the issue which is mentioned in the introduction: using (minimal) semantic change to define enforcement and revision methods.

Acknowledgements This work benefited from the support of the project AMANDE ANR-13-BS02-0004 of the French National Research Agency (ANR), and from the support of the Austrian Science Fund (FWF) under grants P25521 and I1102.

References

- [1] P. Baroni, M. Caminada, and M. Giacomin. An introduction to argumentation semantics. *Knowledge Eng. Review*, 2011.
- [2] R. Baumann. What does it take to enforce an argument? Minimal change in abstract argumentation. In *Proc. ECAI'12*, pages 127–132, 2012.
- [3] P. Besnard, S. Doutre, and A. Herzig. Encoding argument graphs in logic. In *Proc. IPMU'14*, pages 345–354, 2014.
- [4] E. Bonzon, J. Delobelle, S. Konieczny, and N. Maudet. A comparative study of ranking-based semantics for abstract argumentation. In *Proc. AAAI'16*, 2016.
- [5] S. Coste-Marquis, S. Konieczny, J.-G. Mailly, and P. Marquis. On the revision of argumentation systems: Minimal change of arguments statuses. In *Proc. KR'14*, 2014.
- [6] S. Doutre, A. Herzig, and L. Perrussel. A dynamic logic framework for abstract argumentation. In *Proc. KR'14*, pages 62–71, 2014.
- [7] P. M. Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming, and n-person games. *Artif. Intell.*, 77(2):321–357, 1995.
- [8] W. Dvořák and C. Spanring. Comparing the expressiveness of argumentation semantics. *Journal of Logic and Computation*, 2016.
- [9] H. Katsuno and A. O. Mendelzon. Propositional knowledge base revision and minimal change. *Artificial Intelligence*, 52:263–294, 1991.

A Canonical Semantics for Structured Argumentation with Priorities

Phan Minh Dung

AIT, Thailand, email: dung.phanminh@gmail.com

Abstract. Due to a proliferation and diversity of approaches to reasoning with prioritized rules, ordinary properties have been introduced recently for characterization and evaluation of the proposed semantics. While ordinary properties are helpful, a fundamental question of whether they are sufficient to identify a common semantics underlining reasoning with priorities remains open. In this paper we address this question by introducing a new simple and intuitive property of inconsistency-resolving and slightly adapting other ordinary properties to show that they together indeed determine an unique canonical attack relation that could be viewed as defining an uniquely defined common semantics for reasoning with prioritized rules.

Keywords. regular properties, attack relation assignment, semilattice

1. Introduction

Reasoning with prioritized rules is an important and prevalent paradigm in practical reasoning like legal reasoning or commonsense reasoning [3,5,11,14]. Due to a proliferation and diversity of approaches [3,6,5,20,11,24,14,22,23], it is important to establish general principles for characterizing and evaluation of the proposed semantics. Earlier, Brewka and Eiter [6] have proposed two principles for non-argument-based approaches. Caminada and Amgoud [8] have introduced the postulates of consistency and closure that the extensions of argument-based systems should satisfy. A subargument closure postulate stating that any extension should contain all subarguments of its arguments has been studied in [22,21,1]. Though the three proposed postulates are helpful, they are not sufficient to guarantee intuitive semantics as they do not take into account the preferences of defeasible rules. To address this problem, Dung [17,12] has proposed a set of simple properties, referred to as ordinary properties in [12] and argued that they capture the natural intuitions of reasoning with prioritized rules. Still, a fundamental question of whether the proposed properties are sufficient to identify a common semantics underlining reasoning with priorities remains open. In this paper we address this question by introducing a new simple and intuitive property of inconsistency-resolving and showing that this property together with some other ordinary properties indeed determine an unique canonical attack relation that could be viewed as defining a common unique semantics for reasoning with prioritized rules.

The paper is organized as follows. We recall in the next section the key concepts and notions on which the paper is based. We then introduce the important property of inconsistency-resolving in the following section. In section 4, we introduce the new and novel concepts of regular attack relations and regular at-

tack relation assignments. In section 5, we study the semilattice of regular attack relation assignments and propose the canonical semantics. We then conclude.

2. Preliminaries

2.1 Abstract Argumentation and Semilattice

An abstract argumentation framework [16] is defined simply as a pair (AR, att) where AR is a set of arguments and $att \subseteq AR \times AR$ where $(A, B) \in att$ means that A attacks B . A set of argument S attacks (or is attacked by) an argument A (or a set of arguments R) if some argument in S attacks (or is attacked by) A (or some argument in R); S is *conflict-free* if it does not attack itself. A set of arguments S *defends* an argument A if S attacks each attack against A . S is *admissible* if S is conflict-free and defends each argument in it. A *complete extension* is an admissible set of arguments containing each argument it defends. A *stable extension* is a conflict-free set of arguments that attacks every argument not belonging to it.

A partial order (i.e. a reflexive, transitive and antisymmetric relation) \leq on a set S is a **upper-semilattice** (resp. **lower-semilattice**) [10] iff each subset X of S has a supremum denoted by $\sqcup X$ (resp. infimum denoted by $\sqcap X$) wrt \leq . The upper (resp. lower) semilattice is often denoted as a triple (S, \leq, \sqcup) (resp. (S, \leq, \sqcap)). It follows immediately that each upper (resp. lower) semilattice S has a unique greatest (resp. least) element denoted by $\sqcup S$ (resp. $\sqcap S$).

2.2 Defeasible Knowledge Bases

In this section and the following one, we recall the basic notions and notations on knowledge bases from [12,22]. We assume a non-empty set \mathcal{L} of ground atoms (also called a positive literal) and their classical negations (also called negative literals). A set of literals is said to be *contradictory* iff it contains an atom a and its negation $\neg a$. We distinguish between *domain atoms* representing propositions about the concerned domains and *non-domain atoms* of the form ab_d representing the non-applicability of defeasible rule d (even if the premises of d hold).

Following [22,23,19,20,25,12], we distinguish between strict and defeasible rules. A *defeasible* (resp. *strict*) rule r is of the form $b_1, \dots, b_n \Rightarrow h$ (resp. $b_1, \dots, b_n \rightarrow h$) where b_1, \dots, b_n are domain literals and h is a domain literal or an atom of the form ab_d . The set $\{b_1, \dots, b_n\}$ (resp. the literal h) is referred to as the *body* (resp. *head*) of r and denoted by $bd(r)$ (resp. $hd(r)$).

Definition 1 (1.) A **rule-based system** is defined as a triple $\mathcal{R} = (RS, RD, \preceq)$ where 1) RS is a set of strict rules, 2) RD is a set of defeasible rules, and 3) \preceq is a transitive relation over RD representing the preferences between defeasible rules, whose strict core is \prec (i.e. $d \prec d'$ iff $d \preceq d'$ and $d' \not\preceq d$ for $d, d' \in RD$.) (2.) A **knowledge base** is defined as a pair $K = (\mathcal{R}, BE)$ consisting of a rule-based system \mathcal{R} , and a set of ground domain literals BE , the base of evidence of K , representing unchallenged observations, facts ect..

For convenience, knowledge base K is often written directly as a quadruple (RS, RD, \preceq, BE) where RS , RD , \preceq or BE of K are often referred to by RS_K , RD_K , \preceq_K or BE_K respectively.

(3.) A knowledge base K is **basic** if its precedence relation is empty (i.e. $\preceq_K = \emptyset$).

Definition 2 Let $K = (RS, RD, \preceq, BE)$ be a knowledge base. An **argument** wrt K is a proof tree defined inductively as follows:

- (1.) For each $\alpha \in BE$, $[\alpha]$ is an argument with conclusion α .
- (2.) Let r be a rule of the forms $\alpha_1, \dots, \alpha_n \rightarrow / \Rightarrow \alpha$, $n \geq 0$, from $RS \cup RD$ and A_1, \dots, A_n be arguments with conclusions α_i , $1 \leq i \leq n$, respectively. Then $A = [A_1, \dots, A_n, r]$ is an argument with **conclusion** α and **last rule** r denoted by $\text{cnl}(\mathbf{A})$ and $\text{last}(\mathbf{A})$ respectively.
- (3.) Each argument wrt K is obtained by applying the above steps 1, 2 finitely many times.

Example 1 Consider a rule-based system \mathcal{R} (adapted from [6,7,12]) whose sets of rules consisting of three defeasible rules $d_1 : \text{Dean} \Rightarrow \text{Professor}$, $d_2 : \text{Professor} \Rightarrow \text{Teach}$, $d_3 : \text{Administrator} \Rightarrow \neg \text{Teach}$ and two strict rules $r : \text{Dean} \rightarrow \text{Administrator}$, $r' : \neg \text{Administrator} \rightarrow \neg \text{Dean}$ together with a precedence relation consisting of just $d_2 \prec d_3$. Suppose we know some dean who is also a professor. The considered knowledge base is represented by $K = (RS, RD, \preceq, BE)$ with $RS = \{r, r'\}$, $RD = \{d_1, d_2, d_3\}$, $\preceq = \{(d_2, d_3)\}$ and $BE = \{D, P\}$ (D, P, T, A stand for Dean, Professor, Teach and Administrator respectively). Relevant arguments can be found in figure 1 where $A_1 = [[D], d_1]$, $A_2 = [A_1, d_2]$, $A'_2 = [[P], d_2]$, $A_3 = [[[D], r], d_3]$.

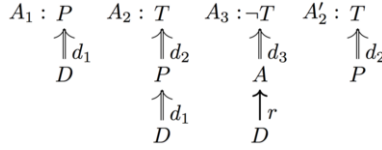


Figure 1. Dean Example

Notation 1 The set of all arguments wrt a knowledge base K is denoted by \mathbf{AR}_K . The set of the conclusions of arguments in a set $S \subseteq \mathbf{AR}_K$ is denoted by $\text{cnl}(S)$.

A **strict argument** is an argument containing no defeasible rule. An argument is **defeasible** iff it is not strict. A defeasible argument A is called **basic defeasible** iff $\text{last}(A)$ is defeasible. For any argument A , the set of defeasible rules appearing in A is denoted by $\mathbf{dr}(\mathbf{A})$. The set of last defeasible rules in A , denoted by $\mathbf{ldr}(\mathbf{A})$, is $\{\text{last}(A)\}$ if A is basic defeasible, otherwise it is equal $\mathbf{ldr}(A_1) \cup \dots \cup \mathbf{ldr}(A_n)$ where $A = [A_1, \dots, A_n, r]$. An argument B is a **subargument** of an argument A iff $B = A$ or $A = [A_1, \dots, A_n, r]$ and B is a subargument of some A_i . B is a proper subargument of A if B is a subargument of A and $B \neq A$.

Definition 3 (1.) The **closure** of a set of literals $X \subseteq \mathcal{L}$ wrt knowledge base K , denoted by $\text{CN}_K(X)$, is the union of X and the set of conclusions of all strict arguments wrt knowledge base $(RS_K, RD_K, \preceq_K, X_{\text{dom}})$ with X_{dom} (the set of all domain literals in X) acting as a base of evidence. X is said to be **closed** iff $X = \text{CN}_K(X)$. X is said to be **inconsistent** iff its closure $\text{CN}_K(X)$ is contradictory. X is **consistent** iff it is not inconsistent. We also often write $X \vdash_K l$ iff $l \in \text{CN}_K(X)$.

(2.) K is said to be **consistent** iff its base of evidence BE_K is consistent.

As the notions of closure, consistency depend only on the set of strict rules in the knowledge base, we often write $X \vdash_{RS} l$ or $l \in CN_{RS}(X)$ for $X \vdash_K l$ or $l \in CN_K(X)$ respectively.

Definition 4 Let $\mathcal{R} = (RS, RD, \preceq)$ be a rule-based system and $K = (\mathcal{R}, BE)$ be a knowledge base.

- (1.) \mathcal{R} and K are said to be closed under transposition [8] iff for each strict rule of the form $b_1, \dots, b_n \rightarrow h$ in RS s.t. h is a domain literal, all the rules of the forms $b_1, \dots, b_{i-1}, \neg h, b_{i+1}, \dots, b_n \rightarrow \neg b_i$, $1 \leq i \leq n$, also belong to RS .
- (2.) \mathcal{R} and K are said to be closed under contraposition [24,23] iff for each set of domain literals S , each domain literal λ , if $S \vdash_{RS} \lambda$ then for each $\sigma \in S$, $S \setminus \{\sigma\} \cup \{\neg \lambda\} \vdash_{RS} \neg \sigma$.
- (3.) \mathcal{R} and K are said to satisfy the self-contradiction property [15] iff for each minimal inconsistent set of domain literals $X \subseteq \mathcal{L}$, for each $x \in X$, it holds: $X \vdash_{RS} \neg x$.

Lemma 1 ([12]) Let \mathcal{R} be a rule-based system that is closed under transposition or contraposition. Then \mathcal{R} satisfies the property of self-contradiction.

Definition 5 (Attack Relation) An attack relation for a knowledge base K is a relation $att \subseteq AR_K \times AR_K$ such that there is no attack against strict arguments, i.e. for each strict argument $B \in AR_K$, there is no argument $A \in AR_K$ such that $(A, B) \in att$.

For convenience, we often say A attacks B wrt att for $(A, B) \in att$.

2.3 Basic Postulates

We recall the postulates of consistency, closure and subargument closure from [8,22,1,21] where we combine the postulate of closure [8] and the postulate of subargument closure [22,1,21] into one.

Definition 6 Let att be an attack relation for a knowledge base K .

- att is said to satisfy the **consistency postulate** iff for each complete extension E of (AR_K, att) , the set $cnl(E)$ of conclusions of arguments in E is consistent.
- att is said to satisfy the **closure postulate** iff for each complete extension E of (AR_K, att) , the set $cnl(E)$ of conclusions of arguments in E is closed and E contains all subarguments of its arguments.

For ease of reference, the above two postulates are often referred to as **basic postulates**.

3. Sufficient Properties for Basic Postulates

As the basic postulates are more about the "output" of attack relations rather than about their structure, we present below two simple properties about the structure of attack relation that ensures the holding of the basic postulates. We first introduce some simple notations.

We say A **undercuts** B (at B') iff B' is basic defeasible and $cnl(A) = ab_{last(B')}$. We also say A **rebuts** B (at B') iff B' is a basic defeasible subargument of B and the conclusions of A and B' are contradictory [8,22].

An argument A is said to be **generated by** a set S of arguments iff all basic defeasible subarguments of A are subarguments of arguments in S . For an example, let $S = \{B_0, B_1\}$ (see figure 2). Let consider A_0 . The set of basic defeasible subarguments of A_0 is $\{[d_0]\}$. It is clear that $[d_0]$ is a subargument of B_0 . Hence A_0 is generated by S . Similarly, A_1 is also generated by S .

We say A *directly attacks* B if A attacks B and A does not attack any proper subargument of B .

Definition 7 (Strong Subargument Structure) *Attack relation att is said to satisfy the property of strong subargument structure for K iff for all $A, B \in AR_K$, followings hold:*

- (1.) *If A undercuts B then A attacks B wrt att .*
- (2.) *A attacks B (wrt att) iff A attacks a basic defeasible subargument of B (wrt att).*
- (3.) *If A directly attacks B (wrt att) then A undercuts B (at B) or rebuts B (at B).*

We present the first result showing that strong subargument property is sufficient to guarantee the postulate of closure.

Lemma 2 *Let att be an attack relation for knowledge base K satisfying the property of strong subargument structure. Then att satisfies the postulate of closure.*

Proof (Sketch) From condition 2 in definition 7, it follows that each attack against an argument generated by complete extension E is an attack against E . The lemma holds obviously. \square

A set S of arguments is said to be *inconsistent* if the set of the conclusions of its arguments, $cnl(S)$, is inconsistent. We introduce below a new simple property of inconsistency resolving, a key result of the paper.

Definition 8 (Inconsistency Resolving) *We say attack relation assignment att satisfies the inconsistency-resolving property for K iff for each finite set of arguments $S \subseteq AR_K$, if S is inconsistent then S is attacked (wrt $att(K)$) by some argument generated by S .*

As we will show later, the inconsistency-resolving property is satisfied by common conditions like closure under transposition, or contradiction or the property of self-contradiction.

Example 2 *Consider the basic knowledge base K consisting of just the rules appearing in arguments in figure 2. The set $S = \{B_0, B_1\}$ is inconsistent. The argument A_0 is generated by S . Let $att = \{(X, Y) \mid X \text{ rebuts } Y\}$. It is obvious that S is attacked by A_0 . It is clear that att is inconsistency-resolving.*

We present now the first important result of this paper.

Theorem 1 *Let att, att' be attack relations for knowledge base K . (1.) If $att \subseteq att'$ and att is inconsistency-resolving for K then att' is also inconsistency-resolving for K ;*

(2.) If att satisfies the strong subargument structure and inconsistency-resolving then att satisfies the postulate of consistency.

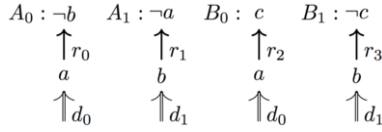


Figure 2. Generated Arguments

Proof (Sketch) Assertion 1 follows easily from the definition of inconsistency-resolving. We only need to show assertion 2. From condition 2 in definition 7, it follows that each argument generated by a complete extension E belongs to E . Therefore, if E is inconsistent then E is conflicting. Since E is not conflicting, E is hence consistent. \square

4. Regular Attack Relation Assignments

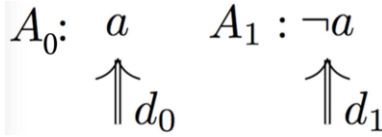


Figure 3. Effective Rebuts

Dung [17,12] has proposed the ordinary properties to capture the intuition of prioritized rules. We recall and adapt them below. We also motivate and explain their intuitions. We then define two new novel concepts of regular attack relations and regular attack relation assignments that lie at the heart of the semantics of prioritized rules.

4.1 A Minimal Interpretation of Priorities

We first recall from [12] the effective rebut property stating a "minimal interpretation" of a preference $d_0 \prec d_1$ that in situations when both are applicable but accepting both d_0, d_1 is not possible, d_1 should be preferred. In figure 3, the effective rebut property dictates that A_1 attacks A_0 but not vice versa.

Definition 9 (Effective Rebut) We say that attack relation att satisfies the effective rebut property for a knowledge base K iff for all arguments $A_0, A_1 \in AR_K$ such that each $A_i, i = 0, 1$, contains exactly one defeasible rule d_i (i.e. $dr(A_i) = \{d_i\}$), and A_0 rebuts A_1 , it holds that A_0 attacks A_1 wrt att iff $d_0 \not\prec d_1$.

4.2 Propagating Attacks

Example 3 Consider the knowledge base in example 1. While the effective rebut property determines that A_3 attacks A'_2 (see figure 1) but not vice versa (because $d_2 \prec d_3$), it does not say whether A_3 attack A_2 .

Looking at the structure of A_2, A'_2 , we can say that A_2 is a weakening of A'_2 as the undisputed fact P on which A'_2 is based is replaced by the defeasible belief P (supported by argument A_1). Therefore if A_3 attacks A'_2 then it is natural to expect that A_3 should attack A_2 too.

The above analysis also shows that attacks generated by the effective rebut property, could be propagated to other arguments based on a notion of weakening of arguments. We recall this notion as well as the associated property of attack monotonicity from [12] below.

Let $A, B \in AR_K$ and $AS \subseteq AR_K$. Intuitively, B is a weakening of A by AS if B is obtained by replacing zero, one or more premises of A by arguments in AS whose conclusions coincide with the premises.

Definition 10 *B is said to be a **weakening** of A by AS iff*

- (1.) $A = [\alpha]$ for $\alpha \in BE$, and $(B = [\alpha] \text{ or } B \in AS \text{ with } \text{cnl}(B) = \alpha)$, or
- (2.) $A = [A_1, \dots, A_n, r]$ and $B = [B_1, \dots, B_n, r]$ where each B_i is a weakening of A_i by AS .

By $A \downarrow AS$ we denote the set of all weakenings of A by AS .

For an illustration, consider again the arguments in figure 1. It is clear that $[P] \downarrow \{A_1\} = \{[P], A_1\}$, $A'_2 \downarrow \{A_1\} = \{A'_2, A_2\}$.

The attack monotonicity property states that if an argument A attacks an argument B then A also attacks all weakening of B . Moreover if a weakening of A attacks B then A also attacks B .

Definition 11 (Attack Monotonicity) *We say attack relation att satisfies the property of attack monotonicity for knowledge base K iff for all $A, B \in AR_K$ and for each weakening C of A , for each weakening D of B , the following assertions hold:*

1. If $(A, B) \in \text{att}$ then $(A, D) \in \text{att}$.
2. If $(C, B) \in \text{att}$ then $(A, B) \in \text{att}$.

We next recall the link-oriented property in [12] which is based on an intuition that attacks are directed towards links in arguments implying that if an argument A attacks an argument B then it should attack some part of B .

Definition 12 (Link-Orientation) *We say that attack relation att satisfies the property of link-orientation for K iff for all arguments $A, B, C \in AR_K$ such that C is a weakening of B by $AS \subseteq AR_K$ (i.e. $C \in B \downarrow AS$), it holds that if A attacks C (wrt att) and A does not attack AS (wrt att) then A attacks B (wrt att).*

In real world conversation, if you claim that my argument is wrong, I would naturally ask which part of my argument is wrong. The link-oriented property could be viewed as representing this intuition.

Example 4 *Consider again arguments in figure 1. Suppose d_2 is now preferred to d_3 (i.e. $d_3 \prec d_2$). The effective rebut property dictates that A_3 does not attack A'_2 . Does A_3 still attack A_2 ? Suppose A_3 attacks A_2 . Since A_3 does not attack A_1 that is a subargument of A_2 , we expect that A_3 should attack some other part of A_2 . In other words, we expect that A_3 attacks A'_2 . But this is a contradiction to the effective rebut property stating that A'_2 attack A_3 but not vice versa. Hence A'_3 does not attack A_2 .*

In other words, the link-orientation property has propagated the "non-attack relation" between A_3, A'_2 to a "non-attack relation" between A_3, A_2 .

We present below a new and novel concept of regular attack relations.

Definition 13 *An attack relation is said to be **regular** if it satisfies to the properties of inconsistency-resolving and strong subargument structure together with the properties of effective rebuts, attack monotonicity and link-orientation.*

4.3 Attack Relation Assignments: Propagating Attacks Across Knowledge Bases

While regular attack relations are natural and intuitive, they are still not sufficient for determining an intuitive semantics of prioritized rules. The example below illustrates this point.

Example 5 Consider a knowledge base K_0 obtained from knowledge base K in example 1 by revising the evidence base to $BE = \{D\}$. It is clear that arguments A_1, A_2, A_3 belong to AR_{K_0} while A'_2 is not an argument in AR_{K_0} .

As A'_2 does not belong to AR_{K_0} , the effective rebuts property does not "generate" any attacks between arguments in AR_{K_0} . How could we determine the attack relation for K_0 .

As both A_2, A_3 belong to AR_K, AR_{K_0} and the two knowledge bases K_0, K have identical rule-based system, we expect that the attack relations between their common arguments should be identical. In other words, because A_3 attacks A_2 wrt K (see example 3), A_3 should attack A_2 also wrt K_0 . This intuition is captured by the context-independence property in [12] linking attack relations between arguments across the boundary of knowledge bases.

The example also indicates that attack relations of knowledge bases with the same rule-based system should be considered together. This motivates the introduction of the attack relation assignment in definitions 14, 15.

Definition 14 Let $\mathcal{R} = (RS, RD, \preceq)$ be a rule-based system. The class consisting of all consistent knowledge bases of the form (\mathcal{R}, BE) is denoted by $\mathcal{C}_{\mathcal{R}}$.

A rule-based system \mathcal{R} is said to be **sensible** iff the set $\mathcal{C}_{\mathcal{R}}$ is not empty. From now on, whenever we mention a rule-based system, we mean a sensible one.

Definition 15 (Attack Relation Assignment) An attack relation assignment $atts$ for a rule-based system \mathcal{R} is a function assigning to each knowledge base $K \in \mathcal{C}_{\mathcal{R}}$ an attack relation $atts(K) \subseteq AR_K \times AR_K$.

We next recall the context-independence property stating that the attack relation between two arguments depends only on the rules appearing in them and their preferences.

Definition 16 (Context-Independence) We say attack relation assignment $atts$ for a rule-based system \mathcal{R} satisfies the property of context-independence iff for any two knowledge bases $K, K' \in \mathcal{C}_{\mathcal{R}}$ and for any two arguments A, B from $AR_K \cap AR_{K'}$, it holds that $(A, B) \in atts(K)$ iff $(A, B) \in atts(K')$

The context-independence property is commonly accepted in many well-known argument-based systems like the assumption-based framework [4, 18], the ASPIC+ approach [24, 22].

We can now present a central contribution of this paper, the introduction of the regular attack relation assignments.

Definition 17 (Regular Attack Relation Assignments) An attack relation assignment $atts$ for a rule-based system \mathcal{R} is said to be **regular** iff it satisfies the property of context-independence and for each knowledge base $K \in \mathcal{C}_{\mathcal{R}}$, $atts(K)$ is regular.

The set of all regular attack relation assignments for \mathcal{R} is denoted by $RAA_{\mathcal{R}}$.

For attack relation assignments $atts, atts'$, define $atts \subseteq atts'$ iff $\forall K \in \mathcal{C}_{\mathcal{R}}, atts(K) \subseteq atts'(K)$.

Minimal Removal Intuition

A key purpose of introducing priorities between defeasible rules is to remove certain undesired attacks while keeping the set of removed attacks to a minimum. The following very simple example illustrates the idea.

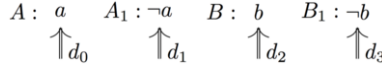


Figure 4. Minimal Removal

Example 6 Consider a knowledge base consisting of just four defeasible rules and four arguments A, A_1, B, B_1 as seen in figure 4. Without any preference between the rules, we have A, A_1 attack each other. Similarly B, B_1 attack each other.

Suppose that for whatever reason d_3 is strictly less preferred than d_2 (i.e. $d_3 \prec d_2$). The introduction of the preference $d_3 \prec d_2$ in essence means that the attack of B_1 against B should be removed, but it does not say anything about the other attacks. Hence they should be kept, i.e. the attacks that should be removed should be kept to a minimum.

Let \mathcal{R} be a rule-based system and $K \in \mathcal{C}_{\mathcal{R}}$. The *basic attack relation assignment* for \mathcal{R} , denoted by $Batts$ is defined by: $\forall K \in \mathcal{C}_{\mathcal{R}}, Batts(K) = \{(A, B) \mid A \text{ undercuts or rebuts } B\}$. Further let $atts$ be a regular attack relation assignment. From the strong subargument structure property, it is clear that $atts \subseteq Batts$. $\forall K \in \mathcal{C}_{\mathcal{R}}$, the set $Batts(K) \setminus atts(K)$ could be viewed as the set of attacks removed from $Batts(K)$ due to the priorities between defeasible rules.

Combining the "minimal-removal intuition" with the concept of regular attack relation assignment suggests that the semantics of \mathcal{R} should be captured by regular attack relations $atts$ such that $\forall K \in \mathcal{C}_{\mathcal{R}}$, the set $Batts(K) \setminus atts(K)$ is minimal, or equivalently the set $atts(K)$ is maximal. As we will see in the next section, such maximal attack relation assignment indeed exists.

5. The Upper Semilattice of Regular Attack Relation Assignments

From now on until the end of this section, we assume an arbitrary but fixed rule-based system $\mathcal{R} = (RS, RD, \preceq)$.

Let \mathcal{A} be a non-empty set of attack relation assignments for $RAA_{\mathcal{R}}$. Define $\sqcup \mathcal{A}$ by: $\forall K \in \mathcal{C}_{\mathcal{R}}: (\sqcup \mathcal{A})(K) = \bigcup \{atts(K) \mid atts \in \mathcal{A}\}$

The following simple lemma and theorem present a deep insight into the structure of regular attack assignments.

Lemma 3 *If the attack relations assignments in \mathcal{A} are regular then $\sqcup \mathcal{A}$ is also regular.*

Proof (Sketch) The proof is not difficult though rather lengthy as we just need to check in a straightforward way that each regular property is satisfied. \square

It follows immediately

Theorem 2 *Suppose the set $RAT_{\mathcal{R}}$ of regular attack relation assignments is not empty. Then $(RAA_{\mathcal{R}}, \subseteq, \sqcup)$ is an upper semilattice. \square*

Definition 18 Suppose the set $RAA_{\mathcal{R}}$ of all regular attack relation assignments for \mathcal{R} is not empty. The **canonical attack relation assignment** of \mathcal{R} denoted by $\mathbf{Att}_{\mathcal{R}}$ is defined by: $\mathbf{Att}_{\mathcal{R}} = \sqcup RAA_{\mathcal{R}}$.

Even though in general, regular attack relation assignments (and hence the canonical one) may not exist (as the example 7 below shows), they exist under natural conditions that we believe most practical rule-based systems satisfy, like the property of self-contradiction or closure under transposition or contraposition (see theorem 3 below).

Example 7 Consider a rule-based system \mathcal{R} consisting of $d_0 : \Rightarrow a$ $d_1 : \Rightarrow b$ $r : a \rightarrow \neg b$ and $d_0 \prec d_1$. Suppose $atts$ be a regular attack relation assignment for $\mathcal{C}_{\mathcal{R}}$. Let $K = (\mathcal{R}, \emptyset)$. The arguments for K are given in figure 5. From the property of effective rebut, it is clear that $(A, B) \notin att(K)$. Hence $att(K) = \emptyset$. The inconsistency-resolving property is not satisfied by att , contradicting the assumption that $atts$ is regular. Therefore there exists no regular attack relation assignment for \mathcal{C}_K .

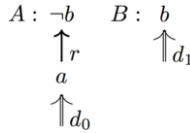


Figure 5. Non-existence of regular assignments

It turns out that a special type of attack relations, the normal attack relations introduced in [12] is regular if the rule-based systems is closed under transposition or contraposition or self-contradiction.

Let K be a knowledge base and $A, B \in AR_K$. We say that A *normal-rebuts* B (at X) iff A rebuts B (at X) and there is no defeasible rule $d \in ldr(A)$ such that $d \prec last(X)$.

The *normal attack relation assignment* [12] $atts_{nr}$ is defined by: For any knowledge base $K \in \mathcal{R}$ and any arguments $A, B \in AR_K$, $(A, B) \in atts_{nr}(K)$ if and only if A undercuts B or A normal-rebuts B .

We present below a central result of this paper.

Theorem 3 Suppose the rule-based system \mathcal{R} satisfies the self-contradiction property. Then the normal attack relation assignment $atts_{nr}$ is regular and the canonical assignment $\mathbf{Att}_{\mathcal{R}}$ exists and $atts_{nr} \subseteq \mathbf{Att}_{\mathcal{R}}$.

Proof (Sketch) From theorem 2 and the definition of the canonical attack relation, we only need to show that $atts_{nr}$ is regular.

It is straightforward to show that for each $K \in \mathcal{C}_{\mathcal{R}}$, the attack relation $atts_{nr}(K)$ satisfies the properties of strong subargument structure, attack monotonicity, effective rebuts and link-orientation. Further it is also obvious that $atts_{nr}$ satisfies the context-independence property. Let $K \in \mathcal{C}_{\mathcal{R}}$. We show that $atts_{nr}(K)$ satisfies the inconsistency-resolving property. Let $S \subseteq AR_K$ s.t. S is inconsistent. Let S' be the set of all basic defeasible subarguments of S and S_0 be a minimal inconsistent subset of S' . Let $A \in S_0$ s.t. $last(A)$ is minimal (wrt \prec) in $\{last(X) \mid X \in S_0\}$. From the self-contradiction property, $cnl(S_0) \vdash \neg hd(last(A))$.

We could then construct an argument B such that B attacks A and all basic defeasible subarguments of B are subarguments of arguments in S_0 . \square

Though the normal and canonical attack relations do not coincide in general, they are equivalent in the sense that they have identical sets of stable extensions.

Theorem 4 *Suppose the rule-based system \mathcal{R} satisfies the property of self-contradiction. Then for each $K \in \mathcal{C}_{\mathcal{R}}$, $E \subseteq AR_K$ is a stable extension wrt $atts_{nr}(K)$ iff E is a stable extension wrt $Att_{\mathcal{R}}(K)$.*

Proof (Sketch) We first show that for each $atts \in RAA_{\mathcal{R}}$, each stable extension of $(AR_K, atts(K))$ is also a stable extension of $(AR_K, atts_{nr}(K))$. Hence each stable extension of $(AR_K, Att_{\mathcal{R}}(K))$ is also stable extension of $(AR_K, atts_{nr}(K))$. The theorem follows then from lemma 4 below. \square

Lemma 4 *Let $atts, atts'$ be regular attack relation assignments for \mathcal{R} such that $atts \subseteq atts'$. Then (1.) each stable extension of $(AR_K, atts(K))$ is a stable extension of $(AR_K, atts'(K))$; and (2.) each stable extension of $(AR_K, atts(K))$ is a stable extension of $(AR_K, Att_{\mathcal{R}}(K))$.*

Proof (Sketch) 1) Let E be a stable extension of $(AR_K, atts(K))$. It is clear that E attacks each argument in $AR_K \setminus E$ wrt $atts'(K)$. If E is not conflict-free wrt $atts'(K)$, E is inconsistent (since both $atts, atts'$ have the same set of undercuts) and hence not conflict-free wrt $atts(K)$ (a contradiction). Hence E is conflict-free (and hence stable) wrt $atts'(K)$. 2) Follows immediately from (1) and the definition of $Att_{\mathcal{R}}$. \square

6. Discussion and Conclusion

The preference-based approaches to argumentation [2,3,24,22,23] define the semantics of defeasible knowledge bases by first defining a preference relation between arguments and then using the preference relation to define attack relation between arguments. We could also define an argument preference assignment for a rule-based system \mathcal{R} as a function assigning to each knowledge base $K \in \mathcal{C}_{\mathcal{R}}$, a relation $\sqsubseteq_K \subseteq AR_K \times AR_K$ representing a preference relation between arguments in AR_K where strict arguments are not strictly less preferred than any other arguments. It is possible to define a lower semilattice over the set of preference relation assignments whose least element corresponds to the canonical semantics (see [13]).

A key property satisfied by many argument-based and non-argument-based approaches to reasoning with prioritized rules is the credulous cumulativity property [12] stating intuitively that if some beliefs in your belief set are confirmed in the reality then your belief set will not change because of it. We show in [13] that credulous cumulativity is satisfied by regular attack relation assignments.

A more liberal notion of rebut, referred to as unrestricted rebut, where a basic defeasible argument could directly attack a non-basic defeasible argument is studied in [9,8]. Intuitively an unrestricted rebut is a rebut against a set of defeasible rules without explicitly rebutting any individual rule in it. It would be interesting to see how this notion of rebut interacts with the regular properties.

References

- [1] L. Amgoud. Postulates for logic-based argumentation systems. *Int J. Approximate Reasoning*, 55(9):2028–2048, 2014.
- [2] L. Amgoud and C. Cayrol. Inferring from inconsistency in preference-based argumentation framework. *Int J. Automated Reasoning*, 29(2):197–215, 2002.
- [3] Trevor J. M. Bench-Capon. Persuasion in practical argument using value-based argumentation frameworks. *J. Log. Comput.*, 13(3):429–448, 2003.
- [4] A. Bondarenko, P.M. Dung, R.A. Kowalski, and F. Toni. An abstract, argumentation-theoretic approach to default reasoning. *Artif. Intell.*, 93:63–101, 1997.
- [5] G. Brewka. Preferred subtheories: An extended logical framework for default reasoning. In *Proc of IJCAI'89*, pages 1043–1048. Morgan Kaufmann, 1989.
- [6] G. Brewka and T. Eiter. Preferred answer sets for extended logic programs. *Artificial Intelligence*, 109:297–356, 1999.
- [7] G. Brewka, I. Niemelä, and M. Truszczynski. Preferences and nonmonotonic reasoning. *AI Magazine*, 29(4):69–78, 2008.
- [8] M. Caminada and L. Amgoud. On the evaluation of argumentation formalisms. *Artificial Intelligence*, 171:286–310, 2007.
- [9] M. Caminada, S. Modgil, and N. Oren. Preferences and unrestricted rebut. In *Proc Comma 2014*, 2014.
- [10] B. A. Davey and H. A. Priestley. *Introduction to Lattices and Order*. Cambridge University Press, 2002.
- [11] J.P. Delgrande, T. Schaub, and H. Tompits. A framework for compiling preferences in logic programs. *Theory and Practice of Logic Programming*, pages 129–187, 2003.
- [12] P. M. Dung. An axiomatic analysis of structured argumentation with priorities. *Artificial Intelligence*, 231, 2016.
- [13] P. M. Dung. Invited lecture, argumentation for practical reasoning: An axiomatic approach. In M. Baldoni, A. K. Chopra, and T. C. Son M. Maes, editors, *PRIMA 2016*, 2016.
- [14] P. M. Dung and G. Sartor. The modular logic of private international law. *Artif. Intell. Law*, pages 233–261, 2011.
- [15] P. M. Dung and P. M. Thang. Closure and consistency and logic-associated argumentation. *J. Artificial Intelligence Research*, 49:79–109, 2014.
- [16] P.M. Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games; acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artif. Intell.*, 77(2):321–358, 1995.
- [17] P.M. Dung. An axiomatic analysis of structured argumentation for prioritized default reasoning. In *Proc of ECAI 2014*, 2014.
- [18] P.M. Dung, R.A. Kowalski, and F. Toni. *Argumentation in AI*, chapter Assumption-based Argumentation. Springer-Verlag, 2009.
- [19] A.J. Garcia and G.R. Simari. Defeasible logic programming: An argumentative approach. *TPLP*, 4(1-2):95–138, 2004.
- [20] M. Gelfond and T. C. Son. Reasoning with prioritized defaults. In *LPKR*, pages 164–223, 1997.
- [21] D. C. Martinez, A. J. Garcia, and G. R. Simari. On acceptability in abstract argumentation frameworks with an extended defeat relation. In T. J. M. Bench-Capon P.E. Dunne, editor, *In proc of Int conference on "Computational models of arguments"*. IOS Press, 2006.
- [22] S. Modgil and H. Prakken. A general account of argumentation with preferences. *Artificial Intelligence*, 197:361–397, 2013.
- [23] S. Modgil and H. Prakken. The aspic+ framework for structured argumentation: a tutorial. *J. Arguments and Computation*, 5:31–62, 2014.
- [24] H. Prakken. An abstract framework for argumentation with structured arguments. *J. Arguments and Computation*, 1, 2010.
- [25] G. Vreeswijk. Abstract argumentation systems. *Artif. Intell.*, 90:225–279, 1997.

Forbidden Sets in Argumentation Semantics

Paul E. DUNNE

Dept. of Computer Science, University of Liverpool, Liverpool, L69 7ZF, UK

Abstract. We consider an alternative interpretation of classical Dung argumentation framework (AF) semantics by introducing the concept of “*forbidden sets*”. In informal terms, such sets are well-defined with respect to any extension-based semantics and reflect those subsets of argument that collectively can never form part of an acceptable solution. The forbidden set paradigm thus provides a parametric treatment of extension-based semantics. We present some general properties of forbidden set structures and describe the interaction between forbidden sets for a number of classical semantics. Finally we establish some initial complexity results in the arena of forbidden set decision problems.

Keywords. abstract argumentation frameworks; extension-based semantics; computational complexity

Introduction

Among the many developments arising from the seminal treatment of argumentation within the abstract graph-theoretic model of argumentation frameworks (AFs) from Dung [7], one of the most prolific areas has been the formulation of alternative “argumentation semantics”: that is the conditions on subsets of a framework’s atomic arguments characterising which such sets present collectively “justified” arguments from which fail to do so. In addition to those presented in [7] one finds ideas such as semi-stable in Caminada [5], ideal from Dung *et al.* [8], together with CF2 semantics arising in Baroni *et al.* [3], the parametric concept of resolution-based semantics described by Baroni and Giacomin [2] together with the analysis of one specific instantiation of this by Baroni *et al.* [1].

Our aim in the present article is *not* to offer yet another semantics of abstract argumentation derived from graph-theoretic considerations within the supporting AF, but rather to examine an alternative view of such that have already been posited and, indeed, may be offered subsequently.

The central conceit underpinning our treatment stems from the property that all such extension-based semantics (as these have come to be generally known) conceptually prescribe solutions via a “*positive*” *enumeration* of “acceptable” subsets of arguments within a framework, e.g. the so-called “conflict-free” solutions are those subsets, S , in which no attack is present between any members of S . Thus, in order to validate a set as acceptable it suffices to find it among the list of allowed solution sets.

Here we examine an alternative view: examining conditions on sets, S , which suffice to eliminate any possibility that S is an acceptable position. These conditions, in a similar style to classical extension bases, may be thought of as described through an enumeration of sets, which we will call the *forbidden sets* (with respect to a given AF and argumentation semantics). In this way if S is a forbidden set with respect to an AF, \mathcal{H} and semantics σ this indicates that *no* σ -extension of \mathcal{H} contains S as a subset.

We provide some basic background in Section 1, proceeding to define formally the concept of forbidden set in Section 2 and prove some generic properties of these. In Section 3 we then consider comparative aspects of forbidden sets defined for some standard semantics and review some questions concerning computational complexity matters within Section 4. Conclusions and open issues are presented in Section 5.

1. Preliminaries

We begin by recalling the concept of abstract argumentation framework and terminology from Dung [7] and outline the main computational problems that have been of interest within this.

Definition 1 We use \mathcal{X} to denote a finite set of arguments with $\mathcal{A} \subseteq \mathcal{X} \times \mathcal{X}$ the so-called attack relationship over these. An argumentation framework (AF) is a pair $\mathcal{H} = \langle \mathcal{X}, \mathcal{A} \rangle$. A pair $\langle x, y \rangle \in \mathcal{A}$ is referred to as ‘ y is attacked by x ’ or ‘ x attacks y ’. Using S to denote an arbitrary subset of arguments for $S \subseteq \mathcal{X}$,

$$\begin{aligned} S^- &=_{\text{def}} \{ p : \exists q \in S \text{ such that } \langle p, q \rangle \in \mathcal{A} \} \\ S^+ &=_{\text{def}} \{ p : \exists q \in S \text{ such that } \langle q, p \rangle \in \mathcal{A} \} \end{aligned}$$

We say that: $x \in \mathcal{X}$ is acceptable with respect to S if for every $y \in \mathcal{X}$ that attacks x there is some $z \in S$ that attacks y . Given $S \subseteq \mathcal{X}$, $\mathcal{F}(S) \subseteq \mathcal{X}$ is the set of all arguments that are acceptable with respect to S , i.e.

$$\mathcal{F}(S) = \{x \in \mathcal{X} : \forall y \text{ such that } \langle y, x \rangle \in \mathcal{A}, \exists z \in S \text{ s.t. } \langle z, y \rangle \in \mathcal{A}\}$$

A subset, S , is conflict-free if no argument in S is attacked by any other argument in S . with \subseteq -maximal conflict-free set referred to as naive extensions. A conflict-free set S is admissible if every $y \in S$ is acceptable w.r.t S . S is a complete extension if S is conflict-free and should $x \in \mathcal{F}(S)$ then $x \in S$, i.e. every argument that is acceptable to S is a member of S , so that $\mathcal{F}(S) = S$. The set of \subseteq -maximal complete extensions coincide with the set of \subseteq -maximal admissible sets these being termed preferred extensions. The set S is a stable extension if S is conflict free and $S^+ = \mathcal{X} \setminus S$. and is a semi-stable extension (Caminada [5]) if admissible and has $S \cup S^+ \subseteq$ -maximal among all admissible sets.

The grounded extension of $\langle \mathcal{X}, \mathcal{A} \rangle$ is defined as the \subseteq -minimal complete extension.

We use σ to denote an arbitrary semantics from

$$\{\text{CF}, \text{NVE}, \text{ADM}, \text{PR}, \text{ST}, \text{COM}, \text{SST}, \text{GR}\}$$

corresponding to conflict-free, naive, admissible, preferred, stable, complete, semi-stable and grounded instances.

For a given semantics σ and AF, $\mathcal{H}(\mathcal{X}, \mathcal{A})$ we use $\mathcal{E}_\sigma(\mathcal{H})$ to denote the set of all subsets of \mathcal{X} that satisfy the conditions specified by σ . We say that σ is a *unique status* semantics if $|\mathcal{E}_\sigma(\mathcal{H})| = 1$ for every AF, \mathcal{H} , denoting the unique extension by $E_\sigma(\mathcal{H})$.

We complete this, brief, overview by describing the three canonical decision problems that may be instantiated for a given semantics: *Verification* (VER), *Credulous Acceptance* (CA) and *Sceptical Acceptance* (SA). Formal definitions of these problems for AFS are presented in Table 1.

Table 1. Decision Problems in AFS

Problem Name	Instance	Question
<i>Verification</i> (VER_σ)	$\mathcal{H}(\mathcal{X}, \mathcal{A}); S \subseteq \mathcal{X}$	Is $S \in \mathcal{E}_\sigma(\mathcal{H})$?
<i>Credulous Acceptance</i> (CA_σ)	$\mathcal{H}(\mathcal{X}, \mathcal{A}); x \in \mathcal{X}$	$\exists S \in \mathcal{E}_\sigma(\mathcal{H})$ for which $x \in S$?
<i>Sceptical Acceptance</i> (SA_σ)	$\mathcal{H}(\mathcal{X}, \mathcal{A}); x \in \mathcal{X}$	$\forall T \in \mathcal{E}_\sigma(\mathcal{H})$ is $x \in T$?

2. Forbidden Sets and Related Structures

In this paper we introduce and explore the properties of a “parametric” operator – the *forbidden set* constructor – and its relationship with the extension-based semantics outlined in the preceding section.

Definition 2 Let $\mathbb{S} \subseteq 2^\mathcal{X}$. A set $T \subseteq \mathcal{X}$ is said to be a forbidden set for \mathbb{S} if for every set $S \in \mathbb{S}$, it is not the case that $T \subseteq S$.

A set, $T \subseteq \mathcal{X}$ is a minimal forbidden set for \mathbb{S} if it is both a forbidden set for \mathbb{S} but no strict subset of T describes a forbidden set for \mathbb{S} . Given \mathbb{S} , the notation $\kappa(\mathbb{S})$ and $\mu(\mathbb{S})$ describe those subsets of \mathcal{X} for which

$$\kappa(\mathbb{S}) = \{ T \subseteq \mathcal{X} : T \text{ is a forbidden set for } \mathbb{S} \}$$

$$\mu(\mathbb{S}) = \{ T \subseteq \mathcal{X} : T \text{ is a minimal forbidden set for } \mathbb{S} \} \subseteq \kappa(\mathbb{S})$$

For k , with $0 \leq k \leq |\mathcal{X}|$, the k -section of \mathbb{S} , denoted $\chi^{(k)}(\mathbb{S})$, is

$$\chi^{(k)}(\mathbb{S}) = \{ P \subseteq \mathcal{X} : |P| = k \text{ and } P \in \kappa(\mathbb{S}) \}$$

In the special case $k = 1$, $\chi^{(1)}$ are those members of \mathcal{X} that do not occur in any set of \mathbb{S} ; while the subsets $\chi^{(2)}$ play an important role in the characterization considered in Dunne *et al.* [10] where these are referred to as “unpaired elements”.

We note that $\kappa(\mathbb{S})$ and, potentially, $\mu(\mathbb{S})$, contains sets which are strict *supersets* of elements in \mathbb{S} . A simple example of such behaviour is given with $\mathcal{X} = \{x_1, x_2\}$ and $\mathbb{S} = \{\{x_1\}, \{x_2\}\}$: in this case $\mu(\mathbb{S}) = \kappa(\mathbb{S}) = \{\{x_1, x_2\}\}$.

Some properties of these operations are exploited in later results such as Lemma 2.

Lemma 1 *Given $\mathbb{S} \subseteq 2^{\mathcal{X}}$ and $\kappa(\mathbb{S})$ as defined in Defn. 2, the set systems $\kappa(\mathbb{S})$, $\mu(\mathbb{S})$ and $\chi^{(k)}(\mathbb{S})$, satisfy*

- a. *If $Q \in \kappa(\mathbb{S})$ there is (at least one) $R \subseteq Q$ with $R \in \mu(\mathbb{S})$.*
- b. *The conditions $\emptyset \in \kappa(\mathbb{S})$, $\mu(\mathbb{S}) = \{\emptyset\}$ and $\mathbb{S} = \emptyset$ are equivalent.*
- c. *$\kappa(\mathbb{S}) = \emptyset$ if and only if $\{x_1, \dots, x_n\} \in \mathbb{S}$, that is to say \mathbb{S} contains the set which comprises all of the arguments in \mathcal{X} .*

Proof: Recall that we assume $\langle \mathcal{X}, \mathcal{A} \rangle$ is a *finite* structure.

For (a) suppose that $Q \in \kappa(\mathbb{S})$. If it is the case that no strict subset of Q is a forbidden set for \mathbb{S} then, by definition, we have $Q \in \mu(\mathbb{S})$. Otherwise we find some $T \subset Q$ for which $T \in \kappa(\mathbb{S})$. Repeating the argument either T is a minimal forbidden set for \mathbb{S} or has some subset which is a forbidden set. Eventually we find some $R \subseteq Q$ which is both forbidden and minimally so.

For (b), that $\mu(\mathbb{S}) = \{\emptyset\}$ if and only if $\emptyset \in \kappa(\mathbb{S})$ follows directly from the definition of $\mu(\mathbb{S})$. To see that $\mu(\mathbb{S}) = \{\emptyset\}$ is only possible when $\mathbb{S} = \emptyset$, again from the definition of forbidden set, were \emptyset to be a forbidden set for \mathbb{S} this indicates that no $S \in \mathbb{S}$ has $\emptyset \subseteq S$. This property can only be satisfied in the degenerate case $\mathbb{S} = \emptyset$.

With (c), $\kappa(\mathbb{S}) = \emptyset$ expresses the property that \mathbb{S} has no forbidden sets at all, so, in particular, the set containing all arguments of \mathcal{X} must belong to \mathbb{S} . Conversely, should $\{x_1, \dots, x_n\} \in \mathbb{S}$ this suffices to rule out any subset of \mathcal{X} as forbidden, i.e. $\kappa(\mathbb{S}) = \emptyset$. \square

3. Comparative Properties of Forbidden Sets in Divers Semantics

Let $\mathcal{H} = \langle \mathcal{X}, \mathcal{A} \rangle$ be an AF. A natural question arising with respect to the forbidden set paradigm, concerns what may be said in general regarding comparisons between distinct extension sets of \mathcal{H} and their associated forbidden sets.

In order to avoid an excess of parentheses, we adopt the following notation when considering a given AF $\mathcal{H} = \langle \mathcal{X}, \mathcal{A} \rangle$.

$$\begin{aligned} \mathcal{E}_\sigma^\mathcal{H} &=_{\text{def}} \mathcal{E}_\sigma(\mathcal{H}) \\ \kappa_\sigma^\mathcal{H} &=_{\text{def}} \kappa(\mathcal{E}_\sigma(\mathcal{H})) \\ \mu_\sigma^\mathcal{H} &=_{\text{def}} \mu(\mathcal{E}_\sigma(\mathcal{H})) \end{aligned}$$

The following results present some basic relationships between forbidden sets and underlying semantics. Noting that Case (b) of Lemma 2 indicates semantics, σ defined as \subseteq -maximal elements of semantics τ have identical forbidden sets determining membership of $S \in \mathcal{E}_\sigma$ cannot be established simply by arguing S has no $R \in \mu(\mathcal{E}_\sigma)$ as a subset. The relationship given in part (c), however, does provide a method by which $S \in \mathcal{E}_\sigma$ can be decided via forbidden set structures.

Lemma 2

- a. If σ, τ are semantics that satisfy, $\mathcal{E}_\sigma^\mathcal{H} \subseteq \mathcal{E}_\tau^\mathcal{H}$ then $\kappa_\tau^\mathcal{H} \subseteq \kappa_\sigma^\mathcal{H}$.
- b. If $\mathcal{E}_\sigma^\mathcal{H}$ is defined to be the (\subseteq) -maximal sets within $\mathcal{E}_\tau^\mathcal{H}$ then $\kappa_\tau^\mathcal{H} = \kappa_\sigma^\mathcal{H}$.
- c. If σ, τ satisfy the condition given in (b) then, for all $S \subseteq \mathcal{X}$ $S \in \mathcal{E}_\sigma^\mathcal{H}$ if and only if

$$(\exists Q \in \mu(\mathcal{E}_\tau^\mathcal{H} \setminus \mathcal{E}_\sigma^\mathcal{H}) : Q \subseteq S) \text{ and } (\forall Q \in \mu_\sigma^\mathcal{H} \neg(Q \subseteq S))$$

Proof: For (a), when σ and τ satisfy $\mathcal{E}_\sigma^\mathcal{H} \subseteq \mathcal{E}_\tau^\mathcal{H}$ no set in $\kappa_\tau^\mathcal{H}$ can be a subset of any set in $\mathcal{E}_\tau^\mathcal{H}$. In particular if $S \subseteq \mathcal{E}_\tau^\mathcal{H}$ then a forbidden set for $\mathcal{E}_\tau^\mathcal{H}$ is perforce also a forbidden set for \mathcal{S} . It follows that any forbidden set for $\mathcal{E}_\tau^\mathcal{H}$ is a forbidden set for $\mathcal{E}_\sigma^\mathcal{H}$, i.e. $\kappa_\tau^\mathcal{H} \subseteq \kappa_\sigma^\mathcal{H}$.

For (b), the maximality premise already ensures $\kappa_\tau^\mathcal{H} \subseteq \kappa_\sigma^\mathcal{H}$ via part (a). Consider any $S \in \kappa_\sigma^\mathcal{H}$ and suppose, for the sake of contradiction, that $S \notin \kappa_\tau^\mathcal{H}$. From the definition of forbidden set this means we can find $T \in \mathcal{E}_\tau^\mathcal{H}$ with $S \subseteq T$. Now, however, we find $R \in \mathcal{E}_\sigma^\mathcal{H}$ with $T \subseteq R$ so that $S \subseteq T \subseteq R \in \mathcal{E}_\sigma^\mathcal{H}$ contradicting $S \in \kappa_\sigma^\mathcal{H}$.

For the relationship in (c), should it be the case that $S \in \mathcal{E}_\sigma^\mathcal{H}$ then $S \notin \mathcal{E}_\tau^\mathcal{H} \setminus \mathcal{E}_\sigma^\mathcal{H}$ so that $S \in \kappa(\mathcal{E}_\tau^\mathcal{H} \setminus \mathcal{E}_\sigma^\mathcal{H})$ and the property of there being some $Q \in \mu(\mathcal{E}_\tau^\mathcal{H} \setminus \mathcal{E}_\sigma^\mathcal{H})$ with $Q \subseteq S$ follows from Lemma 1(a). Similarly the premise $S \in \mathcal{E}_\sigma^\mathcal{H}$ indicates $S \notin \kappa_\sigma^\mathcal{H}$ thus no $Q \in \mu_\sigma^\mathcal{H}$ satisfies $Q \subseteq S$.

Conversely suppose that some $Q \in \mu(\mathcal{E}_\tau^\mathcal{H} \setminus \mathcal{E}_\sigma^\mathcal{H})$ satisfies $Q \subseteq S$ but that no $Q \in \mu_\sigma^\mathcal{H}$ has this property. Then,

$$\begin{aligned} Q \in \mu(\mathcal{E}_\tau^\mathcal{H} \setminus \mathcal{E}_\sigma^\mathcal{H}) \text{ and } Q \subseteq S &\Rightarrow S \notin \mathcal{E}_\tau^\mathcal{H} \setminus \mathcal{E}_\sigma^\mathcal{H} \\ &\Rightarrow S \notin \mathcal{E}_\tau^\mathcal{H} \text{ or } S \in \mathcal{E}_\sigma^\mathcal{H} \end{aligned}$$

In addition,

$$\forall Q \in \mu_\sigma^\mathcal{H} \neg(Q \subseteq S) \Rightarrow S \in \mathcal{E}_\tau^\mathcal{H}$$

Notice that as a consequence of (b) we have $\mu_\sigma^\mathcal{H} = \mu_\tau^\mathcal{H}$ so we cannot directly deduce from $\neg(Q \subseteq S)$ for each $Q \in \mu_\sigma^\mathcal{H}$ that $S \in \mathcal{E}_\sigma^\mathcal{H}$: only $S \in \mathcal{E}_\tau^\mathcal{H}$. Combining $S \notin \mathcal{E}_\tau^\mathcal{H} \setminus \mathcal{E}_\sigma^\mathcal{H}$ and $S \in \mathcal{E}_\tau^\mathcal{H}$ we deduce that $S \in \mathcal{E}_\sigma^\mathcal{H}$ as claimed. \square

Corollary 1 For all $\mathcal{H} = \langle \mathcal{X}, \mathcal{A} \rangle$

- a. $\kappa_{\text{ADM}}^\mathcal{H} = \kappa_{\text{PR}}^\mathcal{H} = \kappa_{\text{COM}}^\mathcal{H}$.
- b. $\kappa_{\text{CF}}^\mathcal{H} = \kappa_{\text{NVE}}^\mathcal{H}$.
- c. $\kappa_{\text{PR}}^\mathcal{H} \subseteq \kappa_{\text{SST}}^\mathcal{H} \subseteq \kappa_{\text{ST}}^\mathcal{H}$.
- d. $\kappa_{\text{CF}}^\mathcal{H} \subseteq \kappa_{\text{ADM}}^\mathcal{H}$.

Proof: Immediate consequence of Lemma 2 and established containment properties of the featured semantics. \square

It is worth noting at this point a distinguishing aspect of the forbidden set paradigm in comparison with the extension-based semantics. It is well known in

the latter formalism, that $\mathcal{E}_{co}^{\mathcal{H}} \subseteq \mathcal{E}_{adm}^{\mathcal{H}}$, i.e. every complete extension is an admissible set. The converse, however, does not hold: one may construct frameworks having $S \in \mathcal{E}_{adm}^{\mathcal{H}}$ but $S \notin \mathcal{E}_{co}^{\mathcal{H}}$.¹ The forbidden set structures for both semantics, however, are identical in consequence of $\mathcal{E}_{pr}^{\mathcal{H}}$ being formed by \subseteq -maximal admissible sets and \subseteq -maximal complete sets.

As a second point Corollary 1(a) offers an interesting point of comparison with recent work of Baumann *et al.* [4]. In this regard if we wish to distinguish $S \in \mathcal{E}_{ADM}^{\mathcal{H}}$ from $S \in \mathcal{E}_{PR}^{\mathcal{H}}$ in order to do so via the forbidden set paradigm the additional information required in terms of Lemma 2 (c) can be used.

We next establish that the relationships from Corollary 1(c)-(d) are exact, i.e we construct instances for which $\mu_{\sigma}^{\mathcal{H}} \not\subseteq \mu_{\tau}^{\mathcal{H}}$ although $\mu_{\sigma}^{\mathcal{H}} \subseteq \kappa_{\tau}^{\mathcal{H}}$, indicating the *minimal* forbidden sets are distinct.

Lemma 3 *There are choices of \mathcal{H} with which,*

- a. $\mu_{CF}^{\mathcal{H}} \not\subseteq \mu_{ADM}^{\mathcal{H}}$.
- b. $\mu_{PR}^{\mathcal{H}} \not\subseteq \mu_{ST}^{\mathcal{H}}$.
- c. $\mu_{PR}^{\mathcal{H}} \not\subseteq \mu_{ST}^{\mathcal{H}}$ and $\mathcal{E}_{st}^{\mathcal{H}} \neq \emptyset$.
- d. $\mu_{PR}^{\mathcal{H}} \not\subseteq \mu_{SST}^{\mathcal{H}}$.
- e. $\mu_{SST}^{\mathcal{H}} \not\subseteq \mu_{ST}^{\mathcal{H}}$.

Proof: Consider the three AFS shown in Fig. 1.

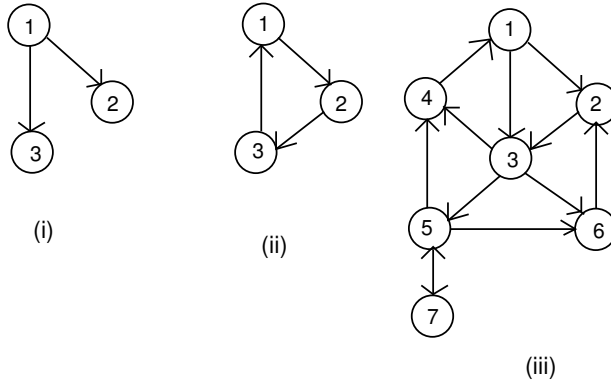


Figure 1. Non-containment properties in minimal forbidden set semantics

The AF shown in Fig. 1(i) has

$$\begin{aligned}\mu_{CF}^{\mathcal{H}} &= \{\{1, 2\}, \{1, 3\}\} \\ \mu_{ADM}^{\mathcal{H}} &= \{\{2\}, \{3\}\}\end{aligned}$$

¹Any AF for which $E_{gr}(\mathcal{H}) \neq \emptyset$ provides such an example: the empty set is always an admissible set but (in these cases) will fail to be a complete extension.

This suffices to establish (a). The AF depicted in Fig. 1(ii) has $\mathcal{E}_{st}^{\mathcal{H}} = \emptyset$ and $\mathcal{E}_{pr}^{\mathcal{H}} = \mathcal{E}_{sst}^{\mathcal{H}} = \{\emptyset\}$. In consequence,

$$\mu_{ST}^{\mathcal{H}} = \{\emptyset\}$$

while

$$\mu_{PR}^{\mathcal{H}} = \mu_{SST}^{\mathcal{H}} = \{\{1\}, \{2\}, \{3\}\}$$

The relationships claimed in (b) and (e) are now immediate.

Finally in the AF shown under Fig. 1(iii) we have,

$$\begin{aligned} \mathcal{E}_{pr}^{\mathcal{H}} &= \{\{1, 5\}, \{7\}\} \\ \mathcal{E}_{st}^{\mathcal{H}} &= \{\{1, 5\}\} \\ \mathcal{E}_{sst}^{\mathcal{H}} &= \{\{1, 5\}\} \\ \mu_{PR}^{\mathcal{H}} &= \{\{2\}, \{3\}, \{4\}, \{6\}, \{1, 5, 7\}\} \\ \mu_{ST}^{\mathcal{H}} &= \{\{2\}, \{3\}, \{4\}, \{6\}, \{7\}\} \\ \mu_{SST}^{\mathcal{H}} &= \{\{2\}, \{3\}, \{4\}, \{6\}, \{7\}\} \end{aligned}$$

From which (c) and (d) are easily deduced. □

Finally we have a select number of instances where the structure of forbidden sets is characterized exactly.

Lemma 4

a. The set $\mu_{CF}^{\mathcal{H}}$ is formed by the \subseteq -minimal sets in

$$\{\{x, y\} : \langle x, y \rangle \in \mathcal{A} \text{ or } \langle y, x \rangle \in \mathcal{A}\} \cup \{\{x\} : \langle x, x \rangle \in \mathcal{A}\}$$

b. A set $S \subseteq \mathcal{X}$ is defenceless in \mathcal{H} if and only if every superset T of S satisfies

$$T \in \mathcal{E}_{cf}^{\mathcal{H}} \Rightarrow \exists r \in T^- : r \notin T^+$$

The set $\mu_{ADM}^{\mathcal{H}}$ is formed by the \subseteq -minimal defenceless sets of \mathcal{H} .

c. For any unique status semantics, σ ,

$$\mu_{\sigma}^{\mathcal{H}} = \{\{x\} : x \notin E_{\sigma}(\mathcal{H})\}$$

Proof: For (a), consider any $Q \in \mu_{CF}^{\mathcal{H}}$ and observe that any such Q has $|Q| \leq 2$: the property of Q being a forbidden set for conflict-free sets is easily seen to be equivalent to $(Q \times Q) \cap \mathcal{A} \neq \emptyset$ so that the corresponding minimal forbidden subsets within Q are formed by those pairs $\{x, y\} \subseteq Q$ linked by an attack in \mathcal{A} together with self-attacking arguments.

For (b) if $S \subseteq \mathcal{X}$ is defenceless in \mathcal{H} not only is S itself not in $\mathcal{E}_{adm}^{\mathcal{H}}$ but also S cannot be extended to an admissible set. Hence S cannot form a subset of any member of $\mathcal{E}_{adm}^{\mathcal{H}}$, i.e. $S \in \kappa_{ADM}^{\mathcal{H}}$ as required.

Part (c) is trivial. □

It is easy to see that for each $0 \leq k \leq n$ ($n = |\mathcal{X}|$) one can construct AFs $\langle \mathcal{X}, \mathcal{A} \rangle$ in which there is some $S \in \mathcal{E}_{pr}(\langle \mathcal{X}, \mathcal{S} \rangle)$ for which $|S| = k$. A similar “hierarchy” is, however, not possible with respect to members of $\mu_{PR}^{\mathcal{H}}$. We present a sub-optimal variant of this claim in,

Theorem 1 For all $n \geq 4$ with $\mathcal{X} = \{x_1, x_2, \dots, x_n\}$

$$\begin{aligned} \forall \mathcal{H}(\mathcal{X}, \mathcal{A}) : \max_{S \in \mu_{PR}^{\mathcal{H}}} |S| &\leq n - \log_2 n \\ \exists \mathcal{H}(\mathcal{X}, \mathcal{A}) : \max_{S \in \mu_{PR}^{\mathcal{H}}} |S| &= \lfloor n/2 \rfloor \end{aligned}$$

Proof: Noting that for $n \in \{2, 3\}$ it is easy to form $S \in \mu_{pr}^{\mathcal{H}}$ having $|S| = 2$ (just use $\mathcal{A} = \{\langle x, y \rangle, \langle x, z \rangle, \langle y, x \rangle\}$ so that $\mu_{pr}^{\mathcal{H}} = \{\{x, y\}, \{x, z\}\}$). The reader may easily verify by inspection that for $n = 3$, no AF having a minimal forbidden set of size 3 can be built.

Thus, assuming $n \geq 4$, we start with the upper bound claim, i.e that

$$\max_{S \in \mu_{PR}^{\mathcal{H}}} |S| \leq n - \log_2 n$$

Let $\mathcal{H} = \langle \mathcal{X}, \mathcal{A} \rangle$ be any AF for which $|\mathcal{X}| = n$ and is such that no other AF, \mathcal{G} of n arguments has

$$\max_{S \in \mu_{PR}^{\mathcal{G}}} |S| > \max_{S \in \mu_{PR}^{\mathcal{H}}} |S|$$

Consider any set S witnessing this behaviour in $\langle \mathcal{X}, \mathcal{A} \rangle$ and without loss of generality assume

$$S = \{x_1, x_2, \dots, x_r\}$$

(where, trivially, $r \geq 3$). It is certainly the case that $S \in \mathcal{E}_{CF}^{\mathcal{H}}$ for otherwise we find a strict subset of S which is in $\kappa_{PR}^{\mathcal{H}}$ in contradiction to $S \in \mu_{PR}^{\mathcal{H}}$.

For each $x_i \in S$, let S_i denote $S \setminus \{x_i\}$. By definition from $S \in \mu_{PR}^{\mathcal{H}}$ we therefore have for every i , $S_i \notin \kappa_{PR}^{\mathcal{H}}$ and hence we can find $T_i \subseteq \mathcal{X} \setminus S$ that satisfies

$$S_i \cup T_i \in \mathcal{E}_{pr}^{\mathcal{H}}$$

Observe that the system $\langle T_1, T_2, \dots, T_r \rangle$ must consist of r *distinct* subsets of $\mathcal{X} \setminus S$, i.e. $T_i = T_j$ if and only if $i = j$. For suppose, without loss of generality, $T_1 = T_2$. Then

$$S_1 \cup T_1 \in \mathcal{E}_{PR}^{\mathcal{H}} \quad \text{and} \quad S_2 \cup T_1 \in \mathcal{E}_{PR}^{\mathcal{H}}$$

so that $S_1 \cup S_2 \cup T_1 = S \cup T_1 \in \mathcal{E}_{ADM}^{\mathcal{H}}$ contradicting $S \in \kappa_{PR}^{\mathcal{H}}$. Notice that admissibility of $S \cup T_1$ follows since the set is conflict-free and should $y \in \mathcal{X} \setminus (S \cup T_1)$ attack $S \cup T_1$ either it attacks some member of T_1 and thence is counterattacked

by both S_1 and S_2 or y attacks some argument in $S = S_1 \cup S_2$ and so is defended by either T_1 or S_1 (if $y \in S_1^-$) or S_2 (should $y \in S_2^-$).

From this argument we obtain the (crude) upper bound claimed on the size of the largest possible set in $\mu_{\text{PR}}^{\mathcal{H}}$: there are r arguments in S and require $n - r$ (the size of $\mathcal{X} \setminus S$) to be such that (at least) r *distinct* sets may be formed. That is we require

$$2^{n-r} \geq r$$

Should $r > n - \log_2 n$ then, $2^{n-r} = 2^{\log_2 n - \varepsilon}$ for some $\varepsilon > 0$, giving $2^{n-r} = n/2^\varepsilon$ and since $2^\varepsilon > 1$ (via $\varepsilon > 0$) it follows that

$$\frac{n}{2^\varepsilon} < n - \log_2 n + \varepsilon$$

as required for the upper bound.

To show that there are AFS, $\mathcal{H}(\mathcal{X}, \mathcal{A})$ for which $\max_{S \in \mu_{\text{PR}}^{\mathcal{H}}} |S|$ is at least $\lfloor |\mathcal{X}|/2 \rfloor$, let $m \geq 2$ and define

$$\mathcal{X} = \begin{cases} \{y_1, y_2, \dots, y_m, z_1, z_2, \dots, z_m\} & \text{if } n = 2m \\ \{y_1, y_2, \dots, y_m, z_1, z_2, \dots, z_m, u\} & \text{if } n = 2m + 1 \end{cases}$$

We construct an AF, $\langle \mathcal{X}, \mathcal{A} \rangle$, for which

$$\{y_1, y_2, \dots, y_m\} \in \mu_{\text{PR}}^{\mathcal{H}}$$

We concentrate on the case $n = 2m$, since the construction for $n = 2m + 1$ is identical. For the arguments, $\{z_1, z_2, \dots, z_m\}$ all of the $m(m-1)$ attacks,

$$\{ \langle z_i, z_j \rangle : 1 \leq i \neq j \leq m \}$$

are added, so that *at most* one z_k can appear in any $P \in \mathcal{E}_{\text{adm}}^{\mathcal{H}}$. The set of attacks is completed with

$$\{ \langle z_i, y_i \rangle : 1 \leq i \leq m \}$$

Consider the set $S = \{y_1, y_2, \dots, y_m\}$. Certainly $S \notin \mathcal{E}_{\text{adm}}^{\mathcal{H}}$ since, although conflict-free, there is no way of defending the attack on y_i arising from z_i . In addition, it is not possible to find a subset T of \mathcal{X} for which $S \cup T \in \mathcal{E}_{\text{adm}}^{\mathcal{H}}$, since the only arguments available to form such a set are with $\{z_1, \dots, z_m\}$ and the resulting $S \cup T$ would fail to be conflict-free.

In total these establish $S \in \kappa_{\text{ADM}}^{\mathcal{H}}$. It is, however, also a *minimal* such set. To see this, let $S_i = S \setminus \{y_i\}$. It is not hard to see that for each i , $S_i \notin \kappa_{\text{ADM}}^{\mathcal{H}}$: the set $S_i \cup \{z_i\}$ being in $\mathcal{E}_{\text{ADM}}^{\mathcal{H}}$ (in fact it is a preferred extension). The argument z_i defends itself from attacks stemming from z_j ($j \neq i$) and, furthermore defends $y_j \in S_i$ from the attack on it by z_j . Thus $S_i \cup \{z_i\}$ is both conflict-free and

defensive, i.e. in $\mathcal{E}_{adm}^{\mathcal{H}}$. This establishes that no strict subset of S belongs to $\kappa_{ADM}^{\mathcal{H}}$ while S itself is in $\kappa_{ADM}^{\mathcal{H}}$. It follows that $S \in \mu_{ADM}^{\mathcal{H}}$ with $|S| = m = \lfloor n/2 \rfloor$. \square

By developing consequences of the idea of “*conflict-sensitivity*” introduced in [10] we can, in fact, show that this lower bound is optimal, i.e. for every $\mathcal{H}(\mathcal{X}, \mathcal{A})$, $\max_{S \in \mu_{PR}^{\mathcal{H}}} |S| \leq \lfloor |\mathcal{X}|/2 \rfloor$. We omit the details on account of limited space.

4. Computational Complexity of Forbidden Set Problems

Given the formal definition of forbidden set it is easy to classify the complexity of membership in $\kappa_{\sigma}^{\mathcal{H}}$ on the basis of results from [6,9,11]. Thus,

Fact 1 *Given $\mathcal{H} = \langle \mathcal{X}, \mathcal{A} \rangle$ and $S \subseteq \mathcal{X}$ deciding if $S \in \kappa_{\sigma}^{\mathcal{H}}$ is*

$$\begin{array}{ll} \text{in P} & \text{if } \sigma \in \{\text{CF, NVE, GR}\} \\ \text{coNP-complete} & \text{if } \sigma \in \{\text{ADM, PR, ST}\} \\ \Pi_2^P\text{-complete} & \text{if } \sigma \in \{\text{SST}\} \end{array}$$

The last two cases holding even if S contains just a single argument.

Proof: Polynomial time methods for $\sigma \in \{\text{CF, NVE}\}$ cases simply involve checking if $S \times S$ has a non-empty intersection with \mathcal{A} , i.e. some attack involves arguments in S . Similarly for grounded semantics $S \in \kappa_{GR}^{\mathcal{H}}$ if and only if S contains an argument not belonging to the grounded extension. This being efficiently computable deciding $S \in \kappa_{GR}^{\mathcal{H}}$ is also so. When $S = \{x\}$ (i.e. a single argument) the decision $S \in \kappa_{\sigma}^{\mathcal{H}}$ is simply a rephrasing of $\neg \text{CA}_{\sigma}(\mathcal{H}, x)$. The complexity classification for $\{\text{ADM, PR, ST}\}$ is now immediate from Dimopoulos and Torres [6] while that of $\{\text{SST}\}$ follows from Dvorak and Woltran [11]. \square

While obtaining exact complexity results for deciding membership of $\kappa_{\sigma}^{\mathcal{H}}$ is straightforward using well-known results, the question of membership of the *minimal* forbidden sets turns out to be rather less so. Although the *single* argument instance $\{x\} \in \mu_{\sigma}^{\mathcal{H}}$ has identical complexity to its general counterpart $\{x\} \in \kappa_{\sigma}^{\mathcal{H}}$ for $\sigma \in \{\text{ADM, PR, SST}\}$ the reason for this is that $\mathcal{E}_{\sigma}^{\mathcal{H}} \neq \emptyset$ for these semantics. From which it follows that

$$(\{x\} \in \mu_{\sigma}^{\mathcal{H}}) \Leftrightarrow (\{x\} \in \kappa_{\sigma}^{\mathcal{H}}) \Leftrightarrow \neg \text{CA}_{\sigma}(\mathcal{H}, x)$$

This argument, however, fails to apply whenever S contains at least two arguments. We can observe, however, that

$$S \in \mu_{\sigma}^{\mathcal{H}} \Leftrightarrow (S \in \kappa_{\sigma}^{\mathcal{H}}) \wedge \left(\bigwedge_{y \in S} S \setminus \{y\} \notin \kappa_{\sigma}^{\mathcal{H}} \right)$$

That is we do not need to test *every* subset of S in order to confirm its membership of $\mu_{\sigma}^{\mathcal{H}}$.

Recalling that the complexity class D^P is defined by those decision problems, Q whose positive instance are both positive instances of some decision problem,

L_1 belonging to NP and positive instances of some decision problem, L_2 , in coNP, the following holds for verifying membership of a given set S in $\mu_\sigma^\mathcal{H}$.

Theorem 2

- a. For $\sigma \in \{\text{PR}, \text{ADM}, \text{COM}\}$, given $\langle S, \langle \mathcal{X}, \mathcal{A} \rangle \rangle$ deciding if $S \in \mu_\sigma^\mathcal{H}$ for the AF $\mathcal{H} = \langle \mathcal{X}, \mathcal{A} \rangle$ is D^p -complete, even for instances $\langle S, \langle \mathcal{X}, \mathcal{A} \rangle \rangle$ in which $|S| = 2$.
- b. For stable semantics deciding $S \in \mu_{\text{ST}}^\mathcal{H}$ is D^p -complete even with instances having $|S| = 1$.

Proof: (Outline) In the case of (a), we recall from Corollary 1(a) that $\kappa_{\text{ADM}}^\mathcal{H} = \kappa_{\text{PR}}^\mathcal{H} = \kappa_{\text{COM}}^\mathcal{H}$ so it suffices to demonstrate the upper bound for $\sigma = \text{ADM}$. Given $\langle S, \langle \mathcal{X}, \mathcal{A} \rangle \rangle$ with $S \subseteq \mathcal{X}$, $S \in \mu_{\text{ADM}}^\mathcal{H}$ requires,

$$\exists \langle T_1, T_2, \dots, T_k \rangle : T_i \subseteq \mathcal{X} \setminus S \text{ and } T_i \cup S \setminus \{y_i\} \in \mathcal{E}_{\text{ADM}}^\mathcal{H}$$

capturing the condition that every strict subset of S can be extended to an admissible set. In addition, S itself must be a forbidden set, i.e.

$$\forall U \subseteq \mathcal{X} \setminus S \quad S \cup U \notin \mathcal{E}_{\text{ADM}}^\mathcal{H}$$

And now defining

$$\begin{aligned} L_1 &= \{ \langle \mathcal{X}, \mathcal{A}, S \rangle : \exists \langle T_1, \dots, T_{|S|} \rangle T_i \cup S \setminus \{y_i\} \in \mathcal{E}_{\text{ADM}}^\mathcal{H} \} \\ L_2 &= \{ \langle \mathcal{X}, \mathcal{A}, S \rangle : \forall U \supseteq S \quad U \notin \mathcal{E}_{\text{ADM}}^\mathcal{H} \} \end{aligned}$$

we see that $S \in \mu_{\text{adm}}^\mathcal{H}$ if and only if $\langle \mathcal{H}, S \rangle \in L_1 \cap L_2$. Since $L_1 \in \text{NP}$ and $L_2 \in \text{coNP}$ we deduce $S \in \mu_{\text{adm}}^\mathcal{H}$ can be decided in D^p .

To establish D^p -hardness we present a reduction to instances $\langle \langle \mathcal{X}, \mathcal{A} \rangle, S \rangle$ from instances $\langle \varphi_1, \varphi_2 \rangle$ of the canonical D^p -complete problem SAT-UNSAT in which these are accepted if and only if the CNF, φ_1 is satisfiable and the CNF φ_2 is unsatisfiable. Given an instance $\langle \varphi_1, \varphi_2 \rangle$ of SAT-UNSAT \mathcal{H} is formed by combining three copies of the “standard translation” of CNF formulae to AFS: two of these with designated arguments φ_1^1 and φ_1^2 capturing the structure of φ_1 ; the other, tied with the argument φ_2 , linked with the structure of φ_2 . The framework uses four additional arguments, $\{p_1, p_2, q_1, q_2\}$ which are configured in a directed cycle

$$\varphi_1^1 \rightarrow p_1 \rightarrow q_1 \rightarrow \varphi_1^2 \rightarrow q_2 \rightarrow p_2 \rightarrow \varphi_1^1$$

Finally the arguments $\{q_1, p_2\}$ are attacked by φ_2 .

It can be shown that $\langle \varphi_1, \varphi_2 \rangle$ is accepted as an instance of SAT-UNSAT if and only if $\{\varphi_1^1, \varphi_1^2\} \in \mu_{\text{PR}}^\mathcal{H}$, i.e. there are admissible sets, S_1 and S_2 for which $\varphi_1^1 \in S_1$ and $\varphi_1^2 \in S_2$, however no admissible set, S , with $\{\varphi_1^1, \varphi_1^2\} \subseteq S$.

We omit the proof of (b) due to space limitations. \square

5. Conclusions

We have presented an alternative view of extension-based semantics within Dung's AF model: rather than describing solutions in terms of (positive) membership of a set we focus on capturing semantics by describing those sets which cannot form part of a solution. We have demonstrated the containment relationships between extension sets determine containments between the corresponding forbidden set structures and derived some preliminary complexity results on verification. To conclude we briefly mention some further directions. In addition to analogues of generic studies of extension based semantics within the forbidden set paradigm (e.g. realizability in the style of Dunne *et. al.* [10]) one has directions specific to the operations κ and μ defined earlier. In particular since $\kappa(\mathbb{S})$ and $\mu(\mathbb{S})$ are themselves sets of subsets, in principle these operations could be iterated. While the structure of $\kappa(\kappa(\mathbb{S}))$ is uninteresting (being either \emptyset or $2^{\mathcal{X}}$) that of $\mu(\mu(\mathbb{S}))$ appears non-trivial.

References

- [1] P. Baroni, P. E. Dunne, and M. Giacomin. On the resolution-based family of abstract argumentation semantics and its grounded instance. *Artificial Intelligence*, 175:791–813, 2011.
- [2] P. Baroni and M. Giacomin. Resolution-based argumentation semantics. In *Proc. 2nd COMMA*, volume 172 of *FAIA*, pages 25–36. IOS Press, 2008.
- [3] P. Baroni, M. Giacomin, and G. Guida. SCC-recursiveness: a general schema for argumentation semantics. *Artificial Intelligence*, 168(1–2):162–210, 2005.
- [4] R. Baumann, T. Linsbichler, and S. Woltran. Verifiability of argumentation semantics. arXiv preprint arXiv:1603.09502, 2016.
- [5] M. Caminada. Semi-stable semantics. In *Proc. 1st COMMA*, volume 144 of *FAIA*, pages 121–130. IOS Press, 2006.
- [6] Y. Dimopoulos and A. Torres. Graph theoretical structures in logic programs and default theories. *Th. Comp. Sci.*, 170:209–244, 1996.
- [7] P. M. Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming, and N -person games. *Artificial Intelligence*, 77:321–357, 1995.
- [8] P. M. Dung, P. Mancarella, and F. Toni. Computing ideal sceptical argumentation. *Artificial Intelligence*, 171:642–674, 2007.
- [9] P. E. Dunne. The computational complexity of ideal semantics. *Artificial Intelligence*, 173(18):1559–1591, 2009.
- [10] P. E. Dunne, W. Dvořák, T. Linsbichler, and S. Woltran. Characteristics of multiple viewpoints in abstract argumentation. *Artificial Intelligence*, 228:153–178, 2015.
- [11] W. Dvořák and S. Woltran. Complexity of semi-stable and stage semantics in argumentation frameworks. *Inf. Process. Lett.*, 110(11):425–430, 2010.

I Heard You the First Time: Debate in Cacophonous Surroundings

Paul E. DUNNE

Dept. of Computer Science, University of Liverpool, Liverpool, L69 7ZF, UK

Abstract. One often finds in debate involving agents strongly committed to their positions, that argument is promoted not through a rational measured exchange of views but rather through stridency and clamour as proponents try to shout down or otherwise suppress their opponents' opinions. While the presence of moderators may go some way to alleviating the effects of such approaches one has the problems of moderators being ignored and the environment being of a nature that makes the appointment of such infeasible. In this article our concern is, in the first instance, to examine the extent to which an environment where argument is pursued through these means can be modelled. Within this model, we briefly review what techniques may be adopted by participants looking to present their own stance with minimal effort and maximal impact.

Keywords. abstract argumentation frameworks; directed graph spectrum; Perron-Frobenius Theory;

Introduction

Consider a debating arena in which numerous different and conflicting opinions are being championed by several protagonists. There are a number of tactics sometimes adopted by participants that are not intended to progress their stance through rational discourse, but rather since those using such means, mistakenly and naïvely believe them to make their point of view more compelling. Thus one finds, for example in playground or nursery debate, techniques such as wearisome repetition of the same point over and over, this sometimes reduced to single word utterances. Repetition as an indicator of logically “weak” argument, has, of course, long been recognized and studied as one class of fallacious reasoning: e.g. the consequences of *eo ipse* moves in the dialogue protocol of Vreeswijk and Prakken [16], the review of “stone-walling” and other non-cooperative tactics from Gabbay and Woods [12,11]. More generally, strategies whose aim is not to advance but rather to stifle or impede discussion underlie several studies, e.g. Dunne [7,8], Sakama [15], Budzynska and Reed [4].

Participants contributing within (supposedly) more “mature” contexts – such as political debates – will usually recognise the futility of constant repetition as an argumentative tool. To compensate, however, (and often not consciously aware that such measures are being used) they may have recourse to another regressive (or at least non-progressive) technique: that of increasing the force

with which their points are delivered. Thus in non-structured debates this will often take the form of increasing vocal volume in an attempt to drown out the arguments of opponents, so rendering them inaudible to neutral observers. This, in turn, may lead to those same opponents adopting identical tactics reiterating their stance at louder and louder volumes. To counteract the deleterious effect on reasoned debate that results from discussions sinking to the level of shambolic shouting contests, in many legislative assemblies a neutral member is recognized as having – among other responsibilities – some authority to intervene and impose a semblance of order. For example, in the U.K. House of Commons, the rôle of Speaker fulfils this function.¹ Nevertheless, despite the presence of a mediator to oversee the conduct of discussions, it can happen (particularly on sensitive issues) that their authority is ignored.² Given that, even within structured settings with a recognized moderator, there is the potential for debate to descend to acrimonious discord, the likelihood of *un-mediated* exchanges degenerating to similar levels is so much the greater.

Our aim in this paper is to consider such settings and a number of questions arising therein. In particular the issue of what forms of model amenable to *analytic* investigation can be used in order to treat,

- a. Synthesis and discovery of strategies that are intended to impose a point through volubility rather than reason.
- b. Differences between moderated and un-moderated discussion, and the susceptibility of the latter to over-strident contributions skewing debate.

We find a basis for our approach by adapting the seminal abstract argumentation frameworks (AFs) of Dung [6]: in their pure form these encapsulate argument interaction as a directed graph structure $\langle \mathcal{X}, \mathcal{A} \rangle$ wherein \mathcal{X} is a (assumed for our purposes to be *finite*) set of *atomic* arguments and $\mathcal{A} \subseteq \mathcal{X} \times \mathcal{X}$ describes an *attack* relationship over these, so that $\langle p, q \rangle \in \mathcal{A}$ captures the concept of the arguments p and q being incompatible by reason of the argument p “attacking” the argument q . For reasons we develop subsequently we augment \mathcal{A} by assigning to each $\langle p, q \rangle \in \mathcal{A}$ a *positive real value*³ which we will refer to as the *volubility* of $\langle p, q \rangle$ and denoted $\nu(\langle p, q \rangle)$. A triple $\langle p, q, r \rangle$ will be referred to as a *discord*, so that one has an implied relationship $\Delta \subseteq \mathcal{X} \times \mathcal{X} \times \mathbb{R}^+$ in which $\langle p, q, r \rangle \in \Delta$ should $\langle p, q \rangle \in \mathcal{A}$ and $\nu(\langle p, q \rangle) = r$. Before developing our approach in depth it is worth observing that the *atomic* (indivisible) view of “argument” taken in Dung’s formalism, although sometimes the source of objections on account of its highly abstract perspective, captures an important aspect relative to the topic of interest in the present article. Specifically it is the forcefulness with which a claim, p is

¹In the UK, the Speaker although having represented one of the major parties as a member of parliament, on assuming this office, is non-partisan. This status being recognized by the fact that in general elections it is a tradition that the (current) Speaker is unopposed when seeking re-election as M.P. for their local constituency. There have been occasions, however, (the 2015 U.K. Parliamentary election being one) when this tradition has been ignored.

²Among (many) such examples in the UK, is the incident of the senior Conservative MP, Michael Heseltine, seizing and waving the symbolic mace at Labour members singing the *Red Flag* in the aftermath of a heated 1976 debate on state ownership (nationalization): the Speaker was forced to suspend the sitting.

³We distinguish *positive* to indicate > 0 as opposed to *non-negative*, i.e. ≥ 0 .

championed over another claim q – the attack $\langle p, q \rangle \in \mathcal{A}$ – as assessed through its volubility that is of interest, rather than any intrinsic merits (or otherwise) of the arguments involved. We note that our model assigns weights to *attacks* rather than to their *source*, i.e. the *argument* from which these originate. There will, of course, be some (implied) relationship between the former (volubility) and the latter (which will be referred to a stridency subsequently). The question of how *exactly* to model the interaction between these two measures is of some interest, however, while we consider some approaches, space does not permit a full consideration of this issue. Our rationale for attack rather than argument weighting is that this explicitly recognises that a *single* argument might be exerted with varying levels of force against different arguments, e.g. an “authority” figure might feel confident enough in pushing a “weak” argument without shouting against an argument of a subordinate while feeling the need to be more forceful when the same argument is used to attack arguments of peers.

Our principal intention is to propose and establish some basic properties of one approach. The formal setting raises a number of questions of interest, however, our discussion of these is largely to emphasize the potential for further development rather than propose specific solutions. In the remainder of the paper, we first reprise background from Dung’s model in Section 1 which gives a foundation for the structures capturing “debate forms” in Section 2. Section 3 offers the main technical development wherein the concept of a debate being “stable” with respect to some underlying criteria is defined. Together with these criteria we present a broad range of contexts through which a moderator may not only determine whether a current state is “acceptable” but also choose or impose rules enforcing stability. Conclusions are presented within Section 4.

1. Preliminaries

We begin by recalling the concept of abstract argumentation framework and terminology from Dung [6]

Definition 1 We use \mathcal{X} to denote a finite set of arguments with $\mathcal{A} \subseteq \mathcal{X} \times \mathcal{X}$ the so-called attack relationship over these. An argumentation framework (AF) is a pair $\mathcal{H} = \langle \mathcal{X}, \mathcal{A} \rangle$. A pair $\langle x, y \rangle \in \mathcal{A}$ is referred to as ‘ y is attacked by x ’ or ‘ x attacks y ’. Using S to denote an arbitrary subset of arguments for $S \subseteq \mathcal{X}$,

$$\begin{aligned} S^- &=_{\text{def}} \{ p : \exists q \in S \text{ such that } \langle p, q \rangle \in \mathcal{A} \} \\ S^+ &=_{\text{def}} \{ p : \exists q \in S \text{ such that } \langle q, p \rangle \in \mathcal{A} \} \end{aligned}$$

In our subsequent treatment it is assumed for every argument $x \in \mathcal{X}$ that $\langle x, x \rangle \notin \mathcal{A}$: our intention being to consider the effect of x on others, i.e. attacks $\langle x, y \rangle$ stemming from x . Participants are considered not to fight against themselves.

Starting from two concepts – those of *conflict-free* sets S and the arguments *acceptable* to such, Dung offers a number of proposals in order precisely to capture the informal notion of “collection of justifiable arguments”. Thus, $x \in \mathcal{X}$ is *acceptable with respect to* S if for every $y \in \mathcal{X}$ that attacks x there is some $z \in S$

that attacks y . Given $S \subseteq \mathcal{X}$, $\mathcal{F}(S) \subseteq \mathcal{X}$ is the set of all arguments that are acceptable with respect to S , i.e.

$$\mathcal{F}(S) = \{x \in \mathcal{X} : \forall y \text{ such that } \langle y, x \rangle \in \mathcal{A}, \exists z \in S \text{ s.t. } \langle z, y \rangle \in \mathcal{A}\}$$

A subset, S , is *conflict-free* if no argument in S is attacked by any other argument in S , with \subseteq -maximal conflict-free sets referred to as *naïve extensions*. A conflict-free set S is *admissible* if every $y \in S$ is acceptable w.r.t S . S is a *complete extension* if S is conflict-free and should $x \in \mathcal{F}(S)$ then $x \in S$, i.e. every argument that is acceptable to S is a member of S , so that $\mathcal{F}(S) = S$. The set of \subseteq -maximal complete extensions coincide with the set of \subseteq -maximal admissible sets, these being termed *preferred extensions*. The set S is a *stable extension* if S is conflict free and $S^+ = \mathcal{X} \setminus S$.

For a given semantics σ and AF, $\mathcal{H}(\mathcal{X}, \mathcal{A})$ we use $\mathcal{E}_\sigma(\mathcal{H})$ to denote the set of all subsets of \mathcal{X} that satisfy the conditions specified by σ .

2. Debate Arenas & Debate Evolution

It was mentioned earlier that an additional component is added to the basic abstract formalism described by AFs.

Definition 2 A debate arena, \mathcal{D} , is formed by a triple $\langle \mathcal{X}, \mathcal{A}, \nu \rangle$ where $\langle \mathcal{X}, \mathcal{A} \rangle$ is an AF and $\nu : \mathcal{A} \rightarrow \mathbb{R}^+$ is the debate volubility function, associating with each $\langle x, y \rangle \in \mathcal{A}$ a positive real value.

The debate volubility function is viewed as describing the force with which its promoter, $\pi(\langle x, y \rangle)$, asserts the argument to its antagonist, $\alpha(\langle x, y \rangle)$.

Of course the idea of augmenting Dung’s *ur*-formalism by allowing quantitative associations with attacks (and, indeed, arguments themselves) has a rich history, being adopted in, amongst others: treatments of so-called “inconsistency tolerance” in Dunne *et al.* [9], Coste-Marquis *et al.* [5]; algorithmic treatments, e.g. Bistarelli and Santini [3]; modelling probabilistic structures, e.g. Li *et al.* [13].

The scenarios of interest to our study involve, however, an aspect which the quantitative formulation of debate arena fails to describe: its treatment of volubility is *static*. In practice, given the context modelled, one would expect the level at which a promoter directs the attack on an antagonist to vary. Such variation need not necessarily be a monotonic increase in $\nu(\langle x, y \rangle)$: hence the often used rhetorical device of reducing the level at which a point is made for emphasis.⁴

Our notion of debate arena can, in essence, be seen as a snapshot within an evolving debate: contributors adjusting their promotion of given arguments over time. In order to reflect *dynamic* elements we formalise this concept via,

⁴For example notice: the contrasting questioning styles in Maximilian Schell’s cross-examination of Montgomery Clift and the underspoken manner in which its final observation is delivered (*Judgement at Nuremberg*, Kramer, 1961); the unvarying level of Olivier’s repetition of the question “Is it safe?” with finality indicated by only a slight drop in tone. (*Marathon Man*, Schlesinger, 1976).

Definition 3 An evolving debate is a sequence,

$$\underline{\mathcal{D}} = \langle \mathcal{D}_0, \mathcal{D}_1, \dots, \mathcal{D}_k, \dots \rangle$$

of debate arenas in which $\mathcal{D}_k = \langle \mathcal{X}, \mathcal{A}_k, \nu_k \rangle$ with $\nu_k : \mathcal{A}_0 \rightarrow \mathbb{R}^+ \cup \{0\}$. This is required to satisfy

$$\forall k \geq 1 \quad \mathcal{A}_k \subseteq \mathcal{A}_{k-1} ; \quad \nu_k : \mathcal{A}_k \rightarrow \mathbb{R}^+$$

Furthermore, should $\langle x, y \rangle \in \mathcal{A}_{k-1}$ but $\langle x, y \rangle \notin \mathcal{A}_k$ then $\nu_k(\langle x, y \rangle) = 0$.

Notice that an evolving debate may, in principle, be an infinite sequence of debate arenas. We can, however, prescribe conditions under which $\underline{\mathcal{D}}$ may be treated as finite. One of these follows directly from the subset condition $\mathcal{A}_k \subseteq \mathcal{A}_{k-1}$, namely: an evolving debate is *terminal* if at some point, $t \geq 0$, we have $\mathcal{A}_t = \emptyset$. This condition is, in fact, an extreme form (and thence implied by), the notion of an evolving debate reaching *stasis*.⁵ Hence an evolving debate has reached stasis if it contains debate arenas \mathcal{D}_k and \mathcal{D}_r with $r > k$ and $\nu_k = \nu_r$ (note that this implies $\mathcal{A}_k = \mathcal{A}_r$). In general we may be interested in the specific case $r = k + 1$, but in principle given that there is no requirement for monotonicity respecting ν_k and ν_r one could reach the situation where identical arenas appear after some interval. Implicitly, by regarding the occurrence of this as indicative of stasis the implication is that should $\nu_k = \nu_r$ then $\nu_{k+1} = \nu_{r+1}$. Notice this is an *assumption* concerning how $\underline{\mathcal{D}}$ would evolve rather than a formal claim of its structure.

3. Stable debates: detection and enforcement

We recall that one issue of interest concerns responding to situations where the force with which arguments are promoted, reaches a level sufficient to obstruct other participants. In fact this scenario has similarities to the well-studied problems of dealing with *power control* in mobile communications, see e.g. Bertoni [2].⁶

A significant distinction from our setting and such as these is the fact that the latter occurs within a rather more “cooperative” context: levels of signal strength being assigned externally having been determined at optimal levels through analysis and, once fixed, no deviation occurs.

To these ends the following factors are relevant: the force with which argument x_j is being pressed upon the promoter of argument x_i ; the stridency with which the champion of argument x_i is proclaiming this to others.

The first of these, which we will denote by F_{ij} is,

$$F_{ij} = \begin{cases} 0 & \text{if } \langle j, i \rangle \notin \mathcal{A} \\ \nu(\langle j, i \rangle) & \text{otherwise} \end{cases}$$

⁵We, intentionally, avoid the, potentially misleading, term “agreement” (which might reflect a specific form of “stasis”) and rather overloaded words such as “equilibrium”.

⁶For example, when there are several competing “mobile phone networks” each using transmitter stations whose signal strength must be high enough to enable good reception by the network users but not at such a level as to cause excessive interference with other networks.

The latter, denoted S_i is described by a positive real value.⁷

For each x_i the quantity S_i represents the overall volume that is being used to press its merits upon others. Conversely, F_{ij} captures the interference with this case being inflicted by the promoter of x_j . One might reasonably claim, therefore, that x_i is being “promoted too forcefully” should its stridency S_i “significantly” exceed the total interference that it must tolerate as inflicted by the other actors in the system. This interpretation raises the following questions: how to assess whether the agent championing x_i is “too strident”, informally, how is it determined if this agent is shouting too loudly? Secondly, *what* level of promotion is considered “excessive”?

For the moment let us assume that each argument has assigned to it a *non-negative* real value, μ_i , that defines (in some sense) the “acceptable” level of force with which x_i can be promoted without this being considered detrimental to the interests of others. Then should the *ratio* between S_i and the amount of interference dealt with, violate the levels set by μ_i then one can conclude that the actor promoting x_i is “shouting too much”. We have, however, one point of detail to consider, namely how to describe what is measured as “the amount of interference dealt with”. In principle one could simply fix this as the total of the forces (F_{ij}) directed against it. The problem with this, however, is its failure to take into account how an agent promoting x_j (with $\langle x_j, x_i \rangle \in \mathcal{A}$) might manipulate the system. Suppose, in a moderated system, the sanction for “shouting too loud” is (perhaps temporary) expulsion. Then measuring “acceptable” noise level via $(S_i / \sum_{j \neq i} F_{ij}) \leq \mu_i$ allows the agent pushing x_j to (for the time being) fix $\nu(\langle j, i \rangle)$ at a “token minimum” whilst compensating, for instance, by increasing the level with which x_j is forced upon other agents. Such manoeuvres lead to an increase in $(S_i / \sum_{j \neq i} F_{ij})$ (even more so if conducted in conjunction with other allied agents) with the possible result that the agent promoting x_i is suspended even though there has been no increase in stridency *from this agent*. Despite this, the agent pressing x_j benefits (x_i is taken out of the system) even though it may be pushing some arguments “harder” (in order to maintain its – presumably considered acceptable – level of stridency). To moderate such manipulative effects (although as we discuss later, it is uncertain whether these can be entirely eliminated), in gauging whether an agent is “shouting too loudly” we view the interference from x_j it must contend with relative to the overall volume with which x_j is being announced. That is to say, the relevant ratio we examine in deciding if x_i is being pushed “too hard” is *not* $S_i / (\sum_{j \neq i} F_{ij})$ but rather

$$\frac{S_i}{\sum_{j \neq i} F_{ij} S_j}$$

This now leads to

Definition 4 Let $\underline{\mu} = \langle \mu_1, \mu_2, \dots, \mu_n \rangle$ and $\mathcal{D} = \langle \mathcal{X}, \mathcal{A}, \nu \rangle$ be a debate arena. We say that \mathcal{D} is stable with respect to stridency $\underline{\mu}$ (or simply $\underline{\mu}$ -stable) if

⁷We defer, for the moment, issues arising in relating S_i to the volubility in promoting x_i .

$$\forall 1 \leq i \leq n \quad \frac{S_i}{\sum_{j \neq i} F_{ij} S_j} \leq \mu_i$$

In other words the debate represented by $\mathcal{D} = \langle \mathcal{X}, \mathcal{A}, \nu \rangle$ is being “harmoniously” conducted should the maximum level of noise (μ_i) set for each participant *not* be exceeded by any.

If we examine the condition described in Defn. 4 then (for the limits defined by $\underline{\mu}$) the debate arena is $\underline{\mu}$ -stable if

$$\forall 1 \leq i \leq n \quad S_i \leq \mu_i \sum_{i \neq j} F_{ij} S_j$$

Now consider the $n \times n$ *force*, \mathbf{F} and *constraint*, \mathbf{C} , matrices defined through

$$\mathbf{F}_{ij} = \begin{cases} 0 & \text{if } i = j \\ F_{ij} & \text{otherwise} \end{cases} \quad \mathbf{C}_{ij} = \begin{cases} \mu_i & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$$

together with the $n \times 1$ column vector, $\underline{\mathbf{S}}$ formed by the transpose of $\langle S_1, S_2, \dots, S_n \rangle$. The n required relations are then expressed as:

$$\mathbf{C} \times \mathbf{F} \times \underline{\mathbf{S}} \geq \underline{\mathbf{S}}$$

or, writing \mathbf{B} for the product of $\mathbf{C} \times \mathbf{F}$:

$$\mathbf{B} \times \underline{\mathbf{S}} \geq \underline{\mathbf{S}} \tag{1}$$

Now, in the scenario we have presented, at any given instant (for \mathcal{D} within an evolving debate $\underline{\mathcal{D}}$): \mathbf{F} is determined by the current volubility function; \mathbf{C} has been fixed (possibly through a moderator); and we have assumed $\underline{\mathbf{S}}$ is within the control of the agents involved.

Thus within a given \mathcal{D} the component $\mathbf{B} = \mathbf{C} \times \mathbf{F}$ of (1) is unchanging and provided that the contribution from $\underline{\mathbf{S}}$ satisfies $\mathbf{B} \times \underline{\mathbf{S}} \geq \underline{\mathbf{S}}$ the debate is $\underline{\mu}$ -stable.

This summary reduces the issue being considered to the following question

Are there conditions for \mathbf{B} that (if satisfied) allow some “suitable” $\underline{\mathbf{S}}$ to be adopted?

This question is simply a rephrasing of a classical linear algebra question, as captured by

Fact 1⁸ For an $n \times n$ real-valued matrix, \mathbf{A} and $\lambda \in \mathbb{R}^+$, the $n \times 1$ (non-zero) vector $\underline{\mathbf{z}}$ satisfies $\mathbf{A}\underline{\mathbf{z}} = \lambda\underline{\mathbf{z}}$ if and only if λ is an eigenvalue of \mathbf{A} and $\underline{\mathbf{z}}$ an associated (right) eigenvector.

Of course, should any $\underline{\mathbf{z}}$ satisfy $\mathbf{A}\underline{\mathbf{z}} = \lambda\underline{\mathbf{z}}$, (with $\lambda \geq 1$) then we find infinitely many such solutions simply by using any scalar multiple of $\underline{\mathbf{z}}$.

Now it is easily seen that \mathbf{B} is non-negative. If it is also *irreducible*⁹ we have,

⁸What is stated as “fact” here, is often used as a formal definition of eigenvalue and eigenvector w.r.t. to a matrix \mathbf{A} .

⁹An $n \times n$ non-negative real-valued matrix \mathbf{A} is said to be *irreducible* if for each $\langle i, j \rangle$ there is some $k \in \mathbb{N}$ for which $[\mathbf{A}^k]_{ij} > 0$.

Theorem 1 (*Perron-Frobenius Theorem [14, 10]*)

If \mathbf{A} is an irreducible $n \times n$ matrix then,

PF1. There is a positive real eigenvalue, λ_{pf}^A , of \mathbf{A} with positive eigenvectors.

PF2. If λ is any other¹⁰ eigenvalue of \mathbf{A} then $|\lambda| < \lambda_{pf}^A$. Notice that, writing $\lambda = x + iy$ with $y \neq 0$ in the case of complex values, $|\lambda|$ is the (positive) square root of $(x^2 + y^2)$.

In total Thm. 1, prescribes *sufficient* conditions for \mathcal{D} to be $\underline{\mu}$ -stable.

Theorem 2 The debate arena $\mathcal{D} = \langle \mathcal{X}, \mathcal{A}, \nu \rangle$ is $\underline{\mu}$ -stable if the product of constraint and force matrices $\mathbf{C} \times \mathbf{F}$ is irreducible and $\lambda_{pf}^{\mathbf{C} \times \mathbf{F}} \geq 1$.

Proof: Immediate from definitions and consequences of Thm. 1. The stridency vector $\underline{\mathbf{S}}$ can be chosen as *any* (positive) eigenvector for $\lambda_{pf}^{\mathbf{C} \times \mathbf{F}}$. These properties and choices ensure

$$(\mathbf{C} \times \mathbf{F}) \times \underline{\mathbf{S}} = \lambda_{pf}^{\mathbf{C} \times \mathbf{F}} \times \underline{\mathbf{S}} \geq \underline{\mathbf{S}}$$

□

The requirement in Thm. 2 that the supporting matrix $\mathbf{C} \times \mathbf{F}$ be irreducible may seem unduly limiting: in fact this is not the case.

Theorem 3 If the structure $\langle \mathcal{X}, \mathcal{A} \rangle$ describes a strongly-connected¹¹ directed graph then,

$$\forall \nu : \mathcal{A} \rightarrow \mathbb{R}^+, \underline{\mu} \in \langle \mathbb{R}^+ \rangle^{|\mathcal{X}|} \quad \mathbf{C} \times \mathbf{F} \text{ is irreducible.}$$

Proof: (Outline) Let \mathbf{B} denote $\mathbf{C} \times \mathbf{F}$. Then,

$$[\mathbf{B}]_{ij} = \sum_{k=1}^n C_{ik} F_{kj} = \mu_i F_{ij}$$

Thus it suffices to establish that $\langle \mathcal{X}, \mathcal{A} \rangle$ being strongly-connected implies \mathbf{F} is irreducible. Consider any $\langle x_i, x_j \rangle \in \mathcal{X}^2$. If $\langle x_i, x_j \rangle \in \mathcal{A}$ then $F_{ij} > 0$ so that the choice $k = 1$ witnesses $[\mathbf{F}^k]_{ij} > 0$. If $\langle x_i, x_j \rangle \notin \mathcal{A}$ (so that $F_{ij} = 0$) let t be the number of arguments in any path from x_i to x_j (where $t > 2$)¹² that is

$$x_i \equiv y_1 \rightarrow y_2 \rightarrow \cdots \rightarrow y_t \equiv x_j$$

¹⁰In general, the *spectrum* $\{\lambda_1, \dots, \lambda_m\}$ of eigenvalues for \mathbf{A} could contain n elements, some of which may be complex.

¹¹A directed graph $\langle V, E \rangle$ is said to be *strongly-connected* if for every pair $\langle v_i, v_j \rangle \in V^2$ there is a directed path of edges in E by which v_j can be reached from v_i .

¹²Note these do *not* have to be distinct, so a path from x_1 to x_1 might be witnessed by $x_1 \equiv y_1 \rightarrow y_2 \rightarrow y_3 \equiv x_1$ in the event of \mathcal{A} containing symmetric attacks $\langle x_1, y_2 \rangle$ and $\langle y_2, x_1 \rangle$.

so that $\langle y_k, y_{k+1} \rangle \in \mathcal{A}$ for all $1 \leq k < t$. We show by induction on $t \geq 2$ that when such a path exists from x_i to x_j then $[\mathbf{F}^{t-1}]_{ij} > 0$. The base ($t = 2$) is already established via $x_i \equiv y_1 \rightarrow y_2 \equiv x_j$, i.e. the case $\langle x_i, x_j \rangle \in \mathcal{A}$. Assuming the property holds for all $t < k$ with $k \geq 3$, i.e. $[\mathbf{F}^{k-1}]_{ij} > 0$, suppose

$$x_i \equiv y_1 \rightarrow y_2 \rightarrow \cdots \rightarrow y_k \equiv x_j$$

is a path linking x_i to x_j and that no path with at least one attack and at most $k - 1$ between the two exists. By definition,

$$[\mathbf{F}^{k-1}]_{ij} = [\mathbf{F}^{k-2} \times \mathbf{F}]_{ij} = \sum_{r=1}^n [\mathbf{F}^{k-2}]_{ir} F_{rj}$$

and (with a slight notational abuse)

$$\sum_{r=1}^n [\mathbf{F}^{k-2}]_{ir} F_{rj} \geq [\mathbf{F}^{k-2}]_{i(t-1)} F_{(t-1)j}$$

Now $F_{(t-1)j} > 0$ since $\langle y_{t-1}, x_j \rangle \in \mathcal{A}$ and (via the Inductive Hypothesis) $[\mathbf{F}^{k-2}]_{i(t-1)} > 0$ (since $x_i \equiv y_1 \rightarrow \cdots \rightarrow y_{k-1}$ is a path from x_i to y_{k-1}). Hence we deduce $[\mathbf{F}^{k-1}]_{ij} > 0$ and \mathbf{F} is irreducible. \square

In combination, Thms. 2 and 3 indicate that for any strongly-connected AF, a moderator is able to carry out some very basic determination of a corresponding debate arena's stability regardless of $\nu : \mathcal{A} \rightarrow \mathbb{R}^+$ and the moderator's desired constraint settings, $\underline{\mu}$.

Thus, given the starting point, \mathcal{D}_0 , in (what will proceed as) an evolving debate $\underline{\mathcal{D}}$ an initial analysis could proceed by:

- a. The moderator decides what they consider to be the maximal acceptable levels of noise, i.e. fixes $\underline{\mu} \in \langle \mathbb{R}^+ \rangle^n$ in such a way that for all $\underline{\tau} \in \langle \mathbb{R}^+ \rangle^n$, should $\tau_i > \mu_i$ (irrespective of other components), then $\underline{\tau}$ is considered to be unreasonable.
- b. Using $\mathbf{C}^{\underline{\mu}}$ and \mathbf{F}^0 (the constraint and force matrices resulting from $\underline{\mu}$ and \mathcal{D}_0) compute $\lambda_{pf}^{0, \underline{\mu}}$ the (unique) maximal positive eigenvalue for $\mathbf{C}^{\underline{\mu}} \times \mathbf{F}^0$.
- c. If $\lambda_{pf}^{0, \underline{\mu}} \geq 1$, set $\langle S_1, S_2, \dots, S_n \rangle$ – the permitted stridency levels – as

$$\left(\frac{\sum_{\langle x_i, x_j \rangle \in \mathcal{A}_0} \nu_0(\langle x_i, x_j \rangle)}{\sum_{i=1}^n w_i} \right) \langle w_1, w_2, \dots, w_n \rangle$$

where \underline{w} is a (transposed) eigenvector of $\lambda_{pf}^{0, \underline{\mu}}$. The multiplicative term preceding this is just a normalizing factor. This gives $w_i > 0$ for all $1 \leq i \leq n$.

- d. Notify agents of the limits on total volubility.

The steps outlined in (a)–(d), raise several further issues. Amongst the most pressing of these we have the following questions.

- Q1 Our derivation did not assume any relations between values of S_i and the volubility used when x_i is promoted. This, however, would not typically be the case, i.e. one would expect to see *some* relationship between S_i and

$$\{ \nu(\langle x_i, x_j \rangle) : \langle x_i, x_j \rangle \in \mathcal{A} \}$$

- Q2 What steps could be taken if, for the choices colouring the computation in (b), the outcome is $\lambda_{pf}^{0, \underline{\mu}} < 1$. In other words \mathcal{D}_0 is “inherently unstable” with respect to $\underline{\mu}$?
- Q3 What effects on overall coordination of debate would arise, should some subset $S \subset \mathcal{X}$ act collectively to exploit some common grounds, e.g. $S \in \mathcal{E}_\sigma(\langle \mathcal{X}, \mathcal{A} \rangle)$ for some semantics σ ?

We consider the first of these in a little more detail here.

Suppose, instead of being an arbitrary positive real, the stridency S_i is *directly* related to $\{ \nu(\langle x_i, x_j \rangle) : \langle x_i, x_j \rangle \in \mathcal{A} \}$, via

$$S_i =_{\text{def}} \sum_{\langle x_i, x_j \rangle \in \mathcal{A}} \nu(\langle x_i, x_j \rangle)$$

That is, the *total* volume emanating in defending x_i is the sum of the efforts put into the individual attacks with x_i as their source. It is easily seen that,

$$S_i = \sum_{\langle x_i, x_j \rangle \in \mathcal{A}} \nu(\langle x_i, x_j \rangle) = \sum_{j \neq i} F_{ji}$$

Recalling that $F_{pq} = 0$ when $\langle x_q, x_p \rangle \notin \mathcal{A}$ the relevant ratio is now,

$$\frac{\sum F_{ji}}{\sum (F_{ij} \sum F_{kj})}$$

How does this affect the matrix representation of the system of inequalities considered earlier? Letting $\underline{\mathbf{1}}$ denote the $n \times 1$ column vector, each of whose elements is 1, it is easy to see that

$$\underline{\mathbf{S}} = \mathbf{F}^T \times \underline{\mathbf{1}}$$

(\mathbf{A}^T denoting the transpose of \mathbf{A} , i.e. the $n \times n$ matrix for which $[\mathbf{A}^T]_{ij} = [\mathbf{A}]_{ji}$).

This now indicates the conditions on \mathbf{C} and \mathbf{F} must satisfy,

$$\mathbf{C} \times \mathbf{F} \times \mathbf{F}^T \times \underline{\mathbf{1}} \geq \mathbf{F}^T \times \underline{\mathbf{1}}$$

In other words *sufficient* conditions for the debate arena, \mathcal{D} , to be $\underline{\mu}$ -stable is that $\mathbf{F}^T \times \underline{\mathbf{1}}$ is an eigenvector for $\lambda_{pf}^{\mathbf{C} \times \mathbf{F}}$ with this eigenvalue being at least 1. In which event,

$$\begin{aligned} \mathbf{C} \times \mathbf{F} \times \mathbf{F}^T \times \underline{\mathbf{1}} &= \lambda_{pf}^{\mathbf{C} \times \mathbf{F}} \times \mathbf{F}^T \times \underline{\mathbf{1}} \\ &\geq \mathbf{F}^T \times \underline{\mathbf{1}} \end{aligned}$$

Regarding our second issue – possible actions in the event that the combination of constraint and force matrices do not allow a suitable stridency assignment to be made – one can posit two approaches: firstly to weaken the desired conditions and adjust \mathbf{C} upwards according to some convention; secondly to consider approaches whereby some subset of existing arguments are “suspended” in the hope that the reduced set-up will allow some degree of harmony. Of course, in both of these approaches a large number of further questions arise. In the first solution approach:

- a. what are good bases for adjusting \mathbf{C} ?
- b. If agents (or a subset of these) view such increased tolerance of noise as an indicator of “weakness” on the part of a moderator, what is to prevent such increasing their contribution to $\underline{\mathbf{S}}$ so that even more generous commitments within \mathbf{C} have no effect?

Similarly the second solution raises,

- c. The “obvious” candidates to remove are those corresponding to agents for which $S_i > \mu_i$. We observed earlier, in choosing $\sum_{j \neq i} F_{ij} S_j$ to measure the degree of interference that x_i is subjected to, that naive mechanisms might allow agents to manipulate the system, (for example if we defined “interference” by $\sum_{j \neq i} F_{ij}$). From the moderator’s perspective such manipulation ought to be ignored. Nevertheless there are many possibilities for choosing the subset of agents to suspend ranging from “the agent for which $(\sum_{j \neq i} F_{ji} - \mu_i)$ is largest”, to all agents exceeding μ_i .
- d. A rather more subtle problem with “brute-force” suspension can, however, appear. Removal of x_i from $\langle \mathcal{X}, \mathcal{A}, \nu \rangle$ will induce a sub-graph of $\langle \mathcal{X}, \mathcal{A} \rangle$. Our discussion of $\mathbf{C} \times \mathbf{F}$ and its properties, was predicated on this being irreducible: property guaranteed in the event of $\langle \mathcal{X}, \mathcal{A} \rangle$ being strongly-connected. Strong-connectivity of $\langle \mathcal{X}, \mathcal{A} \rangle$ does not, however, ensure strong-connectivity of the framework induced by $\mathcal{X} \setminus \{x_i\}$. In principle this may create complications with dominant (i.e. maximal) eigenvalues and existence of positive associated eigenvectors.

As a final issue we, briefly, consider the assumption of “strong-connectivity”. While this is useful in guaranteeing the conditions of Thm. 1 are met, it is not an essential prerequisite of our approach. In particular, by considering the strongly-connected component decomposition of $\langle \mathcal{X}, \mathcal{A} \rangle$ – whose benefits have been studied in Baroni *et al.* [1] – similar analyses of acceptable levels of volubility are possible.

4. Conclusions

The main intention of this paper has been to offer a model (based on Dung’s classical AF formalism) by which problems arising from over-heated debates can be studied. Such models may offer a vehicle for considering divers strategies that could be adopted by moderators in controlling debates with minimal intervention being required. Underpinning the problems of interest is the concern that the *force*

with which an argument is made can seem (to observers) at least as significant factor in gauging its merits as the argument's intrinsic logic and rationale. Our principal aim in this paper has been to highlight an important "non-logical" facet of real-world debate and argument together with a possible modelling approach. It is, of course, the case that this is rather crude and raises a number of directions for future research: a number of these are the focus of work currently in progress.

References

- [1] P. Baroni, M. Giacomin, and G. Guida. SCC-recursiveness: a general schema for argumentation semantics. *Artificial Intelligence*, 168(1–2):162–210, 2005.
- [2] H. L. Bertoni. *Radio propagation for modern wireless systems*. Prentice-Hall, 2000.
- [3] S. Bistarelli and F. Santini. Conarg: A constraint-based computational framework for argumentation systems. In *Proc. ICTAI 2011*, pages 605–612, 2011.
- [4] K. Budzynska and C. Reed. The structure of *ad hominem* dialogues. In *Proc. 4th COMMA*, volume 245 of *FAIA*, pages 410–421. IOS Press, 2012.
- [5] S. Coste-Marquis, S. Konieczny, P. Marquis, and M.A. Ouali. Selecting extensions in weighted argumentation frameworks. In *Proc. 5th COMMA*, volume 245 of *FAIA*, pages 342–349. IOS Press, 2012.
- [6] P. M. Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming, and N -person games. *Artificial Intelligence*, 77:321–357, 1995.
- [7] P. E. Dunne. Prevarication in dispute protocols. In *Proc. 9th ICAIL*, pages 12–21. ACM Press, 2003.
- [8] P. E. Dunne. Suspicion of hidden agenda in persuasive argument. In *Proc. 1st COMMA*, volume 144 of *FAIA*, pages 329–340. IOS Press, 2006.
- [9] P. E. Dunne, A. Hunter, P. McBurney, S. Parsons, and M. Wooldridge. Weighted argument systems: Basic definitions, algorithms, and complexity results. *Artificial Intelligence*, 175(2):457–486, 2011.
- [10] G. Frobenius. Über matrizen aus nicht negativen elementen. *Sitz. Königl. Preuss. Akad. Wiss.*, pages 456–477, 1912.
- [11] D. Gabbay and J. Woods. More on non-cooperation in dialogue logic. *Logic Journal of IGPL*, 9(2):305–324, 2001.
- [12] D. Gabbay and J. Woods. Non-cooperation in dialogue logic. *Synthese*, 127(1):161–186, 2001.
- [13] H. Li, N. Oren, and T. J. Norman. Probabilistic argumentation frameworks. In *Proc. 1st TAFA*, pages 1–16, 2011.
- [14] O. Perron. Zur theorie der matrizen. *Mathematische Annalen*, 64(2):248–263, 1907.
- [15] C. Sakama. Dishonest arguments in debate games. In *Proc. 4th COMMA*, volume 245 of *FAIA*, pages 177–184. IOS Press, 2012.
- [16] G. Vreeswijk and H. Prakken. Credulous and sceptical argument games for preferred semantics. In *Proc. 7th JELIA*, volume 1919 of *LNAI*, pages 224–238. Springer-Verlag, 2000.

Mining Ethos in Political Debate

Rory DUTHIE ^a, Katarzyna BUDZYNSKA ^{a,b}, and Chris REED ^a

^a*Centre for Argument Technology, University of Dundee, UK*

^b*Institute of Philosophy and Sociology, Polish Academy of Sciences, Poland*

Abstract. Despite the fact it has been recognised since Aristotle that ethos and credibility play a critical role in many types of communication, these facts are rarely studied in linguistically oriented AI which has enjoyed such success in processing complex features as sentiment, opinion, and most recently arguments. This paper shows how a text analysis pipeline of structural and statistical approaches to natural language processing (NLP) can be deployed to tackle ethos by mining linguistic resources from the political domain. We summarise a coding scheme for annotating ethotic expressions; present the first openly available corpus to support further, comparative research in the area; and report results from a system for automatically recognising the presence and polarity of ethotic expressions. Finally, we hypothesise that in the political sphere, ethos analytics – including recognising who trusts whom and who is attacking whose reputation – might act as a powerful toolset for understanding and even anticipating the dynamics of governments. By exploring several examples of correspondence between ethos analytics in political discourse and major events and dynamics in the political landscape, we uncover tantalising evidence in support of this hypothesis.

Keywords. Character of speakers; Ethos attack; Ethos support; Natural Language Processing; Parliamentary debates; Sentiment analysis

1. Introduction

Ethos is defined as the character of the speaker [1], i.e. the character of the person who is the participant of communication. It has been extensively studied in disciplines such as rhetoric, epistemology and social psychology for the major role it plays in communication and society.¹ Ethos forms a crucial part of a debate along with two other means of persuasion: pathos which is the audience's emotions; and logos which is the use of reasoning. In [29], arguments containing ethos (argument from expert opinion) are studied from a logos perspective, however pure ethos has been studied less so.

This paper aims to demonstrate that linguistically oriented AI, by making use of a large amount of data, can offer insights and improve our understanding of how ethos influences the interaction between communicating agents and the formation of social structures. In the political sphere, knowing who supports whose ethos (see **Ex. 1**)²; who attacks whom (**Ex. 2**); whether the sentiment is mutual (e.g. Radice used to attack Pawsey and vice versa in the example); which political party the person represents (e.g. Patten supported Ewing twice even though they are from opposite parties); is a powerful tool

¹Although trust attracted quite a lot of attention in AI (cf. [3, 20, 22]), this notion is used differently than the notion of ethos presented in this paper.

²Examples are taken from UK parliament of 1979-1990.

for understanding the dynamics of governments such as the creation of cliques and coalitions, and the rise and fall of rebellious behaviour. However, manual analysis of such large data-sets in broadcast and social media, or parliamentary records is a very labour intensive task. In this paper, we propose a method of automating such analysis.

Example 1 Mr. Chris Patten said, *The hon. Member for Falkirk, East (Mr. Ewing) in his admirable speech, put the position much more clearly than I could.*

Example 2 Mr. Giles Radice said, *In doing so he (Mr. Pawsey) failed to face up to his responsibility both to the House and to the schools of England, Scotland and Wales.*

We use a pipeline of natural language processing techniques to extract the information from the linguistic surface of diplomatic language in expressing opinions during UK parliamentary debates. For example, the phrase, “admirable speech” in **Ex. 1** can suggest support for Mr. Ewing’s ethos,³ while “failed (...) to his responsibility” can be used as a cue that Mr. Pawsey was attacked. This task requires several challenges to be addressed. For example, the dialogical context encourages the use of pronouns (see “he” in **Ex. 2**); reported speech (see **Ex. 3**) includes references to other people which are ethotically neutral; or some phrases, which seem positive such as “honorable”, are in fact a part of political etiquette. Thus, a system we propose is a pipeline of components that deal with these challenges step by step.

Example 3 Mr. Giles Radice said, *The hon. Member for Rugby and Kenilworth (Mr. Pawsey) said that in the United States and Australia this was a local decision.*

Specifically, the contribution of the paper includes: (a) the first freely accessible corpus specifically annotated with tags allowing for the representation of ethotic linguistic structures; (b) a system for ethos mining consisting of existing methods such as Part-of-Speech tagging and SVM-based sentiment classifier as well as new techniques such as anaphora resolution, rule-based expression recognition and a reported speech filter; (c) software for visualisation of the relationships between politicians allowing the analysis to produce insights into data not normally seen in the political science literature; (d) exploratory applications of these visualisation and ethos analytics tools to periods in the historical parliamentary record associated with major political upheaval, demonstrating how the changing political landscape is reflected in, signals in the ethotic interactions in the text.

2. Corpus

Our data is taken from the UK parliamentary record, Hansard, which is an online archive of transcripts of all House of Commons and House of Lords debates dating back to the 1800s (freely available at <http://hansard.millbanksystems.com/>). The archive is organised by the day divided into a number of sessions on different topics. Each turn in the debate consists of the identification of Members of Parliament, MP, followed by their constituency (if this is the first time they have spoken) and their speech.

Corpus	Sessions	Words	Segments	Speakers	Location
Train	30	40,939	387	127	http://arg.tech/Ethan3Train
Test	30	29,178	352	126	http://arg.tech/Ethan3Test
TOTAL	60	70,117	739	253	

Table 1. Summary of the language resources in the EtHan.Thatcher.3 corpus for mining ethos in Hansard.

2.1. Data

The corpus EtHan.Thatcher.3⁴ (see **Table. 1**) was constructed by taking a random sub-sample of Hansard according to the following rubric: select the first two House of Commons debates over 700 words in length from the day closest to the date(s) at the mid-point(s) of the largest uninterrupted date range(s) (initially the midpoint in the range 4th May 1979 and 22nd November 1990 - viz., 11th February 1985; then at the midpoints between 4th May 1979 and 11th February 1985, and between 11th February 1985 and 22nd November 1990, etc.). This avoids bias for annotators and yielded 60 transcripts, the data in each of which was then cleaned such that any titles and section markers were removed to leave only the speakers, organisations or other entities and the statements they made. The transcripts were then split evenly to give a training set and a testing set. The training set formed the training data for the sentiment polarity classifier and was used as the basis for developing domain specific rules for recognising ethotic sentiment expressions.

2.2. Annotation

The annotation was performed by applying four tags (see **Table. 2** for their frequency) according to the following guidelines:

Source-person. Source tag is used to mark a person who utters the statement.

Target-person. Target is a person who is described by the statement.

Ethos support. Ethos support should be identified when: (a) the statement makes explicit mentions of a person, organisation or other entity (excluding groups and assemblages) except when this is reported speech; and (b) it takes the form of supporting a person's credibility or looking to put them in a positive frame through character supports or supports of work; and (c) a support to a person's own ethos should not be analysed as this is deemed to be a fallacy [2]. Compare **Ex. 1**.

Ethos attack. Ethos attack should be identified when: (a) the statement makes explicit mentions of a person, organisation or other entity (excluding groups and assemblages) except when this is reported speech; and (b) it takes the form of attacking a person's credibility or looking to put them into a negative frame; or (c) it may take the form of trying to unbalance authority on a subject giving the attacker more of a right to talk about the subject. Compare **Ex. 2**.

The statements in which speakers refer to other persons are called **Ethotic Sentiment Expressions, ESEs** and the statements which do not contain reference to oth-

³Though such a sentiment can in principle be cancelled or reversed by subsequent linguistic material, in practice in our corpus such situations almost never occur.

⁴The corpus is named as so due to the annotation of session transcripts at different time periods. EtHan.Thatcher.1 containing an original 30 sessions which was extended to EtHan.Thatcher.3 and EtHan.Thatcher.2 containing a subset of EtHan.Thatcher.3 for agreement calculations.

Source-person	243
Target-person	212
Ethos support	179
Ethos attack	560
TOTAL	1,194

Table 2. Occurrences of tags in EtHan_Thatcher_3

ers are denoted as **non-ESEs**. The polarity of these statements is then expressed by the use of abbreviation **+ESE** for positive sentiment (ethos support) and **-ESE** for negative sentiment (ethos attack). The data was analysed according to the standard of argument representation, i.e. Argument Interchange Format (AIF) [24], using the OVA+ annotation tool [9] (freely available at <http://ova.arg-tech.org>) and stored in the AIFdb database [10] (<http://aifdb.org>). Annotation is below sentence level but above word level.⁵

2.3. Evaluation

In order to evaluate annotation, we selected a subset of data used in the EtHan_Thatcher_3 corpus. The selection followed the same method as applied to the whole dataset. The total size of this subset comprises 10% of the EtHan_Thatcher_3 corpus with 6 sessions containing 7,267 words, 91 segments and 30 speakers. Cohen’s kappa for recognising whether the statement is ESE or not gave the value of $\kappa = 0.67$. For ethotic statements, $\kappa = 0.95$, when it is a support or an attack. For source-person of an ethotic statement, $\kappa = 1$ and for target-person it was $\kappa = 0.84$, all for two coders.

3. Automation

3.1. System Architecture

The architecture of the software system for mining ethos consists of three stages, five layers and eight components (see **Fig. 1**). The three stages consist of the ESE / Non-ESE stage, the +/- ESE stage and the network stage. The ESE / Non-ESE stage takes an input of cleaned text transcripts from the EtHan_Thatcher_3 test sub-corpus and classifies each segment as either an ESE or non-ESE. The +/- ESE stage then gives the polarity of ESEs, ESEs with positive sentiment (corresponding to ethos support, as in **Ex. 1**), and ESEs with negative sentiment (corresponding to ethos attack, as in **Ex. 2**). Finally, the network stage provides a visualisation of all ESEs as edges between each participant in the debate.

In the ESE / Non-ESE stage, there are three layers consisting of five components. The parsing layer uses plain text from the EtHan_Thatcher_3 test sub-corpus and applies three different methods to it: Named Entity Recognition (NER), Part-Of-Speech (POS) tagging and a set of domain specific rules. The output is Agent Reference Expressions (AREs) which are any statements referring to another person, organisation or agentive entity. Given the dialogical nature of the material, many statements do not refer to the target-person by their name explicitly, but e.g. by a pronoun (see “he” in **Ex. 1**), by

⁵The annotation is visualised as directed graphs where support is marked as Default Inference, attack as Default Conflict, source-person is in the node with the statement, and target-person – in the node which refers to ethos.

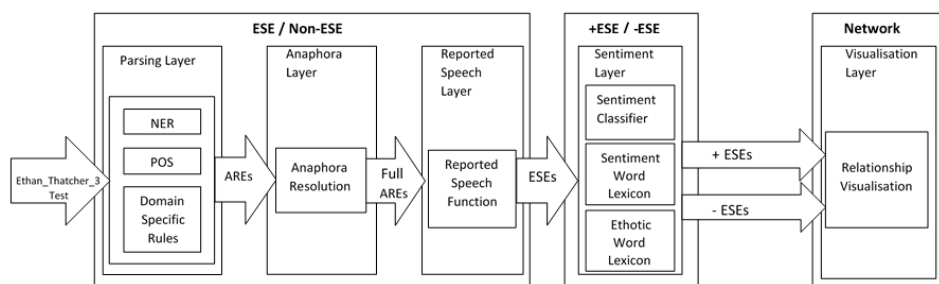


Figure 1. A text analysis pipeline for ethos mining: the extraction, polarisation and networking of ESEs from Hansard sessions in plain text transcripts.

a region MP represents (see “The hon. Member for Falkirk, East” in **Ex. 2**) or by a functional role e.g. “the Prime Minister”. Thus, AREs are then passed to the anaphora layer where both source-person and target-person of the statement are retrieved from the original text. The next challenge is that the repetitions of what has been previously said can be ethotically neutral, especially when an MP wants to remind some thread of the debate which happened many turns earlier (see **Ex. 3**). Therefore, full AREs are passed to the reported speech layer where an ARE is removed if it is not an ethotic expression but a reported speech.

In the +/- ESE stage there is one layer, the sentiment layer, containing two components, the sentiment classifier and the word lexicons. The sentiment classifier and word lexicon components combine to classify ESEs as positive and negative. These two sets are then passed to the Network stage where the visualisation layer displays relationships between people, organisations and other entities. Not attached to **Fig. 1** is the training sub-corpus which is used just for defining domain specific rules and the lexicon for the +/- ESE stage for the sentiment classifier. The techniques of domain specific rules, anaphora resolution, reported speech function and relationship visualisation were developed specifically for the tasks of ethos mining in political debate, and the method of sentiment classification was extended with the development of a lexicon to account for the characteristics of the domain.

3.2. Methods

The ethos mining tool applies existing NLP methods such as Part-of-Speech tagging and an SVM-based sentiment classifier with an existing sentiment word lexicon, and new techniques such as anaphora resolution, rule-based expression recognition, a reported speech filter and an ethotic word lexicon.

Named Entity Recognition (NER). NER, using the Stanford Named Entity Recognizer (Stanford NER) [6] with 92.28% accuracy on the CMU Seminar Announcements information extraction dataset, is performed to extract statements which contain names, organisations and locations from the plain text on the assumption that any specific statement referring to a named entity can in fact be a form of ethotic statement. This is applied to the original text from EtHan.Thatcher_3 test sub-corpus and produces a set of AREs on the assumption that any specific statement made to a named entity can in fact be a form of ethotic statement.

Part-of-Speech (POS) Tagging. POS tagging, using the Stanford POS Tagger [28] with an accuracy of 97.24%, is applied to extract statements which contain pronouns to ac-

count for situations such as in **Ex. 2**. This was applied to the EtHan_Thatcher_3 test sub-corpus and then run against the list of already extracted AREs from the NER to account for any duplicate segments extending the list of AREs.

Domain Specific Rules (DSR). We developed rule-based expression recognition to account for the specific language of the political domain. In the House of Commons, the speaker is not allowed to refer to any other MP by name, but by phrases such as “Honourable Gentleman” or “Honourable Lady”. The constituency name of an MP can be used in the same respect to address an MP such as in **Ex. 1** “The hon. Member for Falkirk, East” Organisations can also be mentioned under a different name, e.g. “the Government” will refer to the party in charge of the government at that time, and “the Opposition” – to the current official opposition. These rules are then extended with the creation of a list of ethotic words to determine if ethos is held in a particular ARE. A list of 326 ethotic statements were compiled from the EtHan_Thatcher_3 training sub-corpus, containing some words not normally used in day-to-day conversation such as “penny-pinching” and “gerrymandering”. These words are common with ethotic attacks. Again the new AREs produced from this component are checked against the list of already extracted AREs to remove duplicates.

Anaphora Resolution (AnaR). We developed this rule-based module with manually defined rules to reconstruct all sources and targets in each AREs. For the source-person, the reconstruction is needed, when a sentence is not the first one in a turn in the dialogue (a turn corresponds to a paragraph in the transcript). In such cases, first the system associates a sentence with a paragraph. Since paragraphs are assigned a source-person, thus this person becomes a source for the sentence. For a target-person, there are two possible cases. First, when the anaphora occurs in situations such as “MP₁ said MP₂ *did this and he did that*”, NER technique is used. In the case of sentences such as “MP₁ said *he did this*”, the system tracks back to the beginning of the paragraph. If nothing is found, then it looks for the speaker of the previous turn.

Reported Speech Function (RSF). We developed a reported speech filter which aims to remove segments containing neutral reports of what previously has been said by other speakers (thus no ethotic sentiment). The technique uses lexical cues such as “says”, “you say” and “told me”, and any segment containing these words is removed from the list of AREs. RSF produces a list of ESEs which are then passed to the sentiment classifier.

Sentiment Classifier (SVM, NB, ME). To perform sentiment analysis three machine learning algorithms were considered: Support Vector Machines (SVM), Naïve Bayes (NB) and Maximum Entropy (ME). A C-Support Vector Classification (C-SVC) algorithm [4] from the LIBSVM library was used to classify ESEs into two sets: positive and negative. To perform NB and ME the Stanford classifier library [13] was used. In selecting these methods, we followed the conclusion formulated in [19] that the discourse approach in sentiment analysis is not satisfactory and that supervised learning techniques are needed (which is demonstrated in [21] with the good performance of SVC of 83%). The lexicon (defined in **Section. 3.2**) was passed to the Stanford CoreNLP [14] library in order to perform lemmatization, allowing the frequency of words in the lexicon to be more accurately calculated.

Lexicon (SWL, EWL). To perform sentiment analysis one existing lexicon was used, the sentiment word lexicon (SWL) [8], and one lexicon created, an ethotic word lexicon (EWL). The SWL, contains 2,006 words tagged as positive and 4,738 words tagged as

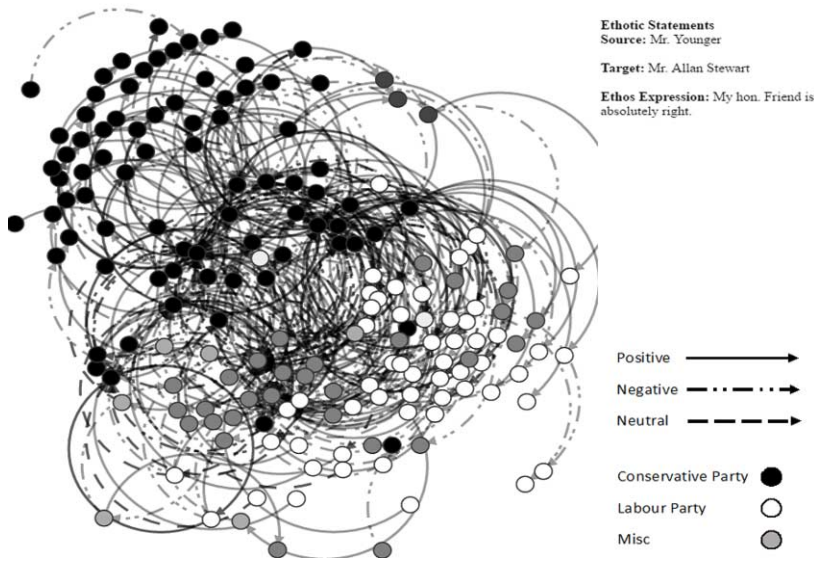


Figure 2. The component of Relationship Visualisation for EtHan_Thatcher_3 test corpus, showing the network of positive and negative relationships in parliament (Available at: <http://arg.tech/EthanVis>).

negative. The EWL is a set of keywords developed using the EtHan_Thatcher_3 training sub-corpus containing 381 tagged sentences with 96 positive and 285 negative from which unigrams, bigrams and trigrams were extracted. Despite the relatively small volume of this set, its advantage lies in its adaptation to sentiment related specifically to ethos in political debate. The removal of non sentiment bearing words and named entities, and the use of n-grams gave 32,858 features overall to be used as training data for machine learning.

Relationship Visualisation. Extracted ESEs with polarity and source and target person were used for visualisation purposes. D3.js a javascript graph visualisation library (available at: <http://d3js.org/>) was used to create force-directed graphs representing positive (coloured as green) and negative (coloured as red) relationships amongst the politicians (see **Fig. 2**). Each edge representing a relationship is associated with a set of ESEs depending on the polarity of the ESE. People are visualised as nodes coloured according to their political party. Nodes are then clustered by political party using a multi-foci technique: nodes which are pulled closer together show that there were either many attacks or supports between them.

3.3. System Results

Results are given for the two stages of ethos mining shown in **Fig. 1**, the ESE / Non-ESE stage and the +/- ESE stage. A result is also given for the combination of these stages. **Table. 3** gives the results of precision, recall and *F*-score for the classification of ESEs as an ESE or Non-ESE. ESEs are defined as correct if the text they hold contains the corresponding segment in the manual analysis, which is true for exact matches and ESEs holding more text than has been manually annotated.

In **Table. 3** we consider a baseline classifier which predicts only the target class (ESE), common machine learning algorithms (ME, NB and SVM) and the ESE / Non-

ESE/Non-ESE	Precision	Recall	F-score
Baseline	0.29	1	0.45
SVM	0.30	0.30	0.30
NB	0.20	0.94	0.32
ME	0.46	0.27	0.34
NER, POS, DSR, AnaR, RSF	0.62	0.77	0.69*
POS, DSR, AnaR, RSF	0.64	0.76	0.70*

Table 3. Results of automatic extraction of ESEs from EtHan.Thatcher_3 Test corpus. We report precision, recall and *F*-score for classifying ESEs as ESE and Non-ESE. The star symbol (*) denotes the classifier above the baseline *F*-score.

+/- ESE	Precision	Recall	F-score
Baseline	0.50	1	0.67
NB, SWL	0.58	0.57	0.57
ME, SWL	0.6	0.65	0.62
SVM, SWL	0.64	0.59	0.62
NB, SWL, EWL	0.74	0.67	0.71*
ME, SWL, EWL	0.71	0.73	0.72*
SVM, SWL, EWL	0.78	0.78	0.78*

Table 4. Results for the sentiment classifier based on a Macro-average of results of both positive and negative classifications. We report precision, recall and *F*-score for a baseline classifier and machine learning classifiers. The star symbol (*) denotes the classifier above the baseline *F*-score.

ESE stage of our system, containing NER and with NER removed. Of these algorithms both of our systems perform above the baseline *F*-score by 53%. To identify people within Hansard it would be logical to perform NER to extract names from text. Although this would be true for most cases of dialogue, due to UK parliamentary rules, the number of instances where names are used explicitly are few. This can cause the problem of many false positives being extracted by the ESE / Non-ESE stage. In removing NER we observe an increase in precision on the ethos mining system with only a slight drop in recall.

In **Table. 4** the results of +/- ESE classification are reported with comparison of common machine learning techniques to a baseline classifier with a macro-averaged precision, recall and *F*-score of the majority class (negative) and the minority class (positive). Comparison is made two different training lexicon, SWL and EWL in **Section. 3.2**. The results indicate that known ethotic words which we developed for the EWL are crucial in obtaining high *F*-score on sentiment classification of ESEs. Using the same set of features an SVM classifier outperforms both a Naïve Bayes Classifier and a Maximum Entropy Classifier with an overall *F*1-score 16% above the baseline.

ESE/Non-ESE & +/- ESE	Precision	Recall	F-score
Baseline	0.14	1	0.25
Full System	0.55	0.65	0.60*

Table 5. Results are provided for the combination of the ESE / Non-ESE stage and the +/- ESE stage.

In **Table. 5** the results of the combination of the ESE/Non-ESE stage and the +/- ESE stage are given. A true value is only given when the system correctly identifies an ESE and gives the correct sentiment polarity, when compared to manual analysis. A drop in overall F -score from **Table. 3** is observed due to the error margin, reported in **Table. 4**. However, when calculating the baseline for the full system this gives F -score 0.25, putting the full system, containing the ESE / Non-ESE stage and SVM +/- ESE stage, 40% ahead of the baseline.

4. Scaling up

In this section, we explore two examples of correspondence between ethos analytics in parliamentary discourse and major political events. In other words, we do not aim here to evaluate the ethos mining tool, but to illustrate its analytical potential by comparing the output of the automatic system not to manual annotation, but to political science publications and news articles from the considered time periods.

February 1st 1997 to April 30th 1997, 53 text transcripts focusing on the final stages of the Conservative government before Labour leader Tony Blair became Prime Minister.⁶ In this time, it was documented that John Major, the then Prime Minister, was struggling to keep his own party on side [7]. This is evident in the analysis with eight ethotic attacks coming from his own party and two attacks coming from Tony Blair, the leader of the opposition at the time, where the average number of attacks is two. Following the loss of the general election to the Labour party a new leader of the Conservatives was elected. Interestingly, in the lead up to the general election, the proposed candidates for the Conservative Leadership election are more prominent in the visualisation as seen in **Table. 6** where the mean for number of supports and attacks for a politician is two. Many supports and attacks of the potential leaders hint at their impending desire to run for party leadership as a high number of either show that the potential leaders are more prominent in debate.

Potential Conservative Leaders	Supports	Attacks
William Hague	33	30
Ian Lang	17	20
Stephen Dorrell	22	10
Michael Howard	4	4
Peter Lilley	3	0
John Redwood	2	0
Kenneth Clarke	0	0
MEAN AVERAGE	2	2

Table 6. Supports and Attacks on ethos of Conservative Leader proposed candidates.

November 30th 1978 to January 20th 1979, 32 text transcripts focusing on a period of time in the UK known as the Winter of Discontent. In this period there were multiple strikes by workers in the UK, putting pressure on the then Labour Prime Minister James Callaghan [27]. This period was characterised by two significant changes in the

⁶Note that ethos analytics was run on a larger set than the EtHan.Thatcher corpus, because we used all transcripts from a given analysed period.

political landscape: first, the growth of mass infighting in the Labour party; and second, Margaret Thatcher becoming Prime Minister on May 4th 1979. These political dynamics are reflected in two ways in the ethos analytics. The Prime Minister James Callaghan had a total of eleven attacks on his ethos, where the mean across all MPs is one. Half of these were from Labour party members, reflecting the deep discontent at his leadership. The infighting which followed is also reflected by the ethos analytics. Shirley Williams, a Labour member at the time, has for example a total of eight attacks on her ethos and seventeen supports (only two come from other Labour members). In the years following the general election, and after the loss of Williams' seat, she became a founding member of the Social Democrat Party (SDP) [5].

5. Related work

Although, as far as we are aware, automatically extracting linguistic expressions of ethos has not previously been explored, ethos mining builds on methods and techniques developed for sentiment analysis and argument mining. The closest approaches to ours include the application of NLP techniques to the UK Hansard to build a database of claims associated with their parliamentary authors [18, 19]; the use of a lexicon based and classification approach in analysing sentiment of UK parliamentary debates [25]; and mining of arguments from the Canadian Hansard parliamentary record [17]. It is however, important to note that these works perform different tasks to ethos mining so the results are not directly comparable.

Sentiment analysis is the classification of documents, sentences or individual words as either positive or negative. Sentiment classification using machine learning can achieve over 80% accuracy [21] when performed on large feature vector sets using only unigrams as features. [12] describes feature-based sentiment analysis using a lexicon of sentiment bearing words to classify text. In [18], a system was developed to extract politicians' statements on specific topics in order to increase the accuracy of queries in UK Hansard. To do this, NLP techniques such as NER and POS tagging were used giving a satisfaction rating of 32% on an ordinal scale. In [19] the approach was extended by applying discourse sentiment analysis, with an accuracy of 44%. In [25], NLP techniques such as POS tagging were applied to parliamentary debates to obtain features for machine learning classifiers. These were then compared to two lexicon based sentiment approaches, an off-the-shelf lexicon approach, SentiWordNet 3.0 and a domain specific lexicon approach. When compared, the machine learning classifiers out performed the lexicon based sentiment approaches, with an accuracy of 61.75%.

Argument mining (also called argumentation mining, see e.g. [16, 23, 26, 30] for an overview) is the automatic extraction of argument from text over many different domains. In [15], legal text is broken down into sentences which then have features extracted. Sentences are then classified as either argumentative or non-argumentative with an accuracy of 68% for legal texts. In [11], claim detection is explored using NLP techniques to extract features of argument. Using an SVM with data extracted using parsing and POS tagging, a precision of 9.8 and a recall of 58.7 were achieved. In [17], a corpus of 138 sentences from Gay Marriage political debates was annotated by three coders with an inter-annotator agreement using weighted kappa of 0.54 for stance (a users stance on a sentence) and 0.46 for frames (pre-existing arguments which highlight an aspect of an argument), with 90% agreement on statements between at least two of three annotators.

An SVM classifier was trained using a bag of words approach, distributed word representations of stance and frames with similarity calculations and the stance of each statement, either pro- or con-, as a feature. Frames were then identified in political speeches with an overall accuracy of 68.9%.

6. Conclusion

Whilst ethos is well-recognised as a critical, load-bearing component in successful communication, it has attracted relatively little attention in AI, particularly with respect to the way in which it is made manifest in language. We have presented the first systematic treatment of ethos from a linguistically-oriented AI perspective, including a simple coding scheme applied to the UK parliamentary record, Hansard, resulting in an annotated corpus which is openly available. We have shown that a text analysis pipeline of hand-crafted domain specific rules, structural linguistic methods and supervised learning techniques, improved by our lexicon of ethotic words, can deliver strong performance on identifying expressions of ethos, when compared to related work in argument mining of political debates. It is important to note that results achieved in argument mining cannot give a definitive comparison due to the difference in logos and ethos. By aggregating the results into visualisation and analytics, it becomes possible to identify patterns in new, unannotated datasets. Indicative and exploratory analysis of historical records suggests that major events and trends in the political landscape are reflected in, or anticipated by, the ethos analytics of the parliamentary record. On the one hand, this opens up an exciting new research programme to understand the relationships between ethos-oriented linguistic interactions amongst politicians and the historical events with which they are associated; but on the other it also raises the intriguing possibility that such techniques may be able to link contemporary ethos dynamics with the political events they presage.

Acknowledgements

This research was supported in part by EPSRC in the UK under grant EP/M506497/1 and in part by the Polish National Science Centre under grant 2015/18/M/HS1/00620. We also thank Marcin Koszowy for providing second annotator analysis.

References

- [1] Aristotle. *On Rhetoric* (G. A. Kennedy, Trans.). New York: Oxford University Press., 1991.
- [2] K. Budzynska. Circularity in ethotic structures. *Synthese*, 190:3185–3207, 2012.
- [3] Chris Burnett, Timothy J. Norman, and Katia Sycara. Trust Decision-Making in Multi-Agent Systems. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence, IJCAI*, 2011.
- [4] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- [5] Liberal Democrats. Shirley Williams, 2016. http://www.libdems.org.uk/shirley_williams [Last Accessed: 02/02/16].
- [6] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL, 2005)*, pages 363–370, 2005.
- [7] Gov.uk. History of Sir John Major - GOV.UK, 2016. <https://www.gov.uk/government/history/past-prime-ministers/john-major> [Last Accessed: 02/02/16].

- [8] M. Hu and B. Liu. Mining and Summarising Customer Reviews. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, Washington, USA, 2004*.
- [9] M. Janier, J. Lawrence, and C. Reed. OVA+: An argument analysis interface. In *Computational Models of Argument (COMMA)*, 2014.
- [10] J. Lawrence, M. Janier, and C. Reed. Working with Open Argument Corpora. In *European Conference on Argumentation (ECA)*, 2015.
- [11] Marco. Lippi and Paolo. Torroni. Context-Independent Claim Detection for Argument Mining. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI*, 2015.
- [12] Bing Liu. Sentiment Analysis and Subjectivity. *Handbook of natural language processing*, 2:627–666, 2010.
- [13] Christopher Manning and Dan Klein. Optimization, maxent models, and conditional estimation without magic. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Tutorials-Volume 5*, pages 8–8. Association for Computational Linguistics, 2003.
- [14] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, 2014.
- [15] M.-F. Moens, E. Boiy, R.M. Palau, and C. Reed. Automatic Detection of Arguments in Legal Texts. In *Proceedings of the International Conference on AI and Law (ICAIL-2007)*, pages 225–230, 2007.
- [16] Marie-Francine Moens. Argumentation Mining: Where are we now, where do we want to be and how do we get there? In *FIRE '13 Proceedings of the 5th 2013 Forum on Information Retrieval Evaluation*, 2013.
- [17] N. Naderi and G. Hirst. Argumentation Mining in Parliamentary Discourse. *Paper presented at 15th Workshop on CMNA, Bertinoro, Italy*, 2015.
- [18] O. Onyimadu, K. Nakata, Y. Wang, T. Wilson, and K. Liu. Entity-Based Semantic Search on Conversational Transcripts Semantic. In *Takeda, H., Qu, Y., Mizoguchi, R. and Kitamura, Y. Semantic Technology. Lecture Notes in Computer Science (7774)*, pages 344–349. Springer, 2013.
- [19] O. Onyimadu, K. Nakata, T. Wilson, D. Macken, and K. Liu. Towards Sentiment Analysis on Parliamentary Debates in Hansard. In *Kim, W., Ding, Y. and Kim, H.-G. Semantic Technology. Lecture Notes in Computer Science (8388)*, pages 48–50. Springer, 2014.
- [20] Fabio Paglieri and Cristiano Castelfranchi. Trust, relevance, and arguments. *Argument & Computation*, 5(2-3):216–236, 2014.
- [21] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up? Sentiment classification using machine learning techniques. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 79–86, 2002.
- [22] Simon Parsons, Katie Atkinson, Zimi Li, Peter McBurney, Elizabeth Sklar, Munindar Singh, Karen Haigh, Karl Levitt, and Jeff Rowe. Argument schemes for reasoning about trust. *Argument & Computation*, 5(2-3):160–190, 2014.
- [23] Andreas Peldszus and Manfred Stede. From argument diagrams to argumentation mining in texts: a survey. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, 7(1):1–31, 2013.
- [24] I. Rahwan, F. Zablith, and C. Reed. Laying the Foundations for a World Wide Argument Web. *Artificial Intelligence*, 171:897–921, 2007.
- [25] Zaher Salah. *Machine learning and sentiment analysis approaches for the analysis of Parliamentary debates*. PhD thesis, University of Liverpool, 2014.
- [26] Jodi Schneider. An Informatics Perspective on Argumentation Mining. In *ArgNLP*, 2014.
- [27] Adam Taylor. Before Thatcher Came To Power, The UK Was Literally Covered In Gigantic Piles Of Garbage, 2013. <http://www.businessinsider.com/thatcher-and-the-winter-of-discontent-2013-4?IR=T> [Last Accessed: 02/02/16].
- [28] Kristina Toutanova, Dan Klein, Christopher Manning, and Yoram Singer. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. *Proceedings of HLT-NAACL*, pages 252–259, 2003.
- [29] Douglas N. Walton. *Fundamentals of Critical Argumentation*. Cambridge University Press, 2006.
- [30] Adam Wyner, Raquel Mochales-Palau, Marie-Francine Moens, and David Milward. Approaches to text mining arguments from legal cases. In *Semantic processing of legal texts*, pages 60–79. Springer Berlin Heidelberg, 2010.

Argumentation as Information Input: A Position Paper

Dov GABBAY^a, Michael GABBAY^b

^aBar Ilan University, King's College London, University of Luxembourg

^bUniversity of Cambridge

Abstract. Given a network (S, R) , with $R \subset S^2$, we view the nodes of S as containing information and view xRy as x transmitting information to y . We argue that such networks provide a more general account of attack and defense as well as being able to simulate the traditional Dung approach.

Keywords. argumentation, information input, bipolar networks

1. Background and Orientation

1.1. Introducing informational input

The traditional view of an argumentation system (S, R) is static. S is a non-empty set of atomic elements and R is a static subset of S^2 and we are looking for subsets E of S satisfying certain properties involving R . Although we view xRy as x ‘attacks’ y , we do not use the dynamic idea of x actually ‘sending’ something to y . The dynamic idea exists, e.g. when we view (S, R) as an Ecology, S as a set of species and R as a predator-prey relation, and the complete extensions are viewed as possible groups of species in equilibrium, see [18]. This view also works for S being a set of arguments, when xRy means that x is a counter argument to y , (in logic this would be x ‘proves’ $\neg y$, see [16,17]). In the informational view, we look at xRy differently, we take xRy to mean that x is actually sending information to y . This information might change y or even ‘kill’ it.

Let us explain this informational idea a bit more formally, but first we must present the traditional notion of what we shall refer to as “argumentation as attack”. There are two ways to present the semantics for argumentation as attack, the traditional set theoretical approach and the Caminada labelling approach. For the mapping connections between the two approaches see [12].

Let us briefly quote the traditional set theoretic approach:

1. We begin with a pair (S, R) where S is a nonempty set of points (arguments) and R is a binary relation on S (the “attack” relation).
2. Given (S, R) , a subset E of S is said to be conflict free if for no x, y in E do we have xRy .
3. E protects an element $a \in S$, if for every x such that xRa , there exists a $y \in E$ such that yRx holds.

4. E is admissible if E protects all of its elements.
5. E is a complete extension if E is admissible and contains every element which it protects.

Various different semantics (types of extensions) can be defined by identifying different properties of E . For example we might define that E is a *stable extension* if E is a complete extension and for each $y \notin E$ there exists $x \in E$ such that xRy or the grounded extension as the unique minimal extension or a preferred extension, being a maximal (with respect to set inclusion) complete extension. The above properties give rise to corresponding semantics (stable semantics, grounded semantics and preferred semantics).

We can also present the complete extensions of $A = (S, R)$, using the Caminada labelling approach, see [12].

A Caminada labelling of S is a function $\lambda : S \mapsto \{\text{in}, \text{out}, \text{und}\}$ such that the following holds

- (C1) $\lambda(x) = \text{in}$ if for all y attacking x , $\lambda(y) = \text{out}$.
- (C2) $\lambda(x) = \text{out}$ if for some y attacking x , $\lambda(y) = \text{in}$.
- (C3) $\lambda(x) = \text{und}$ if for all y attacking x , $\lambda(y) \neq \text{in}$, and for some z attacking x , $\lambda(z) = \text{und}$.

A consequence of (C1) is that if x is not attacked at all, then $\lambda(x) = \text{in}$.

Any Caminada labelling yields a complete extension and vice versa. Any $\{\text{in}, \text{out}\}$ Caminada labelling (i.e. with no “und” value) yields a stable extension and vice versa. Set theoretic minimality or maximality conditions on extensions E correspond to the respective conditions on the “in” parts of the corresponding Caminada labellings. See [12].

We now want to continue and introduce our ideas about argumentation as information input. It would be helpful to have three useful stories in mind.

Story 1, The Party: We are planning a party and we have a set S which is the maximal set of all relatives, friends, colleagues etc. who can be invited to the party. The problem is that some of them do not get along/hate some others. So we have a relation R , where xRy (which we might denote by $x \rightarrow y$) means that if x is invited y must not be invited. We get here a traditional argumentation network with attack relation R .

Story 2, The Debating society: We have a group of people S and we ask each x in S to express an opinion $\mathbf{f}(x)$ about the government. Such opinions can vary from wanting to pay less tax to wanting to accept more refugees. The opinions are made available to all members of S and some members x respond by sending more information to other members y because they think y does not see all the relevant information to his statement $\mathbf{f}(y)$; x does not necessary disagree with y or want to throw y out of the debate, x may merely want to give y more information. Let us denote this information input relation by xRy . Let us denote the information sent by x as $\tau(x)$, (for simplicity, let us assume x sends the same information to all the y such that xRy). We end up with a system (S, R, \mathbf{f}, τ) , and we now have a system of argumentation as information input.

Story 3, The security agency: We have a group of security agents involved in collecting information, say about possible terrorist threats. The relation xRy on S means that x reports to y . This story is different from the debating society Story 2 in that the relation R is substantially well-founded following the hierarchy of the agency. We may have an agent a and another agent b responsible for spying on a foreign country. The agent a

employs a local agent y , who in turn employs several other locals say $x_1 \dots x_n$. We note three properties of this information network:

1. R is fixed, and is external and independent of the information involved.
2. If $x_i R y$ and $y R a$ hold then y waits for all the information from all x_i to arrive, and then y processes it and only afterwards passes it on to a .
3. If we have $a R b$ and $b R a$, then this means that a and b cooperate and share information and we need to determine how they do that.

We now continue more formally with some key motivating examples. We shall abuse conceptual sensitivity and sometimes call the relation “attack” relation instead of “information input” relation.

Consider the simple network F_1 with domain $\{a, b, c\}$. In this network b attacks c and a attacks b . On standard Caminada labelling, a is “in” as it is free from attack, and so it negates the attack on c by b and so c is also free from attack. So we have one grounded extension, namely $a = \text{in}$, $b = \text{out}$ and $c = \text{in}$. The meaning of the attack relation above is taken basically as:

$$x \rightarrow y \text{ means that if } x \text{ is “in” then } y \text{ is “out”} \quad (*)$$

This meaning $(*)$ corresponds to Story 1: the party, it is set theoretical in its nature. We need to define a subset E of S satisfying certain conditions, and $(*)$ is one of them. There is no dynamics involved in the concept of the extension E , except possibly in the case when we give an algorithm for finding E , in which case there will be inductive steps by step “dynamics”. This “dynamics” is external to the argumentation conceptual framework. The $(*)$ interpretation can be formalised in a variety of logics, all meaning basically the above. It can also be instantiated/explained in a variety of ways, for example, in the case of Story 1, instantiating/explaining would mean that instead of just listing abstractly who cannot get along with whom, we can collect statements expressing the reasons for their not getting along and such statements can be used, again, in a manner based on the above. As another example, we may have, in a propositional logic with the language containing $\{\rightarrow, \wedge, \neg\}$, that $x = A \rightarrow e$ and $y = B \wedge \neg e \rightarrow d$ and so we have that $x \rightarrow y$, where the attack considerations are conducted in the specific logic of $\{\rightarrow, \wedge, \neg\}$, where we view y as an argument for d relying on $\neg e$ as an assumption and x is an argument for e , and therefore x attacks y .

In comparison, Stories 2 and 3 involve the transmission of information and not necessarily attack. The perceptive reader might ask what the connection is with argumentation. An information network is more like an electrical network/grid distributing electricity or a water network distributing water, and here we distribute information. What is the connection with argumentation?. The answer is twofold. On the one hand in many debates and arguments in many cases the response to an attacking argument is to give more information to deflect the attack. On the other hand, from the technical point of view, the idea of information input can also be used to actually attack. Our paper [1], suggested a different type of instantiation, using non-monotonic logic and in the context of non-monotonicity, information input can serve as attack. Let \sim be a non-monotonic consequence relation. The non-monotonicity property allows for the following for a theory Δ : $\Delta \sim A$, but $\Delta \cup \Delta' \not\sim A$. Thus if x is instantiated by a theory Δ_x and y by Δ_y , then we may have that $\Delta_y \sim A$ but $\Delta_x \cup \Delta_y \not\sim A$. So we may define the attack of x on y as the input of Δ_x into Δ_y , to form $\Delta_x \cup \Delta_y$ and thus cause A no longer to be derivable.

We thus have a new meaning for the “attack” relation:

$x \rightarrow y$ means that x adds some information $\tau(x)$ to the information of y (**)

Remark 1.1 Note the following about this “information attack” relation:

1. The attacker x and its target y can co-exist in the sense that x and y together can be consistent. In fact, x attacking y just changes y into a new y' . So the “attack” passing of information can even actually be “support”.
2. The traditional notions of re-instatement, admissibility, extensions etc, need no longer apply. We get a new game here.
3. Consider the attack of x on itself, $x \rightarrow x$. According to the $(*)$ reading, x wants itself to be out and so the Caminada labelling can only give x the value out or undecided, it cannot give x the value in. The reading $(**)$ on the other hand, lets x join its information to itself, which can be harmless, or can alter what is derivable via x , depending on the logic governing the information. If we are dealing with a resource logic, for example, such as linear logic, we do have $x \vdash x$ but not $x, x \vdash x$. Of course we can let x send its opposite $\neg x$, in which case x would be mounting a traditional attack on itself, but then x would be sending false information (from x 's point of view).
4. We can go further and generalise and understand $x \rightarrow y$ as $(***)$ below:

x sends an algorithm which revises y to a new y' . (***)

If y is information which yields a conclusion A , x can be additional information which now yields the conclusion $\neg A$, this is $(**)$. However, in practice, one (the supporter of x) may “tell” (the supporter of) y that he (the supporter of) y has not gathered the information correctly and actually y should be replaced by y' , and the conclusion is actually $\neg A$. This is $(***)$.

5. One surprising connection is with bipolar argumentation networks. See for example [13,14,15]. Bipolar networks have the form (S, R'', R') , where S is a non-empty set and R'' and R' are two disjoint binary relations on S . $xR''y$ means x attacks y (our notation, $x \rightarrow y$) and $xR'y$ means x supports y (notation, $x \rightarrow y$). The meaning of attack is the traditional one, as $(*)$ above. As for the meaning of support, there are various approaches almost all compatible with our informational Story 2 and Story 3. There is a lot of discussion and approaches in the literature on how to define extensions for networks with both attack and support. See for example [12,13]. The interest from our point of view is that the “information input” transmission can be either attack or support, depending on what information is being sent, and so we have an opportunity to connect and contribute to the bipolar debate.
6. We elaborate more on the connection with bipolar argumentation. Consider again the geometrical network of $S = \{a, b, c\}$ and $R = \{(a, b), (b, c)\}$. As informational network each element pair (x, y) in R might be attack or support. We cannot tell what it is, because it depends on the information sent from x to y , (in our example, from a to b and from b to c). So depending on the information being sent, we may have a traditional case of $aR''b$ and $bR''c$, or a case of $aR'b$ and $bR'c$. Now the challenge for us is to develop machinery for defining “informa-

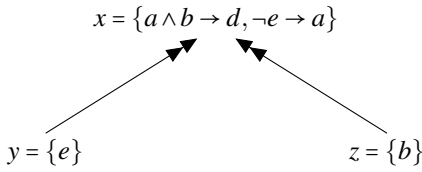


Figure 1. A logic programming example

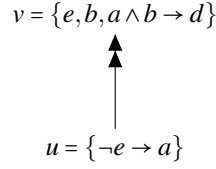


Figure 2. Another logic programming example

tional extensions” in the abstract for (S, R) in such a way that what we get in the abstract for the case of $aR''b$ and $bR''c$ will agree with the traditional approach to it, and what we get for the case of $aR''b$ and $bR'c$ will turn out to be the same as one of the known Bipolar approaches to it.

7. We note that we can show that informational “ R'' attack” networks (to be defined later) can simulate the traditional Dung networks. Given an abstract network (S, R) , we need some specific consequence relation \vdash and some specific initial theories (using S as being included in the atomic propositions of the language) and some specific correspondence theorems which will yield the connection. We use Logic Programming and the idea is illustrated in Example 1.2.

Example 1.2

1. Consider Figure 1

The nodes in Figure 1 are $S = \{x, y, z\}$ and the informational “attack” relation is $R = \{(y, x), (z, x)\}$. The information sets associated with each node in S (which by abuse of notation we also call x, y, z respectively) are logic programming databases where “ \neg ” is negation by failure, “ \wedge ” is conjunction, and “ \rightarrow ” is the logic programming implication. a, b, e and d are atoms. The database x can derive a , we write $x \vdash a$. If x gets attacked by y alone, then it gets the input e and so $\neg e$ no longer succeeds from $x \cup \{e\}$, and so $x \cup y$ cannot derive a . If x is attacked by z alone, then we get that x gets the input b alone so we have $x \cup z$, which derives d as well as e .

If x is attacked by both $y = \{e\}$ and $z = \{b\}$, then it becomes $x \cup y \cup z$ which cannot derive e and can derive just d .

2. So far Figure 1 gives us no more and no less than a geometrical network (S, R) of nodes with which information bases are associated as well as the informational “attacks” flow along R as described in item 1. above. If we want to talk about nodes being “in” or “out” or “undecided” in the Dung sense, we need to allow for a projection function which will give these values for each node. Let $\alpha(x) = a$, $\alpha(y) = b$ and $\alpha(z) = e$.

Define in general that any node u is “in” if the theory at u can prove $\alpha(u)$.

According to the projection defined by α , we have that both y and z are “in”, y supports x (because the information it sends to x strengthens it) while z attacks x .

3. To further our understanding, note that the databases $x' = \{\neg e \rightarrow a, a \wedge b \rightarrow d\}$ and $x = \{a, a \wedge b \rightarrow d\}$ are not the same, even though both derive $\{a, a \wedge b \rightarrow d\}$, because $x \cup \{e\}$ derives a while $x' \cup \{e\}$ does not derive a .

4. Consider now Figure 2. In this figure the node u attacks the node v . The question we ask is what information does u send to v ?
 On the one hand u can derive a . So if it sends $\{a\}$ (i.e. $\tau(u) = \{a\}$), it will enable $v \cup \tau(u)$ to derive d , because v derives $a \wedge b \rightarrow d$.
 But if it sends itself, then $v \cup u$ cannot derive a and so also cannot derive d .

2. Formal presentation of general informational networks

Let us now start with some technical definitions in the spirit of the above discussion.

Definition 2.1 Let \mathbf{L} be a propositional language for a logic. Assume this language contains connectives to define wffs which include classical conjunction and a negation symbol \neg . Also assume that the language has semantics which allows theories Θ (i.e. sets of wffs) to have models. In a model any wffs can be true, false or undecided. There are many possibilities for such logics and semantics, such as Logic Programming, Answer Set Semantics, Default Logics, Kraus-Lehman-Magidor semantics and many more. We use the logic to express our informational theories.

- Definition 2.2**
1. Let (S, R) be a network with $S = \emptyset$, $R \subseteq S \times S$. Assume a logic \mathbf{L} and semantics for it. Assume the elements of S do not appear in \mathbf{L} .
 2. Let \mathbf{f} be a function giving for each $t \in S$ a theory $\mathbf{f}(t) = \Delta_t$ of \mathbf{L} .
 Let $\tau(t)$ be another theory of \mathbf{L} , such that $\mathbf{f}(t) \supseteq \tau(t)$. Let $\alpha(t)$ be another functions on S giving for each t and atom $\alpha(t)$ of \mathbf{L} such that $t \neq s \Rightarrow \alpha(t) \neq \alpha(s)$
 3. We consider the system $(S, R, \mathbf{f}, \tau, \alpha)$ as an information system with a test function as follows:
 - (a) (S, R) is the information flow geometric system.
 - (b) $\mathbf{f}(t)$, $t \in S$, is the information stored at node t .
 - (c) $\tau(t)$ is the information which node t passes to any node s such that tRs holds.
 - (d) Note that the simplest τ is $\tau = \mathbf{f}$.
 - (e) $\alpha(t)$ is a projection function for any node t . Its role is to connect the information system with a possible bipolar argumentation network $\mathbb{B}(\alpha)$ based on (S, R) and defined using \mathbf{f} , τ and α .

Definition 2.3 Let $\mathcal{A} = (S, R, \mathbf{f}, \tau, \alpha)$ be as in Definition 2.2.

1. Let R^* be the reflexive and transitive closure of R .
2. Let $\mathbf{f}^*(t)$, for $t \in S$ be $\mathbf{f}^*(t) = \mathbf{f}(t) \cup \bigcup_{sR^*t} \tau(s)$
3. Let \mathbf{m} be a model of the theory $\Delta^* = \bigcup_{t \in S} \mathbf{f}^*(t)$.
4. Let $E_{\mathbf{m}} = \{t \mid \alpha(t) \text{ holds in } \mathbf{m}\}$.

We say that $E_{\mathbf{m}}$ is an informational extension of \mathcal{A} . The meaning of $E_{\mathbf{m}}$ depends on the underlying logic \mathbf{L} and its models \mathbf{m} . It may not be a Dung complete extension of (S, R) .

Remark 2.4 Note that the informational system concept is much more general than the traditional [23] Dung argumentation concept and its notion of complete extensions. In the full paper [20], we motivate our ideas and compare our concepts with existing related papers such as [21, 22]. We stress that α is not a central informational concept and

is used only to be able to define technically “informational extensions” and use them to show that informational attacks and support can simulate bipolar argumentation.

The following Definition 2.5 and Theorem 2.6 show that traditional argumentation machinery can be simulated by informational networks.

Definition 2.5 Let (S, R) be a finite network, we now define an associated informational network. The Language \mathbf{L} is the logic programming language based on the atoms S . The semantics is the Answer Set Semantics (see [24]).

1. For each $x \in S$ such that $\{y_1, \dots, y_n\}$ are all the attackers of x (i.e. $\{y_1, \dots, y_n\} = \{y \mid yRx\}$) let $C_x = \neg y_1 \wedge \dots \neg y_n \rightarrow x$. Let $\mathbf{f}(x) = \tau(x) = \{C_x\}$. Let $\alpha(x) = x$. Thus \mathbf{f} translates from S into the language \mathbf{LP} of Logic Programs (Recall that in this language, “ \neg ” is negation as failure, see Example 1.2).¹
2. Define Δ to be $\{C_x \mid x \in S\}$

Theorem 2.6 Let (S, R) be a finite network then the informational network of Definition 2.5 can simulate the traditional complete extensions for (S, R) . In fact we can show that this machinery can also deal with bipolar networks with attack and support, see [20].²

Proof. A very long proof is in [20]. We show that the informational extensions, which are in this case the same as all the answer set Programming (ASP) models of Δ of item 2 of Definition 2.5 are the same as all the Dung extensions of (S, R) . ■

3. Conclusion

We showed that information input networks exist in practical argumentation (see for example [2]) and can simulate traditional attack or support or explanation, all depending on the nature of the information. We need to study carefully and extensively how information input is used in real debate and arguments. This is the real challenge here. On the technical side we can compare with logic programming, ASPIC and Assumption Based Argumentation (ABA), but this is done in the full paper [20].

4. Acknowledgements

We thank the COMMA referees for their very valuable comments.

¹Note that this translation has two properties.

- (a) Each $x \in S$ has exactly one clause of which it is the head.
- (b) The elements in the body of each clause are all negated atoms.

²We are talking about bipolar networks (S, R'', R') , with R'' attack and R' support, where we understand support to behave according to the rules:

1. x is in, if all of its attackers are out and all of its supporters are in.
2. x is out, if one attacker is in or one supporter is out.
3. Otherwise x is und.

References

- [1] Gabbay, D. M., and A. Garcez, Logical modes of attack in argumentation networks, *Studia Logica*, 93(2-3): 199-230, 2009.
- [2] Gabbay, D. M., G. Rozenberg, Formal Modelling of Arguments from Expert Opinion, submitted 2016.
- [3] Gabbay, D. M., An Equational Approach to Argumentation Networks, *Argument and Computation*, 2012, vol 3 issues (2-3), pp 87-142
- [4] Wu, Y., Caminada, M., Gabbay, D., Complete extensions in argumentation coincide with 3-valued stable models in logic programming. *Studia Logica* 93(1-2), 383-403 (2009).
- [5] Martin Caminada, Samy Sá, Wolfgang Dvořák and João Alcântara, On the Equivalence between Logic Programming Semantics and Argumentation Semantics, *International Journal of Approximate Reasoning*, Volume 58, March 2015, Pages 87-111.
- [6] Greg Restall, *Introduction to substructural logics*, Routledge 2000
- [7] Pietro Baroni, Massimiliano Giacomin and Giovanni Guida, SCC-recursiveness: a general schema for argumentation semantics, *Artificial Intelligence*, Volume 168, Issues 1-2, October 2005, Pages 162-210
- [8] D. Gabbay, O. Rodrigues and J. Woods, Belief Contraction, Anti-formulas, and Resource Overdraft: Part I, *Logic Journal of the IGPL*, 10, 601-652, 2002.
- [9] D. Gabbay, O. Rodrigues and J. Woods, Belief Contraction, Anti-formulas, and Resource Overdraft: Part II, in *Logic, Epistemology and the Unity of Science*, D. M. Gabbay, S. Rahman, J. Symons and J-P van Bendegem, eds. pp. 291-326, Kluwer, 2004
- [10] Martin Caminada, Samy Sá and João Alcântara, On the Equivalence between Logic Programming Semantics and Argumentation Semantics, in L.C. van der Gaag (Ed.): *ECSQARU 2013*, LNAI 7958, pp. 97-108, 2013.
- [11] D M Gabbay, The handling of loops in argumentation networks, To appear in JLC special issue on Loops, *J Logic Computation*, first published online February 20, 2014 doi:10.1093/logcom/exu007 (83 pages)
- [12] M. Caminada and D. Gabbay. A logical account of formal argumentation. *Studia Logica*, 93:109, 2009.
- [13] C. Cayrol and M.C. Lagasque-Schiex, On the Acceptability of Arguments in Bipolar Argumentation Frameworks, L. Godo (Ed.): *ECSQARU 2005*, LNAI 3571, pp. 378-389, Springer, 2005.
- [14] Dov Gabbay, Logical foundations for bipolar argumentation networks, in special issue of the *Journal of Logic and Computation*, in honour of Arnon Avron, *J. Logic Computation* (2016) 26(1): 247-292. doi: 10.1093/logcom/ext027. First published online: July 22, 2013
- [15] L. Amgoud, C. Cayrol, M. C. Lagasque-Schiex, P. Livet, On Bipolarity in Argumentation Frameworks *International Journal Of Intelligent Systems*, Vol. 23, 1062-1093, Wiley 2008, Doi 10.1002/int.20307.
- [16] D Gabbay and M. Gabbay. The attack as strong negation, Part 1, *Logic Jnl IGPL* (2015) doi: 10.1093/jigpal/jzv033, first published online: September 28, 2015,
- [17] D Gabbay and M Gabbay The Attack as Intuitionistic Negation , to appear in *Logic Journal of IGPL*, Article ID: JIGPAL-jzw012
- [18] H. Barringer, D. M. Gabbay and J. Woods. Temporal dynamics of support and attack networks: from argumentation to zoology, in *Mechanizing Mathematical Reasoning*, pp. 59-98, 2005
- [19] Dunja Šešelja and Christian Straßer, Abstract argumentation and explanation applied to scientific debates, *Synthese* (2013) 190:2195-2217, DOI 10.1007/s11229-011-9964-y
- [20] D Gabbay and M Gabbay, Argumentation as information input, full paper draft.
- [21] Umberto Grandi, Emiliano Lorini and Laurent Perrussel. Propositional Opinion Diffusion. in *Proceedings of the 14th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2015)*, Bordini, Elkind, Weiss, Yolum (eds.), May 4-8, 2015, Istanbul, Turkey.
- [22] Nicolas Schwind, Katsumi Inoue, Gauvain Bourgne and Sebastian Konieczny. *Belief Revision Games*, 2015, Association for the Advancement of Artificial Intelligence (www.aaai.org).
- [23] Phan Minh Dung . On the acceptability of arguments and its fundamental role in non-monotonic reasoning, logic programming and ‘n-person games. *Artificial Intelligence*, Volume 77, Issue 2, September 1995, pages 321-357.
- [24] Gelfond, Michael. Answer sets. In van Harmelen, Frank; Lifschitz, Vladimir; Porter, Bruce. *Handbook of Knowledge Representation*. Elsevier, 2008, pp. 285-316.

Degrees of “in”, “out” and “undecided” in Argumentation Networks

Dov M GABBAY^a and Odinaldo RODRIGUES^{b,1}

^aKing’s College London, Department of Informatics, Ashkelon Academic College and Bar Ilan University, Israel and University of Luxembourg, Luxembourg

^bKing’s College London, Department of Informatics

Abstract. The traditional 3-valued semantics of an argumentation framework $\langle \mathcal{A}, \mathcal{R} \rangle$ identifies arguments that are “in”, “out” and “undecided”. Yet, it has long been recognised by the community that some elements can be at different degrees in each of these categories [1,2,3]. For example, Dung’s semantics can only classify some elements as “out”, but cannot reflect how much “out” they really are or if elements are “in” are they as much “in” as elements which are not attacked at all?

In this paper we shall use a numerical approach to give a measure of “in”, “out” and “undecided” to the nodes of a network. We shall devise equations which allow for solutions that reflect these distinctions.

Keywords. Numerical argumentation, degrees of acceptance, numerical methods

1. Introduction and Preliminary Discussion

Consider the situation depicted in the argumentation network of Figure 1 (L) where we have the set of arguments $\mathcal{A} = \{A, B, C, X, Y, W, Z\}$ with a relative complex geometrical configuration of attacks. In spite of that, all of the traditional argumentation semantics give the network one single extension, namely $\mathcal{E} = \{X, W, Z\}$. This single extension fails to capture a lot of the information in the network. For instance, it does not distinguish between the accepted argument Z , which has one attack, and the accepted arguments X and W which have no attackers and therefore are uncontroversially accepted. From \mathcal{E} and \mathcal{A} , we can also deduce that the nodes Y and A are rejected, but this also fails to capture the fact that A has three attackers (including itself) and therefore is arguably more rejected than Y . The statuses of B and C are undecided, but although B is more attacked than C , the semantics also fails to reflect that. All of these facts can clearly be seen from the geometry of the network, but the traditional three-valued semantics is too coarse to capture them.

Various papers have tried to consider the geometry by looking at a node and the nodes attacking it, and the attackers of these attackers, and so on, until it went back to the top of the network to somehow measure how strongly each node is “in”, “out” or “undecided”. Our own approach is *numerical* using equations describing the node interactions to be able to naturally reflect numerically these geometrical considerations. What this means in principle is that the object-level instrument of traditional extensions cannot be

¹Corresponding Author: Odinaldo Rodrigues, King’s College London, Department of Informatics; email: odinaldo.rodrigues@kcl.ac.uk

solely used to make the kind of distinctions we want to make about the nodes. We need to resort to external meta-level considerations. A recent brilliant numerical approach to tackle this problem was suggested in [3]. The authors however, do not connect with the traditional extensions. We notice that the *equational approach* for obtaining extensions does give numerical values between 0 and 1 for undecided nodes and these values reflect the degree of undecidedness of such nodes [4]. The approach does not distinguish however, between the nodes that are “in” and between the nodes that are “out”. The solution proposed here, which we call the *U-approach using Eq_{inv}* , is conceptually very simple. We make all nodes undecided by having an external additional self-attacking node U attack every node. Solving equations for this augmented network now gives us the degree of “in”, “out” as well as “undecided” whilst still connecting with the traditional extensions as we shall see later in the paper.

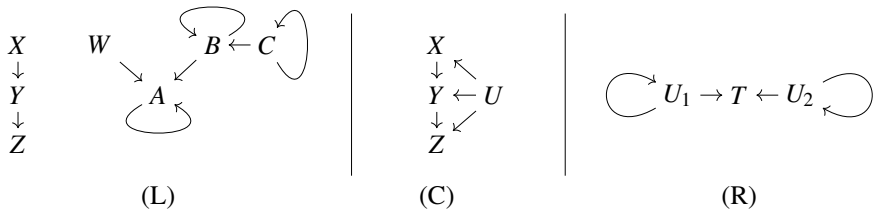


Figure 1. Sample argumentation networks used in the text.

To explain the main idea, consider the sub-network on the left of Figure 1 (L), which has the single extension $\mathcal{E}' = \{X, Z\}$. We would like to say that Z is more controversially accepted than X in \mathcal{E}' , because it is attacked by Y whereas X has no attackers. Adding a new self-attacking node U , which also attacks every other node in the sub-network gives the network in Figure 1 (C). We then write equations for each node, such as the ones that follow:² $U = 1 - U$; $X = 1 - U$; $Y = (1 - U)(1 - X)$; and $Z = (1 - U)(1 - Y)$.

From the figure and solution to the equations, we see that the value of $U = \frac{1}{2}$ is propagated to every node and moreover that the width and the depth of attacks is naturally reflected in the results, since the factor U is applied within each chain of attack: once to X , twice to Y (through X and through Y itself) and three times to Z (twice through Y and once from Z itself). These equations have the solution: $U = X = \frac{1}{2}$, $Y = \frac{1}{4}$, and $Z = \frac{3}{8}$. From this solution, we have $X > Z > Y$. In the context of the extension \mathcal{E}' , X has a higher value than Z and is therefore more “in” than Z . Y is attacked by a node in \mathcal{E}' and is “out”. Applying the same reasoning to the sub-network on the right of Figure 1 (L), we get $A = \frac{3}{19}$, $B = \frac{1}{4}$, $C = \frac{1}{3}$ and $W = \frac{1}{2}$. The reader will note that B and Y will get the same values. This is because the two weaker attackers of B are counterbalanced by the stronger attacker of Y . Geometrically they are indistinguishable, but a second meta-level criteria, such as the status with respect to the extension \mathcal{E} can be used. Looking back at \mathcal{E} for the whole network, we can now distinguish between the arguments in the categories “in”, “out” and “undecided” as follows:

$$\begin{array}{c} \text{More} \\ \updownarrow \\ \text{Less} \end{array} \quad \text{In} \left\{ \begin{array}{c} X, W = \frac{1}{2} \\ | \\ Z = \frac{3}{8} \end{array} \right. \quad \text{Undecided} \left\{ \begin{array}{c} C = \frac{1}{3} \\ | \\ B = \frac{1}{4} \end{array} \right. \quad \text{Out} \left\{ \begin{array}{c} Y = \frac{1}{4} \\ | \\ A = \frac{3}{19} \end{array} \right.$$

²How to arrive at these equations will be explained in detail later.

Note in the solution above that it makes sense for A to be more “out” than Y , because it has the three attackers W , B and itself, and at least one of these is as strong as X . So the calculations take into account the *number* of attackers as well as their *strength*.

The rest of the paper is structured as follows: Section 2 provides the background of the equational approach employed in this paper. Section 3 deals with U -approach using Eq_{inv} in detail. Section 4 compares our solution with the literature, and we conclude in Section 5.

2. Background

The *equational approach* views an argumentation network $\langle \mathcal{A}, \mathcal{R} \rangle$ as a mathematical graph generating equations for functions in the unit interval $[0, 1]$. Any solution f to these equations conceptually corresponds to an extension. Of course, the end result depends on how the equations are generated and we can get different solutions for different equations. Once the equations are fixed, the totality of the solutions to the system of equations is viewed as the totality of extensions via an appropriate mapping. Two equation schema we can possibly use for generating equations are Eq_{max} and Eq_{inv} below, where $f(X)$ is the value of a node $X \in \mathcal{A}$:

$$(Eq_{max}) \quad f(X) = 1 - \max_{Y \in Att(X)} \{f(Y)\} \qquad (Eq_{inv}) \quad f(X) = \prod_{Y \in Att(X)} (1 - f(Y))$$

It is easy to see that according to Eq_{max} the value of any source argument will be 1 (since they have no attackers) and the value of any argument with an attacker with value 1 will be 0. Gabbay has shown that in the case of Eq_{max} the totality of solutions to the system of equations corresponds to the totality of extensions in Dung’s sense. Let $\mathcal{N} = \langle \mathcal{A}, \mathcal{R} \rangle$ be an argumentation framework, the following two theorems, whose proofs can be found in [5], show the relationship between the solutions of Eq_{inv} and extensions of \mathcal{N} .

Theorem 2.1 (Theorem 2.2 in [5]) Every solution f of Eq_{inv} equations written for an argumentation framework \mathcal{N} yields a complete extension for \mathcal{N} .

Theorem 2.2 (Theorem 2.3 in [5]) Every preferred extension of an argumentation framework \mathcal{N} can be obtained from a solution f of Eq_{inv} equations written for \mathcal{N} .

In general terms, the following correspondence will relate a solution f with the traditional semantics: $f(X) = 1 :: X$ is “in”; $f(X) = 0 :: X$ is “out”; and $0 < f(X) < 1 :: X$ is “undecided”.

3. The U -approach using Eq_{inv}

We now ask what equation schema is more appropriate to capture the geometry of a network? Eq_{max} will disregard all but the attacks with maximum value, so it is not ideal. If we think in terms of *probability*, we want the values obtained as solutions to the equations to reflect the probability of being “in”. Thus, 1 is definitely “in” and 0 is definitely **not** “in”, i.e., definitely “out”. $\frac{1}{2}$ is right in the middle of “in” and “out” and hence means definitely “undecided”. Following this reasoning, now consider Figure 1 (R). We ask what is the probability that T is “in”? T is “in”, if both U_1 and U_2 are “out” (i.e., if both have value 0). So it is the product of the probability of each U_i being “out”, which is

the product $(1 - U_1) \times (1 - U_2)$. This motivates the use of the product operation in our computations (i.e., via Eq_{inv}).

An admirable discussion of these issues by Cayrol and Lagasquie-Schiex can be found in [6]. We would like to use Eq_{inv} in a way that responds to these intuitions, without resorting to all kinds of meta-level geometrical analyses and distinctions. However, if we simply use Eq_{inv} we would not be able to distinguish the nodes in the categories “in” and “out”,³ so our idea is to make every node “undecided” in a uniform way thus allowing for a larger spread of values in all categories.

So given $\langle \mathcal{A}, \mathcal{R} \rangle$, we move to $\langle \mathcal{A}_U, \mathcal{R}_U \rangle$ with a new node U making all nodes in \mathcal{A} “undecided”. The relative values in solving Eq_{inv} for $\langle \mathcal{A}_U, \mathcal{R}_U \rangle$ give the relative strength of all nodes in $\langle \mathcal{A}, \mathcal{R} \rangle$.⁴ So for any extension \mathcal{E} of \mathcal{A} and any $X, Y \in \mathcal{E}$ both are “in” in $\langle \mathcal{A}, \mathcal{R} \rangle$ but they are undecided in $\langle \mathcal{A}_U, \mathcal{R}_U \rangle$ and may have different values in a solution \mathbf{f} to the Eq_{inv} equations of $\langle \mathcal{A}_U, \mathcal{R}_U \rangle$. These different values will give us an indication of how much “in” X, Y are in \mathcal{E} (and similarly for nodes that are “out” or “undecided”).

Definition 3.1 Let $f : \mathcal{A} \mapsto [0, 1]$ be an assignment of values to elements of \mathcal{A} . We define the sets $in(f) = \{X \in \mathcal{A} \mid f(X) = 1\}$ and $out(f) = \{X \in \mathcal{A} \mid f(X) = 0\}$.

Definition 3.2 (U-Augmentation of an Argumentation Framework) The U -augmentation of the argumentation network $\langle \mathcal{A}, \mathcal{R} \rangle$ is the network $\langle \mathcal{A}_U, \mathcal{R}_U \rangle$, where $U \notin \mathcal{A}$, $\mathcal{A}_U = \mathcal{A} \cup \{U\}$ and $\mathcal{R}_U = \mathcal{R} \cup \{(U, U)\} \cup \{U\} \times \mathcal{A}$.

The relative degree of membership of each node in the categories “in”, “out” and “undecided” is defined as follows.

Definition 3.3 (Numerical evaluation of the degree of “in”, “out”, and “undecided” in abstract argumentation frameworks) Let $\mathcal{N} = \langle \mathcal{A}, \mathcal{R} \rangle$ be a network and \mathcal{N}_U its U -augmentation. Let \mathcal{E} be an extension for \mathcal{N} . Let \mathbf{f} be a solution to the Eq_{inv} equations for \mathcal{N}_U . Let \leq_f be an ordering on \mathcal{A} defined by $X \leq_f Y$ iff $\mathbf{f}(X) \leq \mathbf{f}(Y)$. Then \leq induces an ordering on the sets $IN = \mathcal{E}$ (“in”); $OUT = \{X \in \mathcal{A} \mid \exists Y \in \mathcal{E} \text{ such that } (Y, X) \in \mathcal{R}\}$ (“out”); and $UND = \mathcal{A} \setminus (\mathcal{E} \cup OUT)$ (“undecided”), giving a degree scale in each category.

Note that Definition 3.3 offers a geometrical ranking of the nodes in a network (\leq_f) which is independent of the notion of extension but can be used in conjunction with an extension to distinguish the nodes in the categories “in”, “out” and “undecided” with respect to that extension.

Remark 3.1 It should be clear that if the network $\mathcal{N} = \langle \mathcal{A}, \mathcal{R} \rangle$ is acyclic, then a solution to the system of Eq_{inv} equations to the U -augmentation of \mathcal{N} exists and is unique. To see this, we simply order the equations in ascending order of the longest chain of attack of each node and solve them in this order. U will solve to $\frac{1}{2}$ as well as every source node in \mathcal{N} . We then propagate these values until all node values are calculated. This will form the unique solution \mathbf{f} .

³In light of Theorem 2.2, all nodes in a preferred extension have value 1 and all nodes attacked by the extension have value 0.

⁴As nicely put by one of the reviewers of this paper, this hypothetical node U could represent an unforeseen future argument attacking all nodes.

In the general case, we believe the solution \mathbf{f} of Definition 3.3 exists and is unique. Empirical results also suggest this. The proof would be similar to the proof of uniqueness in [3], but it has to be written down fully to confirm. Note that because of the introduction of the new node U , all equations involved become contractions.

We now show a number of properties of the solutions to the system of equations. The first one has to do with the upper bound of the value of a node and the second one with the effect of attacks on nodes.

Proposition 3.1 *Let $\mathcal{N} = \langle \mathcal{A}, \mathcal{R} \rangle$ be an argumentation network and \mathbf{f} a solution to the Eq_{inv} equations written for the U -augmentation of \mathcal{N} . For all $A \in \mathcal{A}$, $\mathbf{f}(A) \leq \frac{1}{2}$.*

The above proposition shows that 1) $\frac{1}{2}$ is the upper bound for the values of an U -augmented network; and 2) source nodes get maximum value, i.e., $\frac{1}{2}$. It is also easy to see that the values of nodes decrease proportionally to the number of attackers. In particular:

Proposition 3.2 *Let $\mathcal{N} = \langle \mathcal{A}, \mathcal{R} \rangle$ be an argumentation network and \mathbf{f} a solution to the Eq_{inv} equations written for the U -augmentation of \mathcal{N} . Let $X, Y \in \mathcal{A}$, such that $Att(X) \subseteq Att(Y)$, then $\mathbf{f}(Y) \leq \mathbf{f}(X)$.*

Proposition 3.3 *Let $\mathcal{N} = \langle \mathcal{A}, \mathcal{R} \rangle$ be a finite argumentation network and \mathbf{f} a solution to the Eq_{inv} equations written for the U -augmentation of \mathcal{N} . Take $X \in \mathcal{A}$ and let $|Att(X)| = k$. Then $\mathbf{f}(X) \geq \frac{1}{2^{k+1}}$.*

4. Comparisons with Other Work

We start our comparisons with the approach in [3] in which the relative strengths of arguments in a graph are indirectly calculated in terms of the relative *burden number* of these arguments. Essentially, this technique assigns a unique rank to every node which can be compared with the relative ranking our geometrical interpretation gives:

Definition 4.1 (s_α , [3]) *Let $\alpha \in (0, +\infty)$ and $\mathcal{F} = \langle \mathcal{A}, \mathcal{R} \rangle$ be an argumentation graph. We define the function s_α as follows: $s_\alpha : \mathcal{A} \mapsto [1, +\infty)$ such that $\forall a \in \mathcal{A}$,*

$$s_\alpha(a) = 1 + \left(\sum_{b \in Att(a)} \frac{1}{(s_\alpha(b))^\alpha} \right)^{\frac{1}{\alpha}}$$

If $Att(a) = \emptyset$, then $s_\alpha(a) = 1$. $s_\alpha(a)$ is called the burden number of a .

It is easy to see that, as is the case in our technique, the value of s_α takes into account both the number of attackers of an argument as well as the relative strength of these attackers. An argument with a small burden number is deemed more acceptable than an argument with a greater burden providing what was called a *compensation-based semantics*.

In order to compare the results we will use the sample networks in Figure 2 taken from [3]. The computed s_α values for $\alpha = 1$ of all nodes in the networks are given in Table 1 along with the solutions for the U -augmentation of the networks.

It is easy to see that although the values differ, the rankings of arguments in networks \mathbf{N}_1 , \mathbf{N}_2 , \mathbf{N}_3 and \mathbf{N}_5 are exactly the same. However, in network \mathbf{N}_4 , compensation-based semantics fails to distinguish between arguments P , Q and A when $\alpha = 1$, whereas our formalism considers P and Q equivalent but strictly weaker than A (remember P are Q

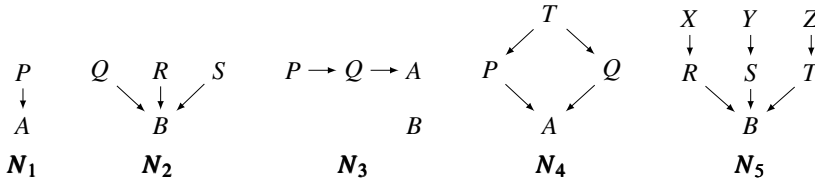


Figure 2. Some networks taken from [3] for comparison.

N_1 :	P	A	N_2 :	Q	R	S	B	N_3 :	P	B	A	Q
$\alpha = 1$	1	2		1	1	1	4		1	1	$\frac{3}{2}$	2
$U(\frac{1}{2})$	$\frac{1}{2}$	$\frac{1}{4}$		$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{16}$		$\frac{1}{2}$	$\frac{1}{2}$	$\frac{3}{8}$	$\frac{1}{4}$
N_4 :	T	A	P	Q	N_5 :	X	Y	Z	R	S	T	B
$\alpha = 1$	1	2	2	2		1	1	1	2	2	2	$\frac{5}{2}$
$U(\frac{1}{2})$	$\frac{1}{2}$	$\frac{9}{32}$	$\frac{1}{4}$	$\frac{1}{4}$		$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{27}{128}$

Table 1. Summary of comparison with compensation-based semantics rankings [3].

are both “out” whereas A is “in” in any traditional semantics). We argue that in N_4 A should be more acceptable than both P and Q , because although it has two attackers, these are both defeated by T . Note that in theirs and ours we still obtain that A is less acceptable than T itself, as expected. In general, we get the following values for the nodes in N_4 in the compensation-based semantics: $s_\alpha(T) = 1$, $s_\alpha(P) = s_\alpha(Q) = 2$ (for $\alpha \geq 1$); and $s_\alpha(A) = 2$, if $\alpha = 1$; and $1 \leq s_\alpha(A) < 2$, if $\alpha > 1$. This shows that within the single network N_4 we can get different rankings for A , P and Q depending on the value of α used. The idea behind this is to fine-tune “the influence of the quality of the attackers”. This has two main problems. 1) It is difficult to know *in advance* which value of α to use. Consider the network N_6 formed by the aggregation of N_1 and N_5 (this is network \mathcal{F}_4 in [3]). In N_6 as well, the relative acceptability ranking between A and B will vary depending on the value of α chosen.⁵ 2) Fixing the relative ranking of some nodes by employing a certain value of α may inadvertently cause the ranking of other nodes to change. In N_4 , if $\alpha > 1$, then again our formalism and theirs will agree on the ranking of all arguments in N_4 except in the limit $\alpha \rightarrow \infty$. We argue that it is simpler to adjust the impact attacks have on the values of the nodes according to the application via the equational approach, because we can separate out the necessary components via an appropriate equation schema. In the case of Eq_{inv} there are two separate components dealing with attacks: the complement-to-1 function (for attack itself) and product (for their aggregation). More sophisticated t -norms can be used instead of product. In Definition 4.1 the two components are intertwined.

In [6], Cayrol and Lagasque-Schiex propose two approaches to evaluate the value of an argument based solely on the attack relation of the argumentation framework to which it belongs. The first approach calculates the value of an argument using only the values of its direct attackers and is called *local*; the second approach takes into account the set of all ancestors of the argument in the attack relation and is called *global*.⁶

In the local approach, the value $v(X)$ of an argument X is obtained via the composition of two functions h and g : h calculates the value of each attacker of X ; and g com-

⁵For $\alpha \approx 1.585$, $s_\alpha(A) = s_\alpha(B)$.

⁶As it turns out, the values of the attackers of an argument are calculated recursively so in effect the local approach also takes into account all ancestors of an argument.

putes the effect on X of the aggregation of the attacks on it. The values of the nodes as calculated by the schema Eq_{inv} can be seen as a *local valuation* of the nodes in the same way as Besnard and Hunter’s h -categoriser valuation [7]. In our case, the function h is the complement to 1 and g is half of the product of the attacks.

In [8], Modgil and Grossi proposed a framework to take into account the degree of justification of the arguments of an argumentation framework. The central idea is the notion of *graded defense* which counts the number of attackers and defenders of a node. This is used to define *graded extensions*, which are parametrised by two integers m for attacks and n for defenders, and in essence select arguments with a particular configuration of attackers and defenders. So the motivation is the same, but the approach is completely different to ours.

Finally, in [9], Thimm and Kern-Isberner propose a *stratified semantics*, which assigns different ranks for the arguments of a network. The ranks are constructed by successively taking the accepted arguments of a network according to a given semantics, assigning them a rank, then considering the network resulting from the removal of such nodes and then re-calculating the nodes in the next rank until all nodes are ranked. This fails to distinguish between the accepted arguments in each rank, but agrees in spirit with our treatment of extensions in that it follows the directionality of attacks.

5. Conclusions, Discussion and Future Work

Given a network $\langle \mathcal{A}, \mathcal{R} \rangle$ and a complete extension $\mathcal{E} \subseteq \mathcal{A}$, our objective was to provide a ranking of the arguments in \mathcal{A} given \mathcal{E} taking into account the geometry of $\langle \mathcal{A}, \mathcal{R} \rangle$.

We offered a solution to the following meta-level problem Π : given an argumentation network $\langle \mathcal{A}, \mathcal{R} \rangle$, we note that some nodes are geometrically attacked more than others or are more “loopy” than others. Can we make this observation more quantitative? Our solution used the *equational approach* in an augmented network with set of nodes $\mathcal{A}^* = \mathcal{A} \cup \{U\}$.

We now discuss the methodological aspect of our solution. The problem Π above is a special case of a general problem: given an object-level system \mathcal{S} and a meta-level property \mathbb{P} of \mathcal{S} how can we express/discuss/quantify this property for \mathcal{S} ?

There are two ways: 1) Construct a meta-level system $\mathbb{P}(\mathcal{S})$ to describe/discuss \mathbb{P} ; and 2) Construct a new system \mathcal{S}^* out of \mathcal{S} , and within \mathcal{S}^* , the property \mathbb{P} can be highlighted. Method 2) is better for the following reasons. It is simpler, using the same machinery used in \mathcal{S} . It is also more robust. If we modify, generalise or apply \mathcal{S} , we do the same for \mathcal{S}^* and thus carry the results for the property \mathbb{P} . If we use $\mathbb{P}(\mathcal{S})$ we may not know what to do for $\mathbb{P}(\mathcal{S}^*)$.

Our approach followed method 2) above. \mathcal{S}^* is simply our U -augmented network and we use for \mathcal{S}^* the same machinery for finding extensions that we use for \mathcal{S} .

Bearing the above considerations in mind, let us summarise what we did. We know that the solutions to classes of equations written for an argumentation framework have a correspondence with the set of extensions of that network. In the case of the equation schema Eq_{max} the totality of the solutions to the system of equations corresponds to the totality of complete extensions of the network. In the case of the schema Eq_{inv} , the solutions only yield preferred extensions. However, for the quantitative measurement of attack we wanted to consider in this paper, Eq_{inv} has a significant advantage over Eq_{max} , because Eq_{max} simply takes the maximum value of the attacks, whereas Eq_{inv} provides a number reflecting the effect of the aggregation of all attacks on a node.

Because of Eq_{inv} 's correspondence with the class of preferred extensions, we cannot simply use it directly, since it will neither differentiate between the nodes in a preferred extension (i.e., the ones with value 1) nor will it differentiate between the nodes attacked by a node in the extension (i.e., the ones with value 0). Eq_{inv} will however differentiate between the nodes in the “undecided” range. So our simple solution is to force all nodes into the “undecided” range and then use the relative ranking of the nodes thus obtained to distinguish between the nodes given the *original* extension. Conceptually, this can be done simply by considering a modified network with a new undecided node attacking all original nodes. Mathematically, this can be seen as yielding a new schema of equations, call it, Eq_{deg} such that the value of a node X is defined as $f(X) = \varepsilon \times \prod_{Y \in Att(X)} (1 - f(Y))$.

We took ε to be $\frac{1}{2}$. This is not simply the same as multiplying a solution to Eq_{inv} by $\varepsilon = \frac{1}{2}$. This is indeed a new class of equations. The reader might then ask why $\varepsilon = \frac{1}{2}$? Conceptually, it makes sense to use $\frac{1}{2}$ as it is arguably the most “undecided” value and given its connection with the U -augmentation it is what a node with a single self-attack resolves to. Some further comparisons with other work on ranking semantics (e.g., [10,11,12]) as well as the effects of using different values of $\varepsilon \in (0, 1)$ is left for a full version of this paper.

References

- [1] C. da C. Pereira, A. G. B. Tettamanzi, and S. Villata. Changing one's mind: erase or rewind? possibilistic belief revision with fuzzy argumentation based on trust. In *Proceedings of the 22nd International joint conference on artificial intelligence : IJCAI'11*, pages 164 – 171, Menlo Park, 2011. AAAI Press.
- [2] P. Dondio. Multi-valued argumentation frameworks. In Antonis Bikakis, Paul Fodor, and Dumitru Roman, editors, *Rules on the Web. From Theory to Applications*, volume 8620 of *Lecture Notes in Computer Science*, pages 142–156. Springer International Publishing, 2014.
- [3] L. Amgoud, J. Ben-Naim, D. Doder, and S. Vesic. Ranking arguments with compensation-based semantics. In *Proceedings of the 15th International Conference on Principles of Knowledge Representation and Reasoning, KR 2016*, pages 12–21, 2016.
- [4] D. M. Gabbay. Equational approach to argumentation networks. *Argument and Computation*, 3:87–142, 2012. DOI: 10.1080/19462166.2012.704398.
- [5] D. M. Gabbay and O. Rodrigues. Probabilistic argumentation: An equational approach. *Logica Universalis*, 9(3):345–382, 2015.
- [6] C. Cayrol and M.-C. Lagasque-Schiex. Graduality in argumentation. *Journal of Artificial Intelligence Research*, 23:245–297, 2005.
- [7] P. Besnard and A. Hunter. A logic-based theory of deductive arguments. *Artificial Intelligence*, 128(1-2):203 – 235, 2001.
- [8] D. Grossi and S. Modgil. On the graded acceptability of arguments. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015*, pages 868–874, 2015.
- [9] M. Thimm and G. Kern-Isberner. On controversiality of arguments and stratified labelings. In *Proceedings of the Fifth International Conference on Computational Models of Argumentation (COMMA'14)*, September 2014.
- [10] E. Bonzon, J. Delobelle, S. Konieczny, and N. Maudet. A comparative study of ranking-based semantics for abstract argumentation. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA.*, pages 914–920, 2016.
- [11] L. Amgoud and J. Ben-Naim. Ranking-based semantics for argumentation frameworks. In *Scalable Uncertainty Management - 7th International Conference, SUM 2013, Washington, DC, USA, September 16-18, 2013. Proceedings*, pages 134–147, 2013.
- [12] P.-A. Matt and F. Toni. A game-theoretic measure of argument strength for abstract argumentation. In *Logics in Artificial Intelligence, 11th European Conference, JELIA 2008, Dresden, Germany, September 28 - October 1, 2008. Proceedings*, pages 285–297, 2008.

Formalizing Balancing Arguments

Thomas F. GORDON ^{a,1}, and Douglas WALTON ^b

^a*Fraunhofer FOKUS, Berlin, Germany*

^b*University of Windsor, Windsor, Canada*

Abstract. Dung intended his abstract argument frameworks to be used for modeling a particular form of human argumentation, where arguments attack each other and are evaluated following the principle summarized by “The one who has the last word laughs best.” However this form does not fit a wide class of arguments, which is arguably more prototypical and common in human argumentation, namely arguments where pros and cons are balanced to choose among alternative options. Here we present a formal model of structured argument which generalizes Dung abstract argumentation frameworks to also handle balancing. Unlike most other models of structured argument, this model does not map structured arguments to abstract arguments. Rather it generalizes abstract argumentation frameworks, allowing them to be simulated using structured arguments. The model can handle cumulative arguments (“accrual”) without causing an exponential blowup in the number of arguments and has been fully implemented in Version 4 of the Carneades Argumentation System.

Keywords. structured argumentation, argument evaluation, argument accrual, cumulative arguments, balancing arguments

1. Introduction

A wide class of human argumentation involves the balancing of pros and cons to choose among a set of options:

- Practical argumentation to choose a course of action involves balancing the costs and benefits of alternative actions, taking into consideration multiple-criteria, in addition to arguing about the preconditions and effects of the actions. Value-based practical reasoning [2] evaluates costs and benefits relative to particular audiences, in terms of the degree to which each action promotes or demotes some value. Arguing about governmental policies is of this type [9].
- Theoretical argumentation, both in natural science and in the humanities, including law, involves constructing, comparing and choosing among alternative theories, taking into consideration multiple evaluation criteria, such as the extent to which the theories explain the evidence or, in the law, precedent cases, and their simplicity, in line with Occam’s razor, among other factors, to choose the most coherent theory.

¹Corresponding Author: Prof. Dr. Thomas F. Gordon, Fraunhofer FOKUS, Kaiserin-Augusta-Allee 31, Berlin, Germany; E-mail: thomas.gordon@fokus.fraunhofer.de.

- Factual argumentation about whether or not some event occurred, involves the acquisition and weighing of evidential arguments, such as witness testimony, as well as the construction, comparison and balancing of the properties of alternative narratives (“stories”), to choose the most coherent narrative [4].
- Arguing about whether some concrete fact situation can or should be subsumed under some abstract concept, for example whether or not users have a “reasonable expectation of privacy” regarding the personal data on their smart phones, which would then be protected from “unreasonable searches and seizures” by the Fourth Amendment of the US Constitution. This can involve balancing different methods of constitutional interpretation (e.g. literal, historical and teleological). Moreover teleological interpretation can involve the balancing of interests in a way which respects the balance (“proportionality”) set by the founders in the constitution.

The leading computational model of argument, abstract argumentation frameworks [8], was not designed to handle balancing arguments, but rather another form of argumentation, where arguments are viewed as attacking arguments and evaluated according to the following principle:

The goal of this paper is to give a scientific account of the basic principle “The one who has the last word laughs best” of argumentation, and to explore possible ways for implementing this principle on computers. [8, pg 322]

Most of the leading models of structured argumentation [3,12,17] are defined as pre-processors for an argument evaluator for abstract argumentation frameworks, following the methodology proposed by Dung in [8, pg 348]:

Any argumentation system is composed of two essential components: One for generating the arguments together with the attack-relationship between them. The other is for determining the acceptability of arguments. So we can think of an argumentation system as consisting of two units, an argument generation unit, AGU, and an argument processing unit, APU.

Dung illustrated this process with a pipeline, where the AGU generates an argumentation framework (arguments and attacks) and the APU then evaluates this framework to determine which of these arguments are acceptable. In practice, structured models of argument have extended this pipeline with an additional process at the end, for labeling the *statements* (propositions) in the structured model of argument. Typically, a statement is acceptable (in) if and only if it is supported by an acceptable argument.

The linearity of this pipeline presents a problem when one wants to model balancing arguments, since the value (weight) of a balancing argument can depend on the acceptability of statements and, recursively, the acceptability of statements depends, as in the extended Dung methodology, on the value the arguments supporting them. When balancing, pro and con arguments are weighed against each other. An out premise can reduce (or increase!) the weight of an argument, without defeating it completely.

It is unclear whether it would be possible in principle to define a model of structured argumentation for balancing arguments using Dung’s pipeline methodology. Since there are no limits on the structure of arguments or the attack relation used to instantiate the abstract framework, some clever encoding of balancing arguments may be possible, but one has to wonder how straightforward or intuitive such a model of balancing arguments could be. In this paper we take a more direct, requirements-driven approach to modeling

structured argumentation with support for both attacks and balancing, which preserves the recursivity of the balancing process, without worrying about trying to find some way to model this process as a linear pipeline to comply with Dung's methodology.

2. The Formal Model

2.1. Structure

Let \mathcal{L} be a logical language for expressing statements (propositions). As in ASPIC+ [14], this model is a "framework". It can be instantiated with any logical language.

An *argumentation scheme* is an abstract structure in this framework providing functions for generating, validating and weighing arguments. The framework can be instantiated with various models of argumentation schemes. For our purpose here of modeling balancing arguments, only the weighing functions of argumentation schemes are relevant. See Definition 5 for the signature and further details of weighing functions.

Definition 1 (Argument) *An argument is a tuple (s, P, c, u) , where:*

- s is the scheme instantiated by the argument
- P , the premises of the argument, is a finite subset of \mathcal{L}
- c , a member of \mathcal{L} , is the conclusion of the argument, and
- u , a member of \mathcal{L} , is the undercutter of the argument.

This model of argument closely fits the usual conception of an argument in informal logic and argumentation theory in philosophy [18]. Notice that an argument here, unlike in ASPIC+, is not a complete proof tree, but rather only a single inference step in such a proof tree. Undercutters here are modeled in the same way as in ASPIC+, with a proposition in \mathcal{L} for each undercutter. In practice, these propositions will typically be constructed by applying some predicate to a term naming the argument, such as $\text{undercut}(a_1)$. But this is a detail to be worked out when instantiating the framework. We also call arguments undercutters which have undercutter statements as their conclusion. Notice that the argument includes a reference to the scheme used to construct (or reconstruct) the argument. This will be used to weigh the argument.

Example 1 *Following the tradition of [5], let us use as our running example a practical reasoning task about choosing a car to buy. Let us assume that a domain-dependent argumentation scheme for car buying has been defined, where the premises express the claimed properties of a particular car, one for each of the criteria to be considered, and the weighing function of the scheme computes a weighed sum of the proven (not claimed) properties of the car, where the weight assigned to each property by the scheme is chosen to reflect the relative importance of the criterion, relative to the other criterion, in the manner of multi-criteria decision analysis. Here is an example of an argument for a particular auto, applying this scheme:*

Let $a_1 = (s, P, c, u)$ be an argument for buying a Porsche, where:

- s is a car buying scheme, described in more detail in Example 4
- P , the premises, are:
 1. $\text{type}(\text{porsche}, \text{sports})$

2. $price(porsche, high)$
 3. $safety(porsche, medium)$
 4. $speed(porsche, fast)$
- c , the conclusion, is $buy(porsche)$, and
 - u , the undercutter, is $undercut(a_5)$

Definition 2 (Issue) An issue is a tuple (O, f) , where:

- O , the options (also called positions) of the issue, is a finite subset of \mathcal{L} .
- f , the proof standard of the issue, is a function which tests whether an option satisfies the standard. See Definition 6.

Issues are inspired by Issue-Based Information Systems (IBIS) [11]. They extend the concept of a “contrary” in the ASPIC+ model of structured argument, from a binary relation to an n -ary relation. Allowing more than two options is important for two reasons:

1. To allow more than two alternative options in deliberation dialogues and other decision-making contexts.
2. To avoid false dilemmas, by allowing alternatives other than true or false (or yes or no) for issues representing questions of the kind “Have you stopped beating your spouse?”.

Proof standards of issues are borrowed from the 2007 version of Carneades [10]. Associating proof standards with issues is designed to assure that the same proof standard applies to every position of the issue.

Definition 3 (Argument Graph) An argument graph is a tuple (S, A, I, R) , where:

- S , the statements of the argument graph, is a finite subset of \mathcal{L} .
- A , the assumptions, is a subset of S assumed to be provable.
- I , the issues of the argument graph, is a finite set of issues, where every position of every issue is a member of S and no $s \in S$ is a position of more than one $i \in I$, and
- R , the arguments of the argument graph, is a finite set of arguments, where all conclusions, premises and undercutters are members of S .

These structures are called graphs for historical reasons. Admittedly this a bit of an abuse of terminology. But every argument graph (S, A, I, R) can be easily mapped to a directed graph (V, E) as follows:

- The vertices, V , of the graph consist of the statements (S), issues (I) and arguments (R) of the argument graph.
- The edges, E , of the graph are constructed by linking arguments in A to their premises, conclusions and undercutters in S , and issues in I to their options in S , in the obvious way.

In most other models of structured argument, argument graphs for structured arguments are not formally defined. In [3], Besnard and Hunter use the term “argument graph” as a synonym for abstract argumentation frameworks. In ASPIC+ arguments are proof trees. Sets of such arguments are often visualized in ASPIC+ presentations as an

argument graph, where each argument is a subgraph of the argument graph, but the argument graph per se is not a part of the formal ASPIC+ model.

Example 2 Figure 1 shows an argument graph for the car buying example, with an argument for buying a Porsche and another argument for buying a Volvo. The labels of the statement nodes, displayed with colors, and arguments, displayed as numbers (weights) on the edges from the arguments to their conclusions, are explained in Section 2.2. The proof standard “PE” used by both issues, means “preponderance of the evidence” and is also defined in Section 2.2. Undercutters are visualized with dashed edges.

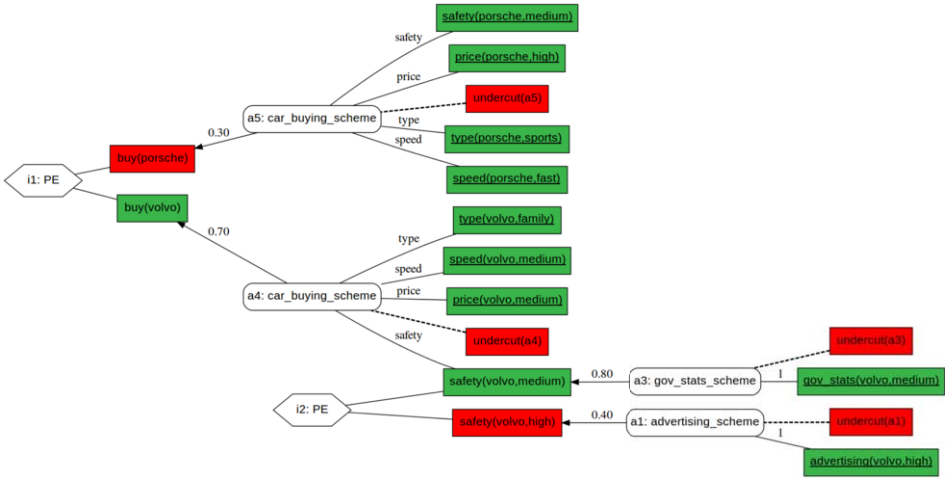


Figure 1. Example Argument Graph

2.2. Semantics

The semantics of argument graphs is defined here in a way inspired by and analogous to the labeling semantics of abstract argumentation frameworks [1], but without mapping argument graphs to abstract argumentation frameworks.

Definition 4 (Labeling) A labeling is a total function from \mathcal{L} to $\{\text{in}, \text{out}, \text{undecided}\}$.

Notice that *statements*, not arguments, are labeled **in**, **out**, or **undecided** here, unlike the labeling semantics for abstract argumentation frameworks. Arguments here are labeled by their weights, as described below.

Example 3 In the argument map shown in Figure 1, the statements shown in green and red are labeled **in** and **out**, respectively. All other statements in \mathcal{L} are, by default, labeled **undecided**.

Now we are in position to define weighing functions of argumentation schemes more precisely.

Definition 5 (Weighing Function) A weighing function maps a (labeling \times argument graph \times argument) tuple to normalized weights, real numbers in the range of 0.0 to 1.0. Every weighing function must assign the weight of 0.0 to an argument in a labeling if its undercutter is *in* in the labeling.

Notice that the weight of an argument can depend on:

- the labeling
- properties of the argument graph, including but not limited to properties of other arguments about the same issue
- properties of the argument, such as the scheme applied

It is the potential dependence of the weight of an argument on the labeling of statements in the argument graph which makes it unclear how this model could be mapped to the pipelined evaluation model of abstract argumentation frameworks, where the labeling of all statements takes place at the end of the pipeline, after all the arguments have been labeled. Not all weighing functions may be sensible. An interesting project for future work might be to define further rationality constraints for weighing functions, in addition to assuring that undercut arguments weigh 0.0.

Example 4 In the example shown in Figure 1, argument a_4 applies the domain-dependent car buying argumentation scheme. The weighing function of this scheme computes a weighted sum of the proven (not claimed) properties of the option supported by the argument, in the manner of multi-criteria decision analysis. In the example, an issue has been made out of the premise of a_4 stating that Volvo's have medium safety, in issue i_2 . Had this issue been resolved in favor of the other position of the issue, claiming that Porsche's are not merely medium safe, but rather highly safe, then the argument for buying the Porsche, a_4 , would have weighed more than it does, 0.7. This illustrates that the failure of a premise can not only weaken an argument, but also strengthen it, a fortiori. The example also illustrates how the weight of an argument, a_4 , can depend on the label of a statement, $\text{safety}(\text{volvo}, \text{medium})$, which in turn (recursively) depends on weights of other arguments, a_1 and a_3 .

Definition 6 (Proof Standard) A proof standard is a mapping from (labeling \times argument graph \times statement) to $\{\text{true}, \text{false}\}$. A statement s satisfies a proof standard, f , given a labeling l and argument graph AG , iff $f(l, AG, s) = \text{true}$. Since proof standards are used to justify decisions, a proof standard may allow at most one position of an issue to satisfy the standard.

Example 5 The preponderance of evidence proof standard can be defined as follows: a position of an issue satisfies the preponderance of evidence standard in an argument graph AG , if and only if there exists an argument in AG for this position (i.e. having this position as its conclusion) which weighs more than every argument in AG for every other position of the same issue, where the weight of an argument, a_i , is derived by applying the weighing function of the argumentation scheme of a_i to (l, AG, a_i) .

Definition 7 (Applicable Argument) An argument $r \in R$ is applicable in a labeling l if and only if:

- The undercutter of r is not *undecided* in l and

- Every premise of r is not **undecided** in l .

Notice that premises of an argument need not be **in** for the argument to be applicable. Premises that are **out** can weaken or strengthen the argument, without causing it to become inapplicable. Also, somewhat unintuitively, an argument can be applicable even if its undercutter is **in**. Undercut arguments have zero weight. (See Definition 5.)

Example 6 In the argument graph shown in Figure 1, all of the arguments are applicable, since all of their undercutters are **out** and none of their premises are **undecided**.

Definition 8 (Supported Statement) In a labeling l , a statement s is supported by an argument graph (S, A, I, R) iff there exists an argument $r \in R$ such that

- s is the conclusion of r ,
- r is applicable in l , and
- $w(l, AG, r) > 0.0$, where w is the weighing function of the scheme of r .

In other words, a statement is *supported* if it is the conclusion of an applicable argument weighing greater than 0.0. Note that a supported statement is not necessarily labeled **in** in l .

Example 7 In the argument graph shown in Figure 1 several statements are supported, including *safety(volvo, medium)*, *safety(volvo, high)*, *buy(volvo)* and *buy(porsche)*.

Definition 9 (Unsupported Statement) Let l be a labeling, (S, A, I, R) be an argument graph, and P be the subset of the arguments R having a statement s as their conclusion. s is unsupported by the argument graph iff

- P is empty or
- for every argument $r \in P$: r is applicable in l but the weight of r in l is 0.0, i.e. $w(l, AG, r) = 0.0$, where w is the weighing function of the scheme of r .

That is, a statement is *unsupported* if every argument for this statement (i.e. having this statement as its conclusion) is either undercut or applicable but with a weight of 0.0. Note that supported and unsupported are not duals: A statement can be neither supported nor unsupported.

Definition 10 (Resolvable Issue) An issue i is resolvable in a labeling l , if for every position p of i : every argument $r \in R$ with the conclusion p is applicable in l or the undercutter of r is **in** in l .

The basic intuition here is that an issue in an argument graph is ready to be resolved in a labeling, if the labeling provides enough information to evaluate every argument for every position of the issue. It may be that no position of a resolvable issue satisfies its proof standard. Thus being resolvable does not imply that some position of the issue is **in**.

Example 8 Both issues of the argument graph shown in Figure 1 are resolvable.

Definition 11 (Conflict Free Labeling) Let AG be an argument graph (S, A, I, R) . A labeling l is conflict free with respect to AG iff, for every statement $s \in S$:

- if $s \in A$ then $l(s) \neq \mathbf{out}$
- if $s \notin A$ and s is unsupported in l then $l(s) \neq \mathbf{in}$
- if s is not a position of some issue $i \in I$ and s is supported in l then $l(s) \neq \mathbf{out}$
- if s is a position of some issue $i \in I$ such that i is resolvable in l and s does not satisfy the proof standard of i then $l(s) \neq \mathbf{in}$
- if s is a position of some issue $i \in I$ such that i is resolvable in l and s satisfies the proof standard of i then $l(s) \neq \mathbf{out}$

The concept of conflict-freeness here is analogous to conflict-freeness in abstract argumentation frameworks. The purpose is to define constraints which must be satisfied by every labeling of an argument graph. The constraints tell us what the labels may not be, but do not tell us what they must be. Labeling a statement **undecided** is always permitted. So, more precisely, the constraints tell us when a statement may not be **in** or **out**:

- Assumptions may not be **out**.
- An unsupported statement which is not an assumption may not be **in**.
- If a supported statement is not at issue, it may not be **out**.
- If an issue is resolvable and some position of the issue does not satisfy the proof standard of the issue, then the position may not be **in**.
- If an issue is resolvable and some position of the issue satisfies the proof standard of the issue, then the position may not be **out**.

Inspired also by abstract argumentation frameworks, we define the semantics of argument graphs using fix-points of a characteristic function:

Definition 12 (Characteristic Function) Let AG be an argument graph (S, A, I, R) . The characteristic function of argument graphs, $f : \text{labeling} \rightarrow \text{labeling}$, is defined as follows:

$f(l) =$
 let m be the resulting labeling
 for each $s \in S$:
 if $l(s) \neq \mathbf{undecided}$ then $m(s) = l(s)$
 else if $s \in A$ then $m(s) = \mathbf{in}$
 else if s is unsupported in l
 then $m(s) = \mathbf{out}$
 else if s is not a position of some issue and s is supported in l
 then $m(s) = \mathbf{in}$
 else if s is a position of some issue $i \in I$ such that
 i is resolvable in l and s does not satisfy the proof standard of i
 then $m(s) = \mathbf{out}$
 else if s is a position of some issue $i \in I$ such that
 i is resolvable in l and s satisfies the proof standard of i
 then $m(s) = \mathbf{in}$
 else $m(s) = l(s)$

The basic intuition behind this characteristic function is that it is intended to complete a labeling of an argument graph, relabeling some or all **undecided** statements to **in**

or **out**, as much as possible in a “single step”. The characteristic function can be applied repeatedly (iteratively) until a fix-point is found, i.e. where $f(l) = l$.

Fix-point semantics requires the characteristic function to be monotonic:

Definition 13 (In and Out Statements of a Labelling; Extensions) *Given an argument graph (S, A, I, R) and a labeling l , let $i(l)$, called the extension of the argument graph in l , denote the subset of S labeled **in** in l and $o(l)$ denote the subset of S labeled **out** in l .*

Conjecture 1 (Monotonicity of the Characteristic Function) *Let us overload \subseteq to also denote a preorder on labelings, where $l_1 \subseteq l_2$ iff $i(l_1) \subseteq i(l_2)$ and $o(l_1) \subseteq o(l_2)$. The characteristic function f is monotonic, preserving this order: for every labeling l_1 and l_2 , if $l_1 \subseteq l_2$ then $f(l_1) \subseteq f(l_2)$.*

Finally, assuming the monotonicity conjecture is true, we can define various fix-point semantics of argument graphs, in a way analogous to the semantics of abstract argumentation frameworks:

Definition 14 (Fix-Point Semantics) *Given an argument graph (S, A, I, R) , a labeling l is:*

- admissible if and only if l is conflict-free.
- complete if and only if l is admissible and $f(l) = l$, i.e. l is a fix-point of f .
- grounded if and only if l is complete and minimal, i.e. there does not exist a labeling l' such that $l' \subset l$.
- preferred if and only if l is complete and maximal, i.e. there does not exist a complete labeling l' such that $l' \supset l$.

Example 9 *The grounded labeling of the argument graph of the running example is shown in Figure 1. The **in** and **out** labels of statements are shown by filling the boxes of the statements with green and red color, respectively. (No statements are **undecided** in the grounded extension of this argument graph.)*

We are developing a version of this formal model in Higher-Ordered Logic (HOL) for the Isabelle proof assistant². The Isabelle version of the formalization is available online³. We plan to use Isabelle to help us prove properties of the model, in future work, including Conjecture 1, about the monotonicity of the characteristic function.

The formal model has been fully implemented in Version 4 of the Carneades Argumentation System. Carneades is open source software, published using the MPL 2.0 license.⁴ Carneades can be used as a command line program or as a web application. You can try out the web version online using the Carneades server.⁵

3. Related Work

This formal model of structured argument has been clearly inspired by Dung’s work [8], even if we have chosen to not follow his recommended pipeline methodology by trying

²<https://isabelle.in.tum.de>

³<https://github.com/carneades/caes2-formalization>

⁴<https://github.com/carneades/carneades-4>

⁵<http://carneades.fokus.fraunhofer.de/carneades>

to map argument graphs to abstract argumentation frameworks. Rather, we have used Dung's approach, in particular its use of fix-point semantics, as a model and adapted it to the purpose of handling balancing arguments, in addition to attack relations among arguments. Some parts of [8] suggest that Dung intended abstract argumentation frameworks to be expressive enough for evaluating all kinds of human argumentation, but all of his examples were from computer science, nonmonotonic (defeasible) logic and logic programming. He did not consider how to model balancing arguments, which are widespread in human argumentation.

We conjecture that it is possible, indeed straightforward, to simulate abstract argumentation frameworks with the model of argument graphs presented here. An example explaining how this can be done can be found online.⁶ Both arguments and attacks of the abstract framework are mapped to structured arguments. Attacks of the abstract framework are modeled as undercutters. If the abstract framework has m arguments and n attacks, the resulting argument graph has at most $2 * m$ statements and $m + n$ arguments. Thus the simulation has polynomial complexity.

Another source of inspiration for this work was ASPIC+ [14]. All three kinds of attack relations supported by ASPIC+ (premise attacks, rebuttals and undercutters) are also supported in our model. We have successfully reconstructed many of the examples used to illustrate ASPIC+. These examples are available online.⁷

In future work we would like to show formally how to simulate ASPIC+ using our model. We conjecture our model is both simpler and more expressive than ASPIC+, considered as a whole, despite some elements of our model being more complex. Both models are frameworks which can be instantiated in various ways (e.g. logical language, priority relation over arguments), but only our system can handle balancing arguments. ASPIC+ can handle a special case, argument accrual, but only at the cost of replacing each accrued argument with multiple arguments, one for each subset of its premises, causing an exponential blow-up in the number of arguments, which negatively impacts both on the efficiency of argument evaluation and the comprehensibility of argument maps used to visualize and explain the evaluation. In [13], Prakken defined three principles of argument accrual, including the principle that accrued arguments can be weaker than arguments with subsets of their premises. We conjecture that these principles are satisfied by the model presented here, but this remains to be formally proved.

We considered basing this model on Abstract Dialectical Frameworks (ADFs) [6], since they provide a convenient platform for defining a wide variety of graph-based formalisms. The nodes of ADFs can in principle model anything, not just arguments, including presumably also statements and issues, as we need. However, ADFs evaluate and label nodes using functions attached to nodes which depend only upon the parents of the nodes, i.e. the immediate predecessors of the node in the directed graph. This does not appear to be general enough for our purposes, as can be seen in the running example used here, where the weighing function of the car buying scheme needs to consider not only the premises of the argument, but also alternative positions of each premise at issue, since the weight of the argument depends on the proven properties of the car being considered, not only its claimed properties. These other positions are three links away in the argument graph from the argument being weighed.

⁶<https://github.com/carneades/carneades-4/blob/master/examples/AGs/YAML/dung-attack-cycle.yml>

⁷<https://github.com/carneades/carneades-4/tree/master/examples/AGs/YAML>

Finally, of course the model presented here is derived from our own prior work on structured argumentation [10] and preserves all of its features, including its support for variable proof standards and its support for modeling the two kinds of critical questions of argumentation schemes, using assumptions and exceptions. However the new model is simpler and more general in several ways:

1. Con arguments and rebuttals are now modeled as arguments pro other positions (options) of issues.
2. There is now only one kind of premise, instead of three (ordinary, exception, assumption). Assumptions are now a subset of the statements of the argument graph. Exceptions are now modeled using undercutters, which are more general, since an undercutter can have more than one premise.
3. All premises are positive. (Previously, premises could be positive or negative.)
4. The new model lifts the restriction to cycle-free argument graphs, thanks to its Dung-inspired fix-point semantics.
5. Argument weights are now derived, by applying weighing functions attached to argumentation schemes, rather than asserted.

The main additional complexity in the new model is its introduction of a third node type for issues, in addition to statements and arguments.

One important advantage of the new formal model is that argument graphs are now much closer to the conceptual model underlying the argument diagrams typically used in informal logic textbooks, such as [18]. This conceptual model underlies several argument mapping tools, including Araucaria [16], and is also the basis for the Argument Interchange Format (AIF) [15]. Version 4 of Carneades, based on the new model presented here, can import and evaluate AIF files.

4. Conclusion

This paper has presented an original formal model of structured argument with support for both attack relations among arguments (premise defeat, rebuttals and undercutters) as well as balancing arguments, using argument weighing functions. The model has been illustrated using a practical reasoning example about which car to buy, where the weighing function computes a weighted sum of the proven properties of the proposed options, in the style of multi-criteria decision analysis. This model can handle cumulative arguments [19] and argument accrual [13] without causing an exponential blow-up in the number of arguments. While the model does not map structured arguments to abstract arguments, it is inspired by the fix-point semantics of abstract argumentation frameworks and uses comparable methods to handle and resolve cycles in argument graphs. The formal model has been fully implemented in Version 4 of the Carneades argumentation system, for grounded semantics. Many examples from the literature on structured argumentation have been successfully reconstructed, and several new examples have been developed to illustrate the model's features for balancing arguments.

In future work we would like to formally prove Conjecture 1 about the monotonicity of the characteristic function, as well proving the conjecture that the model can simulate abstract argumentation frameworks, for common semantics (e.g. complete, grounded, preferred) and formally investigating relationships between this model and other models

of structured argument, in particular ASPIC+. We also plan to investigate whether or not Caminada's rationality postulates for structured argumentation [7], e.g. closure, direct consistency and indirect consistency, are meaningful in the context of this model and, if so, whether they are satisfied. We plan to use the version of the formal model in Higher-Order Logic for the Isabelle proof assistant to facilitate this future work.

References

- [1] Pietro Baroni, Martin Caminada, and Massimiliano Giacomin. An introduction to argumentation semantics. *The Knowledge Engineering Review*, 26(4):365–410, 2011.
- [2] Trevor Bench-Capon. Persuasion in practical argument using value-based argumentation frameworks. *Journal of Logic and Computation*, 13(3):429–448, 2003.
- [3] Philippe Besnard and Anthony Hunter. Constructing argument graphs with deductive arguments: A tutorial. *Argument and Computation*, 5(1):5–30, 2014.
- [4] Floris J. Bex. *Arguments, Stories and Criminal Evidence: A Formal Hybrid Theory*. Springer, Dordrecht, 2011.
- [5] Gerhard Brewka and Thomas F Gordon. How to buy a porsche: An approach to defeasible decision making. In *Working Notes of the AAAI-94 Workshop on Computational Dialectics*, pages 28–38, Seattle, Washington, 1994.
- [6] Gerhard Brewka and Stefan Woltran. Abstract Dialectical Frameworks. In *Proceedings of the Twelfth International Conference on the Principles of Knowledge Representation and Reasoning*, pages 102–111. AAAI Press, 2010.
- [7] Martin Caminada and Leila Amgoud. On the evaluation of argumentation formalisms. *Artificial Intelligence*, 171(5-6):286–310, April 2007.
- [8] Phan Minh Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77(2):321–357, 1995.
- [9] Isabela Fairclough and Norman Fairclough. *Political discourse analysis: A method for advanced students*. Routledge, 2013.
- [10] Thomas F. Gordon, Henry Prakken, and Douglas Walton. The Carneades model of argument and burden of proof. *Artificial Intelligence*, 171(10-11):875–896, 2007.
- [11] Werner Kunz and Horst W.J. Rittel. Issues as elements of information systems. Technical report, Institut für Grundlagen der Planung, Universität Stuttgart, 1970.
- [12] Sanjay Modgil and Henry Prakken. The ASPIC+ framework for structured argumentation: A tutorial. *Argument and Computation*, 5(1):31–62, 2014.
- [13] Henry Prakken. A Study of Accrual of Arguments, with Applications to Evidential Reasoning. In *Proceedings of the Tenth International Conference on Artificial Intelligence and Law*, pages 85–94, New York, 2005. ACM Press.
- [14] Henry Prakken. An abstract framework for argumentation with structured arguments. *Argument & Computation*, 1:93–124, 2010.
- [15] I. Rahwan and Chris Reed. The Argument Interchange Format. In I. Rahwan and Chris Reed, editors, *Argumentation in Artificial Intelligence*. Springer, 2009.
- [16] Chris A Reed and Glenn W A Rowe. Araucaria: Software for argument analysis, diagramming and representation. *International Journal of AI Tools*, 13(4):961–980, 2004.
- [17] Francesca Toni. A tutorial on assumption-based argumentation. *Argument and Computation*, 5(1):89–117, 2014.
- [18] Douglas Walton. *Fundamentals of Critical Argumentation*. Cambridge University Press, Cambridge, UK, 2006.
- [19] Douglas N. Walton, Christopher W. Tindale, and Thomas F. Gordon. Applying recent argumentation methods to some ancient examples of plausible reasoning. *Argumentation*, 28(1):85–119, 2014.

Assigning Likelihoods to Interlocutors' Beliefs and Arguments

Seyed Ali HOSSEINI^{a,1}, Sanjay MODGIL^a and Odinaldo RODRIGUES^a

^a *Department of Informatics, King's College London*

Abstract. This paper proposes mechanisms for agents to model other agents' beliefs and arguments, thus enabling agents to anticipate their interlocutors' arguments in dialogues, which in turn facilitates strategising and the use of enthymemes. In contrast with existing works on "opponent modelling" that treat arguments as abstract entities, the likelihood that an interlocutor can construct an argument is derived from the likelihoods that it possesses the beliefs required to construct the argument. We therefore address how a modelling agent can quantify the certainty that its interlocutor possesses beliefs, based on the modeller's previous dialogues, and the membership of its interlocutor in communities.²

Keywords. Second-order belief, Second-order argument, Community of agents, Argumentation-based dialogue

1. Introduction

Context and Contributions In argumentation-based dialogues [2], the ability of agents to model their interlocutors' arguments enables the strategic choice of arguments that are less susceptible to attack, and the use of enthymemes (i.e. arguments with incomplete logical structures [3,4]) so as to avoid sending information already known to interlocutors. Agents therefore need to not only construct *first-order* arguments from their own knowledge-bases, but also maintain models of their interlocutor's arguments, referred to here as *second-order* arguments.

In existing works on opponent modelling (e.g. [5,6]), an agent assigns an *uncertainty value* $[0, 1]$ to an abstract argument, representing the likelihood that another agent can construct this argument. However, these models of second-order abstract arguments are incomplete in the sense that they do not account for all second-order arguments that can be constructed from their constituent beliefs. Hence in this paper we provide an account of opponent modelling that is distinctive in its consideration of arguments' internal structures. Thus, uncertainties associated with second-order arguments are derived from uncertainties associated with their constituents; that is to say, quantitative valuations of uncertainty as-

¹Correspondence to: Seyed Ali Hosseini, Department of Informatics, King's College London, WC2R 2LS, UK. E-mail: ali.hosseini@kcl.ac.uk.

²This paper is a substantially extended version of [1]

sociated with a modeller's belief that his interlocutor possesses the premises and inference rules for constructing arguments. This then begs the question as to the provenance of these latter uncertainty valuations, which most existing works on opponent modelling do not address. Our primary contribution is to therefore propose two sources for these uncertainty values. The first source is the information that is exchanged in the dialogues an agent participates in. The second, applying when dialogical data is insufficient, is a quantitative measure of similarity amongst all agents, based on their membership in *agent communities*.

Outline of the paper In Section 2 we recall a general framework for structured argumentation – ASPIC+ [7] – which we choose as the underlying argumentation framework due to its generality in accommodating existing argumentation systems. We then illustrate the need to account for uncertainty valuations over second-order beliefs when establishing uncertainty values over second-order arguments. Section 3 then describes how dialogical evidence and community-based estimates are used by agents to assign uncertainty values to second-order beliefs. Finally Section 4 concludes by discussing applications of our model.

2. Preliminaries

In order to assign uncertainty values to arguments and their constituents, explicit access to the structure of arguments is required. We base our model on the ASPIC+ framework [7] which offers a structural account of argumentation that is both general in accommodating existing approaches to argumentation (e.g. [8,9,10]), and is shown to satisfy rationality postulates [11]. In what follows, we recall key concepts of ASPIC+, with some modifications necessary for this work.

We assume all agents are equipped with an ASPIC+ *argumentation theory*, a tuple $\langle \mathcal{S}, \mathcal{K} \rangle$, where \mathcal{S} is an *Argumentation System* capturing the reasoning capability of an agent, and \mathcal{K} is a knowledge-base. \mathcal{S} is a tuple $\langle \mathcal{L}, \mathcal{R}, -, n \rangle$ where \mathcal{L} is a logical language, \mathcal{R} is a set of strict (\mathcal{R}^s) and defeasible (\mathcal{R}^d) inference rules, where the latter are assigned names (wff in \mathcal{L}) by the naming function n , and “-” is a conflict function generalising the notion of negation. A *knowledge-base* \mathcal{K} consists of two disjoint subsets of axiom \mathcal{K}^n and ordinary premises \mathcal{K}^p , where \mathcal{K}^p and \mathcal{R}^d represent (respectively infer) fallible information. On the other hand, axiom premises \mathcal{K}^n and strict rules \mathcal{R}^s are non-fallible, thus cannot be challenged. Typical examples include axioms and inference rules of a deductive logic (see [12] for more detail), and so we assume a unique set of axiom premises and strict inference rules shared amongst all agents. Furthermore, we assume that all agents share the same language \mathcal{L} , conflict function ‘-’ and naming function n .

Given an argumentation theory \mathcal{T} , arguments are constructed by iterative applications of inference rules on premises from \mathcal{K} . The following is a tree-based definition for an argument that is equivalent to the ASPIC+ definition but in which inference rules are explicitly represented:

Definition 1. [Argument] An *argument*, based on a knowledge-base \mathcal{K} and an argumentation system $\langle \mathcal{L}, \mathcal{R}, -, n \rangle$, is a tree where each node is either a formula from \mathcal{L} , or a rule from \mathcal{R} , and the leaves are premises from \mathcal{K} . For every node x :

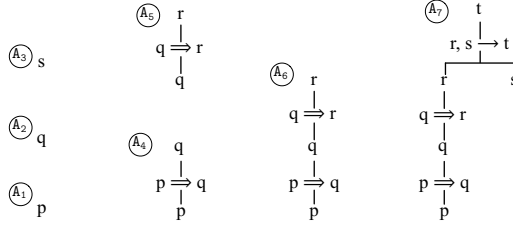


Figure 1. Arguments corresponding to Example 2. Here inference rules using \rightarrow are strict and those using \Rightarrow are defeasible.

a) if x is an inference rule of the form $\phi_1, \dots, \phi_n \rightarrow / \Rightarrow \psi$ then x has a parent ψ , and for every ϕ_i in x 's antecedent, x has a child ϕ_i ; b) if x is a wff ϕ that is not the root, then x 's parent is an inference rule with ϕ in its antecedent; c) if x is a wff ϕ that is not a leaf, then x 's child is an inference rule with ϕ as its conclusion.

Henceforth, we will assume [7]'s notation **Prem(A)** and **Rules(A)** to respectively denote **A**'s premises and inference rules.

Example 2. [Running Example] Let i, j be two agents with argumentation theories $\mathcal{T}_i, \mathcal{T}_j$ respectively. Now let $\mathcal{K}_j^n = \{\}$, $\mathcal{K}_j^p = \{p, q, s\}$, $\mathcal{R}_j^s = \{r, s \rightarrow t\}$ and $\mathcal{R}_j^d = \{p \Rightarrow q, q \Rightarrow r\}$. All arguments that are constructable on the basis of \mathcal{T}_j (i.e. A_1 to A_7) are shown in Figure 1. For argument A_7 , **Prem(A₇)** = $\{p, s\}$ and **Rules(A₇)** = $\{r, s \rightarrow t, q \Rightarrow r, p \Rightarrow q\}$.

In this work, we are concerned with how an agent i can evaluate the likelihood that another agent j can construct a certain argument. Existing works on second-order arguments [6] treat arguments as abstract entities. Therefore, once an agent j commits to a set of arguments $\{A_1, \dots, A_n\}$ in a dialogue, agent i will only consider A_1, \dots, A_n as arguments j can construct, without taking into account all other arguments that can be constructed from A_1, \dots, A_n 's constituents.

Example 3. [Cont. Example 2] Suppose agent j submitted only arguments A_3, A_4, A_5 in Figure 1, in dialogues with i . If i treats arguments as abstract entities, it would believe that j only has arguments A_3, A_4, A_5 . It is however clear that j can also construct A_6 as it has the required beliefs to do so. The same judgement can be made regarding A_7 , as it additionally contains the shared strict rule $r, s \rightarrow t$.

The above example illustrates the need for accessing arguments' internal structures when determining the likelihood that an agent has a certain argument. One common approach [13], though studied in the context where uncertainty values denote likelihoods of truth, is to derive the values associated with arguments from those associated with their constituent beliefs, which [13] considers to be arguments' premises. This is because in [13]'s deductive setting, the set of inference rules, corresponding to classical inferences, is assumed unique and shared by everyone. In our context, this means that any uncertainty as to whether an agent can construct an argument is a function of the uncertainty that it has the necessary beliefs to do so, which in addition to ordinary premises include defeasible rules (since the latter may vary from agent to agent). Therefore, for every two agents i, j , we will assume a function $u_{ij} : \mathcal{L} \cup \mathcal{R} \mapsto [0, 1]$ such that for any wff or rule α (henceforth referred to as a belief), $u_{ij}(\alpha)$ is the likelihood given by agent i that agent j has α . In case the argumentation formalism enforces that

the set of axiom premises and strict inference rules be shared amongst agents, we will have the following conditions: C_1 : if $r \in \mathcal{R}_i^s$ then $u_{ij}(r) = 1$, and C_2 : if $\phi \in \mathcal{K}_i^n$ then $u_{ij}(\phi) = 1$ for all agents i, j . In the next section, we will propose two complimentary mechanisms for evaluating uncertainty over second-order beliefs.

3. Uncertainties over Second-Order Beliefs

In the previous section we established that the uncertainty of a second-order argument is a function of the uncertainties associated with its constituent beliefs (premises and inference rules). We now show how an agent i exploits its dialogues with other agents to assign uncertainty values to these second-order beliefs.

3.1. Dialogical Evidences (DE)

Agents engage in dialogues, which in addition to satisfying a dialogue's primary purpose (e.g. persuading, deliberating), also increases the participants' awareness of each other's states of belief. Note that the information exchanged in dialogues are not necessarily beliefs that agents consider to be 'true' i.e. claims of justified arguments, rather they indicate the beliefs that agents can construct (not necessarily justified) arguments for. The "experience" gained by an agent from its dialogues with other agents is captured by the assignment \mathbf{d} defined below.

Definition 4. For any two agents i, j , a *direct dialogical evidence assignment* $\mathbf{d}_{ij} : \mathcal{L} \cup \mathcal{R}_i \longrightarrow [0, 1] \cup \{\perp\}$ represents the likelihood i assigns to j 's having a premise or inference rule, based on direct dialogical evidence.

A concrete specification of \mathbf{d}_{ij} , including how to consolidate different dialogical evidences can only be provided within a specific dialogue framework. For the purposes of this paper, it suffices to assume that $\mathbf{d}_{ij}(\alpha) = \perp$ indicates that i has some dialogical evidence suggesting that j does *not* believe in α . If i has dialogical evidence that j believes in α , then $\mathbf{d}_{ij}(\alpha)$ gives a value in $[0, 1]$ representing i 's degree of confidence that j believes in α based on i 's dialogical data. Initially, $\mathbf{d}_{ij}(\alpha) = 0$, indicating the absence of any dialogical evidence. Examples of how $\mathbf{d}_{ij}(\alpha)$ is updated each time i obtains an evidence include: when j commits to α as part of an argument in a dialogue with i , $\mathbf{d}_{ij}(\alpha)$ is set to 1; when i gets informed of j 's belief in α through another agent k , in which case $\mathbf{d}_{ij}(\alpha)$ could correspond to i 's level of trust in k ;³ in failed information-seeking or inquiry dialogues with j initiated by i , in which case $\mathbf{d}_{ij}(\alpha)$ could be set to \perp ; and so forth.

Using \mathbf{d} , agents can build models of other agents beliefs and subsequently arguments by harnessing the information they directly obtain through dialogues. Naturally, these models rely on communication and the more frequent that takes place, the more accurate the models become. However, in many cases an agent i may need to determine whether another agent k is able to construct an argument A without any dialogical data directly supporting its decision. In these situations, i must use a different mechanism to estimate k 's ability to construct A . In the next section, we describe how this can be done via the concept of *agent communities*.

³As well as trust valuations, there are other mechanisms from which a value between 0 and 1 for $\mathbf{d}_{ij}(\alpha)$ could be obtained e.g. [5].

3.2. Community-based Estimates (CE)

In a multi-agent environment agents may have various properties (e.g. organisational roles). An *agent group* g can be defined as a set of agents who share a specific property. Logicians and lawyers are both real-world examples of agent groups. We use \mathcal{G} to denote the set of all agent groups. One can also see a group g as a predicate specifying the property that the members of g possess.

A general assumption underpinning our framework is that the shared property between members of a group licenses their sharing of a specific set of beliefs. For example, logicians are all assumed to be aware of the basics of logic. As agents may have multiple properties, agent groups may intersect, and each of these intersections may themselves license the sharing of a separate set of beliefs between its members. For example, consider A and B to be two groups of agents, $AB = A \cap B$ a third group, and for any group G , let \mathcal{B}_G be the set of beliefs shared by agents in G . By assuming a monotonic relationship between group membership and beliefs, we have $\mathcal{B}_{AB} \supseteq \mathcal{B}_A \cup \mathcal{B}_B$ where the set $\mathcal{B}_{AB} \setminus \{\mathcal{B}_A \cup \mathcal{B}_B\}$ is the set of beliefs shared exclusively between AB 's members due to their membership to both A and B .

Therefore, given the set of all groups \mathcal{G} , we consider its powerset $2^{\mathcal{G}}$, call each member of $2^{\mathcal{G}}$ a *community*, and associate it with a distinct set of beliefs that is shared between its members.

Notation 5. Henceforth we assume a finite set of agents AG , a finite set of groups $\mathcal{G} \subseteq 2^{AG}$, and a finite set of communities $\mathcal{C} = 2^{\mathcal{G}}$. Let A , B and C be groups of agents. To simplify notation, we will represent the community $\kappa = \{A, B, C\}$ as the string ABC , and given a community κ , we will use $ag \in \kappa$ instead of $ag \in \cap \kappa$.

Remark 6. The community AB is considered more *specific* than the community A due to their members having more properties, and A is considered to be more *general* than AB . As such, agents in the community \emptyset do not need to have any specific properties – essentially this community contains all agents in the environment – and the beliefs shared amongst them is just *common knowledge*.

We now describe the process of estimating whether an agent has a premise or inference rule, based on its membership to communities. Here, the goal for an agent i is to analyse the data it obtains through dialogues regarding other agents' beliefs, and determine the correlation between having specific premises and rules, and community membership. The idea is to allow i to estimate the likelihood that an agent j has a certain belief based on the communities j belongs to.

Definition 7. Let $ag \in AG$. Then $\text{Gr}(ag) = \{g \in \mathcal{G} \mid ag \in g\}$ is the set of groups to which agent i belongs, and $\text{Cm}(ag) = 2^{\text{Gr}(ag)}$.

Example 8. [Running Example] Let L and P respectively denote “lawyers” and “paralegals” and $\mathcal{G} = \{L, P\}$. Let α be some technical legal information. The experience of an agent i after consulting with several legal firms is summarised in Figure 2, which shows agents' community memberships and whether i assumes they believe (+) or do not believe (−) α . In this context, the community \emptyset , containing all agents, represents “anyone working in a legal firm”.

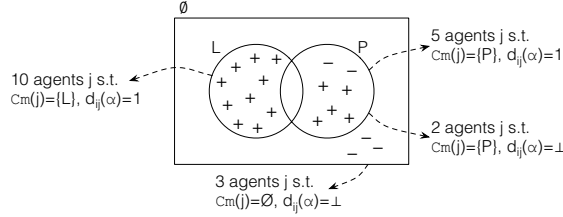


Figure 2. The figure corresponding to Example 2

In order to identify the correlation between believing α and being in a community, each community must be assigned a value representing the likelihood that a member of that community has α .

Definition 9. A *community estimate* for α is a tuple $\langle \kappa, p \rangle$ where $\kappa \in \mathcal{C}$, and $p \in [0, 1]$ is called the *p-score* of κ w.r.t. α . Let S be a set of community estimates for a belief α . Then: 1) $\mathcal{C}(S) = \{\kappa \mid \langle \kappa, p \rangle \in S\}$; 2) $\mathcal{P}(S) = \{p \mid \langle \kappa, p \rangle \in S\}$; and 3) $\max_p(S) = \{\langle \kappa, p \rangle \in S \mid \forall \langle \kappa', p' \rangle \in S, p \not\leq p'\}$.

For every agent i , we will consider a *community-based estimate function* $c_i : \mathcal{C} \times \{\mathcal{L} \cup \mathcal{R}_i\} \mapsto [0, 1]$, assigning a p-score to every community κ w.r.t. a belief α , where $c_i(\kappa, \alpha)$ denotes the likelihood agent i assigns to members of κ having α . This assignment will be defined in two stages to highlight some of the issues that arise in its construction. First, we define the p-score of a community as a standard conditional probability: the probability that members of a community κ believe α based solely on their membership to κ .

Definition 10. The *basic p-score assignment* F_b for an agent i regarding α is defined as follows:

$$F_b^i(\alpha) = \{\langle \kappa, p \rangle \mid \kappa \in \mathcal{C}, \{x \in \kappa \mid d_{ix}(\alpha) \neq 0\} \neq \emptyset\} \text{ where } p \stackrel{\text{def}}{=} \frac{\sum_{x \in \kappa, d_{ix}(\alpha) > 0} d_{ix}(\alpha)}{|\{x \in \kappa \mid d_{ix}(\alpha) \neq 0\}|}$$

Example 11. [Continuing Example 8] Let us calculate the community estimate of \emptyset using $F_b^i(\alpha)$. From amongst the 20 members of \emptyset (i.e., all agents), $d_{ij}(\alpha)$ assigns 1 to 15 agents and \perp to the remaining 5. Thus, the p-score of \emptyset is $15/20 = 0.75$. Similar calculations will yield the following: $F_b^i(\alpha) = \{\langle L, 1 \rangle, \langle P, 0.71 \rangle, \langle \emptyset, 0.75 \rangle\}$.

Remark 12. Note that a p-score as given in Definition 10 is normalised by dividing the sum of the positive dialogical evidences regarding members of the relevant community by the number of all members of that community about whom some dialogical evidence (either positive or \perp) is held (i.e. all agents k s.t. $d_{ik}(\alpha) \neq 0$).

F_b gives a p-score to communities as long as there is some dialogical evidence for at least some of their members. However, there are several issues with F_b . Firstly, the values returned by F_b are not accurate (we call this issue **(P1)**). In Example 11, the community \emptyset gets a p-score of 0.75 w.r.t. α . This means that upon encountering an agent j working in a legal firm, an agent i should rationally expect j to believe α with 0.75 certainty. However, ‘working in a legal firm’ is not in and of itself necessarily relevant to believing α . The problem is that F_b simply takes into account the frequency of agents who belong to \emptyset and believe α , without requiring that those agents believe α *due to their membership in* \emptyset . In Example 8, 10 out of the 15 agents who have α and are members of \emptyset , are also members of L . Thus, in addition to \emptyset ’s p-score, these agents also contribute to L ’s p-score, and it

may well be that these agents believe α exclusively because of their membership in L , rendering their membership of \emptyset irrelevant w.r.t. believing α .

Identifying relevant communities with regard to any belief α can be achieved using the p-scores that are returned by F_b itself. Intuitively, if an agent is in communities X and Y with p-scores p^X and p^Y , and $p^X > p^Y$, then community X is identified as the more likely reason why the agent has α . As a consequence, this agent should be excluded in the calculation of Y 's p-score. In Example 8, the agents who are in \emptyset are also in L which have F_b values 0.75 and 1, respectively (see Example 11). Therefore for these agents, membership to L is identified as the reason for having α , and in the more refined p-score assignment of \emptyset defined below (which we call simply F), these agents are excluded from the calculation.

The new p-score assignment F is defined iteratively. At each step we ensure that for every community c : a) only those agents who belong to c contribute to c 's p-score; and b) membership to c is identified as the most likely reason for the belief of these agents in α (according to the rationale described above). Initially, we use F_b to calculate the p-score of all communities. We then set the p-scores of the communities with the highest p-score. The agents who contributed to these p-scores are implicitly assigned only to these communities, as they have the highest p-score. On each subsequent iteration, we then re-calculate the p-score of the remaining communities and set the p-scores of those with the highest value as before, except that we now exclude from the calculations those agents who have already been previously assigned to a community.

Another issue (referred to as **(P2)**) is that according to Definition 10, F_b does not return a p-score for communities κ for which an agent i has no dialogical data, i.e., when $\{k \in \kappa \mid d_{ik}(\alpha) \neq 0\} = \emptyset$. To illustrate, in Example 11 the p-score of LP w.r.t. α is undefined. To resolve **(P2)**, note that any agent j who belongs to LP ($LP \in \mathbf{Cm}(j)$), also belongs to the communities of lawyers ($L \in \mathbf{Cm}(j)$) and paralegals ($P \in \mathbf{Cm}(j)$). Thus, although agent i has no dialogical experience regarding members of LP^4 , i can appeal to its dialogical experience regarding members of the more general communities L and P to estimate the likelihood that j has α . Specifically i assigns the higher of the F_b values for L and P (i.e. 1).

Finally, we have the issue **(P3)** of when an agent i has no dialogical data for any agent w.r.t. a belief α . In these cases, F assigns 0 to all communities ((1) in the definition below), reflecting that for i , there is as yet no evidence that any agent has α .

Definition 13. Let $i, j \in AG$, α a premise or inference rule, and \mathcal{C} be the set of all communities. The *probability assignment* F is inductively defined as follows:⁵

$$F_0^i(\alpha) = \begin{cases} \{\langle \kappa, 0 \rangle \mid \kappa \in \mathcal{C}\} & \text{if } \max_p(S_0) = \emptyset \\ \max_p(S_0) & \text{otherwise} \end{cases} \quad \begin{matrix} \text{(P3)} & (1) \\ \text{(P1)} & (2) \end{matrix} \quad \text{where}$$

$$S_0 = \{\langle \kappa, p \rangle \mid \kappa \in \mathcal{C}, \{j \in \kappa \mid d_{ij}(\alpha) \neq 0\} \neq \emptyset\} \text{ and } p \stackrel{\text{def}}{=} \frac{\sum_{j \in \kappa, d_{ij}(\alpha) > 0} d_{ij}(\alpha)}{|\{j \in \kappa \mid d_{ij}(\alpha) \neq 0\}|}$$

⁴ L and P could be mutually exclusive, or i 's dialogical data could be incomplete.

⁵Note that \subset_{\max} represents maximal proper subset

and for all $x > 0$

$$F_x^i(\alpha) = \begin{cases} F_{x-1}^i(\alpha) \cup \max_p(S_x) & \text{if } \max_p(S_x) \neq \emptyset \quad (\mathbf{P1}) \quad (3) \\ F_{x-1}^i(\alpha) \cup S'_x & \text{otherwise} \quad (\mathbf{P2}) \quad (4) \end{cases} \quad \text{where}$$

$$S_x = \{ \langle \kappa, p \rangle \mid \kappa \in \mathcal{C}/\mathcal{C}(F_{x-1}^i(\alpha)), \{k \in \kappa \mid d_{ik}(\alpha) \neq 0\} \neq \emptyset \};$$

$$S'_x = \{ \langle \kappa, p' \rangle \mid \kappa \in \min_{\subseteq}(\mathcal{C}/\mathcal{C}(F_{x-1}^i(\alpha))) \};$$

$$\sum d_{ij}(\alpha) \\ p \stackrel{\text{def}}{=} \frac{j \in (\kappa / \cup \mathcal{C}(F_{x-1}^i(\alpha))), d_{ij}(\alpha) > 0}{|\{j \in \kappa \mid d_{ij}(\alpha) \neq 0\}|}; \text{ and } p' \stackrel{\text{def}}{=} \max(\mathcal{P}(\{ \langle \kappa', p' \rangle \in F_{x-1}^i(\alpha) \mid \kappa' \subset_{\max} \kappa \}))$$

Example 14. [Continuing Example 11] Let us calculate the p-score of all communities w.r.t. α , using F given in Definition 13. Since $\max_p(S_0) \neq \emptyset$, we use (2). Here, $S_0 = F_b^i(\alpha) = \{ \langle L, 1 \rangle, \langle \emptyset, 0.75 \rangle, \langle P, 0.71 \rangle \}$. Therefore, $F_0^i(\alpha) = \max_p(S_0)$ so $F_0^i(\alpha) = \{ \langle L, 1 \rangle \}$. We now consider the remaining communities in the second iteration. Since, $\max_p(S_1) \neq \emptyset$, then case (3) is triggered and $S_1 = \{ \langle P, 0.71 \rangle, \langle \emptyset, 0.25 \rangle \}$, thus $F_1^i(\alpha) = \{ \langle L, 1 \rangle, \langle P, 0.71 \rangle \}$. Continuing with the iteration yields $F_2^i(\alpha) = \{ \langle L, 1 \rangle, \langle P, 0.71 \rangle, \langle \emptyset, 0 \rangle \}$. At the next iteration $F_3^i(\alpha)$, since $S_3 = \emptyset$ and thus $\max_p(S_3) = \emptyset$, case (4) is activated. At this stage, $\min_{\subseteq}(\mathcal{C}/\mathcal{C}(F_2^i(\alpha))) = \text{LP}$ whose p-score is the maximum of the p-scores of communities which are one level more general than LP i.e. L with p-score 1 and P with 0.71. Thus, $S'_3 = \{ \langle \text{LP}, 1 \rangle \}$, and $F_3^i(\alpha) = \{ \langle L, 1 \rangle, \langle P, 0.71 \rangle, \langle \emptyset, 0 \rangle, \langle \text{LP}, 1 \rangle \}$. At the next iteration, case (4) is still active since $\max_p(S_4) = \emptyset$. Here, $\min_{\subseteq}(\mathcal{C}/\mathcal{C}(F_3^i(\alpha))) = \emptyset$, hence $S'_4 = \emptyset$. Therefore, $F_4^i(\alpha) = F_3^i(\alpha) \cup \emptyset$, thus: $F_4^i(\alpha) = \{ \langle L, 1 \rangle, \langle P, 0.71 \rangle, \langle \emptyset, 0 \rangle, \langle \text{LP}, 1 \rangle \}$. It is clear that for all other iterations $x > 4$, $F_x^i(\alpha) = F_{x-1}^i(\alpha) \cup \emptyset = F_{x-1}^i(\alpha)$.

Given any agent i , let us now consider some of F^i 's properties.

Proposition 1. Let α be a premise or inference rule held by an agent i : 1) For every iteration x , $F_x^i \subseteq F_{x+1}^i$ (Monotonicity). 2) There is an iteration x s.t. $\mathcal{C}(F_x^i(\alpha)) = \mathcal{C}$ (Exhaustion). 3) There is an iteration x s.t. $F_x^i = F_{x+y}^i$, for $y \geq 0$ (Fixed-point).

Proof. (Sketch) The function by construction satisfies 1-3. For 1) observe that for all iterations $x > 1$, F_x^i is the result of a union operation. For 2), because of the condition $\mathcal{C}/\mathcal{C}(F_{x-1}^i(\alpha))$ in S_x and S'_x , the function assigns a value to a unique community, and since \mathcal{C} is finite, it is eventually exhausted. For 3), due to exhaustion, at some iteration x , the function will run out of communities to assign a value to, thus, $F_x^i = F_{x-1}^i$, and trivially $F_x^i = F_{x+y}^i$ ($y \geq 0$). \square

Proposition 2. For all beliefs α , if $F_x^i(\alpha) = F_{x+1}^i(\alpha)$, then $F_x^i(\alpha)$ is a function assigning a unique p-score to every community w.r.t α .

Proof. (Sketch) Because of $\mathcal{C}/\mathcal{C}(F_{x-1}^i(\alpha))$ in S_x and S'_x , at each iteration the function assigns a unique value to each community. Hence, given 2) and 3) in Proposition 1, the fixed point of $F^i(\alpha)$ which exhausts \mathcal{C} , is a function. \square

We define an agent i 's community-based estimate of the likelihood that a member of κ believes α , denoted $c_i(\kappa, \alpha)$, as the fixpoint of F^i .

Definition 15. Let $i \in AG$, and \mathcal{C} be the set of all communities. Agent i 's community-based estimate function $c_i : \mathcal{C} \times \{\mathcal{L} \cup \mathcal{R}_i\} \mapsto [0, 1]$ is defined such that $c_i(\kappa, \alpha) = p$ where $\langle \kappa, p \rangle \in F_x^i$, and x is an iteration such that $F_x^i = F_{x+1}^i$.

Example 16. [Continuing Example 14] The earliest iteration x such that $F_x^i = F_{x+1}^i$ is 3. Hence: $c_i(L, \alpha) = 1$, $c_i(P, \alpha) = 0.71$, $c_i(\emptyset, \alpha) = 0$, $c_i(LP, \alpha) = 1$.

It is useful for an agent i to know the likelihood of a specific agent j believing in α (denoted by $c_{ij}(\alpha)$), given j 's membership to communities. This is defined as the p-score of the most specific community that j belongs to (trivially $\text{Gr}(j)$).

Definition 17. Let $i, j \in AG$, and α a premise or rule. Then, $c_{ij}(\alpha) = c_i(\text{Gr}(j), \alpha)$.

Example 18. [Continuing Example 8] Suppose agent i encounters agent j and identifies that $\text{Gr}(j) = \{L\}$. We have that $\text{Cm}(j) = \{\emptyset, L\}$ and the agent i 's community-based estimate regarding j 's belief in α is: $c_{ij}(\alpha) = c_i(L, \alpha) = 1$.

Remark 19. The complexity introduced by the number of communities is exponential relative to the overall number of properties that agents in the environment could have. Though this may be problematic with human agents, for computational agents, the actual number of communities considered may well be less, due to a) agents' operation in specialized domains, limiting the number of properties to consider, and b) possibility of using certain heuristics to limit the number of properties one needs to consider (e.g. certain property combinations may be mutually exclusive, thus eliminating communities containing those combinations).

We now combine the dialogical (DE) and community (CE) based estimates (respectively obtained by assignments d and c) to compute the overall likelihood that an agent j believes α . One option is to prioritise dialogical evidence over community-based estimates. Thus, to derive the likelihood that an agent j has an argument A , i considers each of A 's constituents beliefs (i.e. premises and inference rules) α , using $d_{ij}(\alpha)$ if available, and $c_{ij}(\alpha)$ otherwise. Thus, u_{ij} would be defined as follows:

Definition 20. Let d and c be defined according to Definitions 4 and 17, respectively. Then for any two agents i, j and premise or inference rule α : $u_{ij}(\alpha) = d_{ij}(\alpha)$, if $d_{ij}(\alpha) > 0$; $u_{ij}(\alpha) = 0$, if $d_{ij}(\alpha) = \perp$; and $u_{ij}(\alpha) = c_{ij}(\alpha)$, if $d_{ij}(\alpha) = 0$.

3.3. Uncertainties over Second-Order Arguments

As discussed in Section 2, the uncertainty that is associated with second-order arguments, is a function of the uncertainties that are associated with their constituent beliefs. For this purpose, we will define a function U , where for any two agents i, j and argument A , $U_{ij}(A)$ is the likelihood that agent j can construct A according to agent i .

There are a number of techniques in the literature for propagating uncertainty values in arguments, e.g. the weakest link principle (using Min) [14], and [15]. For the purpose of this work, we do not need to commit to any specific method, and assume a general function F that propagates uncertainty values from premises and rules, to arguments composed thereof.

Definition 21. Let \mathcal{A}_i be the set of all arguments defined by agent i 's argumentation theory \mathcal{T}_i . Let j be an agent and F a t-norm. Then

$$U_{ij}(A) = F(\{u_{ij}(\alpha) \mid \alpha \in \text{Prem}(A) \cup \text{Rules}(A)\})$$

is the likelihood that j can construct argument A from i 's point of view.

Consider a complete example deriving the uncertainty of a second-order argument using U and the propagation function $F = \text{Min}$.

Example 22. [Continuing Example 2] Assume agent j moves argument A_5 in a dialogue with agent i , and that this yields $d_{ij}(q) = 1$ and $d_{ij}(q \Rightarrow r) = 1$.⁶ Suppose later that i is informed, by another agent k , that j has argument A_3 , and i 's trust in k yields $d_{ij}(s) = 0.5$. Also assume that through dialogues with other agents, i makes the following assignments $c_{ij}(p) = 1$, $c_{ij}(p \Rightarrow q) = 0.8$. Hence, given $d_{ij}(p) = 0$, $d_{ij}(p \Rightarrow q) = 0$, then by Definition 20:

$u_{ij}(p) = c_{ij}(p) = 1$; $u_{ij}(s) = d_{ij}(s) = 0.5$; $u_{ij}(r, s \rightarrow t) = 1$ (by condition C_1); $u_{ij}(p \Rightarrow q) = c_{ij}(p \Rightarrow q) = 0.8$; and $u_{ij}(q \rightarrow r) = d_{ij}(q \rightarrow r) = 1$.

By Definition 21, and using Min as the propagation function F , the likelihood i assigns to j having argument A_7 is: $U_{ij}(A_7) = \text{Min} \bigcup_{\alpha \in \{p, s, (r, s \rightarrow t), (p \Rightarrow q), (q \Rightarrow r)\}} u_{ij}(\alpha) = \text{Min}(\{1, 0.5, 1, 0.8, 1\}) = 0.5$.

The above example illustrates how the likelihood that an agent i assigns to another agent j being able to construct an argument A can be derived from the likelihoods that i assigns to j having A 's constituent beliefs, which are in turn based on dialogical evidence and j 's membership in communities.

4. Discussion

In this work, we proposed a mechanism that enables agents to model other agents' arguments. We began by highlighting the inadequacy of modelling other agents' arguments as abstract entities, so proposed that a modeller derive the likelihood that another agent can construct an argument based on the likelihood that the arguments' constituent premises and inference rules are held by that agent. We then addressed the provenance of uncertainty values over the constituents of arguments in dialogical settings – something that is not addressed in other works on “opponent modelling” (e.g. [5,16]) – by harnessing a modelling agent's previous dialogues and utilising the notion of agent communities.

Our work has a number of applications, including the strategic choice of arguments in dialogues. Consider persuasion dialogues [17] in which an agent can advantageously anticipate its interlocutor's arguments [18]. For example, suppose i attempts to persuade j to accept ϕ , by communicating an argument claiming ϕ . From amongst all of i 's arguments claiming ϕ (denoted $\text{Poss}(\phi)$), i can strategically choose that which is least susceptible to being attacked by j . That is, for each $A \in \text{Poss}(\phi)$, i must first identify every possible counter-argument to A along with the likelihoods associated with j being able to construct each such

⁶In Section 4 we will comment further on how uncertainty values are propagated from arguments to their constituent beliefs.

counter-argument, and then use this information to select from amongst $\text{Poss}(\phi)$ the argument which is least likely to be attacked by j .

Another application area is the use of *enthymemes*, i.e. arguments with incomplete logical structure [4,3]. Enthymemes are a ubiquitous feature of human dialogue and there are a number of motivations for their use, e.g. to avoid revealing parts of arguments which are susceptible to attack, or to avoid the exchange of information already believed by the dialogue's participants, making their inclusion in arguments redundant in terms of furthering a dialogue's goal. To avoid sending parts (i.e. sub-arguments) of an argument, one needs to determine whether these sub-arguments are known by the recipients of the enthymeme. Therefore, for i to construct an enthymeme from argument A for sending to agent j , i needs to examine all sub-arguments A' of A in descending order of size, and remove A' from A if $U_{ij}(A')$ is higher than a predefined threshold. The reconstruction of the original argument by j would then involve building all complete arguments from which the received enthymeme can be constructed, such that according to j , i is highly likely able to construct the removed sub-arguments using its beliefs. Of course more sophisticated construction and reconstruction procedures would be possible with a move to a higher order modelling, when i can model the arguments that j believes i has. However, this type of modelling is outside the scope of this paper.

There remains a number of open challenges and opportunities for further work. Firstly, as illustrated in Example 22, we have not in this paper formally defined a function that propagates uncertainty values from received arguments to their constituent beliefs, when defining the assignment d_{ij} to those beliefs. Ideally, such a function would be the inverse \bar{U} of the function U that propagates uncertainties from beliefs to arguments. As Example 22 illustrates, \bar{U} makes the assignment $d_{ij}(\alpha) = x$ (α a premise or inference rule in A), where x is the likelihood associated with A (e.g., x maybe 1 if A is directly communicated by j , or $x \leq 1$ where x is the degree of trust in the agent k who informs i that j can construct A). If we assume U makes use of $F = \text{Min}$, then trivially U will assign x to A when propagating $d_{ij}(\alpha)$ to the argument A reconstructed from its constituent α s. Clearly then, the choice of how F and \bar{U} are defined needs to be carefully made if we require that the latter is the inverse of U .

To illustrate, assume that an agent i receives dialogical evidence regarding j having A_4 (in Figure 1) with 0.6 certainty. Assuming that \bar{U} makes the assignment $d_{ij}(\alpha) = 0.6$ to all premises and inference rules α in A_4 , we would have $d_{ij}(p) = 0.6$ and $d_{ij}(p \Rightarrow q) = 0.6$, thus $u_{ij}(p) = 0.6$ and $u_{ij}(p \Rightarrow q) = 0.6$. Then later when A_4 is reconstructed, its uncertainty will be derived from the values assigned to its constituents using U_{ij} . For $F = \text{Min}$, we would have $U_{ij}(A_4) = \text{Min}(u_{ij}(p), u_{ij}(p \Rightarrow q)) = 0.6$, which is the original value i assigned to A_4 upon receipt.

Secondly, we can integrate our work with existing models of probabilistic argumentation (e.g. [13,16]) in which the acceptability of arguments are determined using probabilities. This would imply that not only can agents anticipate other agents' arguments, but also what arguments they deem acceptable, which, for example, allows for devising more sophisticated strategies in dialogues.

Moreover, in this work we have focused on scenarios in which an agent wishes to determine the likelihood that another agent can construct a specific argument. However, another possible scenario is when i wants to determine whether j be-

believes some ϕ in general, regardless of the specific argument justifying that belief. For example, i might want to know whether j can construct A_5 in Figure 1 (i.e. believes r) but is indifferent as to the reasons why j believes q (i.e., whether j believes q as a premise or as the claim of another argument such as A_4). To address these types of questions, some of the underlying formalisations, especially the community-based estimates, need to be updated to take into account every possible argument that can be constructed for a given well-formed formula.

Finally given that our proposed formalism models the use of arguments by computational and human agents, an interesting direction to pursue would be the evaluation using human subjects.

References

- [1] S. A. Hosseini, S. Modgil, and O. Rodrigues. Estimating second-order arguments in dialogical settings. In *Proceedings of the 15th International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS '16*, 2016.
- [2] P. McBurney and S. Parsons. Chapter 13: Dialogue games for agent argumentation. In I. Rahwan and G. Simari, editors, *Argumentation in AI*, pages 261–280. Springer, 2009.
- [3] E. Black and A. Hunter. A relevance-theoretic framework for constructing and deconstructing enthymemes. *Journal of Logic and Computation*, 22(1):55–78, 2012.
- [4] S. A. Hosseini, S. Modgil, and O. Rodrigues. Enthymeme construction in dialogues using shared knowledge. In *Computational Models of Argument*, volume 266 of *Frontiers in Artificial Intelligence and Applications*, pages 325 – 332. IOS Press, 2014.
- [5] C. Hadjinikolis, Y. Siantos, S. Modgil, E. Black, and P. McBurney. Opponent modelling in persuasion dialogues. In *Proceedings of IJCAI '13*. AAAI Press, August 2013.
- [6] T. Rienstra, M. Thimm, and N. Oren. Opponent models with uncertainty for strategic argumentation. In *Proceedings of IJCAI '13*, pages 332–338, 2013.
- [7] S. Modgil and H. Prakken. A general account of argumentation with preferences. *Artificial Intelligence*, 195(0):361 – 397, 2013.
- [8] A. Bondarenko, P. M. Dung, R. A. Kowalski, and F. Toni. An abstract, argumentation-theoretic approach to default reasoning. *Artificial Intelligence*, 93(1,2):63 – 101, 1997.
- [9] N. Gorogiannis and A. Hunter. Instantiating abstract argumentation with classical logic arguments: Postulates and properties. *Artificial Intelligence*, 175:1479–1497, 2011.
- [10] D. N. Walton. *Argumentation Schemes for Presumptive Reasoning*. Lawrence Erlbaum Associates, 1996.
- [11] M. W. A. Caminada and L. Amgoud. On the evaluation of argumentation formalisms. *Artificial Intelligence*, 171(5-6):286 – 310, 2007.
- [12] S. Modgil and H. Prakken. The ASPIC+ framework for structured argumentation: a tutorial. *Argument & Computation*, 5(1):31–62, 2014.
- [13] A. Hunter. A probabilistic approach to modelling uncertain logical arguments. *International Journal of Approximate Reasoning*, 54(1):47 – 81, 2013.
- [14] J. L. Pollock. Defeasible reasoning with variable degrees of justification. *Artificial Intelligence*, 133(1-2):233 – 282, 2001.
- [15] Bram Roth, Antonino Rotolo, Regis Riveret, and Guido Governatori. Strategic argumentation: A game theoretical investigation. In *Proceedings of the Eleventh International Conference on Artificial Intelligence and Law*, pages 81–90. ACM Press, 2007.
- [16] H. Li, N. Oren, and T. J. Norman. Probabilistic argumentation frameworks. In *Theorie and Applications of Formal Argumentation*, volume 7132 of *Lecture Notes in Computer Science*, pages 1–16. Springer Berlin Heidelberg, 2012.
- [17] D. N. Walton and E. C.W. Krabbe. *Commitment in Dialogue: Basic Concepts of Interpersonal Reasoning*. State University of New York Press, 1995.
- [18] E. Black, A. Coles, and S. Bernardini. Automated planning of simple persuasion dialogues. In *Computational Logic in Multi-Agent Systems*, volume 8624 of *Lecture Notes in Computer Science*, pages 87–104. Springer International Publishing, 2014.

A System for Dispute Mediation: The Mediation Dialogue Game

Mathilde JANIER ^{a,1}, Mark SNAITH ^a, Katarzyna BUDZYNSKA ^{a,b},
John LAWRENCE ^a and Chris REED ^a

^aCentre for Argument Technology, University of Dundee, UK

^bPolish Academy of Sciences, Poland

Abstract. We propose a dialogue game for mediation and its formalization in DGDL. This dialectical system is available as software through Arvina for automatic execution. This work expands the literature in dialectical systems, in particular those for more than two players, and shows the practical impact on mediation activity through the opportunity offered to mediators once implemented.

Keywords. dispute mediation, dialogue game, formal dialectical systems, DGDL, DGEP, Arvina

1. Introduction

In dialectical systems, dialogical interaction is viewed as a game with rules that players must follow. Rules of a game are explicated in a dialogue protocol which specifies how the discussion can or should unfold. These rules depend on the type of dialogue the participants are involved in (e.g. persuasion, negotiation or inquiry), and a variety of dialogue games has been proposed (e.g. [1,2]) that help us understand, improve or replicate argumentative interactions. Despite a large number of dialectical systems in the literature (see [3] for an overview), none has, to our knowledge, ever been developed specifically for dispute mediation. This is the challenge taken up here, with a motivation which is two-fold: first, formalizing mediation discourse promises a theoretical framework and a normative view of argumentative interactions; second, executing the game offers the opportunity to deliver a practical tool to support mediation, a conflict resolution process that has enjoyed little computational attention.

Research focused on discourse in dispute mediation such as [4] has proven the major role that arguments play in this activity. In mediation, disputants try to resolve a conflict with the help of a third-neutral, the mediator, who makes sure their discussion is efficient and reasonable so that agreement can be quickly reached. To do so, mediators encourage parties to make their positions clear and redirect the discussion whenever parties do not manage to agree on particular issues. Mediation discourse therefore possesses its own dialogical and argumentative character. Of crucial importance is the mediator's central place in the discussions. Most of their contributions in the discussion consists in asking parties to explicitly deliver and explain their position regarding an issue via pure

¹Corresponding Author: School of Computing, University of Dundee, UK; E-mail: m.janier@dundee.ac.uk.

questions and challenges. A certain type of question also allows mediators to directly seek the parties' agreement or disagreement on some issues: defined in [5] as assertive questioning, this type of question, is a convenient way of making them agree or disagree on propositions. Also, the neutrality of mediators does not prevent them from being assertive but, in contrast to parties who assert their points of view, mediators usually summarize or clarify the discussion [4]. As we will see in Section 2.2, these types of moves can be seen as *restating* (or reframing) the parties' positions.

If we consider that a typical mediation is a discussion in which parties must argue for or against a proposition and the mediator redirects the discussion or restates the disputants' standpoints whenever agreement cannot be quickly reached, then we can take advantage of a general framework where the dialogue can be easily modeled and formalized to define a mediation dialogue game. Implementing it in a system also promises a real application that could be used by trainee mediators to practice their skills.

In Section 2 we present the rules of the Mediation Dialogue Game (MDG), formalize and implement it in Sections 3 and 4, and then compare MDG to other existing dialectical systems in Section 5. We finally discuss future work in Section 6.

2. Specifying a Mediation Dialogue Game: MDG

In this section, we specify the rules of a generic mediation dialogue game (MDG). The definition of the rules relies on empirical knowledge of mediation interactions (such as [4]) and close analyses of the Dispute Mediation Corpus (DMC)² [6]. The rules capture the minimal characteristics of mediation dialogues. Keeping in mind that this game can be executed and used for mediation training, the rules provide strategic moves to the mediator e.g. tackling new issues (see e.g. rule SR9.3 in Section 2.4 below). Moreover, the game offers a normative framework guaranteed by rules that assure parties' reasonableness: they cannot have inconsistent commitments and are obliged to answer to questions and challenges (see e.g. rule SR7 below).

2.1. Players, Domain and General Considerations

MDG captures the opening and argumentative stages of a dispute which involves three players: P_1 and P_2 , who play the role of disputing participants (or parties), and M who plays the role of the mediator. We also use P_x and P_y , where $x, y \in \{1, 2\}$ and $x \neq y$ when we are not interested in a party in particular, but nevertheless need to make a distinction between them. In MDG, P_1 , P_2 and M engage in a dialogue to resolve a dispute on topic t by advancing a set of propositions p , q , and so on, that pertain to the domain t (e.g. divorce, child custody). t can be any topic that is tackled in civil case mediations, and propositions p , q etc. are any proposition about the dispute at stake.

2.2. Locution Rules

Locution rules define the types of moves that players can perform during the game. They are composed of two elements: the proposition (or propositional content) symbolized by lower-case letters (e.g. p) and its illocutionary force [7], forming a function of the type Illoc-Force(p). The locution rules of our game are given in Table 1.

²Corpus available at arg.tech/DMC

Table 1. Locution rules

LR1	<p>M can only question (Q), challenge (Ch) or restate (R):</p> <ol style="list-style-type: none"> 1. $PQ(p)$ when he asks whether p is the case, i.e. if P_x believes p 2. $AQ(p)$ when he seeks P_x's agreement on p 3. $PCh(p)$ when he seeks P_x's ground for stating p 4. $R(p)$ when he reuses P_x's proposition p
LR2	<p>P_x cannot question or challenge but will respond to Qs and Chs in one of the following ways:</p> <ol style="list-style-type: none"> 1. $A(p)$ when he states an opinion 2. $W(p)$ when he retracts p 3. $Agr(p)$ when he agrees on p 4. $Disagr(p)$ when he disagrees on p

Mediator's typical moves i.e. questioning (Q) and restating (R) participants' locutions (see Section 1) must be available in our dialogue game rules: this is provided by LR1. We also constrain P_1 and P_2 's moves by forbidding Q, Ch and R. If this game is indeed intended to mediators for practicing their techniques, M should be the only one to have 'strategic' moves available: PQs (pure questions) to launch the discussion and new issues to broach, AQs (assertive questions) to seek other players (dis-)agreement, PChs (pure challenges) to foster argumentation and, most importantly, R to be able to go back on a previous proposition; furthermore, we prevent M from asserting (A) to comply with the mediator's principle of neutrality.

P_1 and P_2 can make assertions (A) that allow them to give their opinion (LR2.1). With LR2.2 parties can withdraw (W) a proposition, a feature needed in particular to keep commitments updated and which usefulness is elicited by structural rules (see Sections 2.3 and 2.4). Finally, P_x can Agr (agree) and Disagr (disagree) to show his position regarding claims that he did not introduced (LR2.3 and LR2.4).

It is important to note that we do not specify a locution rule to permit players to *argue*. As stated in [8] and [9], 'arguing' is a complex illocutionary force that takes shape only by virtue of the interrelation between locutions: one can build an argument by asserting p and q and showing that there is an inference between p and q , e.g. " p because q ". Hence, *arguing* is automatically created when support for a proposition is given and, in MDG, PCh allows for triggering inference.

2.3. Commitment Rules

Integrating commitment-stores is a convenient way for detecting when consensus on an issue is reached [2]. They allow for keeping track of which propositions speakers are committed to. Propositions are thus updated in function of the developments of the dialogue. In Table 2, Com_x symbolizes P_x 's commitment-store. Note that only P_1 and P_2 have commitment-stores; this is because we want to reflect the mediator's neutrality. Updating a store therefore only happens when P_x moves. As in most formal dialogue systems (e.g. DC [10], CB [11], or PPD [2]), MDG allows players to retract propositions: if a proposition is withdrawn, it is assumed that the players are no more in conflict about this proposition and consensus is reached on that particular proposition (CR2). Commitment rules in MDG however differ from those in other dialogue games in that propositions are added only if they have been asserted or agreed on: we do not assume that a proposition is accepted by all players until it is retracted. This is defined in CR1 and

Table 2. Commitment rules

CR1	After $A(p)$, performed by P_x , p is added to Com_x
CR2	After $W(p)$, performed by P_x , p is removed from Com_x
CR3	After $Agr(p)$, performed by P_x , p is added to Com_x
CR4	After $Disagr(p)$, performed by P_x , $\neg p$ is added to Com_x

CR3. CR4 specifies that if a proposition p is disagreed on, then the opposite proposition ($\neg p$) is added to a store.

2.4. Structural Rules

Structural rules regulate how the dialogue can proceed i.e. which move is permitted, by which player, after a particular move. These are presented in Table 3.

The beginning of the dialogue aims at revealing P_1 and P_2 's respective standpoints w.r.t. the topic of the dispute [4], that is why M must ask both parties about the topic t (SR3). To reflect the argumentative function of the dialogue game, P_1 and P_2 must argue but, given constraint SR1 and LR2, argumentation can only be performed by M advancing PCh and P_1 and P_2 answering the challenge, specified in SR4. SR5 specifies that M can ask a player whether she also believes p , agrees on p , or ask to the player whose commitment-store contains p grounds for stating such a proposition. SR6 specifies that P_1 and P_2 must make their positions clear on a proposition p when M poses a PQ: they are either committed to p (SR6.1) or not (SR6.2). After an AQ, a player can withdraw p or (dis-)agree on p (SR7). SR8 allows a player to argue for a standpoint (SR8.1) or retract a proposition (SR8.2). If a player withdraws a proposition p , M can ask whether the player is then committed to $\neg p$ (SR9.1) or, he can explore new issues by asking questions on other propositions (SR9.2 and SR9.3). M can also explore other propositions with SR10. If a player disagrees on a proposition p , M can redirect the discussion on another issue (SR11.1), or check if the player is then committed to $\neg p$ by restating $\neg p$ (SR11.2), and either trigger the player's (dis-)agreement on $\neg p$ (SR12.1) or ask him grounds for $\neg p$ (SR12.2). With the last three rules, we can see the importance of the technique of restating: we have seen that when a player disagrees on a proposition p , the opposite proposition is added to its commitment store (rule CR4). This proposition $\neg p$, however, has never been asserted by the player, and M may want to make sure that the player actually believes $\neg p$. There are two possibilities for this: either seek for (dis-)agreement on $\neg p$ via an AQ, or challenging $\neg p$, in which case the player will give a support for $\neg p$ or withdraw it. These rules therefore allow M to clarify the players' standpoints: if they disagree on a proposition p , it does not necessarily mean that they believe the opposite, and this must be made clear in the game so that all positions are explicitly provided.

2.5. Termination and Outcome Rules

Termination rules define how and when the dialogue must end. In mediation, the process ends when a final agreement between disputants has been reached or when, after a certain time, disputants and mediators reckon that agreement is not possible. In MDG, the dialogue can terminate at any point, provided that the last player to move is not M i.e. when M 's questions or challenges have been responded to.

Outcome rules should specify, at the end of a dialogue, who wins and who loses. In MDG, only P_1 and P_2 can win. At the start of the game, P_1 is committed to p and P_2 to

Table 3. Structural rules

SR1	P_1 and P_2 can only perform one move per turn
SR2	M can perform a maximum of two moves per turn iff the first move consists of restating (R)
SR3	The dialogue starts with M seeking P_1 and P_2 's respective points of view regarding t , therefore: 1. M moves first with $PQ(t)$ addressed to P_1 2. After that, P_1 must answer with $A(p)$ 3. Then, M moves with $PQ(t)$ addressed to P_2 4. Next, P_2 must answer with $A(q)$
SR4	The second step of the opening stage is to discover P_1 and P_2 's grounds for p and q , therefore: 1. M performs $PCh(p)$ addressed to P_1 2. After that, P_1 must answer with $A(r)$ 3. Then, M performs $PCh(q)$ addressed to P_2 4. Next, P_2 must answer with $A(s)$
SR5	After P_x performed $A(p)$, M can perform: 1. $PQ(p)$ addressed at P_y 2. $AQ(p)$ addressed at P_y 3. $PCh(p)$ addressed at P_x
SR6	After M performed $PQ(p)$ addressed at P_x , P_x can perform: 1. $A(p)$ 2. $A(\neg p)$
SR7	After M performed $AQ(p)$ addressed at P_x , P_x can: 1. $W(p)$ 2. $Agr(p)$ 3. $Disagr(p)$
SR8	After M performed $PCh(p)$ to P_x , P_x can: 1. $A(q)$ 2. $W(p)$
SR9	After P_x performed $W(p)$, M can: 1. $AQ(\neg p)$ addressed to P_x 2. $PQ(q)$ addressed either to P_x or P_y 3. $AQ(q)$ addressed either to P_x or P_y
SR10	After P_x performed $Agr(p)$, M can: 1. $PQ(q)$ addressed either to P_x or P_y 2. $AQ(q)$ addressed either to P_x or P_y
SR11	After P_x performed $Disagr(p)$, M can, 1. $PQ(q)$ addressed to any player 2. $R(\neg p)$ addressed to P_x and P_y
SR12	After M performed $R(\neg p)$, M must either: 1. $AQ(\neg p)$ addressed to P_x i.e. the player who previously disagreed on p , or 2. $PCh(\neg p)$ addressed to P_x i.e. the player who previously disagreed on p

q and, in order to win, the players must be committed to their initial proposition, and: (i) have this proposition accepted by the opponent or, (ii) have the opponent retract his initial proposition or, (iii) have the opponent committed to no proposition at all. In all other cases the winner of the game is left undecided. The 12 different final situations are summarized in Table 4.

Table 4. Final situations in MDG

Situation	P ₁ is committed to	P ₂ is committed to
P ₁ wins if	p	\emptyset
	p	$\neg q$
	p	p
P ₂ wins if	\emptyset	q
	$\neg p$	q
	q	q
undecided	p	q
	$\neg p$	$\neg q$
	$\neg p$	\emptyset
	\emptyset	$\neg q$
	\emptyset	\emptyset
	q	p

3. Formal Specification in DGDL

The Dialogue Game Description Language (DGDL) [12] is a language developed to cope with the diversity of dialectical systems, allowing for a standardized formalization of games. The formal specification of MDG consists in translating the rules presented in Section 2 so that the game can be executed. We do not include it here, however it is available to the reader at: `arg.tech/MDG`. In our DGDL specification, the first line explains that the system described is the mediation dialogue game, where there is not a predefined number of turns (line 2). Lines 3-11 specify the number of players, their role and identification (see Section 2.1), and their commitment stores (see Section 2.3). The *Interactions* (line 13 onwards) are the moves that each participant in the dialogue can make, along with the associated effects. Line 15 explains that the dialogue starts with M asking a PQ to P₁. Lines 20-22, 26-29 and 33-35 correspond to structural rules SR6, SR7 and SR4 respectively. Lines 38-51 specify rules SR5 and CR1 together, and the obligation for M to move next. Lines 57-67 correspond to rules SR9 and CR2. Lines 62-82 and 87-97 specify SR10 and CR3, as well as SR11 and CR4. Finally SR12 is given in lines 102-104.

4. Implementation and Product

The Dialogue Game Execution Platform (DGEP) was created to handle any DGDL specifications in order to implement a variety of systems, giving us the opportunity to automatically execute our game in a system to play it. Arvina is a dialogical support system for the execution of games [13,14] relying on both DGDL and DGEP. It allows users to play a dialogue game with virtual agents and or other humans on a user-friendly interface. The advantages of using Arvina in public deliberation contexts has been shown in [13], and additional dialogue games (e.g. for debates) have been implemented. This flexibility therefore ensures the possibility to execute our MDG.

Figure 1 is a screenshot of MDG executed in Arvina. The users (three human players) advanced propositions that were extracted from a dialogue taken from DMC³. We

³Available at `arg.tech/map9373`

can see the Mediator asking the first mandatory PQ and PCh to Viv (playing the role of a party) following the other party (Eric)'s response to the same questions. The bottom banner with "Select a move: No moves available" shows that after the PCh, Mediator is not authorized to perform a move until Viv answers. This figure shows that the game matches up reasonably well with natural discourse.

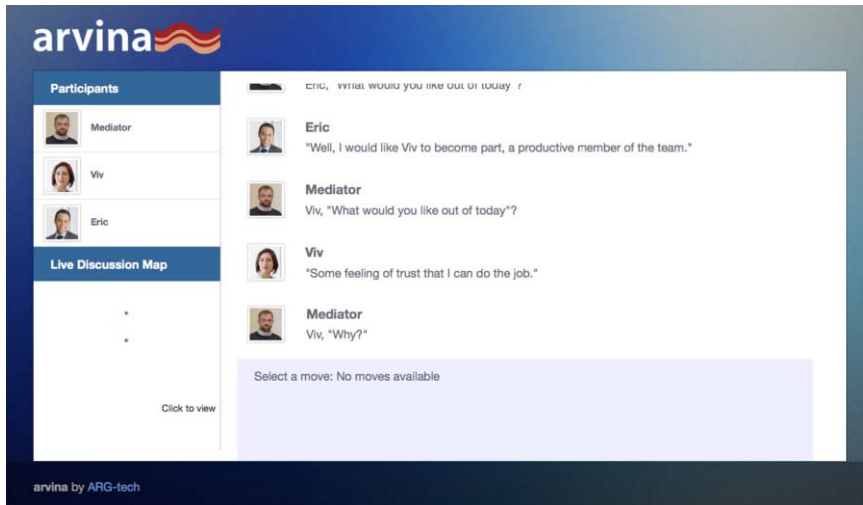


Figure 1. MDG in Arvina

5. Related Work

In [15], Prakken was one of the first to develop a formal system involving three players: he introduced an ‘adjudicator’ to persuasion dialogue systems to reflect the role of participants in legal settings. MDG is similar to Prakken’s model in that we provided a specification for three players, including the mediator (M) whose role is in some way similar to Prakken’s adjudicator in the argumentation phase. Similarly to Prakken’s system, our game allows a fair and efficient resolution of the conflict. Structural rules are designed to encourage fairness thanks to a balance between P_1 and P_2 ’s contributions (e.g. the first PQ and PCh are asked alternately to both players) and efficiency is facilitated by AQ that permits M to seek agreement on several points.

A significant difference between MDG and the state of the art lies in the way it handles argumentation. In [16] and [15] players argue via locutions of the type ϕ since S or argue A . In our system, argumentation is implicit and is the result of the interactions rather than an action per se. This more closely matches evidence from empirical studies that show that arguments are created by dialogical interactions [9].

6. Conclusion and future work

In this paper we proposed a dialectical system for dispute mediation dialogues: MDG. This game aims at providing a minimal and generic framework that can be derived to

grasp other mediation subtleties. As an example, in [17], the authors identified three types of discussions in mediation (critical, bargaining and therapeutic). It would be possible to further specify MDG to play these three different types of games. Also, it would be interesting to further constrain our game by allowing strategic moves to parties; that would not only make the mediator's task tougher, but would also be more representative of what mediation discussions actually look like. After exploring these tracks and bringing improvements to our game, it will be possible to deliver the tool to mediation practitioners for evaluation.

In conclusion, this paper offers advances on both theoretical and practical sides. It extends knowledge on dialectical systems and mediation discourse, while at the same time finding a real utility in supporting the ever-growing practice of dispute mediation.

7. Acknowledgments

We gratefully acknowledge the support of the Leverhulme Trust under grant RPG-2013-076.

References

- [1] C. L. Hamblin, *Fallacies*. Vale Press, 1970.
- [2] D. N. Walton and E. C. W. Krabbe, *Commitment in dialogue: Basic concepts of interpersonal reasoning*. State University of New York Press, 1995.
- [3] P. McBurney and S. Parsons, "Dialogue game protocols," in *Communication in Multiagent Systems*, pp. 269–283, Springer, 2003.
- [4] S. Greco Morasso, *Argumentation in dispute mediation*. John Benjamins Publishing Company, 2011.
- [5] M. Janier and C. Reed, "Towards a theory of close analysis for dispute mediation discourse," *Argumentation*, vol. 0.1007/s10503-015-9386-y, 2015.
- [6] M. Janier and C. Reed, "Corpus resources for dispute mediation discourse," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pp. 1014–1021, 2016.
- [7] J. R. Searle and D. Vanderveken, *Foundations of illocutionary logic*. Cambridge University Press, 1985.
- [8] F. H. van Eemeren and R. Grootendorst, "The speech acts of arguing and convincing in externalized discussions," *Journal of Pragmatics*, vol. 6, no. 1, pp. 1–24, 1982.
- [9] K. Budzynska, M. Janier, J. Kang, C. Reed, P. Saint Dizier, M. Stede, and O. Yaskorska, "Towards argument mining from dialogue," in *Proceedings of COMMA*, vol. 266, pp. 185–196, 2014.
- [10] J. D. Mackenzie, "Question-begging in non-cumulative systems," *Journal of Philosophical Logic*, vol. 8, no. 1, pp. 117–133, 1979.
- [11] D. N. Walton, *Logical dialogue-games and fallacies*. University Press of America Inc., 1984.
- [12] F. Bex, J. Lawrence, and C. Reed, "Generalising argument dialogue with the Dialogue Game Execution Platform," in *Proceedings of COMMA*, vol. 266, pp. 141–152, 2014.
- [13] M. Snaith, J. Lawrence, and C. Reed, "Mixed initiative argument in public deliberation," in *Proceedings of Online Deliberation, Fourth International Conference, OD2010*, pp. 2–13, 2010.
- [14] J. Lawrence, F. Bex, and C. Reed, "Dialogues on the Argument Web: Mixed initiative argumentation with Arvina," in *Proceedings of COMMA*, vol. 245, pp. 513–514, 2012.
- [15] H. Prakken, "A formal model of adjudication dialogues," *Artificial Intelligence and Law*, vol. 16, no. 3, pp. 305–328, 2008.
- [16] H. Prakken, "Formal systems for persuasion dialogue," *The Knowledge Engineering Review*, vol. 21, no. 02, pp. 163–188, 2006.
- [17] S. Jacobs and M. Aakhus, "What mediators do with words: Implementing three models of rational discussion in dispute mediation," *Conflict resolution quarterly*, vol. 20, no. 2, pp. 177–203, 2002.

On ASPIC⁺ and Defeasible Logic

Ho-Pun LAM¹, Guido GOVERNATORI and Régis RIVERET

*Data61, CSIRO | NICTA, Australia*²

Abstract. Dung-like argumentation framework ASPIC⁺ and Defeasible Logic (DL) are both well-studied rule-based formalisms for defeasible reasoning. We compare the two frameworks and establish a linkage between an instantiation of ASPIC⁺ and a DL variant, which leads to a better understanding and cross-fertilization – in particular our work sheds light on features such as ambiguity propagating/blocking, team defeat and strict rules for argumentation, while emphasizing the argumentation-theoretic features of DL.

Keywords. ASPIC⁺, Defeasible Logic, argumentation

1. Introduction

The argumentation framework ASPIC⁺ and Defeasible Logic (DL) support, from different perspectives, rule-based inferences pertaining to defeasible reasoning.

ASPIC⁺ [20, 16, 17] originates from a project aiming at integrating and consolidating well-studied approaches to structured argumentation. ASPIC⁺ develops the instantiation of Dung’s abstract framework [8] provided in [1]. to give a general structured account of argumentation that is intermediate in its level of abstraction between concrete logics and the fully abstract level, providing guidance on the structure of arguments, the nature of attacks, and the use of preferences, accommodating at the same time a broad range of instantiating logics and allowing for the study of conditions under which the various desirable properties are satisfied by these instantiations.

DL [18, 3] is a simple, efficient but flexible non-monotonic formalism capable of dealing with many different intuitions of non-monotonic reasoning. DL has a very distinctive feature: the logic was designed to be easily implementable right from the beginning, and has linear complexity [13]; DL is a framework hosting different variants of DL; within this framework DL can be “tuned” in order to obtain a logic with desired properties, such as ambiguity blocking/propagation and team defeat.

Dung [8] presented an abstract argumentation framework, and different works showed that several well-known nonmonotonic reasoning systems are concrete instances of the abstract framework. Although DL can be described informally in terms of arguments, the various variants have been formalized in a proof-theoretic setting in which arguments play no role. For this reason, [11] gave an argumentation semantics for the variants of DL. They showed that Dung’s grounded semantics characterizes the ambiguity propagation defeasible logic without team defeat.

¹Corresponding Author: Ho-Pun Lam E-mail: brian.lam@data61.csiro.au

²NICTA is funded by the Australian Government through the Department of Communications and the Australian Research Council through the ICT Centre of Excellence Program.

In this paper, we establish close connections between ASPIC⁺ and DL variants, and highlight their differences. Such connections are meant to lead to a better understanding of each framework, and cross-fertilization. For example, the interpretation of DL proofs in terms of argument interplays shall lead to a more intuitive understanding of DL proof theory, while discussions on ambiguity blocking/propagation in DL shall suggest possible developments in ASPIC⁺. Since there are already very flexible and efficient implementations of DL, our research may lead to the implementations of argumentation systems on the basis of DL.

This paper is structured as follows. In the next two sections we outline the key concepts of ASPIC⁺ and DL. The similarities and differences of the two formalisms will be discussed in Section 4. In Section 5, we propose a mapping between an instantiation of ASPIC⁺ and DL, followed by the conclusions.

2. ASPIC⁺

ASPIC⁺ [20, 16, 17] develops Amgoud et al.'s [1] instantiation of Dung's [8] abstract frameworks with accounts of the structure of arguments, the nature of attack and the use of preferences. In the remainder, we will mostly refer to the version given in [17]. The framework posits an unspecified logical language \mathcal{L} , and defines arguments as inference trees formed by applying strict or defeasible inference rules to premises that are well formed formulae (wff) in \mathcal{L} . A strict rule means that if one accepts the antecedents, then one must accept the consequent no matter what. A defeasible rule means that if one accepts all antecedents, then one must accept the consequent if there is insufficient reason to reject it.

In order to define attacks in the context of a general language \mathcal{L} , one needs an appropriately general notion of conflict (i.e., one that does not commit to specific forms of negation). Thus, some minimal assumptions on \mathcal{L} are made; namely that certain wff are a contrary or contradictory of certain other wff. Apart from this, the framework is still abstract: it applies to any set of strict and defeasible inference rules, and to any logical language with a defined contrary relation.

Definition 1. An *argumentation system* is a triple $AS = (\mathcal{L}, \mathcal{R}, n)$ where: (i) \mathcal{L} is a logical language closed under negation (\neg). (ii) $\mathcal{R} = \mathcal{R}_s \cup \mathcal{R}_d$ is a set of strict (\mathcal{R}_s) and defeasible (\mathcal{R}_d) inference rules of the form $\varphi_1, \dots, \varphi_n \rightarrow \varphi$ and $\varphi_1, \dots, \varphi_n \Rightarrow \varphi$ respectively (where φ_i, φ are meta-variables ranging over wff in \mathcal{L}), and such that $\mathcal{R}_s \cap \mathcal{R}_d = \emptyset$. (iii) n is a partial function such that $n : \mathcal{R}_d \rightarrow \mathcal{L}$.

Informally, $n(r)$ means that r is applicable. To ease the comparison with DL, we assume in the remainder that \mathcal{L} is a language of propositional literals composed from a set of propositional atoms. Given a literal l , $\sim l$ denotes the complement of l , that is, $\sim l = \neg m$ if $l = m$ and $\sim l = m$ if $l = \neg m$.

In ASPIC⁺, a knowledge base \mathcal{K} is used to specify the premises from which an argument can be built, which is the union of two disjoint kinds of formulae: the axiom \mathcal{K}_n (which cannot be defeated), and the ordinary premises \mathcal{K}_p (which can be defeated).

Definition 2. An *argumentation theory* is a tuple $AT = (AS, \mathcal{K})$ where AS is an argumentation system and \mathcal{K} is a knowledge base in AS .

On the basis of an argumentation theory, arguments can be built. An argument is basically the chain applications of the inference rules starting with elements from the knowledge base. We give here a more compact variant of the definition given in [17].

Definition 3. An *argument* A on the basis of an argumentation theory with a knowledge base \mathcal{K} and an argumentation system $(\mathcal{L}, \mathcal{R}, n)$ is:

- φ if $\varphi \in \mathcal{K}$, with: $\text{Prem}(A) = \{\varphi\}$; $\text{Conc}(A) = \{\varphi\}$; $\text{Sub}(A) = \{A\}$; $\text{Rules}(A) = \emptyset$; $\text{DefRules}(A) = \emptyset$, $\text{TopRule}(A) = \text{undefined}$.
- $A_1, \dots, A_n \rightarrow / \Rightarrow \psi$ if A_1, \dots, A_n are arguments such that there exists a strict/defeasible rule $\text{Conc}(A_1), \dots, \text{Conc}(A_n) \rightarrow / \Rightarrow \psi$ in $\mathcal{R}_s/\mathcal{R}_d$, with:
 - * $\text{Prem}(A) = \text{Prem}(A_1) \cup \dots \cup \text{Prem}(A_n)$,
 - * $\text{Conc}(A) = \psi$,
 - * $\text{Sub}(A) = \text{Sub}(A_1) \cup \dots \cup \text{Sub}(A_n) \cup \{A\}$. Note that A_1, \dots, A_n are referred to as the proper sub-arguments of A ,
 - * $\text{DefRules}(A) = \{r \mid r \in \text{Rules}(A), r \in \mathcal{R}_d\}$
 - * $\text{StRules}(A) = \{r \mid r \in \text{Rules}(A), r \in \mathcal{R}_s\}$
 - * $\text{TopRule}(A) = \text{Conc}(A_1), \dots, \text{Conc}(A_n) \rightarrow / \Rightarrow \psi$

where Prem returns the set of formula from \mathcal{K} (premises) that used to build A , Conc returns its conclusion, Sub returns all its sub-arguments, DefRules and StRules respectively return the set of defeasible and strict rules in A , and TopRule returns the last inference rule applied in A .

Definition 4. An argument A is *strict* if $\text{DefRules}(A) = \emptyset$ and *defeasible* otherwise; *firm* if $\text{Prem}(A) \subseteq \mathcal{K}_n$; *plausible* if $\text{Prem}(A) \not\subseteq \mathcal{K}_n$; *fallible* if A is plausible or defeasible.

ASPIC⁺ emphasises that (i) attacks should only be targeted at fallible elements of the attacked argument, (ii) a distinction between preference dependent and preference independent attacks, which leads to the following definition for attacks and defeats.

Definition 5. Argument A *attacks* B iff A undercuts, rebuts or undermines B . Argument A *undercuts* argument B (on B') iff $\text{Conc}(A) = \neg n(r)$ for some $B' \in \text{Sub}(B)$ such that B' 's top rule r is defeasible. Argument A *rebuts* argument B (on B') iff $\text{Conc}(A) = \sim \varphi$ for some $B' \in \text{Sub}(B)$ of the form $B'_1, \dots, B'_n \Rightarrow \varphi$. Argument A *undermines* B (on φ) iff $\text{Conc}(A) = \sim \varphi$ for an ordinary premise φ of B .

Note that an attack originating from an argument A requires that its conclusion $\text{Conc}(A)$ is in conflict with some fallible elements – i.e., non-axiom premises, or defeasible rules or conclusions of defeasible rules – in the attacked argument.

The attack relation tells which arguments are in conflict with each other: if two arguments are in conflict then they cannot both be justified. In ASPIC⁺, it is assumed that an argument A can be used as a counter-argument to B , if A successfully attacks, i.e., defeats, B . Whether an attack from A to B (on its sub-argument B') succeeds as a defeat, may depend on the relative strength of A and B' , i.e., whether B' is strictly stronger than, or strictly preferred to A . So, the preferences amongst arguments are specified by a binary ordering \leq over arguments³.

Notice that while several methods to assign preference orderings have been proposed in ASPIC⁺, ASPIC⁺ as a framework does not make any assumption on the argument ordering. To facilitate the comparison with DL we consider the following “last-link” inspired ordering:

- from amongst all the defeasible rules in B there exists a rule which is weaker than (strictly less than according to \leq) all the last defeasible rules in A , and

³As is usual, its strict counterpart $<$ is defined as $X < Y$ iff $X \leq Y$ and $Y \not\leq X$.

- from amongst all the ordinary premises in B there is an ordinary premise which is weaker than (strictly less than according to \leq') all the last ordinary premises in A .

On the basis of the preferences over arguments, successful attacks (defeats) are defined.

Definition 6. Let A and B be arguments. A *successfully rebuts* B if A rebuts B on B' and $A \not\prec B'$. A *successfully undermines* B if A undermines B on φ and $A \not\prec \varphi$. A *defeats* B iff A undercuts or successfully rebuts or successfully undermines B .

Let us recap. Based on an argumentation theory (see Def. 2), we can build arguments (Def. 3), attack (Def. 5) and defeat relations (Def. 6), and finally a Dung's argumentation framework [8] can be built. ASPIC⁺ addresses such constructions by considering the concept of structured argumentation framework.

Definition 7. Let AT be an argumentation theory (AS, KB) . A *structured argumentation framework* (SAF) defined by AT , is a triple (\mathcal{A}, C, \leq) where \mathcal{A} is the smallest set of all finite arguments constructed from KB in AS satisfying Def. 3; \leq is an ordering on \mathcal{A} ; $(X, Y) \in C$ iff X attacks Y .

Notice that a structured argumentation framework is defined with respect to finite arguments, and thus infinite arguments are ignored. Eventually, a Dung framework can then be instantiated with ASPIC⁺ arguments and the ASPIC⁺ defeat relation.

Definition 8. An *abstract argumentation framework* corresponding to a SAF = (\mathcal{A}, C, \leq) is a pair (A, D) such that D is the defeat relation on \mathcal{A} determined by (\mathcal{A}, C, \leq) .

From this argumentation graph, the justified arguments can be computed, using standard definition of arguments, acceptable arguments and extensions in Dung's abstract argumentation semantics [8]. In this paper, we will focus on the grounded extension. A conclusion φ is *justified* if, and only if, at least one argument, whose conclusion φ is in the grounded extension.

3. Defeasible Logic (DL)

Knowledge in DL is a triple $(F, R, >)$ where F is a finite set of facts, R is a finite set of rules and $>$ is a binary relation on R called superiority relation. In expressing the proof we consider only propositional rules. Rules containing free variables are interpreted as the set of their variable-free instances. There are three types of rules: (i) *Strict rules*, (ii) *Defeasible rules*, and (iii) *Defeaters*. The definition of *Strict rules* and *defeasible rules* in DL are essentially the same as in ASPIC⁺; *defeaters* are a special kind of rules that can only prevent some conclusions, but not actively support them. For example, the rule $heavy(X) \rightsquigarrow \neg flies(X)$ states that an animal being heavy is not sufficient enough to conclude that it does not fly. It is only evidence against the conclusion that a heavy animal flies.

A *superiority relation* on R is a relation $>$ on R . Where $r_1 > r_2$, then r_1 is called superior to r_2 and r_2 is inferior to r_1 , which express that r_1 may override r_2 . The superiority relation indicates the relative strength of two rules. For example, given the defeasible rules:

$$r_1: bird(X) \Rightarrow flies(X) \quad r_2: brokenWing(X) \Rightarrow \neg flies(X)$$

which contradict one another. If the superiority relation is empty we are not able to determine which of the two rules prevails over the other. Hence, no conclusive decision can

be made about whether a bird with broken wings can fly. But if we introduce a superiority relation $>$ with $r_2 > r_1$, with the intended meaning that r_2 is strictly stronger than r_1 , then we can indeed conclude that the bird cannot fly.

In the following, we use $A(r)$ to denote the set of literals that appears in the body of the rule r , R_s (respectively, R_d) denotes the set of strict (defeasible) rules in R , and $R[q]$ denotes the set of rules with head q .

We now give a short informal presentation of how conclusions are drawn in DL. Let D be a theory in DL (described above). A *conclusion* of D is a tagged literal and can have one of the following four forms: (i) $+\Delta l$ meaning that we have a definite derivation of l ; (ii) $-\Delta l$ meaning that we do not have a definite derivation of l ; (iii) $+\partial l$ meaning that we have a defeasible derivation of l ; (iv) $-\partial l$ meaning that we do not have a defeasible derivation of l .

Provability is defined below grounded on the concept of a derivation (or proof, which is a finite sequence of tagged literals) in a DL theory D . Given a proof P we use $P(n)$ to denote the n -th element of the sequence, and $P(1..n)$ denotes the first n elements of P . Given a DL theory D and a proof tag $\#$, $\#(D)$ denotes the set of literals provable with tag $\#$ in D .

Definition 9. Given a DL theory D and a proof tag $\# \in \{\Delta, \partial\}$, we have the following:

- A rule $r \in R$ is $\#$ -applicable at $n + 1$ if, and only if, $\forall l \in A(r)$, $+\#l \in P(1..n)$.
- A rule $r \in R$ is $\#$ -discarded at $n + 1$ if, and only if, $\exists l \in A(r)$, $-\#l \in P(1..n)$.

The definition above means that a rule is applicable (at $n + 1$) if all its antecedent are provable; or discarded if at least one of the literals in the antecedent has been rejected in the derivation.

Strict (or definite) derivations are obtained by forward chaining of facts and strict rules while a defeasible conclusion q can be derived if there is a rule whose conclusion is q , whose prerequisites (antecedent) have either already been proved or given in the case at hand (i.e. facts), and any stronger rule whose conclusion is $\sim q$ has prerequisites that fail to be derived. In other words, a conclusion q is (defeasibly) derivable when:

- $+\partial$) If $P(n + 1) = +\partial q$ then either
- (1) $+\Delta q \in P(1..n)$; or
 - (2) $-\Delta \sim q \in P(1..n)$, and
 - (1) $\exists r \in R_{sd}[q]$ such that r is ∂ -applicable, and
 - (2) $\forall s \in R[\sim q]$ either
 - (1) s is ∂ -discarded; or
 - (2) $\exists t \in R_{sd}[q]$ such that t is ∂ -applicable and $t > s$.

The inference conditions for negative proof tags ($-\Delta$ and $-\partial$) are derived from the inference conditions for the corresponding positive proof tags by applying the Principle of Strong Negation introduced by [2]: the strong negation of a formula is closely related to the function that simplifies it by moving all negations to an innermost position in the resulting formula and replace the positive tags with the respective negative tags and vice-versa.

3.1. Ambiguity Blocking and Ambiguity Propagation

A conclusion is ambiguous if there are arguments for it and arguments for its opposite and there are no ways to solve the conflict. Consider, for instance, the following example.

Example 1 (Presumption of Innocence). Consider a **DL theory** with the following rules.

$$\begin{array}{ll} r_1 : \text{evidence}A \Rightarrow \neg \text{responsible} & r_3 : \text{responsible} \Rightarrow \text{guilty} \\ r_2 : \text{evidence}B \Rightarrow \text{responsible} & r_4 : \quad \quad \Rightarrow \neg \text{guilty} \end{array}$$

and there is no additional information to determine the strength of the rules (i.e., in **DL** the superiority relation is empty, and there are no preferences on the rules in **ASPIC⁺**). Given both *evidenceA* and *evidenceB*, the literal *responsible* is ambiguous since the rules r_1 and r_2 , each supporting the negation of the other, are applicable and of the same strength. As a consequence, r_3 is not an applicable rule supporting the *guilty* verdict. We refer to this behaviour as *ambiguity blocking* since the support of *guilty* is blocked by *responsible*, which is the default semantics of **DL**. Accordingly, we obtain $+\partial\neg\text{guilty}$.

Notice that there are no justified arguments in the grounded extension, thus $\neg\text{guilty}$ is not a skeptical conclusion in **ASPIC⁺**.

However, in some cases, it may be preferable for ambiguity to be propagated from *responsible* to *guilty* since we are reserving the judgment of whether the literal *responsible* is provable or not, but possibly it could be. Consequently the literals *guilty* and $\neg\text{guilty}$ are ambiguous; hence an undisputed conclusions cannot be drawn, and we refer to this behaviour as *ambiguity propagation*. Considering the example above, is it appropriate to say that we have reached a not guilty verdict without any reasonable doubt given the fact that the defendant was responsible has not been refuted?⁴

The ambiguity propagation variant of **DL**, for which we use δ as defeasible proof tag, can be easily achieved by making minor changes to the inference conditions for $+\partial$, as shown below [2].

$$\begin{array}{ll} +\delta) \text{ If } P(n+1) = +\delta q \text{ then either} & +\sigma) \text{ If } P(n+1) = +\sigma q \text{ then either} \\ (1) +\Delta q \in P(1..n); \text{ or} & (1) +\Delta q \in P(1..n) \text{ or} \\ (2) -\Delta\sim q \in P(1..n), \text{ and} & (2) (1) -\Delta\sim q \in P(1..n), \text{ and} \\ (1) \exists r \in R_{sd}[q], r \text{ is } \delta\text{-applicable, and} & (2) \exists r \in R_{sd}[q] \text{ such that} \\ (2) \forall s \in R[\sim q] \text{ either} & (1) r \text{ is } \sigma\text{-applicable, and} \\ (1) s \text{ is } \sigma\text{-discarded; or} & (2) \forall s \in R[\sim q] \text{ either} \\ (2) \exists t \in R_{sd}[q] \text{ such that} & s \text{ is } \delta\text{-discarded or } s \not\prec r. \\ t \text{ is } \delta\text{-applicable and } t > s. & \end{array}$$

Their explanation is similar to that of $+\partial$. The major difference is that to prove q this time we make it easier to attack it (clause 2.2). Instead of asking that the arguments attacking it are justified arguments, we just ask for defensible arguments, that is rules whose premises are just supported (i.e., there is a valid chain of reasoning leading to it), denoted by $+\sigma$.

Example 1 (continued). Under *ambiguity propagation*, we obtain $+\sigma\text{guilty}$ and $+\sigma\neg\text{guilty}$ as they are all supported in the theory. Hence, we obtain $-\delta\text{guilty}$ and $-\delta\neg\text{guilty}$.

The question of the example above is whether it is appropriate to say that we have reach a not guilty verdict without any reasonable doubt. The evidence supporting that the defendant was responsible has not been refuted.

Example 2. Let us extend the previous example. Suppose that the legal system allows for compensation for wrongly accused people. A person (defendant) has been wrongly accused

⁴For an in-depth discussion of ambiguity propagation and ambiguity blocking in the context of legal reasoning and their relationships with proof standards see [9].

if the defendant is found innocent, where innocent is defined as \neg guilty. In addition, by default, people are not entitled to compensation. The additional elements of this scenario are modelled by the following rules:

$$\begin{array}{ll} r_5 : \neg\text{guilty} \Rightarrow \text{innocent} & r_6 : \text{innocent} \Rightarrow \text{compensation} \\ r_7 : & \Rightarrow \neg\text{compensation} \end{array}$$

where $r_6 > r_7$.

So, if we adopt **ambiguity blocking**, then we have that despite there is some doubt about responsibility and, consequently, we cannot rule out that the defendant was wrongly accused, the conclusion is that the defendant is entitled to be compensated for having been wrongly accused. **Ambiguity propagation** does not allow us to draw the same conclusion; in fact we have $-\delta\text{compensation}$.

3.2. Team Defeat

The proof conditions above incorporate the idea of team defeat. That is, an attack on a rule with head l by a rule with head $\neg l$ may be defeated by a different rule with head l .

Example 3. Suppose that a crusader has been given the order by his captain not to kill the enemy, and by his general to kill the enemy. Moreover his priest told that they should not kill the enemy, but the bishop told them to kill the enemy. The theory modelling this scenario contains the rules:

$$\begin{array}{ll} r_1 : \text{general} \Rightarrow \text{kill} & r_2 : \text{bishop} \Rightarrow \text{kill} \\ r'_1 : \text{captain} \Rightarrow \neg\text{kill} & r'_2 : \text{priest} \Rightarrow \neg\text{kill} \end{array}$$

the facts are *general*, *bishop*, *captain* and *priest*; and the superiority relation is $r_1 > r'_1$ and $r_2 > r'_2$. All rules are applicable, so we can argue pro *kill* using r_1 , then we have to consider all possible attacks to it. r'_1 is defeated by r_1 itself and r'_2 is defeated by r_2 . So *kill* is justified (i.e., $+\partial\text{kill}$) since for every reason against this conclusion there is a stronger reason defeating it (r_1 and r_2 respectively).

Alternatively, we can say that there are two distinct hierarchies of rules both converging to the same conclusion. It is easy to verify that there are no justified arguments concluding *kill* in the grounded extension of the theory when the preference over the rules is the same as the superiority relation in **DL**, thus *kill* is not a skeptical conclusion in **ASPIC⁺** under the grounded semantics.

Even though the idea of team defeat is natural, it is worth noting that it is not adopted by many related systems and concrete systems of argumentation. On the other hand, the notion of accrual of arguments [19] is gaining more prominence, and team defeat is a form of accrual (albeit one that can only strengthen the arguments in the team).

In case this feature is not desired, **DL** provides variants of the proof conditions given so far to reject it. The proof conditions for the variants without team defeat can be obtained from the corresponding proof conditions given above with the following changes [6]:

- For $+\partial$ and $+\delta$, clause (2.2.2) is replaced by $r > s$; we use $+\partial^*$ and $+\delta^*$, for the proof tags thus obtained.
- For $+\sigma$ and $-\sigma$, the occurrences of $+\delta$ and $-\delta$ is replaced by $+\delta^*$ and $-\delta^*$, for the proof tags thus obtained we use $+\sigma^*$ and $-\sigma^*$, respectively.

Accordingly, to prove a conclusion we must have an applicable rule which is stronger than all applicable/non discarded rules for the negation of the conclusion we want to prove.

The logical properties of **DL** have been thoroughly investigated [3, 6]; in particular the relationships between the various proof tags are stated in the following theorem.

Theorem 1 (Inclusion Theorem). [6, 9] Given a *DL theory* D , we have:

- $+\Delta(D) \subseteq +\delta^*(D) \subseteq +\delta(D) \subseteq +\partial(D) \subseteq +\sigma(D) \subseteq +\sigma^*(D)$;
- $+\delta^*(D) \subseteq +\partial^*(D) \subseteq +\sigma^*(D)$.

There are theories where all the inclusions are proper.

Notice that the conditions for team defeat are more general than the corresponding conditions where this feature does not hold. Besides the set of conclusions we can derive under ambiguity blocking are, in general, different. However, this is not the case for *ambiguity propagation* where one set of conclusions is included in the other as we have two chains of proof conditions, and that the set of conclusions we can derive from one proof tags in one chain are different from the set of conclusions we can derive from a proof tag in the other chain.

As *DL* is skeptical in nature, unless otherwise specified, the discussion below will be focused on skeptical semantics.

4. Acceptability of Arguments: ASPIC⁺ vs Defeasible Logic

As we have seen in the previous section, while *ASPIC⁺* and *DL* share many similarities in both the set of features and inference processes, there are several substantial differences. In this section, we are going to describe some of them.

Both formulae are *relative consistent* (or *indirect consistent* in *ASPIC⁺* term [7]). That is, a theory cannot conclude that both a proposition p and its negation are justified unless they are both supported by the monotonic part (strict rules) of the theory [3]. Researchers on both sides do not consider this notion as a weakness of the logics [5, 21]. Instead, [21] believe that this is a strength to *ASPIC⁺* as this makes a wide range of alternative logical instantiations of *ASPIC⁺* possible. However, both researchers agree that undesirable conclusions could be inferred if inconsistency appears in the monotonic part of the theories.

In general, we have two variants of argumentation semantics of *DL*, namely (i) ambiguity blocking (which corresponds to the semantics of *DL*), (ii) ambiguity propagation (which corresponds to the grounded semantics of Dung's argumentation framework) [11]. *DL* is neutral about *ambiguity blocking* and *ambiguity propagation*. It is possible to justify both views on ambiguity and that both views have their own sphere of applicability. There are applications where *ambiguity blocking* is counterintuitive and there are applications where *ambiguity propagation* is counterintuitive, and there are applications that need both.

The outcome of the discussions here is that a (skeptical) non-monotonic formalism should be able to accommodate both. Through varying the semantics of the proof conditions, *DL* allows us to use the same language without the need to modify a rule/knowledge base to capture different intuitions under different scenarios. Indeed, several variants [2], including support of well-founded semantics, have been defined to cater for the needs of different situations.

Unlike *ASPIC⁺* that support *negation as failure* (NAF), *DL* is an early approach to skeptical non-monotonic reasoning without NAF. That is, *DL* does not support NAF by default. However, it is possible for us to capture this behavior in *DL*. For instance, consider the rule below.

$$r: B, \text{nafa} \Rightarrow q$$

where B is a set of positive literals.

We can transform the weak negated literal *nafa* to *not_a* and introduce new propositions and rule below to simulate the effect of **NAF**.

$$\begin{aligned} r &: B, not_a \Rightarrow q \\ r_a^- &: a \Rightarrow \neg not_a \\ r_a^+ &: \Rightarrow not_a \\ r_a^- &> r_a^+ \end{aligned}$$

In **DL**, conclusions with negative proof tags are generated when the literals is rejected by the theory, and no conclusions will be inferred if the literal is *undecidable* [14]. However, in **ASPIC⁺**, there is no general notion of rejected conclusion. Even though one could say that a conclusion is credulously/skeptically rejected if one of its contraries is credulously/skeptically accepted, then this notion of rejection would again be based on arguments. Consider the theory containing only the rules below.

$$\begin{aligned} p &\Rightarrow p \\ p &\Rightarrow q \\ &\Rightarrow \neg q \end{aligned}$$

DL cannot infer any conclusions as *p* is undecidable unless we reason on the theory using *well-founded* semantics [15]. In such case, *p* will be rejected, subsequently inferring the conclusions $-\delta^*p$, $-\delta^*q$ and $+\delta^*\neg q$. For decisive theories, i.e., theories without undecided literals, the negative extension of a theory (i.e., $\{l : D \vdash -\partial l\}$) is the complement of the positive extension (i.e., $\{l : D \vdash +\partial l\}$). In other terms, if one extends the in/out labelling from arguments to conclusion (see, [4]), then $out(\mathcal{L}) = \mathcal{L} \setminus in(\mathcal{L})$, where $in(\mathcal{L})$ and $out(\mathcal{L})$ are the set of literals in \mathcal{L} labeled in and out, respectively.

On the other hand, since **ASPIC⁺** does not support infinite arguments, there are no arguments about *p* and the state of its conclusion is the same as “before” the argumentation process. Then the question is: what is the default state of *p* that when the argumentation process does not classify as accepted (nor rejected)? It seems that this definition is missing in **ASPIC⁺**.

DL contains a feature called *defeater* (\rightsquigarrow), which can be used to prevent some conclusions from inferred, while **ASPIC⁺** does not. However, this difference is not that significant under (normal) logic programming as we can always transform a **DL theory** with defeater to an equivalent **DL theory** without defeater using the transformation described in [3]. However, this may make a difference in **Modal Defeasible Logic** as defeaters may be used to capture the notion of permission [12].

5. Mapping ASPIC⁺ to DL

In this section we are going to establish a formal relationship between an instantiation of **ASPIC⁺** and **DL**. In particular, we assume: (i) the contrariness relation in **ASPIC⁺** is an involutive negation, (ii) the last-link ordering discussed in Section 2, and (iii) the preference ordering over ordinary premises is empty, i.e., $\leq' = \emptyset$. We prove that **ASPIC⁺** under ground semantics corresponds to the **ambiguity propagation** no team defeat variant of **DL**. To begin with, let's consider the example below which shows the differences of the two formalisms.

Example 4. (extracted from [21]) Consider an argument *A* with a strict top rule for *x* and an argument *B* with a defeasible top rule for $\neg x$, as shown below.

$$\begin{aligned} A &: \Rightarrow p, p \Rightarrow q, q \Rightarrow r, r \rightarrow x \\ B &: \rightarrow d, d \rightarrow e, e \rightarrow f, f \Rightarrow \neg x \end{aligned}$$

It can be observed that A asymmetrically attacks B . So, in ASPIC⁺, x is concluded instead of $\neg x$.

However, the case in DL is a bit different. DL concerns only whether a literal is supported in the inference process, irrespective of the type of rule(s) being used. So, if we infer the above arguments in DL, we have the following conclusions:

$$\begin{array}{ll} D \vdash_{DL} -\Delta x & D \vdash_{DL} -\Delta \neg x \\ D \vdash_{DL} +\sigma^* x & D \vdash_{DL} +\sigma^* \neg x \end{array}$$

That is, both x and $\neg x$ are supported by the DL theory D containing only the rules above (used in arguments A and B) and attack each others with the same strength. Hence, both will be rejected (i.e., $-\delta^* x$ and $-\delta^* \neg x$) in DL. However, if we specify that $r \rightarrow x > f \Rightarrow \neg x$, we are able to conclude $+\delta^* x$.

Hence, despite the similarities, it is not possible to use directly an ASPIC⁺ knowledge base as a DL theory and the other way around. This is due to the treatment of (defeasible) arguments in DL which involve strict rules.

To establish the correspondence between ASPIC⁺ and DL we introduce a mapping from ASPIC⁺ theories to DL theories, based on the ambiguity propagation variant of DL without team defeat, as shown in the definition below. We assume that the same propositional language \mathcal{L} has been used in both ASPIC⁺ and DL.

Definition 10. Let $AT = ((\mathcal{L}, \mathcal{R}, n), \mathcal{K})$ be an ASPIC⁺ theory and $D = (F, R, >)$ be a DL theory. An argument mapping is a function $D = T(AT)$ that map an argument in AT to rules in DL, such that:

$$\begin{aligned} F &= \mathcal{K}_n \\ R &= \{r : \Rightarrow q \mid q \in \mathcal{K}_p\} \cup \mathcal{R} \\ > &= \{r > s \mid (s \leq r) \in \leq\} \cup \\ &\quad \{r > s \mid r \in \mathcal{R}_s[q], s \in \mathcal{R}_d[\sim q]\} \cup \\ &\quad \{r > s \mid r \in \mathcal{R}[\sim q], s \in R[q] \text{ such that } q \in \mathcal{K}_p\} \end{aligned}$$

In the transformation, knowledge in \mathcal{K} has been transformed into different features in DL according to their nature. For instance, axioms (\mathcal{K}_n) are information that cannot be defeated and will be mapped into facts directly (in DL) without any transformation; while ordinary premises (\mathcal{K}_p) are information that can be defeated when arguments with stronger support appear, and are transformed into defeasible rules.

Regarding the preference order, note that besides including all preference order that appears in \leq , the transformation includes also the superiority relations between defeasible rules and their conflicting strict rules in \mathcal{R} , and those rules that are generated (in the transformed theory) based on the ordinary premises (\mathcal{K}_p).⁵ The former is used to ensure that the support of literal in the defeasible rule can be blocked (under superiority relation) when applicable conflicting strict rules are appeared during the inference process; whereas the latter is used to defeat ordinary premises with a stronger argument.

We are now prepared to give the relationship between ASPIC⁺ and DL.

Theorem 2. Let $AT = ((\mathcal{L}, \mathcal{R}, n), \mathcal{K})$ be an ASPIC⁺ argumentation theory and $p \in \mathcal{L}$,

- (i) $AT \vdash_{A+}^{GS} p \iff T(AT) \vdash_{DL} +\Delta p$
- (ii) $AT \vdash_{A+}^{GS} p \iff T(AT) \vdash_{DL} +\delta^* p$

⁵Note that the $R[q]$ in the last case of $>$ refers to the rules introduced due to the ordinary premises \mathcal{K}_p .

where $AT \vdash_{A+}^{GS} p$ and $AT \vdash_{A+}^{GS} p$ means that p is *strictly* and *defeasibly* justified in the argumentation theory AT using the grounded semantics in ASPIC⁺, respectively.

Proof. (sketch) The proof is by induction on the length of a derivation in DL and the number of iterations of the application of the characteristic function \mathcal{F}_G in the construction of the fixed-point of the set of acceptable arguments. The inductive base is straightforward given that the base of acceptability for ASPIC⁺ is whether a literal is an axiom in \mathcal{K}_n or not, and for DL is being a fact or not. But facts in the DL theory corresponds to the axioms in ASPIC⁺ argumentation system. For the inductive step we first notice that $+\sigma p$ means that, in ASPIC⁺, there is an undefeated argument for p , and that the argument is not undercut (all the antecedents are under the inductive hypothesis) and the last step is to see that there are not attacking (undefeated) arguments for $\sim p$. \square

As can be seen, an ASPIC⁺ argumentation system can be transformed into a DL theory by applying the transformations above. It is immediate to see that the mapping from a ASPIC⁺ argumentation theory to the corresponding DL theory is, in the worse case, *quadratic*, given that we have to consider the relationship between conflicting rules and arguments to derive the superiority relations. Hence, given that the complexity of computing the extensions of DL is linear w.r.t. the size of the theory [13], we have the following result.

Corollary 3. *Acceptability of a proposition in ASPIC⁺ under grounded semantics can be computed in polynomial time.*

6. Conclusions

In this paper we addressed the question of how to instantiate ASPIC⁺ in DL. For the other direction, it is possible to capture the ambiguity propagation no team defeat variant of DL in ASPIC⁺, given that such a variant of DL is characterised by the grounded semantics and, the two formalisms share the same language. Thus a theory in DL is indistinguishable from an argumentation theory in ASPIC⁺. Moreover, other variants are characterised by skeptical argumentation semantics different from grounded semantics, and, to the best of our knowledge, the relationships between such semantics and ASPIC⁺ have not been studied.

While it is possible to adopt different argumentation semantics to be applied on top of ASPIC⁺, this step alone might not be enough to model defeasible logic as an instance of ASPIC⁺. For example, DL with ambiguity blocking would require to introduce a second “attack” relation on arguments (see [11]) with a ripple down effects on the ASPIC⁺ definitions setting the various statuses of the argument. Similarly, DL with team defeat would require changes in the definition of what arguments are: an argument would be a set of proof trees instead of a single proof tree [10]. In this paper, we do not address such issues. However, they show that there is potential for cross-fertilization for research on the relationship between ASPIC⁺ and DL.

References

- [1] L. Amgoud, L. Bodenstaff, M. Caminada, P. McBurney, S. Parsons, H. Prakken, J. van Veenen, and G. Vreeswijk. Final review and report on formal argumentation systems. Technical Report Deliverable D2.6, ASPIC IST-FP6-002307, 2006.

- [2] G. Antoniou, D. Billington, G. Governatori, and M. J. Maher. A Flexible Framework for Defeasible Logics. In *Proceedings of the 17th National Conference on Artificial Intelligence*, pages 405–410. AAAI Press / The MIT Press, 2000.
- [3] G. Antoniou, D. Billington, G. Governatori, and M. J. Maher. Representation Results for Defeasible Logic. *ACM Transactions on Computational Logic*, 2(2):255–286, 2001.
- [4] P. Baroni, G. Governatori, H.-P. Lam, and R. Riveret. On the Justification of Statements in Argumentation-based Reasoning. In J. Delgrande and F. Wolter, editors, *Proc KR 2016*. AAAI Press, Cape Town, South Africa, Apr. 2016.
- [5] D. Billington. Defeasible Logic is Stable. *Journal of Logic and Computation*, 3(4):379–400, 1993.
- [6] D. Billington, G. Antoniou, G. Governatori, and M. J. Maher. An Inclusion Theorem for Defeasible Logics. *ACM Transactions on Computational Logic*, 12(1):6:1–6:27, Nov. 2010.
- [7] M. Caminada and L. Amgoud. On the evaluation of argumentation formalisms. *Artificial Intelligence*, 171(5-6):286 – 310, 2007.
- [8] P. M. Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77(2):321–358, 1995.
- [9] G. Governatori. On the relationship between Carneades and defeasible logic. In *Proc ICAIL 2011*, pages 31–40. ACM, 2011.
- [10] G. Governatori and M. J. Maher. An Argumentation-Theoretic Characterization of Defeasible Logic. In W. Horn, editor, *Proc ECAI 2000*, pages 469–474, 2000.
- [11] G. Governatori, M. J. Maher, G. Antoniou, and D. Billington. Argumentation Semantics for Defeasible Logic. *Journal of Logic and Computation*, 14(5):675–702, 2004.
- [12] G. Governatori, F. Olivieri, A. Rotolo, and S. Scannapieco. Computing Strong and Weak Permissions in Defeasible Logic. *Journal of Philosophical Logic*, 42(6):799–829, 2013.
- [13] M. J. Maher. Propositional Defeasible Logic has Linear Complexity. *Theory and Practice of Logic Programming*, 1(6):691–711, 2001.
- [14] M. J. Maher. Relative expressiveness of defeasible logics. *Theory and Practice of Logic Programming*, 12(4-5):793–810, 2012.
- [15] M. J. Maher and G. Governatori. A Semantic Decomposition of Defeasible Logics. In *Proc AAAI-99*, pages 299–305, Menlo Park, CA, USA, 1999. AAAI Press.
- [16] S. Modgil and H. Prakken. A General Account of Argumentation with Preferences. *Artificial Intelligence*, 195:361–397, Feb. 2013.
- [17] S. Modgil and H. Prakken. The ASPIC⁺ framework for structured argumentation: a tutorial. *Argument & Computation*, 5:31–62, 2014.
- [18] D. Nute. Defeasible Logic. In D. Gabbay and C. Hogger, editors, *Handbook of Logic for Artificial Intelligence and Logic Programming*, volume III, pages 353–395. Oxford University Press, 1994.
- [19] H. Prakken. A Study of Accrual of Arguments, with Applications to Evidential Reasoning. In *Proc ICAIL 2005*, pages 85–94, New York, NY, USA, 2005. ACM.
- [20] H. Prakken. An abstract framework for argumentation with structured arguments. *Argument & Computation*, 1(2):93–124, 2010.
- [21] H. Prakken and S. Modgil. Clarifying some misconceptions on the ASPIC⁺ framework. In B. Verheij, S. Szeider, and S. Woltran, editors, *Proc COMMA 2012*, pages 442–453. Vienna, Austria, Sept. 2012.

Argument Analytics

John LAWRENCE ^a, Rory DUTHIE ^a, Katarzyna BUDZYNSKA ^{a,b}, and Chris REED ^a

^a*Centre for Argument Technology, University of Dundee, UK*

^b*Institute of Philosophy and Sociology, Polish Academy of Sciences, Poland*

Abstract. Rapid growth in the area of argument mining has resulted in an ever increasing volume of analysed argument data. Being able to store information about arguments people make in favour or against different opinions, decisions and actions is a highly valuable resource, yet extremely challenging for sense-making. How, for example, can an analyst quickly check whether in a corpus of citizen dialogue people tend to rather agree or disagree with new policies proposed by the department of transportation; how can she get an insight into the interactions typical of this specific dialogical context; how can the general public easily see which presidential candidate is currently winning the debate by being able to successfully defend his arguments? In this paper, we propose Argument Analytics – a suite of techniques which provide interpretation of, and insight into, large-scale argument data for both specialist and general audiences.

Keywords. Argument Interchange Format, corpus resources, argument visualisation

1. Introduction

Over the last decade, a lot of effort has been made to provide various tools for argument analysis, evaluation and visualisation (*cf.* [6]). Systems cover various domains of applications such as analysis and visualisation of reasoning structure (e.g. Carneades [5], Rationale [13]), visualisation of debates [10] or evaluation of argument acceptability [4, 12]. The Argument Interchange Format (AIF) [2], supports exchange and data reuse between these tools, and is currently the only mechanism that handles linguistic, structural and abstract facets of argumentation. Though AIF representations are scalable, analytical tools struggle at scale: for example, an OVA analysis of a single, 45-minute episode of the BBC Radio 4 program *Moral Maze*¹ contains over 200 statements with a similar number of connections between them²; comparable large-scale analyses can be created with Rationale, Carneades and so on. Whilst such maps enable us to follow chains of reasoning, and answer questions about the relationships between individual points, they fall short of providing clear insight into the nature of what is taking place within the argument. With rapid growth of corpora of analysed arguments resulting in part from increased coherence in analysis techniques, and in part from improving results from argument mining systems, the task of making sense out of argument data is becoming increasingly important. Something more than mere visualisation is required.

¹<http://www.bbc.co.uk/programmes/b006qk11>

²<http://www.arg-tech.org/AIFdb/argview/789>

Argument Analytics provides a suite of techniques for analysing AIF data, with components ranging from the detailed statistics required for discourse analysis or argument mining, to infographic-style representations, offering insights in a way that is accessible to a general audience. The extensible set of modules currently comprises: simple statistical data (Section 3), which provides both an overview of the argument structure and frequencies of patterns such as argumentation schemes; comparative data (Section 4) providing a range of measures describing the similarity of two analyses; dialogical data (Section 5) highlighting the behaviour of participants of the dialogue; and real-time data (Section 6) allowing for the graphical representation of a developing over time argument structure. Together these analytics open an avenue to giving feedback on live debates, producing summaries of deliberative democracy, mapping citizen science, and more.

2. Foundations

The Argument Analytics platform is designed specifically for making sense out of argument data represented according to the Argument Interchange Format(AIF) [2] such as the data stored in the AIFdb³ database [7]. The Social Layer [11] is used to enrich this data, providing details on participants such as biographies. AIFdb Corpora enables Argument Analytics to display the interpretations of data, whether on a single AIF argument map (stored in AIFdb as a NodeSet), or a large corpus containing hundreds or thousands of such AIF representations.

The AIF was developed as a means of describing argument networks that would provide a flexible, yet semantically rich, specification of argumentation structures. These networks are comprised of seven types of node:

Node Type	Description
I	propositional information contained in an argument, such as a conclusion, premise, data etc.
L	subset of I-nodes referring to propositional reports specifically about discourse events
RA	application of a scheme of reasoning or inference
CA	application of a scheme of conflict
MA	application of a scheme of rephrasing
YA	application of a scheme of illocution describing communicative intentions which speakers use to introduce propositional contents
TA	application of a scheme of interaction or protocol describing relations between locutions

3. Simple Statistics

The simple statistics modules allows an analyst to quickly make sense of a large amount of annotated argument data. Although these calculations are straightforward and relatively easy to automate, they nevertheless provide interesting insights into the data.

³<http://www.aifdb.org>

The **overview** page shows a range of statistics, offering a rapidly digested summary of the overall argumentative structure. The number of Information nodes provides an indication of the overall size of the analysis. The average number of words per Information Node illustrates the complexity of the ideas presented, and how succinctly they are expressed. The numbers of inference (RA) and conflict (CA) nodes give a suggestion as to the nature of the dialogue, which is further expanded by showing the ratios of RA to CA (capturing how diverse are the perspectives in the debate) and RA to I (how dense the argumentation is).

The **Pattern Count** modules expand on the overview to give detailed statistics suitable for more in-depth argument and discourse analysis. They provide the frequencies of commonly occurring patterns, split into two categories. Firstly, argumentative and illocutionary patterns which describe both the nature of the interactions, for example levels of agreement and disagreement, and the way in which participants have expressed themselves and interacted with each other, such as how frequently a participant questions the statements of others compared to how frequently they assert their own views. The second category, dialogical patterns, illustrates the flow of the discourse and gives an indication of any dialogical rules, either explicit or implicit, to which the participants are conforming. Such dialogical patterns are also useful, for instance, to show cross-cultural differences in dialogue, or differences in the formality and setting of dialogues.

4. Comparative Statistics

The comparative statistics modules [3] allow for the validation of both manual and automatic analysis for argument mining (*cf.* [8, 9]). Such calculations enable comparison between two manual analyses to determine the efficacy of annotation guidelines via inter-annotator agreement, or the comparison of results from automatic techniques to a manually created gold standard. The examples given in this section refer to two human annotators, but in each case the same calculations could be applied with one of these being an annotation produced by an automatic system.

There are a number of considerations that must be taken into account when calculating agreement or results, such as what effect a differing segmentation of the original text, in two separate annotations, may have on the assignment of inference and conflict in an argument structure. To account for this, the agreement and results calculations were split into smaller sub-calculations covering segmentation similarity, propositional contents (inference and conflict) and dialogical contents (locutions). Calculating agreement for segmentation of argumentative units is a challenging task [14]. The modular architecture of Argument Analytics allows for a range of measures to be displayed, and currently differences are accounted for using various segmentation similarity algorithms, which give an overall normalised score for the similarity. Propositional contents are compared by separating nodes from the text and instead using the Levenshtein distance for the matching of nodes. Dialogical contents are compared in the same way with word ordering added to the Levenshtein distance for node matching and with the addition of added calculations for the intricacies of dialogue (see [3] for an in-depth description of the comparative statistics module).

5. Dialogically Oriented Statistics

For those argument analyses where there is a dialogue taking place between multiple participants, a range of dialogically oriented, analytics modules are able to provide insights into the dynamics of the discourse, and make these complex interactions accessible to a general audience. There is growing demand to present complex argumentative structures to a broad audience in ways which are both intuitive and interactive. Whilst there is some progress towards this goal, for example, the Election Debate Visualisation Project [10], many of these approaches rely on custom, genre-specific interfaces for both the elicitation and display of argumentative structure. Dialogically oriented, analytics modules make use of both the locution details stored in AIFdb, as well as the participant details provided by the Argument Web social layer.

Each of the modules in this section are illustrated using data from an episode of the BBC Radio 4 program *Moral Maze*⁴. These examples show how such graphical displays of information can take the technical details captured in the argumentative structure of a complex debate, and present them in ways which are easily processed by a general audience.

5.1. Structural Statistics

The structural statistics modules extract particular facets of the argumentative structure in order to display data such as who is speaking most, which pairs of participants are interacting most and who is making the most well supported arguments. As such, they provide a greater insight into the argumentative structure than that which is afforded by looking at a simple argument map of the same data.

Participation: For each participant, the number of locutions they have made is counted and represented in a bar chart. This provides an easy way of identifying which participants were most, and least, dominant within a dialogue. An example can be seen in Figure 1, which shows that Jan Macvarish was the most active participant in this dialogue with twenty-three locutions, whereas Matthew Taylor was least active with only one locution made.

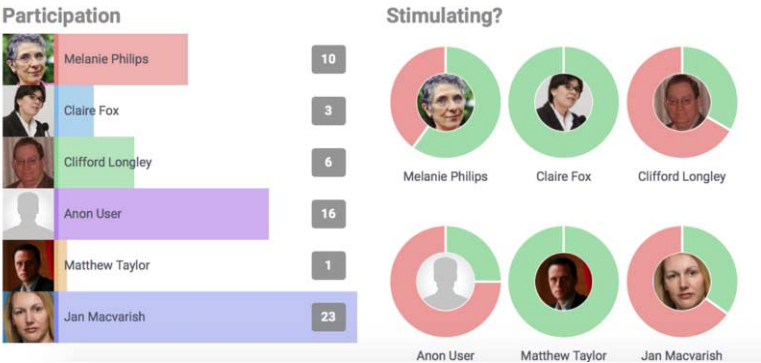


Figure 1. Graphical representations of the relative involvement of each participant in a dialogue, and how stimulating the points made by each participant are.

⁴<http://www.bbc.co.uk/programmes/b006qk11>

Stimulating: A point of debate is stimulating if it receives responses, either to agree or disagree. From the analysed argument structure, we count the number of locutions which each participant has made that have at least one response, and those which have been ignored by the other participants. The example in Figure 1 shows that whilst Claire Fox has only made three locutions, they have all been responded to in some way, whereas, of the six locutions made by Clifford Longley, only two received any attention from the other participants.

Chord Diagram: The chord diagram shows the interaction between participants. A chord diagram is a graphical method of displaying the inter-relationships between data in a matrix. The data is arranged radially around a circle with the relationships between the points drawn as arcs connecting the data together. In this case, the arcs represent interaction between participants, with the width of the arc at each end representing the number of locutions made by that participant to which the connected participant has responded. Viewing the interactions in this way makes it easy to identify, for example, cliques. An example chord diagram can be seen in Figure 2. Clicking on a specific participant emphasises their connections with other participants. For example, with Melanie Philips selected (as shown on the right of the figure), we can see that the majority of her interactions were with Jan Macvarish, reflecting the fact that, for a period of the dialogue, Melanie was questioning Jan.

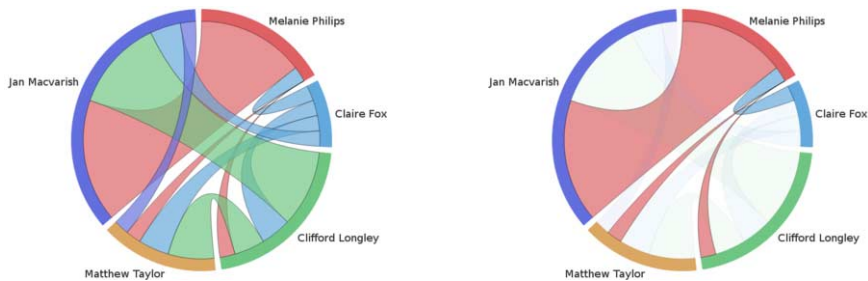


Figure 2. Chord diagrams showing the frequency of interactions between participants. The diagram on the right shows Melanie Philips selected, highlighting just those interactions in which she is involved.

Verbosity: Similar to the average number of words per I-node presented in the overview, verbosity shows a comparison of the average length of locutions made by each participant. By comparing in this way, we are able to see not just the overall complexity of the ideas expressed, but also how prolix or concise each participant is in presenting their ideas.

5.2. Temporal Statistics

Temporal statistics use the time-stamping of locutions provided by AIFdb to show how the state of a dialogue has altered as it has progressed. These statistics provide clues, not easily discernible from an argument map, as to when individual participants have been most involved in the dialogue, when conflict has arisen, and changes in topic that have occurred as the dialogue progresses.

Turn Structure: Using the timestamping of locutions provided by AIFdb, a graphical representation of the turn structure in a dialogue is created. This visualisation pro-

vides a quick overview of when each participant has been most active, suggesting details of any pre-defined turn-taking rules. The example shown in Figure 3 reflects the turn structure in a *Moral Maze* episode. As the episode begins, each of the four regular panelists speak briefly about the topic being discussed. A guest witness is then introduced, and, after providing their own views on the topic, are then questioned by first one of the panelists and then by a second.

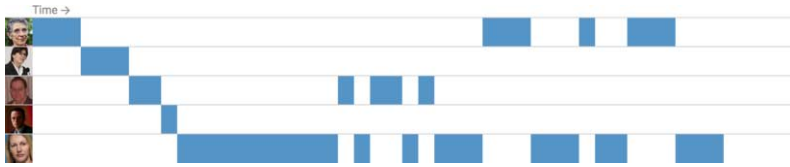


Figure 3. Graphical representation of the turn structure in a dialogue, highlighting the way in which each participant introduces themselves, followed by direct interactions between two pairs of participants.

5.3. Semantics-based Statistics

Semantics-based analytics use Dung-style semantics to determine the acceptability of a participant's arguments. An AIF graph is translated into ASPIC⁺ then, using TOAST, a Dung-style abstract argumentation framework is derived and evaluated.

Defended: The defended points in a dialogue, are those where conflicting points have been made, but these conflicting points have, in turn been attacked. It is easy in a broad ranging and complex dialogue for points to be made which are not challenged either due to going unnoticed, or being simply dismissed. By looking at those points which are challenged and then later defended we gain an insight into both the validity of a point, and how crucial it is to the argument which a participant is making.

Sway: Where one participant has more acceptable arguments than another, the former is said to carry more sway. This value is calculated for each participant, and displayed as the relative balance in sway between each pair of the most commonly interacting participants. This can, to some extent, be viewed as who is winning in a debate; best supporting their own points and best attacking the points made by the other participants in the dialogue.

6. Real-time Statistics

Many of the modules used in Argument Analytics have the ability to not only display data on a fixed, pre-analysed argument structure, but to update in real-time as the structure evolves. This functionality has been used, for example, in a tool developed for the *Built Environment for Social inclusion through the Digital Economy (BESiDE)* project⁵, to facilitate round table discussions between architects working on the design of care environments, and the various stakeholders involved in the design process.

As the discussion is taking place, the audio is recorded and an analyst uses a custom-designed interface to segment the dialogue when either the topic or the speaker changes.

⁵<http://beside.ac.uk/>

A simple dialogue protocol is used, allowing participants to make moves of various types (e.g. asking questions, agreeing with another participant, and offering their own opinion), and relating to a set of pre-defined topics relevant to the design project.

Throughout the discussion, the dialogue overview shown in Figure 4 is displayed for all participants to see. This overview includes a transcript of the dialogue on the right hand side, and analytics modules displaying how much each participant has spoken, and which topics have been discussed on the left. In testing these interfaces, it is interesting to see that they serve not only an informative function, but actually impact the dynamics of the dialogue. When a participant can see that they are talking more than everyone else, they tend to let others speak more. When someone hasn't spoken yet, the other participants notice this, and make an effort to direct questions at them. And, when one topic has been less explored than the others, there is a noticeable shift towards that area in both the questions asked and the points raised.

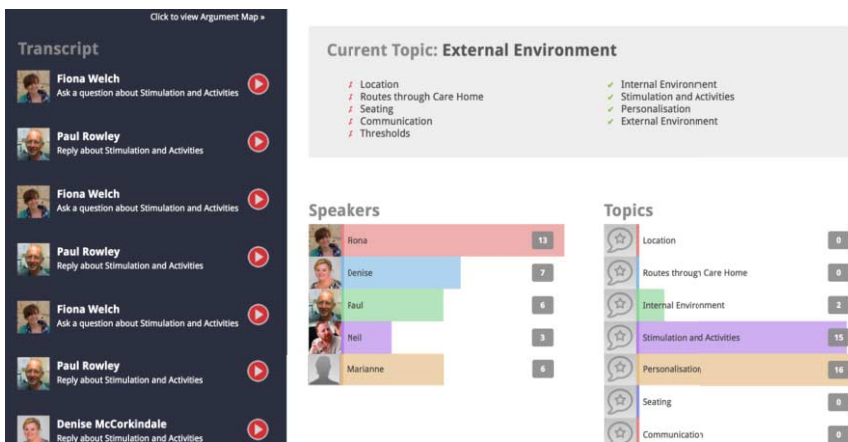


Figure 4. Real-time Argument Analytics highlighting the involvement of individual participants and the topics discussed.

This ability for the argumentative and dialogical structure to, not only represent the outcome of a discussion, but to inform the participants and help ensure that all areas are fully explored has wide ranging potential applications. The current limitation to providing this kind of interface more widely is the ability to perform real-time analysis, but as tools, such as the AnalysisWall [1] which has been used to analyse several hour-long radio programmes in real time, improve, and automatic argument mining techniques develop, it is easy to imagine such a live display accompanying activities such as debates, meetings and media coverage.

7. Conclusions

The Argument Analytics suite provides a comprehensive range of analytic tools from the detailed statistics required for discourse analysis, to graphic visual representations making the same data accessible to a general audience. The existing modules which we have described offer solutions to a broad range of potential user groups, including those involved in argument analysis and critical discourse analysis, those working on argument

mining applications, people performing political or social studies, and members of the general public who wish to get a greater understanding of the issues and dynamics of a complex debate.

There are a range of existing tools which provide argument analysis and visualisation capabilities, but, by using the ability of AIFdb to translate the output from many of these tools into an Argument Interchange Format compliant representation, Argument Analytics allows their output to be displayed in a far broader range of ways and with a broader range of potential applications than any one of these tools currently provides.

Acknowledgments

We would like to acknowledge that the work reported in this paper has been supported in part by EPSRC in the UK under grants EP/M506497/1, EP/N014871/1 and EP/K037293/1.

References

- [1] Floris Bex, John Lawrence, Mark Snaith, and Chris Reed. Implementing the Argument Web. *Communications of the ACM*, 56(10):66–73, Oct 2013.
- [2] Carlos Chesñevar, Sanjay Modgil, Iyad Rahwan, Chris Reed, Guillermo Simari, Matthew South, Gerard Vreeswijk, Steven Willmott, et al. Towards an Argument Interchange Format. *The Knowledge Engineering Review*, 21(04):293–316, 2006.
- [3] Rory Duthie, John Lawrence, Katarzyna Budzyska, and Chris Reed. The CASS Technique for Evaluating the Performance of Argument Mining. *to appear in Proceedings of the Third Workshop on Argument Mining. Association for Computational Linguistics*, 2016.
- [4] Alejandro J García and Guillermo R Simari. Defeasible logic programming: An argumentative approach. *Theory and practice of logic programming*, 4(1+ 2):95–138, 2004.
- [5] Thomas F Gordon, Henry Prakken, and Douglas Walton. The Carneades model of argument and burden of proof. *Artificial Intelligence*, 171(10):875–896, 2007.
- [6] Paul A Kirschner, Simon J Buckingham-Shum, and Chad S Carr. *Visualizing argumentation: Software tools for collaborative and educational sense-making*. Springer, 2003.
- [7] John Lawrence, Floris Bex, Chris Reed, and Mark Snaith. AIFdb: Infrastructure for the Argument Web. In *Proceedings of the Fourth International Conference on Computational Models of Argument (COMMA 2012)*, pages 515–516, 2012.
- [8] Marie-Francine Moens. Argumentation mining: Where are we now, where do we want to be and how do we get there? In *FIRE '13 Proceedings of the 5th 2013 Forum on Information Retrieval Evaluation*, 2013.
- [9] Andreas Peldszus and Manfred Stede. From argument diagrams to argumentation mining in texts: a survey. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, 7(1):1–31, 2013.
- [10] Brian Plüss and Anna De Liddo. Engaging citizens with televised election debates through online interactive replays. In *Proceedings of the ACM International Conference on Interactive Experiences for TV and Online Video*, pages 179–184. ACM, 2015.
- [11] Mark Snaith, Rolando Medellín, John Lawrence, and Chris Reed. Arguers and the Argument Web. In *Proceedings of the 13th workshop on Computational Models of Natural Argument (CMNA 13)*, Rome, Italy, 2013. Springer.
- [12] Mark Snaith and Chris Reed. TOAST: Online ASPIC+ implementation. In *Proceedings of the Fourth International Conference on Computational Models of Argument (COMMA 2012)*, pages 390–397, Vienna, 2012. IOS Press.
- [13] Tim van Gelder. The rationale for Rationale. *Law, probability and risk*, 6(1-4):23–42, 2007.
- [14] Nina Wacholder, Smaranda Muresan, Debanjan Ghosh, and Mark Aakhus. Annotating multiparty discourse: Challenges for agreement metrics. *LAW VIII*, page 120, 2014.

Argument Mining Using Argumentation Scheme Structures

John LAWRENCE and Chris REED

Centre for Argument Technology, University of Dundee, UK

Abstract. Argumentation schemes are patterns of human reasoning which have been detailed extensively in philosophy and psychology. In this paper we demonstrate that the structure of such schemes can provide rich information to the task of automatically identify complex argumentative structures in natural language text. By training a range of classifiers to identify the individual proposition types which occur in these schemes, it is possible not only to determine where a scheme is being used, but also the roles played by its component parts. Furthermore, this task can be performed on segmented natural language, with no prior knowledge of the text's argumentative structure.

Keywords. Argumentation Schemes, Argument Mining, Natural Language Processing

1. Introduction

The continuing growth in the volume of data which we produce has driven efforts to unlock the wealth of information this data contains. Automatic techniques such as Opinion Mining and Sentiment Analysis [12] allow us to determine the views expressed in a piece of textual data, for example, whether a product review is positive or negative. Existing techniques struggle, however, to identify more complex structural relationships between concepts. By identifying the argumentative structure and its associated premises and conclusions, we are able to tell not just *what* views are being expressed, but also *why* those particular views are held. In this paper, we use argumentation schemes [22], common patterns of human reasoning, to automatically determine instances where such a pattern is being used, as well as the roles played by its component parts.

1.1. Argumentation Schemes

Argumentation schemes capture structures of (typically presumptive) inference from a set of premises to a conclusion and represent stereotypical patterns of human reasoning. As such, argumentation schemes represent a historical descendant of the topics of Aristotle [1] and, much like Aristotle's topics, play a valuable role in both the construction and evaluation of arguments.

Several attempts have been made to identify and classify the most commonly used schematic structures [6,16,9,17,20,5,8,22]. Although these sets of schemes overlap in many places, the number of schemes identified and their granularity can be quite different. As such, most argument analyses tend to contain examples from only one scheme

set, with the Walton set being the most commonly used. Several examples of Walton’s argumentation schemes can be seen in Table 1.

Analogy (AN) <i>Premise [SimilarityOfCases]:</i> Generally, case C1 is similar to case C2 <i>Premise [Precedent]:</i> A is true (false) in case C1 <i>Conclusion:</i> A is true (false) in case C2
CauseToEffect (CE) <i>Premise [Causal]:</i> Generally, if A occurs, then B will (might) occur <i>Premise [Occurrence]:</i> In this case, A occurs (might occur) <i>Conclusion:</i> Therefore, in this case, B will (might) occur
PracticalReasoning (PR) <i>Premise [Goal]:</i> I have a goal G <i>Premise [GoalPlan]:</i> Carrying out this action A is a means to realise G <i>Conclusion:</i> Therefore, I ought (practically speaking) to carry out this action A
VerbalClassification (VC) <i>Premise [ContainsProperty]:</i> a has a property F <i>Premise [ClassificationProperty]:</i> For all x, if x has a property F, then x can be classified as having a property G <i>Conclusion:</i> a has property G

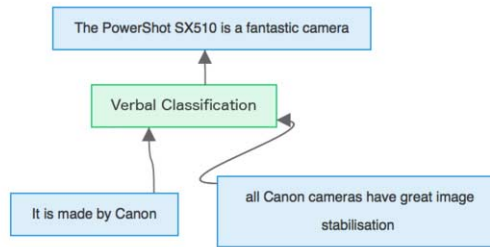
Table 1. Argumentation schemes

Understanding the argumentative structure being expressed in a piece of natural language text can help us gain a deeper understanding of what is being said compared to many existing techniques for extracting meaning. If we consider the product review shown below, then sentiment analysis techniques allow us to understand at a high level what views are being presented, for example, that this review is positive, but are unable to provide details on exactly why the reviewer likes the product.

The PowerShot SX510 is a fantastic camera. It is made by Canon and all Canon cameras have great image stabilisation.

Looking at the argumentative structure contained within this review, we can see that the propositions “It is made by Canon” and “all Canon cameras have great image stabilisation” are working together to support the conclusion “The PowerShot SX510 is a fantastic camera”. Furthermore, we can see that the link between the premises and conclusion is a form of Verbal Classification¹. A graphical representation of the argument structure can be seen in Figure 1.

¹In fact, the example here does not exactly conform to the Verbal Classification scheme. In a more thorough analysis, an enthymeme would be added showing that the premises actually support the fact that the camera has great image stabilisation and that this in turn is a feature of a fantastic camera.

Figure 1. Argument analysis of a product review, showing an example of the Verbal Classification scheme

The features of these common patterns of argument provide us with a way in which to both identify that an argument is being made and determine its structure. By using the specific nature of each component proposition in a scheme, we can identify where a particular scheme is being used and classify the propositions accordingly, thereby gaining a deeper understanding of the argumentative structure which a piece of text contains.

1.2. Argument Mining

Argument Mining² is the automatic identification and extraction of argument components and structure. One of the first attempts to automate this process was presented in [13,18], where a text is first split into sentences and then features of each sentence are used to classify them as “Argument” or “Non-Argument”. This approach was built upon in [14], where each argument sentence is additionally classified as either a premise or conclusion, a method referred to as “Argument proposition classification”. A Context-Free Grammar is then used to determine the internal structure of each individual argument.

The majority of the more recent developments in Argument Mining have followed a similar approach to this early work, applying a range of techniques to uncover the argumentative sections of a text, identifying premises and conclusions and attempting to link these together to determine the overall argument structure. This methodology has been applied to a range of domains including online user comments [15], social media [4] and essays [19], with the results obtained being generally encouraging. However, such attempts do not consider exactly how the discovered premises are working together to support the conclusion.

The concept of automatically identifying argumentation schemes was first discussed in [21] and [3]. Walton proposes a six-stage approach to identifying arguments and their schemes. The approach suggests first identifying the arguments within the text and then fitting these to a list of specific known schemes. A similar methodology was implemented by Feng & Hirst, who produced classifiers to assign pre-determined argument structures as one in a list of the most common argumentation schemes. Another possible approach is suggested in [2], where the connection between argumentation schemes and discourse relations is highlighted, however, this requires these discourse relations to be accurately identified before scheme instances can be determined.

²Sometimes also referred to as Argumentation Mining

The main challenge faced by these approaches is the need for some prior analysis of the text to have taken place. By instead looking at the features of each component part of a scheme, we are able to overcome this requirement and identify parts of schemes in completely unanalysed text. Once these scheme components have been identified, we are able to group them together into specific scheme instances and thus obtain a complete understanding of the arguments being made.

2. Identifying Scheme Components

Being able to determine the argumentation scheme structure contained within a piece of text gives us a much deeper understanding of both what views are being expressed and why those views are held, as well as providing a route to the automatic reconstruction of certain types of enthymeme [7]. However, existing approaches to automatically identifying scheme instances have relied on the basic argumentative structure being previously identified.

By training a range of classifiers to identify the individual components of a scheme, we are able to identify not just the presence of a particular scheme, but also the roles which each of the premises play within a particular scheme instance. Furthermore, we are able to perform this based only on a list of the propositions contained within the text, requiring no previous analysis to have been performed. In Section 2.1 we look at using one-against-others classification to identify propositions of each type from a set of completely unstructured propositions. Being able to successfully perform this task for even one of the proposition types allows us to discover areas of the text where the corresponding scheme likely to be being used. This can be viewed as a first step in obtaining the argument structure following the extraction of propositions from natural text using a technique such as *Proposition Boundary Learning* [11], a specialised type of Elementary Discourse Unit identification.

In Section 2.2, we also consider the situation where some of the argumentative structure has already been determined. If we know that we have a set of premises supporting a conclusion and that a particular scheme is being used, then we wish to determine what role each premise is playing in the scheme. In order to achieve this, we implemented pairwise classifiers for each scheme type capable of classifying each premise into their respective role.

In order to accomplish these tasks, a range of classifiers for each proposition type was implemented using the *scikit-learn*³ Python module for machine learning, with the features described in Table 2. Part Of Speech (POS) tagging was performed using the Python NLTK⁴ POS-tagger and the frequencies of each tag added as individual features. The similarity feature was added to extend the information given by unigrams to include an indication of whether a proposition contains words similar to a pre-defined set of keywords. The keywords used for each type are shown in Table 3. Similarity scores were calculated using WordNet⁵ to determine the maximum similarity between the synsets of the keywords and each word in the proposition. The maximum score for the words in the

³<http://scikit-learn.org/stable/>

⁴<http://www.nltk.org/>

⁵<http://wordnet.princeton.edu/>

proposition was then added as a feature value, indicating the semantic relatedness of the proposition to the keyword.

Feature	Description
Unigrams	Each word in the proposition
Bigrams	Each pair of successive words
Length	The number of words in the proposition
AvgWLength	The average length of words in the proposition
POS	The parts of speech contained in the proposition
Punctuation	The presence of certain punctuation characters, for example “ ” indicating a quote
Similarity	The maximum similarity of a word in the proposition to pre-defined words corresponding to each proposition type

Table 2. Features used for classification

Type	Keywords
AN Similar	similar, generally
AN Precedent	be (to be)
AN Conc	be (to be)
CE Causal	generally, occurs
CE Occurance	occurs
CE Conc	occurs
PR Goal	goal
PR GoalPlan	action
PR Conc	ought, perform
VC Property	be (to be)
VC Class	all, if
VC Conc	be (to be)

Table 3. Keywords used for each proposition type

Both of these tasks were carried out using annotated scheme data from AIFdb [10]. Although there are a number of argument analysis tools (such as Araucaria, Carneades, Rationale and OVA) which allow the analyst to identify the argumentation scheme related to a particular argumentative structure, the vast majority of analyses which are produced using these tools do not include this information. For example, less than 10% of the OVA analyses contained in AIFdb include any scheme structure. AIFdb contains the complete Araucaria corpus [18] used by previous argumentation scheme studies and, supplemented by analyses from other sources, offers the largest annotated dataset available.

The data available comes from a range of different domains, with analyses including details of schemes, and the types of scheme premises, from the Walton scheme set. Although there are over 500 examples of schemes identified in AIFdb, not all of these include complete annotation of the premise types.

Limiting the data to those schemes with at least twenty instances that are fully defined leaves us with four schemes to consider (the number of examples for each scheme type is shown in Table 4.)

Scheme	Number of Examples
Analogy (AN)	31
Cause To Effect (CE)	89
Practical Reasoning (PR)	68
Verbal Classification (VC)	38

Table 4. Number of example instances of each scheme type

2.1. One-against-others classification

For each of the scheme types previously discussed, the conclusions and each type of premise were classified using three different types of classifier (Multinomial Naïve Bayes, Support Vector Machines (SVMs) and Decision Trees) against a random selection of argument propositions from AIFdb.

Table 5 shows the precision, recall and F-score obtained using 10-fold cross validation for each proposition type with each classifier. For each proposition type, the F-Score of the best performing classifier is highlighted in bold.

As can be seen from the table, the Multinomial Naïve Bayes classifiers perform best in most cases, and even for those proposition types where one of the other methods perform better, the results are comparable. In particular, the results for SVMs are lower than those for the other types of classifier. This can be explained by the fact that our feature set is considerably larger than the sample, a situation in which SVMs generally perform less well.

Notably, the results for Analogy (Conclusion) and Cause To Effect (Occurrence) are quite weak in comparison to the other proposition types. In the case of Analogy, the conclusion often does not include details of the specific case being discussed, but instead refers to the general situation being discussed, for example “Invading Iraq has been a foolish action”. Because of this, many of these conclusions take the form of very simple factual statements that are often hard to distinguish from other propositions. With Cause To Effect the Occurrence premise again suffers from a similar lack of complete specificity and details of the specific situation are often omitted.

The results for the remaining proposition types are more promising and, even for those schemes where the classification of one proposition type is less successful, the results for the other types are better. If we consider being able to correctly identify at least one proposition type, then our results give F-scores between 0.78 and 0.91 for locating an occurrence of the different scheme types. The results also show that in many cases it would be possible to not only determine that a scheme is being used, but to accurately classify all of its component propositions.

2.2. Pairwise Classification

For pairwise classification, we assume that identification of a specific argumentation scheme instance (along with its associated premises and conclusion) has previously been

Type	Naïve Bayes			SVM			Decision Tree		
	p	r	f1	p	r	f1	p	r	f1
AN Similar	0.58	1.00	0.74	0.60	0.43	0.50	0.56	0.71	0.63
AN Precedent	0.64	1.00	0.78	0.75	0.43	0.55	0.29	0.29	0.29
AN Conc	1.00	0.29	0.44	0.38	0.43	0.40	0.57	0.57	0.57
CE Causal	0.57	0.89	0.70	0.58	0.61	0.59	0.94	0.89	0.91
CE Occurance	0.50	0.72	0.59	0.40	0.22	0.29	0.38	0.33	0.35
CE Conc	0.73	0.89	0.80	0.54	0.78	0.64	0.57	0.72	0.63
PR Goal	0.65	0.79	0.71	0.55	0.86	0.67	0.59	0.71	0.65
PR GoalPlan	0.65	0.93	0.76	0.76	0.93	0.84	0.75	0.86	0.80
PR Conc	0.90	0.64	0.75	0.55	0.43	0.48	0.76	0.93	0.84
VC Property	0.88	0.88	0.88	1.00	0.50	0.67	0.75	0.75	0.75
VC Class	0.58	0.88	0.70	0.67	0.75	0.71	0.75	0.75	0.75
VC Conc	1.00	0.50	0.67	0.62	0.62	0.62	1.00	0.38	0.55

Table 5. Results of one vs others proposition classification using 10-fold cross validation (The highest f-score for each scheme component is highlighted in bold)

carried out, and look at classifying proposition types for each premise against the other premise proposition types. Being able to successfully perform this task would enable us to determine the full schematic structure of any argument previously analysed at the structural level, be it a manual analysis or one performed by another argument mining technique.

This task was firstly performed using the same approach as the one-vs-others classification, with a Naïve Bayes classifier created for each proposition type, but in this case using only the other premises from the same scheme to test against. The resulting probabilities for each premise type were then compared and assignment to each type was made. The precision, recall and F-score for these classifications can be seen in Table 6.

Type	p	r	f1
PR Goal/GoalPlan	1.00	0.79	0.88
CE Causal/Occurance	0.75	0.50	0.60
AN Similar/Precedent	1.00	0.43	0.60
VC Property/Class	0.75	0.75	0.75

Table 6. Results of pairwise premise classification

In order to take further advantage of the fact that each proposition is already known to belong to a certain scheme and that all of the other premises are also available, we also implemented comparative versions of some of the features. It can be seen from the scheme descriptions that the different premises in each scheme may often contain many of the same words. However, to differentiate between them we want to consider how the vocabulary used for each premise type differs. In order to help us understand this, uni-grams were calculated using words appearing only in the proposition being considered and not in any of the other scheme instance's premises. Additionally, as each scheme we

Type	p	r	f1
PR Goal/GoalPlan	1.00	0.79	0.88
CE Causal/Occurance	0.82	0.50	0.62
AN Similar/Precedent	1.00	0.43	0.60
VC Property/Class	0.78	0.88	0.82

Table 7. Results of pairwise premise classification with additional comparative features

are considering has only two premise types, we were able to use the comparative length of the premises, giving an indication of whether one type of premise is generally longer or shorter than the other.

The results from adding these comparative features are shown in Table 7. The values highlighted in bold show where the addition of these features gave an improvement in the results (all of the other results remained unchanged.)

The difference caused by adding comparative features is particularly notable for the Verbal Classification scheme. This is suggested by the structure of this scheme as described in Table 1. Although the length of both premises may vary depending, for example, on the property that the scheme instance is discussing, the *ClassificationProperty* premise will very often be longer than the *ContainsProperty* premise.

In both sets of results, the performance when classifying the premises of Practical Reasoning schemes and Verbal Classification schemes is considerably greater than that for Analogy and Cause To Effect. It can be seen from the descriptions of these schemes that the premises for the latter pair have more in common than those for the former and as such it is unsurprising that these are harder to distinguish. These results provide a positive indication that being able to determine which of the premises in a pre-identified scheme instance are which, is at least feasible.

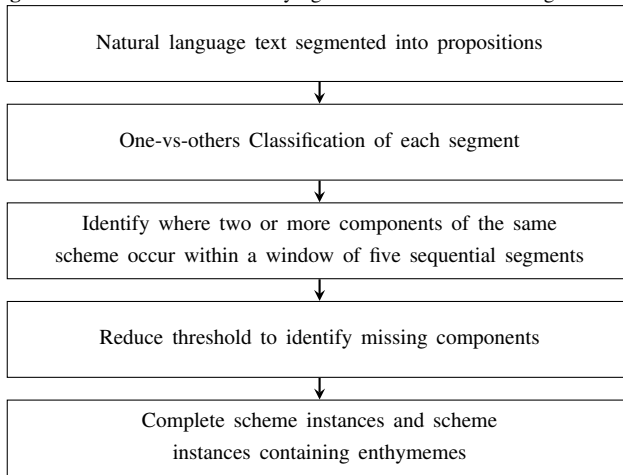
3. Identification of Scheme Instances

The one-against-others results suggest that it is feasible to classify propositions by type. Performing this classification on a piece of text would enable us to identify places where a particular scheme is being used. We now move on to look at how well these classifiers are able to identify not just individual occurrences of a proposition type but complete scheme instances. The ability to successfully perform this task would enable us to take a sample of natural language and understand a large amount of the argument structure it contains. In order to investigate this, we used the proposition corpus created for the Digging by Debating project⁶. This corpus contains over 1,000 sequential propositions extracted from three chapters of “THE ANIMAL MIND: A Text Book of Comparative Psychology” by Margaret Floy Washburn.

The aim of this experiment is not to identify the complete argumentative structure represented by the text, but to illustrate that, even considering the difference in language and methods of expression employed in a 19th century philosophy text, it is possible to use the classifiers that we have produced to extract complete scheme instances.

Our aim here is to identify complete occurrences of a particular scheme within a piece of natural language text. In order to accomplish this, we first perform one-vs-others

⁶<http://diggingbydebating.org/>

Figure 2. Process used for identifying scheme instances from segmented text

classification of each segment using the Multinomial Naïve Bayes classifiers discussed in Section 2.1. We then look at each group of five sequential segments, and identify places where two or more components of the same scheme type occur together. In cases where there is still a missing component, we reduce the threshold for the classifier corresponding to the missing piece. If reducing the threshold still does not offer a candidate for the missing scheme component, we assume that this is unstated enthymematic content in the argument. By performing these steps, we are able to take segmented text and identify either complete scheme instances, or partial scheme instances which have some enthymematic component. The process followed is illustrated in Figure 2.

The classification process identified 9 possible occurrences of Analogy, 14 of Cause To Effect, 18 of Practical Reasoning and 23 of Verbal Classification. The Animal Mind corpus is not annotated for scheme instances, however we can see that, although some instances may have been missed, many of those identified are a close match to the scheme descriptions. For example, the structure in Figure 3 was identified as an occurrence of Practical Reasoning. In this case, the proposition “Thorndike’s aim in this research was to place his animals (chicks, cats, and dogs) under the most rigidly controlled experimental conditions” was identified as a goal and “The cats and dogs, reduced by fasting to a state of ‘utter hunger,’ were placed in boxes, with food outside” as a plan for achieving that goal. Although these two propositions fit the scheme well, the suggested conclusion (“the process whereby they learned to work the various mechanisms which let them out was carefully observed”) does not follow the required pattern.

An example of an identified instance of Verbal Classification can be seen in Figure 4. Again, in this case, the premises fit the scheme quite well (*Classification Property*: “If it is argued that we have no direct, but only an inferential, knowledge of the processes in an animal’s mind, the argument is equally valid against human psychology” and *Contains Property*: “the psychologist has only an inferential knowledge of his neighbour’s mind”), but the conclusion does not follow.

A final example, this time showing an identified instance of Cause To Effect, is shown in Figure 5. Once more, the premises fit the scheme description, but the conclusion again does not follow. This difficulty in discovering the conclusions may be due to the

Figure 3. Automatically identified Practical Reasoning instance

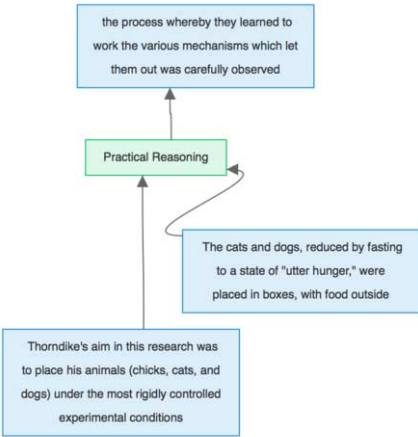
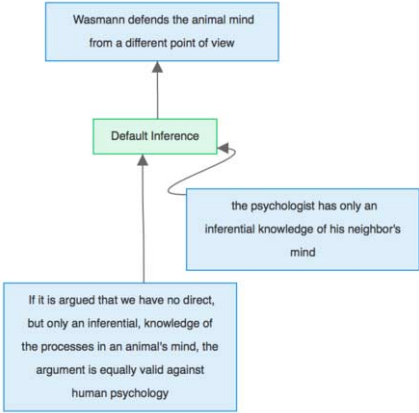
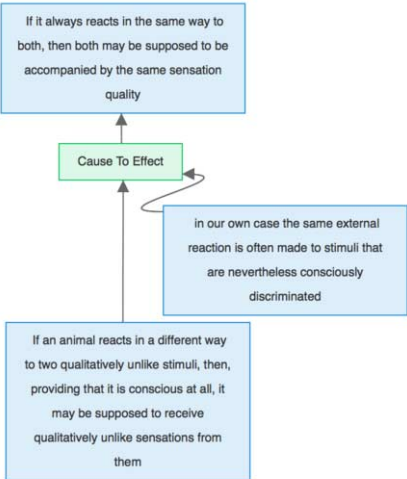


Figure 4. Automatically identified Verbal Classification instance



fact that generally conclusions are not as clearly stated and may be the general topic being discussed as opposed to a clearly expressed proposition located close to the supporting premises. This can be seen even in the example of Verbal Classification from section 1.1 and suggests that an amount of reconstruction may be necessary to fully identify all parts of a scheme.

Figure 5. Automatically identified Cause To Effect instance



Although these examples are not perfect identifications of scheme instances, it is clear that even with the limitations involved, we have come close to being able to identify at least where a scheme is occurring, and to correctly assign at least some of the propositions.

4. Conclusion

Whilst argumentation schemes have been detailed extensively in philosophy and psychology, perhaps due to the relative complexity of these structures, they have received little attention in argument mining. In [3], instances of particular schemes are classified from text which has previously been annotated for its argumentative structure, a process which could be considered as the second step in the six-stage approach to identifying arguments and their schemes suggested by [21].

Here, we have shown that by considering the features of the individual types of premise and conclusion that comprise a scheme, it is possible to reliably classify these scheme components. Despite the differing goals, our results are comparable results to those of Feng & Hirst, where the occurrence of a particular argumentation scheme was identified with accuracies of between 62.9% and 90.8%. Our results show that, on the same dataset, it is possible to identify individual scheme components with similar performance (F-scores between 0.78 and 0.91) can be achieved in identifying argumentation schemes in unanalysed text.

Furthermore, by searching for groupings of these proposition types, we have shown it is possible to determine not just that a particular scheme is being used, but to correctly assign assign propositions to their schematic roles. In future work accuracy of these techniques could be further improved by considering domain specific schemes, such as the Consumer Argumentation Scheme (CAS) [23] aimed specifically at product reviews.

Our results also compare favourably with those presented in [14] where sentences were classified as either premise (F-score, 0.68) or conclusion (F-score, 0.74). For each of the schemes we considered, we were able to classify conclusions with F-scores between 0.71 and 0.91, and premises with F-scores between 0.59 and 0.88. Although these values are not quite as high for all premise types, we are able to determine not only that something is a premise, but also what role it plays in the scheme, showing that scheme component identification offers valuable information that could play an instrumental role in determining the full argumentative structure, be it as a stand-alone method, a source of feature data for more complex classifiers or part of a larger ensemble approach.

Acknowledgments

We would like to acknowledge that the work reported in this paper has been supported in part by EPSRC in the UK under grants EP/N014871/1 and EP/K037293/1.

References

- [1] Aristotle. *Topics*. Oxford University Press, 1958.
- [2] E. Cabrio, S. Tonelli, and S. Villata. From discourse analysis to argumentation schemes and back: Relations and differences. In *International Workshop on Computational Logic in Multi-Agent Systems*, pages 1–17. Springer, 2013.
- [3] V. W. Feng and G. Hirst. Classifying arguments by scheme. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 987–996. Association for Computational Linguistics (ACL), 2011.
- [4] T. Goudas, C. Louizos, G. Petasis, and V. Karkaletsis. Argument extraction from news, blogs, and social media. In *Artificial Intelligence: Methods and Applications*, pages 287–299. Springer, 2014.

- [5] W. Grennan. *Informal Logic: Issues and Techniques*. McGill-Queen's Press-MQUP, 1997.
- [6] A. C. Hastings. *A Reformulation of the Modes of Reasoning in Argumentation*. PhD thesis, Northwestern University, 1963.
- [7] D. Hitchcock. Enthymematic arguments. *Informal Logic*, 7(2):289–98, 1985.
- [8] J. Katzav and C. Reed. On argumentation schemes and the natural classification of arguments. *Argumentation*, 18(2):239–259, 2004.
- [9] M. Kienpointner. *Alltagslogik: struktur und funktion von argumentationsmustern*. Frommann-Holzboog, 1992.
- [10] J. Lawrence, F. Bex, C. Reed, and M. Snaith. AIFdb: Infrastructure for the argument web. In *Proceedings of the Fourth International Conference on Computational Models of Argument (COMMA 2012)*, pages 515–516, 2012.
- [11] J. Lawrence, C. Reed, C. Allen, S. McAlister, and A. Ravenscroft. Mining arguments from 19th century philosophical texts using topic based modelling. In *Proceedings of the First Workshop on Argumentation Mining*, pages 79–87, Baltimore, Maryland, June 2014. Association for Computational Linguistics (ACL).
- [12] B. Liu. Sentiment analysis and subjectivity. *Handbook of natural language processing*, 2:627–666, 2010.
- [13] M.-F. Moens, E. Boiy, R. M. Palau, and C. Reed. Automatic detection of arguments in legal texts. In *Proceedings of the 11th international conference on Artificial intelligence and law*, pages 225–230. ACM, 2007.
- [14] R. M. Palau and M.-F. Moens. Argumentation mining: the detection, classification and structure of arguments in text. In *Proceedings of the 12th international conference on artificial intelligence and law*, pages 98–107. ACM, 2009.
- [15] J. Park and C. Cardie. Identifying appropriate support for propositions in online user comments. In *Proceedings of the First Workshop on Argumentation Mining*, pages 29–38, Baltimore, Maryland, June 2014. Association for Computational Linguistics (ACL).
- [16] C. Perelman and L. Olbrechts-Tyteca. *The New Rhetoric: A Treatise on Argumentation*. University of Notre Dame Press, 1969.
- [17] J. L. Pollock. *Cognitive carpentry: A blueprint for how to build a person*. MIT Press, 1995.
- [18] C. Reed, R. Mochales Palau, G. Rowe, and M.-F. Moens. Language resources for studying argument. In *Proceedings of the 6th Language Resources and Evaluation Conference (LREC-2008)*, pages 91–100, Marrakech, 2008.
- [19] C. Stab and I. Gurevych. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 46–56, Doha, Qatar, October 2014. Association for Computational Linguistics (ACL).
- [20] D. Walton. *Argumentation schemes for presumptive reasoning*. Lawrence Erlbaum Associates, Mahwah, New Jersey, 1996.
- [21] D. Walton. Argument mining by applying argumentation schemes. *Studies in Logic*, 4(1):38–64, 2011.
- [22] D. Walton, C. Reed, and F. Macagno. *Argumentation Schemes*. Cambridge University Press, 2008.
- [23] A. Wyner, J. Schneider, K. Atkinson, and T. Bench-Capon. Semi-automated argumentative analysis of online product reviews. In *Proceedings of the Fourth International Conference on Computational Models of Argument (COMMA 2012)*, pages 43–50, 2012.

A Specialized Set Theoretic Semantics for Acceptability Dynamics of Arguments¹

Martín O. MOGUILLANSKY² and Guillermo R. SIMARI

*CONICET, Institute for Research in Computer Science and Engineering (ICIC),
Universidad Nacional del Sur (UNS), ARGENTINA*

Abstract

Inspired by the standard set theoretic Tarskian semantics, we propose a novel interpretation structure for studying the acceptability dynamics of arguments (i.e., the eventual changes on their acceptability condition) for logic-based argumentation. Interpretation structures identify possible scenarios in which a given argument would be accepted, or not, according to some standard extension-based argumentation semantics. These scenarios are configured in accordance to the consideration or inconsideration of other arguments from the given argumentation framework. Thereafter, it would be possible to ensure the acceptability of an argument by handling the evolution of the argumentation framework throughout the use of argumentative change operations. Hence, an interpretation structure which is a model of a given argument specifies a possible epistemic state to which the argumentation framework could evolve towards the argument's positive acceptability. Moreover, the analysis of several models of a given argument brings the opportunity of satisfying additional restrictions towards the evolution of a framework. Finally, we propose a revision operator whose rationality is ensured through postulates and a corresponding representation theorem.

Keywords. Argumentation, Belief Revision, Argumentation Dynamics

1. Introduction

We propose a new perspective for studying acceptability dynamics of arguments upon logic-based argumentation. Although an argumentation change operator is presented at the end, the main objective of this article is not precisely its proposal, but the introduction of a new way of reasoning about dynamics in argumentation by considering the changes on the acceptability status of arguments. We intend to deal with the question of which arguments provide, or interfere with, the acceptability of others, studying the interaction between acceptance and rejection through the consideration of sub-frameworks, and facilitating a theoretical analysis of the implications of change on a framework in advance, before the formal application of some argumentation change operation. We get inspiration from the

¹Partially supported by UNS (PGI 24/NO40).

²Corresponding author: M.O. Moguillansky. CONICET-ICIC. E-mail: mom@cs.uns.edu.ar.

standard set theoretic Tarskian semantics and the idea of constructing interpretation structures for reasoning about dynamics in argumentation. The cornerstone for such structures relies on the notions of core and remainder sets [11], two different constructions for recognizing acceptance and rejection of arguments. An interpretation structure proposes a “further” epistemic state in which the acceptability of a formula is analyzed in contrast with its acceptability status on the current epistemic state, *i.e.*, the current framework. Since that formula is supported by the claim of certain arguments, the interpretation structure ends up analyzing their acceptability as well. Afterwards, an interpretation ensuring the acceptability on the further epistemic state is referred as a model. Since several different models may appear, we obtain alternatives of change for analyzing and deciding which should be the most appropriate according to rationality conditions. On its basis, we propose thereafter an acceptance revision operator which deals with the matter of incorporating a new argument while ensuring its acceptance. Finally, its rationality is guaranteed through the axiomatic characterization and corresponding representation theorem according to classic belief revision [2] literature and an argument-based belief revision model like Argument Theory Change (ATC) [12].

2. Fundamentals for Reasoning on Logic-based Frameworks

We refer to the *argument domain set* $\mathbb{A}_{\mathcal{L}}$ for identifying (*logic-based*) *arguments* $a \in \mathbb{A}_{\mathcal{L}}$ built with formulae $\vartheta \in \mathcal{L}$, where \mathcal{L} is some underlying logic. Arguments are expressed through a pair $\langle S, \vartheta \rangle$ where $S \subseteq \mathcal{L}$ is the argument’s *support*, and $\vartheta \in \mathcal{L}$ its *claim*. The functions $\mathbf{cl} : \mathbb{A}_{\mathcal{L}} \rightarrow \mathcal{L}$ and $\mathbf{sp} : \mathbb{A}_{\mathcal{L}} \rightarrow \wp(\mathcal{L})$ are used for identifying the claim $\mathbf{cl}(a) \in \mathcal{L}$ and support set $\mathbf{sp}(a) \subseteq \mathcal{L}$, of an argument $a \in \mathbb{A}_{\mathcal{L}}$. The function \mathbf{sp} will be overloaded to apply over sets of arguments, $\mathbf{sp} : \wp(\mathbb{A}_{\mathcal{L}}) \rightarrow \wp(\mathcal{L})$, such that $\mathbf{sp}(\Theta) = \bigcup_{a \in \Theta} \mathbf{sp}(a)$ will identify the *base* determined by the set of supports of arguments in $\Theta \subseteq \mathbb{A}_{\mathcal{L}}$. The logic \mathcal{L} will be considered along with its corresponding inference operator \models . Thus, we can say that an argument $a \in \mathbb{A}_{\mathcal{L}}$ *supports*, or *is a supporter of* ϑ , to specify that $\mathbf{cl}(a) \models \vartheta$ holds. In order to avoid multiple representation of arguments with a same support set, we will restrict their construction to the *canonical form* [4], in which for any argument a , its claim is $\mathbf{cl}(a) = \bigwedge \mathbf{sp}(a)$. Hence, we will assume $\mathbb{A}_{\mathcal{L}}$ as the domain of canonical arguments. In consequence, for any pair $a, b \in \mathbb{A}_{\mathcal{L}}$, $a = b$ *iff* if $\mathbf{sp}(a) = \mathbf{sp}(b)$ then $\mathbf{cl}(a) = \mathbf{cl}(b)$. We write $a \sqsubseteq b$ for expressing that an argument $a \in \mathbb{A}_{\mathcal{L}}$ is a *sub-argument* of argument $b \in \mathbb{A}_{\mathcal{L}}$ (and also that b is a *super-argument* of a), implying that $\mathbf{sp}(a) \subseteq \mathbf{sp}(b)$ holds. When $\mathbf{sp}(a) \subset \mathbf{sp}(b)$, we say that a is a *strict sub-argument* of b , by writing $a \sqsubset b$. Arguments with no strict sub-arguments inside are referred as *atomic arguments*, thus $a \in \mathbb{A}_{\mathcal{L}}$ is *atomic* *iff* $|\mathbf{sp}(a)| = 1$. The *atoms function* $\mathbf{at} : \mathbb{A}_{\mathcal{L}} \rightarrow \wp(\mathbb{A}_{\mathcal{L}})$ identifies the set $\mathbf{at}(a) \subseteq \mathbb{A}_{\mathcal{L}}$ of all the atomic arguments of $a \in \mathbb{A}_{\mathcal{L}}$. The atoms function will be overloaded as $\mathbf{at} : \wp(\mathbb{A}_{\mathcal{L}}) \rightarrow \wp(\mathbb{A}_{\mathcal{L}})$ to apply over sets $\Theta \subseteq \mathbb{A}_{\mathcal{L}}$ such that $\mathbf{at}(\Theta) = \bigcup_{a \in \Theta} \mathbf{at}(a)$. The set $\mathbf{R}_{\Theta} \subseteq \mathbb{A}_{\mathcal{L}} \times \mathbb{A}_{\mathcal{L}}$ identifies the *defeat relation* between pairs of arguments from $\Theta \subseteq \mathbb{A}_{\mathcal{L}}$. A pair $(a, b) \in \mathbf{R}_{\Theta}$ implies that $a \in \Theta$ *defeats* $b \in \Theta$, or equivalently, a is a *defeater* of b , meaning that $\mathbf{sp}(a) \cup \mathbf{sp}(b) \models \perp$ and $a \succ b$, where $\succ \subseteq \mathbb{A}_{\mathcal{L}} \times \mathbb{A}_{\mathcal{L}}$ is an abstract *preference relation* assumed to be *total*

—thus, for any pair of arguments $a, b \in \mathbb{A}_{\mathcal{L}}$ we know either $a \succ b$ or $b \succ a$ (or both). This is a necessary condition to ensure a functional construction of the defeat relation $\mathbf{R} : \wp(\mathbb{A}_{\mathcal{L}}) \rightarrow \wp(\mathbb{A}_{\mathcal{L}} \times \mathbb{A}_{\mathcal{L}})$, verifying $\mathbf{sp}(a) \cup \mathbf{sp}(b) \models \perp$ iff $(a, b) \in \mathbf{R}_{\Theta}$ or $(b, a) \in \mathbf{R}_{\Theta}$, for any pair $a, b \in \Theta$. In addition, for guaranteeing $\mathbf{sp}(\Theta) \models \perp$ iff $\mathbf{R}_{\Theta} \neq \emptyset$, we will rely upon *closed sets of arguments*: a set containing all the sub- and super-arguments that can be constructed from its arguments. We provide such implementation through an *argumentation closure operator* \mathbb{C} such that for any $\Theta \subseteq \mathbb{A}_{\mathcal{L}}$, $\mathbb{C}(\Theta) = \{a \in \mathbb{A}_{\mathcal{L}} \mid \text{at}(a) \subseteq \Theta \text{ or } a \sqsubseteq b, \text{ for any } b \in \Theta\}$. Thus, we will say $\mathbf{A} \subseteq \mathbb{A}_{\mathcal{L}}$ is *closed* iff $\mathbf{A} = \mathbb{C}(\mathbf{A})$, and will usually note as \mathbf{A} any closed set.

Example 1 Assuming a propositional logic \mathcal{L} and a set $\Theta \subseteq \mathbb{A}_{\mathcal{L}}$ such that $\Theta = \{a, b, c\}$ where $a = \langle \{p\}, p \rangle$, $b = \langle \{q\}, q \rangle$, and $c = \langle \{-p \vee \neg q\}, \neg p \vee \neg q \rangle$; the functional construction of the defeat relation will trigger a set $\mathbf{R}_{\Theta} = \emptyset$, although $\mathbf{sp}(\Theta) \models \perp$ holds. However, the argumentation closure renders a closed set $\mathbf{A} = \mathbb{C}(\Theta) = \{a, b, c, d, e, f\}$, where $d = \langle \{p, q\}, p \wedge q \rangle$, $e = \langle \{p, \neg p \vee \neg q\}, p \wedge (\neg p \vee \neg q) \rangle$, and $f = \langle \{q, \neg p \vee \neg q\}, q \wedge (\neg p \vee \neg q) \rangle$. Afterwards, the defeat relation ends up as $\mathbf{R}_{\mathbf{A}} = \{(a, f), (b, e), (c, d), (d, e), (d, f), (e, d), (f, d)\}$, for a preference relation prioritizing arguments in Θ over others, being symmetric otherwise.

A (canonical logic-based) *argumentation framework* (AF) is identified through the structure $\langle \Theta, \mathbf{R}_{\Theta} \rangle$, where $\Theta \subseteq \mathbb{A}_{\mathcal{L}}$, and whenever $\mathbf{A} \subseteq \mathbb{A}_{\mathcal{L}}$ is known to be closed, the structure $\langle \mathbf{A}, \mathbf{R}_{\mathbf{A}} \rangle$ identifies a *closed AF*. Since the defeat relation is a function over $\mathbb{A}_{\mathcal{L}}$ -arguments, we refer to an operator \mathbb{F}_{Θ} as the *AF generator from Θ* iff $\mathbb{F}_{\Theta} = \langle \Theta, \mathbf{R}_{\Theta} \rangle$. Note that \mathbb{F}_{Θ} is the AF constructed from Θ . Finally, we refer to an AF $\mathbb{F}_{\mathbf{A}}$, implying that $\mathbb{F}_{\mathbf{A}}$ is the closed AF $\langle \mathbf{A}, \mathbf{R}_{\mathbf{A}} \rangle$, and thus, $\mathbf{A} = \mathbb{C}(\mathbf{A})$. Given an AF $\mathbb{F}_{\mathbf{A}}$, for any not necessarily closed set $\Theta \subseteq \mathbf{A}$, it is possible to construct the *sub-framework* \mathbb{F}_{Θ} . In such a case, we overload the sub-argument operator ‘ \sqsubseteq ’ by also using it for identifying sub-frameworks, writing $\mathbb{F}_{\Theta} \sqsubseteq \mathbb{F}_{\mathbf{A}}$. Observe that, if $\mathbb{C}(\Theta) = \mathbf{A}'$ and $\mathbf{A}' \subset \mathbf{A}$, then $\mathbb{F}_{\mathbf{A}'}$ is a closed *strict sub-framework* of $\mathbb{F}_{\mathbf{A}}$, i.e., $\mathbb{F}_{\mathbf{A}'} \subset \mathbb{F}_{\mathbf{A}}$. Our intention is to simplify AFs for concentrating on acceptability dynamics of arguments. Consequently, for an AF τ , we refer to its set of arguments through the set $\mathbf{A}(\tau)$ and to its set of defeats through $\mathbf{R}(\tau)$.

Given an AF $\mathbb{F}_{\mathbf{A}}$, as usual in abstract argumentation [8], for any $\Theta \subseteq \mathbf{A}$ we say that Θ *defeats* an argument $a \in \mathbf{A}$ iff there is some $b \in \Theta$ such that b defeats a ; Θ *defends* an argument $a \in \mathbf{A}$ iff Θ defeats every defeater of a ; Θ is *conflict-free* iff $\mathbf{R}_{\Theta} = \emptyset$; and Θ is *admissible* iff it is conflict-free and defends all its members. However, as seen before, a logic-based framework should be closed to ensure that all sources of conflict are identified through the defeat relation. For instance, in Ex. 1, $\Theta \subseteq \mathbf{A}$ is admissible given that it is conflict-free and defends all its members, however $\mathbf{sp}(\Theta) \models \perp$. This is undesirable since an admissible set could trigger an inconsistent set of supports. Thus, we reformulate the classic notion of admissibility for abstract argumentation into *logic-based admissibility* [11]:

Definition 1 (Logic-based Admissibility [11]) For any $\Theta \subseteq \mathbf{A}$ we say that Θ is *admissible* iff Θ is closed, conflict-free, and defends all its members.

We will just say admissibility to refer to logic-based admissibility. (In Ex. 1, Θ cannot be admissible since it is not closed, thus, the only admissible sets are

$\{a\}$, $\{b\}$, and $\{c\}$.) The *extension semantics*, which rely upon admissibility, will also be affected by the notion of logic-based admissibility without inconvenience. We will only refer to the *complete semantics* in some examples, however, any of the extension semantics could also be applied. Thus, given an AF $\tau = \mathbb{F}_{\mathbf{A}}$, a set $\mathbf{E} \subseteq \mathbf{A}$ is a *complete extension* if \mathbf{E} is admissible and contains every argument it defends. Afterwards, the set $\mathbb{E}_{\mathfrak{s}}(\tau) \subseteq \wp(\mathbf{A})$ identifies the *set of \mathfrak{s} -extensions* \mathbf{E} from τ , where an *\mathfrak{s} -extension* is an extension in τ according to some specific extension semantics \mathfrak{s} . Observe that any extension $\mathbf{E} \in \mathbb{E}_{\mathfrak{s}}(\tau)$ is admissible and thus, it contains a consistent support base, *i.e.*, $\mathfrak{sp}(\mathbf{E}) \not\models \perp$ holds.

We refer as *acceptance criterion* to the determination of acceptance of arguments in either a *sceptical* or *credulous* way. Several postures may appear. For instance, a *sceptical set* may be obtained by intersecting every \mathfrak{s} -extension $\bigcap \mathbb{E}_{\mathfrak{s}}(\tau)$, while a *credulous set* may arise from the selection of a single extension $\mathbf{E} \in \mathbb{E}_{\mathfrak{s}}(\tau)$ according to some specific preference. For instance, selecting “the best” extension among those of maximal cardinality. We will abstract the implementation of any acceptance criterion by referring to an *acceptance function* $\delta : \wp(\wp(\mathbf{A})) \rightarrow \wp(\mathbf{A})$ where $\delta(\mathbb{E}_{\mathfrak{s}}(\tau))$ determines the outcome of the adopted criterion. In addition, we refer as (*argumentation*) *semantics specification* \mathcal{S} to a tuple $\langle \mathfrak{s}, \delta \rangle$, where \mathfrak{s} stands for identifying some extension semantics and δ for an acceptance function implementing some acceptance criterion. Afterwards, we refer to the set $\mathcal{A}_{\mathcal{S}}(\tau) \subseteq \mathbf{A}$ as the *acceptable set* of τ according to \mathcal{S} *iff* $\mathcal{A}_{\mathcal{S}}(\tau) = \delta(\mathbb{E}_{\mathfrak{s}}(\tau))$. Finally, for any $a \in \mathbf{A}$, a is *\mathcal{S} -accepted* in τ (resp. of, *\mathcal{S} -rejected*) *iff* $a \in \mathcal{A}_{\mathcal{S}}(\tau)$ (resp. of, $a \notin \mathcal{A}_{\mathcal{S}}(\tau)$).

3. Acceptability Analysis through Core and Remainder Sets

We rely upon the notions of *admissible* and *core sets* ([11]) of an argument as the fundamentals for recognizing the sources of an argument’s acceptability condition, and upon *rejecting sets* for the argument’s rejecting condition.

Definition 2 (Admissible Sets of an Argument) *Given an AF $\tau = \mathbb{F}_{\mathbf{A}}$ and an argument $a \in \mathbf{A}$; for any $\Theta \subseteq \mathbf{A}$, we say that: 1) Θ is an *a-admissible set* in τ *iff* Θ is an admissible set³ such that $a \in \Theta$, and 2) Θ is a **minimal a-admissible set** in τ *iff* Θ is a-admissible and for any $\Theta' \subset \Theta$, it follows that Θ' is not a-admissible.*

Definition 3 (Core Sets) *Given an AF $\tau = \mathbb{F}_{\mathbf{A}}$ and an argumentation semantics specification \mathcal{S} , for any $\mathcal{C} \subseteq \mathbf{A}$, we say that \mathcal{C} is an *a-core* in τ , noted as *a-core _{\mathcal{S}}** *iff* \mathcal{C} is a minimal a-admissible set and a is \mathcal{S} -accepted in τ .

Definition 4 (Rejecting Sets of an Argument) *Given an AF $\mathbb{F}_{\mathbf{A}}$, a semantics specification \mathcal{S} , and an argument $a \in \mathbf{A}$; for any $\Theta \subseteq \mathbf{A}$, we say that Θ is a *\mathcal{S} -a-rejecting set* in $\mathbb{F}_{\mathbf{A}}$ *iff* a is \mathcal{S} -rejected in $\mathbb{F}_{\mathbf{A}}$ but it is \mathcal{S} -accepted in $\mathbb{F}_{\mathbf{A} \setminus \Theta}$.*

We have defined rejecting sets in an intuitive manner. For constructing rejecting sets of an argument a (see [11]) we need to identify those arguments that interpose to the construction of an *a-core _{\mathcal{S}}* set. This is the seed for further con-

³Recall that from now on by admissibility we refer only to its logic-based definition.

structuring *remainder sets*. However, the acceptability analysis must apply upon closed frameworks for avoiding inconveniences as described in Ex. 1. We only can be sure that a is \mathcal{S} -accepted in $\mathbb{F}_{\mathbf{A} \setminus \Theta}$ if we can ensure that $\mathbb{F}_{\mathbf{A} \setminus \Theta}$ is a closed AF. An *expansive closure* is a sort of “complementary closure operator” which ensures that removing an *expanded set* from a closed set delivers a closed set.

Definition 5 (Expansive Closure) Given $\Theta \subseteq \mathbf{A}$, \mathbb{P} is an **expansive closure** iff $\mathbb{P}(\Theta) = \{a \in \mathbf{A} \mid b \sqsubseteq a, \text{ for every } b \in \text{at}(\mathbb{P}_0(\Theta))\}$, where $\mathbb{P}_0(\Theta) = \{a \in \Theta \mid \text{there is no } b \in \Theta \text{ such that } b \sqsubset a\}$. We say that Θ is **expanded** iff it holds $\Theta = \mathbb{P}(\Theta)$.

Example 2 Suppose $\{a_1, a_2, a, b_1, b, c\} \subseteq \mathbf{A}$, where $a \sqsubset b$ and $b \sqsubset c$, $\text{at}(a) = \{a_1, a_2\}$, and $\text{at}(b) = \{a_1, a_2, b_1\}$; and $\Theta = \{a, b\}$. This means that $\mathbb{P}_0(\Theta) = \{a\}$. Removing a from \mathbf{A} should prevent its construction, thus, a_1 and a_2 should not be simultaneously present (since $a \in \mathbb{C}(\{a_1, a_2\})$, $\mathbf{A} = \mathbb{C}(\mathbf{A} \setminus \{a\})$). The expanded set ends up being $\mathbb{P}(\Theta) = \{a_1, a_2, a, b, c\}$, which ensures that $\mathbf{A} \setminus \mathbb{P}(\Theta)$ is a closed set. Observe however that $\mathbb{P}(\Theta)$ is not a minimal expanded set for the removal of arguments a and b : for instance if $\Theta' = \{a_1, a, b\}$ then $\mathbb{P}(\Theta') = \{a_1, a, b, c\}$, which is a minimal alternative for such purpose.

Proposition 1 Given two sets $\mathbf{A} \subseteq \mathbb{A}_{\mathcal{L}}$ and $\Theta \subseteq \mathbb{A}_{\mathcal{L}}$, where \mathbf{A} is closed; if $\Theta \subseteq \mathbf{A}$ then $\mathbf{A}' = \mathbf{A} \setminus \mathbb{P}(\Theta)$ is a closed set, i.e., $\mathbf{A}' = \mathbb{C}(\mathbf{A}')$.

Remainder sets identify “responsible” arguments for the non-acceptability of an argument. Intuitively, an a -remainder is a minimal expanded \mathcal{S} - a -rejecting set.

Definition 6 (Remainder Sets) Given an AF $\mathbb{F}_{\mathbf{A}}$ and a semantics specification \mathcal{S} , for any $\mathcal{R} \subseteq \mathbf{A}$, \mathcal{R} is an a -remainder in $\mathbb{F}_{\mathbf{A}}$, noted as $a\text{-remainder}_{\mathcal{S}}$ iff \mathcal{R} is a minimal expanded \mathcal{S} - a -rejecting set: 1) \mathcal{R} is a \mathcal{S} - a -rejecting set, 2) $\mathcal{R} = \mathbb{P}(\mathcal{R})$, and 3) for any set $\Theta \subset \mathcal{R}$ such that $\Theta = \mathbb{P}(\Theta)$, it holds a is \mathcal{S} -rejected in $\mathbb{F}_{\mathbf{A} \setminus \Theta}$.

Example 3 Assume \mathcal{L} as the propositional logic and $\mathbb{A}_{\mathcal{L}}$ as the domain of canonical arguments. Let $\Theta = \{a, b, c, d\} \subseteq \mathbb{A}_{\mathcal{L}}$ be a set of canonical arguments such that $\Theta = \{a, b, c, d\}$, where $a = \langle \{p \wedge q_1\}, p \wedge q_1 \rangle$, $b = \langle \{p \wedge q_2\}, p \wedge q_2 \rangle$, $c = \langle \{\neg p\}, \neg p \rangle$, and $d = \langle \{\neg q_2\}, \neg q_2 \rangle$. The argumentation closure renders the closed set of arguments $\mathbf{A} = \mathbb{C}(\Theta) = \{a, b, c, d, e, f, g\}$, where:

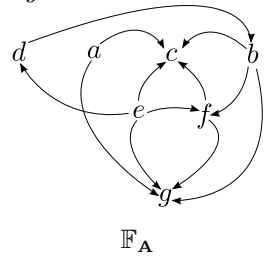
$e = \langle \{p \wedge q_1, p \wedge q_2\}, p \wedge q_1 \wedge q_2 \rangle$ ($a \sqsubseteq e$, $b \sqsubseteq e$)

$f = \langle \{p \wedge q_1, \neg q_2\}, p \wedge q_1 \wedge \neg q_2 \rangle$ ($a \sqsubseteq f$, $d \sqsubseteq f$)

$g = \langle \{\neg p, \neg q_2\}, \neg p \wedge \neg q_2 \rangle$ ($c \sqsubseteq g$, $d \sqsubseteq g$)

Thus, $\mathbb{F}_{\mathbf{A}}$ is closed and due to some preference relation:

$\mathbf{R}_{\mathbf{A}} = \{(a, c), (b, c), (d, b), (e, c), (e, d), (b, f), (f, c), (a, g), (b, g), (e, f), (e, g), (f, g)\}$. Assuming $\mathcal{S} = \langle \mathfrak{s}, \delta \rangle$, where \mathfrak{s} is a complete semantics and δ selects “the best” \mathfrak{s} -extension of higher cardinality (credulous), a b -core $_{\mathcal{S}}$ $\mathcal{C}_b = \{a, b, e\}$ is constructed by $\mathbb{C}(\{b, e\})$. Since c and d are \mathcal{S} -rejected, we have remainders for both of them: a c -remainder $_{\mathcal{S}}$ $\mathcal{R}_c = \{a, e, f\}$ and two d -remainder $_{\mathcal{S}}$ sets $\mathcal{R}_d = \{a, e, f\}$ and $\mathcal{R}'_d = \{b, e\}$. Note $\{a, b, e, f\} = \mathbb{P}(\{e\})$ is not a d -remainder $_{\mathcal{S}}$ since it is not minimal given that it contains $\mathbb{P}(\{a, e\}) = \mathcal{R}_d$ and $\mathbb{P}(\{b, e\}) = \mathcal{R}'_d$.



Proposition 2 *Given an AF $\mathbb{F}_{\mathbf{A}}$, it holds: 1) $a \in \mathcal{A}_S(\mathbb{F}_{\mathbf{A}})$ iff there is some $a\text{-core}_S$ in \mathbf{A} , 2) $a \notin \mathcal{A}_S(\mathbb{F}_{\mathbf{A}})$ iff there is some $a\text{-remainder}_S$ in \mathbf{A} , and 3) there is some $a\text{-core}_S$ in \mathbf{A} iff there is no $a\text{-remainder}_S$ in \mathbf{A} .*

4. Argumentation Dynamics Contra-Semantics

The idea behind the theory of *Argumentation Dynamics Contra-Semantics* is to analyze the current epistemic state determined by an AF $\tau = \mathbb{F}_{\mathbf{A}}$ for answering whether a formula $\vartheta \in \mathcal{L}$ is \mathcal{S} -accepted in τ , and in the case ϑ is \mathcal{S} -rejected in τ , whether it is possible, and how, to provoke the evolution of τ to reach a *further epistemic state* in which ϑ would end up \mathcal{S} -accepted. The language \mathcal{L} is interpreted in terms of a specialized set theoretic semantics à la Tarski, through an *argumentation dynamics interpretation structure* $\mathcal{I}_S = \langle \Delta^{\mathcal{I}}, \Gamma^{\mathcal{I}}, \cdot^{\mathcal{I}} \rangle$, which considers an (*acceptability dynamics*) *interpretation function* $\cdot^{\mathcal{I}}$ and two different domains containing the “constants” of the AF which would be the lower-case letters naming arguments from \mathbf{A} . The positive domain $\Delta^{\mathcal{I}}$ referred as *interpretation domain*, describes which arguments are considered in the proposed interpretation, and the negative domain $\Gamma^{\mathcal{I}}$ referred as *interpretation contra-domain*, describes undesired arguments. Through the interpretation function, the acceptability dynamics of a formula $\vartheta \in \mathcal{L}$ are interpreted as $\vartheta^{\mathcal{I}} \subseteq \mathbb{W}_{\mathcal{L}}$, where $\mathbb{W}_{\mathcal{L}} \subseteq \mathbb{A}_{\mathcal{L}} \times \wp(\mathbb{A}_{\mathcal{L}}) \times \wp(\mathbb{A}_{\mathcal{L}})$.

Definition 7 (Dynamics Interpretation Structure) *Given an AF $\mathbb{F}_{\mathbf{A}}$ and a semantics specification \mathcal{S} ; a structure $\mathcal{I}_S = \langle \Delta^{\mathcal{I}}, \Gamma^{\mathcal{I}}, \cdot^{\mathcal{I}} \rangle$, where $\Delta^{\mathcal{I}} \subseteq \mathbf{A}$, $\Gamma^{\mathcal{I}} \subseteq \mathbf{A}$, and $\cdot^{\mathcal{I}} : \mathcal{L} \rightarrow \wp(\mathbb{W}_{\mathcal{L}})$; is referred as **(argumentation dynamics) interpretation structure** iff 1) $\Delta^{\mathcal{I}} \cup \Gamma^{\mathcal{I}} = \mathbf{A}$, 2) $\Delta^{\mathcal{I}} \cap \Gamma^{\mathcal{I}} = \emptyset$, 3) $\Delta^{\mathcal{I}}$ is closed, and 4) for any $\vartheta \in \mathcal{L}$, $\vartheta^{\mathcal{I}}$ is the set of **interpretation triples** $(a, X, Y) \in \mathbb{W}_{\mathcal{L}}$ verifying:*

- a) $a \in \mathbf{A}$, is a support for ϑ , i.e., $\text{cl}(a) \models \vartheta$
- b) $X \subseteq \Delta^{\mathcal{I}}$, is an $a\text{-core}_S$ in $\mathbb{F}_{\Delta^{\mathcal{I}}}$ or else the empty set
- c) $Y \subseteq \Gamma^{\mathcal{I}}$, is an $a\text{-remainder}_S$ in $\mathbb{F}_{\mathbf{A}}$ or else the empty set

We refer to $\Delta^{\mathcal{I}}$ as the **interpretation domain**, $\Gamma^{\mathcal{I}}$ as the **interpretation contra-domain**, and $\cdot^{\mathcal{I}}$ as the **(acceptability dynamics) interpretation function**.

The interpretation function related to an interpretation \mathcal{I}_S brings the possibility to understand the consequences of discarding arguments (those in the contra-domain) from the acceptability analysis. \mathcal{I}_S proposes a further epistemic state, i.e., an evolutive step of the AF, in which we can see the concrete implications by observing the corresponding interpreted formula and its triples. As being specified above, the interpretation structure counts with two mutually exclusive domains and with an interpretation triple composed by an argument $a \in \mathbf{A}$ which is a support of the interpreted formula ϑ , and two sets of arguments, an $a\text{-core}_S$ and an $a\text{-remainder}_S$. The $a\text{-remainder}_S$ indicates which are the arguments contained in the contra-domain that interfere with the acceptability of a and the $a\text{-core}_S$ specifies which is the core set of arguments that will ensure a positive acceptability for a in the interpretation’s proposed evolutive step, i.e., the AF determined by the interpretation domain: $\mathbb{F}_{\Delta^{\mathcal{I}}}$. Note that the evolved AF is closed.

Example 4 (From Ex. 3) Assume an interpretation $\mathcal{I}_S = \langle \Delta^\mathcal{I}, \Gamma^\mathcal{I}, \cdot^\mathcal{I} \rangle$ for the AF \mathbb{F}_A , where $\Delta^\mathcal{I} = A \setminus \Gamma^\mathcal{I}$ and $\Gamma^\mathcal{I} = \{a, b, e, f\}$. The interpretation function $\cdot^\mathcal{I}$ applied over a formula $\vartheta = \neg p \vee \neg q_2$ ends up containing four triples, $\vartheta^\mathcal{I} = \{(c, \{c, d\}, \{a, e, f\}), (d, \{d\}, \{a, e, f\}), (d, \{d\}, \{b, e\}), (g, \{g\}, \{a, b, e, f\})\}$. For a formula like $p \wedge q_2$, the function $\cdot^\mathcal{I}$ brings some empty triples: $(p \wedge q_2)^\mathcal{I} = \{(b, \emptyset, \emptyset), (e, \emptyset, \emptyset)\}$ since both b and e are part of the contra-domain, and moreover, for a formula like $\neg q_1$, the function $\cdot^\mathcal{I}$ ends up empty, i.e., $(\neg q_1)^\mathcal{I} = \emptyset$, since there is no argument in A supporting $\neg q_1$.

A dynamics interpretation may be accurate for the positive acceptability of some formulae. This is captured by the notion of *interpretation model*.

Definition 8 (Interpretation Model) Given an AF τ , an interpretation \mathcal{I}_S , and a formula $\vartheta \in \mathcal{L}$; we say \mathcal{I}_S is a **model** of ϑ in τ , noted $\mathcal{I}_S \models \vartheta$ iff there is some triple $(a, X, Y) \in \vartheta^\mathcal{I}$ such that $X \neq \emptyset$ holds. On the contrary, if there is no such triple, we say \mathcal{I}_S is not a model of ϑ , writing $\mathcal{I}_S \not\models \vartheta$.

Proposition 3 if $\mathcal{I}_S \models \vartheta$ then for any $(a, X, Y) \in \vartheta^\mathcal{I}$, it holds 1) a is \mathcal{S} -accepted in $\mathbb{F}_{\Delta^\mathcal{I}}$ and 2) a is \mathcal{S} -accepted in $\mathbb{F}_{A \setminus Y}$.

The previous proposition shows that when $Y \subset \Gamma^\mathcal{I}$, the argument's acceptability is unaffected by the additional contra-domain's arguments given that they do not belong to the associated remainder set Y . This brings about the need for restricting the contra-domain only to the exclusively necessary arguments to achieve a further positive acceptability. This can be understood as a pathway to minimal change (discussed later). We look for a construction which minimizes the contra-domain. This means that a *minimal model* will ensure that each argument in the contra-domain is necessarily there in order for ϑ to be accepted.

Definition 9 (Minimal Model of a Formula) Given an AF τ , an interpretation \mathcal{I}_S , and a formula $\vartheta \in \mathcal{L}$ such that $\mathcal{I}_S \models \vartheta$, we say \mathcal{I}_S is a **minimal model** of ϑ in τ iff there is no interpretation \mathcal{I}'_S such that $\mathcal{I}'_S \models \vartheta$, where $\Gamma^{\mathcal{I}'} \subset \Gamma^\mathcal{I}$ holds.

Proposition 4 \mathcal{I}_S is a minimal model of ϑ iff for every $(a, X, Y) \in \vartheta^\mathcal{I}$, $Y = \Gamma^\mathcal{I}$.

Example 5 (From Ex. 4) Although \mathcal{I}_S models ϑ , it is not a minimal model since its contra-domain $\Gamma^\mathcal{I}$ strictly contains $\{a, e, f\} \subseteq \Gamma^\mathcal{I}$ and $\{b, e\} \subseteq \Gamma^\mathcal{I}$, defining two minimal models $\mathcal{I}_S^1 = \langle \{b, c, d, g\}, \{a, e, f\}, \cdot^{\mathcal{I}^1} \rangle$ and $\mathcal{I}_S^2 = \langle \{a, c, d, f, g\}, \{b, e\}, \cdot^{\mathcal{I}^2} \rangle$. Observe that ϑ is supported by c, d , and f ; several alternatives are available. Through the minimal model \mathcal{I}_S^1 , ϑ is \mathcal{S} -accepted through the acceptance of c and d , while in the case of the second minimal model \mathcal{I}_S^2 , ϑ is \mathcal{S} -accepted through the acceptance of d and f . Finally, $\vartheta^{\mathcal{I}^1} = \{(c, \{c, d\}, \{a, e, f\}), (d, \{d\}, \{a, e, f\})\}$ and $\vartheta^{\mathcal{I}^2} = \{(d, \{d\}, \{b, e\}), (f, \{a, d, f\}, \{b, e\})\}$.

It is observable a fine distinction between what an interpretation says about the acceptability of a modeled formula in the current epistemic state and its acceptability dynamics on a further epistemic state. As we have already discussed, an interpretation \mathcal{I}_S proposes an alternative framework $\mathbb{F}_{\Delta^\mathcal{I}}$ in which a modeled formula ϑ would end up accepted. However, much more than this can be said.

We refer as *dynamic model* to that which proposes an alternative of change for modifying the current acceptability state of ϑ , from rejected to accepted. Such a situation can be observed when for every triple $(a, X, Y) \in \vartheta^{\mathcal{I}}$ both X and Y are always non empty sets, showing that in the current epistemic state there is always an *a-remainder_S* set, *i.e.*, a set of arguments blocking the acceptability of the supporter of ϑ , *i.e.*, a . On the other hand, when there is some triple $(a, X, Y) \in \vartheta^{\mathcal{I}}$ where $X \neq \emptyset$ and $Y = \emptyset$, we have that there is no *a-remainder_S* given the *a-core_S* X , which means that ϑ is already accepted through argument a in the current epistemic state $\mathbb{F}_{\mathbf{A}}$, and also in the further epistemic state $\mathbb{F}_{\Delta^{\mathcal{I}}}$ where the *a-core_S* X is built. We refer to such interpretation as *static model*. In summary, if $\mathcal{I}_{\mathcal{S}}$ statically models ϑ we can ensure ϑ is accepted in the current framework $\mathbb{F}_{\mathbf{A}}$ as well as in $\mathbb{F}_{\Delta^{\mathcal{I}}}$, on the other hand, if $\mathcal{I}_{\mathcal{S}}$ dynamically models ϑ we can ensure ϑ is rejected in the current framework $\mathbb{F}_{\mathbf{A}}$ whereas it is accepted in $\mathbb{F}_{\Delta^{\mathcal{I}}}$.

As we have seen before, when for every triple $(a, X, Y) \in \vartheta^{\mathcal{I}}$, X is an empty set, we would be considering an interpretation which does not model ϑ . However, this can be still meaningful. Whenever, Y is a non-empty set, we have an interpretation which ensures that ϑ is rejected in the current epistemic state, given that it is possible to identify a set of arguments blocking its acceptability (the *a-remainder_S* set Y), but since no *a-core_S* set can be built considering only arguments from the interpretation domain $\Delta^{\mathcal{I}}$, we can infer that the contra-domain contains arguments that are needed for constructing an *a-core_S* set. We refer to such an interpretation as a *contra-model*. On the other hand, when both X and Y are empty, we know ϑ will not be accepted in $\mathbb{F}_{\Delta^{\mathcal{I}}}$, however the interpretation does not tell anything about ϑ 's acceptability in $\mathbb{F}_{\mathbf{A}}$ –given that it may be the case (or not) that some *a-remainder_S* set is constructible but not from the arguments in the contra-domain– and therefore, the acceptability dynamics of ϑ will be unknown. Such an interpretation will be referred as a *failure* for ϑ .

$X \neq \emptyset$	$Y \neq \emptyset$	$a \in \mathcal{A}_{\mathcal{S}}(\mathbb{F}_{\mathbf{A}})$	$a \in \mathcal{A}_{\mathcal{S}}(\mathbb{F}_{\Delta^{\mathcal{I}}})$	Referred as
✓	✓	×	✓	Dynamic Model
✓	×	✓	✓	Static Model
×	✓	×	×	Contra-Model
×	×	?	×	Failure

Example 6 (From Ex. 5) Both $\mathcal{I}_{\mathcal{S}}^1$ and $\mathcal{I}_{\mathcal{S}}^2$ are dynamic models for ϑ . Let $\vartheta' = p \wedge q_2$, and two interpretations $\mathcal{I}_{\mathcal{S}}^3 = \langle \{a, b, d, e, f\}, \{c, g\}, \cdot^{\mathcal{I}^3} \rangle$, where $(b, \{a, b, e\}, \{\}) \in \vartheta'^{\mathcal{I}^3}$, and $\mathcal{I}_{\mathcal{S}}^4 = \langle \{a, b, c, e\}, \{d, f, g\}, \cdot^{\mathcal{I}^4} \rangle$, where $(b, \{b\}, \{\}) \in \vartheta'^{\mathcal{I}^4}$. Both interpretations are static models for ϑ' since b is \mathcal{S} -accepted in $\mathbb{F}_{\mathbf{A}}$ and its positive acceptability is maintained in each evolved AF since *b-core_S* sets are identified in each case. Note that the “canonical interpretation” $\langle \mathbf{A}, \emptyset, \cdot^{\mathcal{I}^e} \rangle$ is always a static model for any \mathcal{S} -accepted formula, like ϑ' . A contra-model can be seen by considering $\mathcal{I}_{\mathcal{S}}^5 = \langle \{b, c\}, \{a, d, e, f, g\}, \cdot^{\mathcal{I}^5} \rangle$ where the formula $(\neg p)$ is interpreted as $(c, \{\}, \{a, e, f\}) \in (\neg p)^{\mathcal{I}^5}$. Here we have no *c-core_S* set since a part of it is in the contra-domain (d would be required), which implies $\mathcal{I}_{\mathcal{S}}^5 \not\models \neg p$ and thus $(\neg p)$ will be \mathcal{S} -rejected in $\mathbb{F}_{\Delta^{\mathcal{I}^5}}$. However, since we can build a *c-remainder_S*, we know that there is a set of arguments responsible for the non-acceptability of c and therefore, we can also ensure that $(\neg p)$ is also \mathcal{S} -rejected in $\mathbb{F}_{\mathbf{A}}$. Two cases of failure can be referred to Ex. 4 through the interpretation $\mathcal{I}_{\mathcal{S}}$ for $(p \wedge q_2)$ and $(\neg q_1)$.

We will refer to the special notational convention $\mathcal{I}_S, a \approx \vartheta$ for specifying that $a \in \mathbf{A}$ is an specific argument by which $\mathcal{I}_S \approx \vartheta$ holds. In this sense, it will be also possible to restrict the construction of an interpretation model of a formula ϑ through the acceptability of a specific argument a by requiring $\mathcal{I}_S, a \approx \vartheta$ to be satisfied. Notice that it will be possible to have an interpretation model for a formula ϑ , that is, $\mathcal{I}_S \approx \vartheta$ which does not model ϑ through the acceptability of a particular argument a , i.e., $\mathcal{I}_S, a \not\approx \vartheta$. It is clear that, if $\mathcal{I}_S, a \approx \vartheta$ then $a \in \mathcal{A}_S(\mathbb{F}_{\Delta^T})$ and $\mathbf{cl}(a) \models \vartheta$, and moreover, if \mathcal{I}_S is a static model, we also know that $a \in \mathcal{A}_S(\mathbb{F}_{\mathbf{A}})$. For the specific case in which $\vartheta = \mathbf{cl}(a)$ we will make a slight abuse of notation (for simplicity), writing $\mathcal{I}_S \approx a$ instead of $\mathcal{I}_S, a \approx \mathbf{cl}(a)$. In such a case, we may say that \mathcal{I}_S is a *model of argument* a , although, its formal meaning is more likely to correspond to: \mathcal{I}_S models $\mathbf{cl}(a)$ through the acceptability of argument a . Afterwards, with a slight abuse of notation, it will also be possible to write $(X, Y) \in a^T$ as a shortcut for $(a, X, Y) \in \mathbf{cl}(a)^T$. Finally, we say \mathcal{I}_S is a *minimal model of argument* a iff $\mathcal{I}_S \approx a$ and for every $(X, Y) \in a^T$, $Y = \Gamma^T$ holds.

5. Argumentation Dynamics through Contra-Semantics

We say that an argument $a \in \mathbb{A}_{\mathcal{L}}$ is *external to the AF* $\mathbb{F}_{\mathbf{A}}$ (or just, *external*) iff $a \notin \mathbf{A}$. An *expansion operation* incorporates an external argument ensuring a new resulting closed AF. Thus, given an AF $\mathbb{F}_{\mathbf{A}}$ and an external argument $a \in \mathbb{A}_{\mathcal{L}}$, the operator $+$ stands for an *expansion* iff $\mathbb{F}_{\mathbf{A}} + a = \mathbb{F}_{\mathbf{C}(\mathbf{A} \cup \{a\})}$.

We identify the domain of all interpretation structures of a given AF τ through the set $\mathbb{I}_S^T \subseteq \wp(\mathbb{A}_{\mathcal{L}}) \times \wp(\mathbb{A}_{\mathcal{L}}) \times \mathbb{W}_{\mathcal{L}}$, in addition, we identify the *set of all minimal models* of an argument $a \in \mathbb{A}_{\mathcal{L}}$ in τ through the operator $\mathcal{M}_S(a, \tau) \subseteq \mathbb{I}_S^T$. Next we define a *selection function* for identifying “the best” minimal model.

Definition 10 (Minimal Model Selection) *Given an AF $\tau = \mathbb{F}_{\mathbf{A}}$, a semantics specification \mathcal{S} , and an argument $a \in \mathbf{A}$; a **minimal model selection** is obtained by a **selection function** $\gamma : \wp(\mathbb{I}_S^T) \rightarrow \mathbb{I}_S^T$ applied over the set $\mathcal{M}_S(a, \tau)$ for selecting some minimal model of a in τ , where $\gamma(\mathcal{M}_S(a, \tau)) \in \mathcal{M}_S(a, \tau)$ is such that for every $\mathcal{I}_S \in \mathcal{M}_S(a, \tau)$ it holds $\gamma(\mathcal{M}_S(a, \tau)) \preceq_{\gamma} \mathcal{I}_S$, where \preceq_{γ} is a **selection criterion** by which it is possible to select the best representative minimal model.*

The selection criterion can be any method for ordering sets of arguments which takes in consideration any possible *perspective of minimal change*. Probably, the simplest perspective is to prefer the models of smaller contra-domain in order to remove as less arguments as possible (for instance, in Ex. 5, \mathcal{I}_S^2 should be preferred over \mathcal{I}_S^1), however, the criterion should look deeper into the set for deciding among several models with contra-domains of identical cardinality. A different perspective of minimal change could be to prefer those minimal models whose proposed evolutive step removes as less as possible conflicts between pairs of arguments, thus looking for a minimal change regarding the morphology of the graph of arguments. But probably, the most powerful and distinctive advantage of relying upon contra-semantics, for analyzing and selecting minimal models, is that we can study the impact of change directly over the resulting acceptable set. Not only for deciding to reduce as less as possible the acceptable set, but also for

making a selective change operation which could value the positive acceptability of certain arguments more than others, or even for analyzing the classification of models in order to keep as controlled as possible the number of dynamic models, and to avoid reducing the number of static models while keeping as low as possible the cardinality of contra-models. Such a discussion deserves to be deepened with more space, and is part of the ongoing work about this theory. An *acceptance revision* will incorporate an external argument a to the AF ensuring the positive acceptability of a by referring to a minimal model selection.

Definition 11 (Acceptance Revision) *Given an AF τ , a semantics specification \mathcal{S} , and an external argument $a \in \mathbb{A}_{\mathcal{L}}$; the operator \otimes stands for an **acceptance revision** iff $\tau \otimes a = \mathbb{F}_{\Delta^{\mathcal{I}}}$, where $\Delta^{\mathcal{I}}$ is the interpretation domain of the selected minimal model $\mathcal{I}_{\mathcal{S}} = \gamma(\mathcal{M}_{\mathcal{S}}(a, \tau + a))$. When necessary we will write $\tau \otimes_{\gamma} a$ to identify the minimal model selection γ by which the revision $\tau \otimes a$ is obtained.*

The axiomatization of the acceptance revision is achieved by analyzing the different characters of revisions from classical belief revision [2,10] and from ATC revision [12], for adapting the classical postulates to argumentation. For space reasons, we will not discuss the intuitions motivating each postulate. For a detailed discussion on this matter, the interested reader may refer to [11,12].

- (closure) if $\mathbf{A}(\tau) = \mathbb{C}(\mathbf{A}(\tau))$ then $\mathbf{A}(\tau \otimes a) = \mathbb{C}(\mathbf{A}(\tau \otimes a))$
- (success) a is \mathcal{S} -accepted in $\tau \otimes a$
- (consistency) $\mathcal{A}_{\mathcal{S}}(\tau \otimes a)$ is conflict-free
- (inclusion) $\mathbf{A}(\tau \otimes a) \subseteq \mathbf{A}(\tau + a)$
- (vacuity) If a is \mathcal{S} -accepted in $\tau + a$ then $\mathbf{A}(\tau + a) \subseteq \mathbf{A}(\tau \otimes a)$
- (core-retainment) If $b \in \mathbf{A}(\tau) \setminus \mathbf{A}(\tau \otimes a)$ then exists an AF τ' such that $\mathbf{A}(\tau') \subseteq \mathbf{A}(\tau)$ and a is \mathcal{S} -accepted in $\tau' + a$ but \mathcal{S} -rejected in $(\tau' + b) + a$
- (uniformity) if $a \equiv b$ then $\mathbf{A}(\tau) \cap \mathbf{A}(\tau \otimes a) = \mathbf{A}(\tau) \cap \mathbf{A}(\tau \otimes b)$

The uniformity postulate makes reference to an *equivalence relation* “ \equiv ” for arguments (see [12]) to ensure that the revisions $\tau \otimes a$ and $\tau \otimes b$ have equivalent outcomes when arguments a and b are equivalent. For any pair of arguments $a, b \in \mathbb{A}_{\mathcal{L}}$, we say that a and b are *equivalent arguments*, noted as $a \equiv b$ iff $\text{cl}(a) \models \text{cl}(b)$ and $\text{cl}(b) \models \text{cl}(a)$ and for any $a' \sqsubset a$ there is $b' \sqsubset b$ such that $a' \equiv b'$. Inspired by smooth incisions in Hansson’s Kernel Contractions [10], we introduce an additional condition on minimal models selection functions for guaranteeing uniformity. Under the consideration of two equivalent arguments a and b , the idea is to ensure that the selection function will trigger one minimal model for each argument (a and b) whose interpretation domains are identical except for the presence of a or b in each corresponding case. Note that we refer to the expansion closure operator \mathbb{P} for looking at the common base of each interpretation domain.

Definition 12 (Smooth Minimal Model Selection) *Given an AF τ and two external arguments $a, b \in \mathbb{A}_{\mathcal{L}}$. If $a \equiv b$ then $\Delta^{\mathcal{I}^a} \setminus \mathbb{P}(\{a\}) = \Delta^{\mathcal{I}^b} \setminus \mathbb{P}(\{b\})$, where $\mathcal{I}^a_{\mathcal{S}} = \gamma(\mathcal{M}_{\mathcal{S}}(a, \tau + a))$ and $\mathcal{I}^b_{\mathcal{S}} = \gamma(\mathcal{M}_{\mathcal{S}}(b, \tau + b))$.*

An operation $\tau \otimes_{\gamma} a$ is a *smooth acceptance revision* iff $\tau \otimes_{\gamma} a$ is an acceptance revision obtained through a smooth minimal model selection ‘ γ ’.

Representation Theorem 1 *Given an AF τ , a semantics specification \mathcal{S} , and an external argument $a \in \mathbb{A}_{\mathcal{L}}$; $\tau \circledast a$ is a smooth acceptance revision iff ' \circledast ' satisfies closure, success, consistency, inclusion, vacuity, core-retainment, and uniformity.*

6. Related Work & Conclusions

A revision approach in an AGM spirit is presented in [6] through revision formulae that express how the acceptability of some arguments should be changed. As a result, they derive argumentation systems which satisfy the given revision formula, and are such that the corresponding extensions are as close as possible to the extensions of the input system. The revision presented is divided in two subsequent levels: firstly, revising the extensions produced by the standard semantics. This is done without considering the attack relation. Secondly, the generation of argumentation systems fulfilling the outcome delivered by the first level. Minimal change is pursued in two different levels, firstly, by ensuring as less change as possible regarding the arguments contained in each extension, and secondly, procuring as less change as possible on the argumentation graph. The methods they provide do not provoke change upon the set of arguments, but only upon the attack relations. Their operator is more related to a distance based-revision which measures the differences from the actual extensions with respect to the ones obtained for verifying the revision formula. They give a basic set of rationality postulates in the very spirit of AGM, but closer to the perspective given in [9]. They only show that the model presented satisfies the postulates without giving the complete representation theorem for which the way back of the proof, *i.e.*, from postulates to the construction, is missing. However, the very recent work [7], which is in general a refinement of [6] and [5], proposes a generic solution to the revision of argumentation frameworks by relying upon complete representation theorems. In addition, the revision from the perspective of argumentation frameworks is also considered. A different approach, but still in an AGM spirit was presented in [3], where authors propose expansion and revision operators for Dung's abstract argumentation frameworks (AFs) based on a novel proposal called *Dung logics* with the particularity that equivalence in such logics coincides with strong equivalence for the respective argumentation semantics. The approach presents a reformulation of the AGM postulates in terms of monotonic consequence relations for AFs. They finally state that standard approaches based on measuring distance between models are not appropriate for AFs.

The aforementioned works differ from ours in the perspective of dealing with the argumentation dynamics. This also renders different directions to follow for achieving rationality. To our knowledge, [12] was the first work to propose AGM postulates for rationalizing argumentation dynamics, providing also complete representation theorems for the proposed revision operations built upon logic-based argumentation. The rationalization done here is mainly inspired by such results.

The main objective of the *dynamics contra-semantics* is to bring a new theoretical structure conceived from scratch to deal with acceptability dynamics of arguments. The expected virtue of this theory is to ease the proposal and rationality analysis of new models of argumentative change. We believe that it could be

simpler to show that the outcome of a “rational” change operator coincides with an interpretation model than showing the complete rationality through a representation theorem. If this hypothesis is true, the full rationality of new change operators could be achieved by means of the representation theorem here presented. In this sense, the intuitions behind the notions of core and remainder sets exceed the scope of the standard argumentation semantics. Their constructions can be redefined for being applied over other kind of frameworks like abstract and dialectical argumentation. For instance, the concept of remainders could match well as a generalization of the idea proposed in ATC [13,12] about selectable con-arguments from a set of attacking lines in a dialectical tree [14] (argumentation lines whose parity interfere with the possibility of acceptance of the root argument). The study of the dynamics contra-semantics upon dialectical argumentation seems to be possible also, given that the reference to standard argumentation semantics in this work has been parametrized, thus allowing the modeling of marking criteria for trees of arguments. The intention would be to bring a formal methodology for studying acceptability dynamics upon an argumentation which fits better for reasoning about a main issue in dispute, *i.e.*, a root argument, as done in dialogues and legal reasoning. This is part of the ongoing work.

References

- [1] IJCAI 2015, Buenos Aires, Argentina, 2015. AAAI Press (2015)
- [2] Alchourrón, C., Gärdenfors, P., Makinson, D.: *On the Logic of Theory Change: Partial Meet Contraction and Revision Functions*. Journal of Symbolic Logic 50, 510–530 (1985)
- [3] Baumann, R., Brewka, G.: AGM Meets Abstract Argumentation: Expansion and Revision for Dung Frameworks. In: IJCAI 2015, Buenos Aires, Argentina, 2015 [1], pp. 2734–2740
- [4] Besnard, P., Hunter, A.: *Elements of Argumentation*. The MIT Press (2008)
- [5] Coste-Marquis, S., Konieczny, S., Mailly, J., Marquis, P.: A Translation-Based Approach for Revision of Argumentation Frameworks. In: JELIA 2014, Madeira, Portugal, 2014. Proceedings. LNCS, vol. 8761, pp. 397–411. Springer (2014)
- [6] Coste-Marquis, S., Konieczny, S., Mailly, J., Marquis, P.: On the Revision of Argumentation Systems: Minimal Change of Arguments Statuses. In: KR 2014, Vienna, Austria, 2014. AAAI Press (2014)
- [7] Diller, M., Haret, A., Linsbichler, T., Rümmele, S., Woltran, S.: An Extension-Based Approach to Belief Revision in Abstract Argumentation. In: IJCAI 2015, Buenos Aires, Argentina, 2015 [1], pp. 2926–2932
- [8] Dung, P.M.: On the Acceptability of Arguments and its Fundamental Role in Nonmonotonic Reasoning and Logic Programming and n -person Games. Artif. Intell. 77, 321–357 (1995)
- [9] Gärdenfors, P.: *Knowledge in Flux: Modelling the Dynamics of Epistemic States*. The MIT Press, Bradford Books, Cambridge, Massachusetts (1988)
- [10] Hansson, S.O.: *A Textbook of Belief Dynamics. Theory Change and Database Updating*, Kluwer Academic (1999)
- [11] Moguillansky, M.O.: A Study of Argument Acceptability Dynamics Through Core and Remainder Sets. In: FoIKS 2016, Austria. LNCS, vol. 9616, pp. 3–23. Springer (2016)
- [12] Moguillansky, M.O., Wassermann, R., Falappa, M.A.: Inconsistent-tolerant base revision through Argument Theory Change. Logic Journal of the IGPL 20(1), 154–186 (2012)
- [13] Rotstein, N.D., Moguillansky, M.O., Falappa, M.A., García, A.J., Simari, G.R.: Argument Theory Change: Revision Upon Warrant. In: Proc. of COMMA. pp. 336–347 (2008)
- [14] Rotstein, N.D., Moguillansky, M.O., Simari, G.R.: Dialectical Abstract Argumentation: A Characterization of the Marking Criterion. In: Boutilier, C. (ed.) IJCAI. pp. 898–903 (2009)

Construction and Strength Calculation of Threats

Mariela MORVELI-ESPINOZA ^a, Ayslan T. POSSEBOM ^b and Cesar A. TACLA ^a

^aCPGEI- Federal University of Technology - Paraná, Brazil

^bFederal Institute of Parana - Paranaíba, Brazil

Abstract. Threats make part of the set of rhetorical arguments, which are used in negotiation dialogues when a proponent agent tries to persuade his opponent to accept a proposal more readily. When more than one threat is generated, the proponent must evaluate each and select the most adequate. One way of evaluation is calculating the strength of threats, since a strong threat may quickly convince an opponent.

The contribution of this paper is twofold. On the one hand, we present a mechanism of generation of threats and, on the other hand, we propose a model for calculating the strength of threats, which is based on the goal processing inside the mental state of the opponent. We propose two ways for calculating the strength of threats depending on the kind of negotiation the agent is participating. The first proposal is to be used when the agent negotiates only with one opponent, and the second when the agent negotiates with more than one opponent.

Keywords. threats, rhetorical arguments, arguments strength, persuasive negotiation

1. Introduction

Persuasive negotiation involves negotiating using rhetorical arguments (such as threats, rewards, or appeals), which act as persuasive elements that aim to force or convince an opponent to accept a given proposal [1]. This work is focused on threats, which carry out sanctions when the opponent does not agree with the proposal sent by the proponent. According to Sycara [2] threats are the rhetorical arguments with most persuasive power, thus when an agent wants to achieve an own goal, threatening an important goal of his opponent is the most effective of arguments.

Let's see the following scenario where *mom* is a proponent agent, *son* an opponent agent and the goal of *mom* is that *son* does his homework¹. Taking into account the knowledge base of agent *mom*, the following threats can be generated:

- *mom*: if you do not do your homework, you will not go to the party the weekend.
- *mom*: if you do not do your homework, I will not buy the Nintendo.

The question is which of these threats will *mom* choose to persuade *son* to do the homework? One way of knowing this is by calculating the strength of threats. According to Sarvapali et al. [1] a strong argument is one that quickly convinces an opponent to do a

¹This scenario is inspired by the example presented in [3].

proposal, while a weak argument is less persuasive. Therefore, calculating the strength of threats is important in persuasive negotiation dialogues, since the quickness of persuasion depends on it. Thus, the stronger a threat is, the quicker an opponent is persuaded.

Threats are built using the goals of both the proponent and the opponent (going to a party this weekend, getting a Nintendo). Some studies on this topic take into account the importance of the goal of the opponent and the certainty level of the beliefs that make up an argument [3,4], however, other criteria are necessary for a more exact calculation. In the following we present some examples of situations that show this:

1. Agent *mom* knows that “getting a Nintendo” (go_2) is a more important goal (for *son*) than “going to the party” (go_1). Considering only the importance, *mom* would use go_2 for generating a threat. However, what happens if *mom* also knows that go_2 is not an achievable goal since the family budget is not enough? (and *son* also knows it).

2. Agent *mom* has already threatened *son* before and rarely she has fulfilled it, and obviously *son* knows about it.

In the first case, it does not matter how important a goal is if it is not possible to be achieved, and in the second case, the strength of a threat is also influenced by the execution credibility that the proponent has from the point of view of his opponent (i.e. the execution level the proponent believes the opponent has about him)). Thus, the aim of this article is to propose a model for calculating the strength of threats by taking into account new criteria, which will lead to a more accurate calculation.

In order to determine how achievable a goal is, we will use the belief-based goal processing (BBGP) model proposed by Castelfranchi and Paglieri [5], which can be considered an extension of the belief-desire-intention (BDI) model [6]. In BBGP model, the processing of goals is divided in four stages: (i) activation, (ii) evaluation, (iii) deliberation, and (iv) checking; and the states a goal can adopt are: (i) active (=desire), (ii) pursuable, (iii) chosen, and (iv) executable (=intention). The state of a goal changes when it passes from one stage to the next. Depending on this state, a goal can be considered more or less threatened, it is considered more threatened when it is closer of the executable state and less threatened when its state is active.

This paper is organized as follows: Section 2 presents the BBGP model in which is mainly based our strength calculation model. In Section 3 a negotiating agent architecture that considers the necessary mental states and functions that support our proposal is defined. A formal definition of threat and the mechanism for the its generation are presented in Section 4. Section 5 shows an analysis of the elements of a threat and our proposed strength calculation model. In Section 6, the main related works are compared with our proposal. Finally, Section 7 is devoted to some conclusions and future work.

2. Belief-based goal processing model

In this section, the four stages of the BBGP model of Castelfranchi and Paglieri are presented². The aim of this section is not to present in detail the beliefs used in each stage. We focus on the states of the goals and make clear when a goal is considered active, pursuable, chosen, and executable, because these states will be used in the strength calculation model that is proposed in this work. Following a brief description of each stage:

²A more detailed version of this model is presented in [5].

1. Activation stage: In this stage, goals are activated by means of motivating beliefs. For example, if the agent has the belief that today is Thursday, it activates the goal of going to the French class, or the motivating belief that today is sunny activates the goal of playing football. When a motivating belief is satisfied, the supported goal becomes *active*. An active goal can also be seen as a desire.

2. Evaluation stage: In this stage, goals are evaluated using assessment beliefs. When there are no assessment beliefs for a certain goal, it becomes *pursuable*. Three types of assessment beliefs were defined: (i) those that control that there is no impossibility for a goal be pursued; (ii) those that control goals that are realized in the world autonomously and without the direct intervention of the agent; and (iii) those that control goals that have already realized, and that will remain as such.

3. Deliberation stage: The aim of this stage is to act as a filter on the basis of incompatibilities and preferences among pursuable goals. Goals that pass this stage are called *chosen* goals. These beliefs are concerned with the different forms of incompatibility among goals that lead an agent to choose among them. For dealing with incompatibilities, an agent uses preference beliefs.

4. Checking stage: The aim of this stage is to evaluate if the agent knows how to achieve a goal and if it is capable of performing the required actions to achieve a chosen goal, in other words if the agent has a plan and he is capable of executing it. Goals that pass this stage are called *executable* goals and have the same characteristics of intentions. These can be executed immediately or saved in the agent's agenda.

3. The agent

In this section, we define the main structures and functions an agent should have in order to be able to generate threats and calculate their strengths³. This architecture is based on the BBGP model, however it is a small fragment as a complete formalization is out of the scope of this article.

Let \mathcal{L} be a first order logical language which will be used to represent the goals and beliefs of the agent. \wedge, \vee and \neg denote the logical connectives conjunction, disjunction and negation, and \vdash stands for the classical inference.

Definition 3.1. (Basic structures) An agent has five basic structures:

- \mathcal{K} is the knowledge base of the agent;
- \mathcal{O}_p stores the opponents of the agent;
- $\mathcal{G} = \mathcal{G}_a \cup \mathcal{G}_p \cup \mathcal{G}_c \cup \mathcal{G}_e$ is the set of goals of the agent, such that \mathcal{G}_a is the set of active goals, \mathcal{G}_p of pursuable goals, \mathcal{G}_c of chosen goals and \mathcal{G}_e of executable goals. It holds that $\mathcal{G}_x \cap \mathcal{G}_y = \emptyset$, for $x, y \in \{a, p, c, e\}$ and $x \neq y$;
- $\mathcal{GO} = \mathcal{GO}_a \cup \mathcal{GO}_p \cup \mathcal{GO}_c \cup \mathcal{GO}_{agd}$ is the set of the goals of the opponent, such that \mathcal{GO}_a is the set of active opponent goals, \mathcal{GO}_p of pursuable ones, \mathcal{GO}_c of chosen ones and \mathcal{GO}_{agd} ⁴ are the goals scheduled in the opponent's agenda. It holds that $\mathcal{G}_x \cap \mathcal{G}_y = \emptyset$ for

³We assume that the agent has in advance the necessary information for generating and calculating the strength of threats. Some interesting works about opponent modelling related to argumentation are [7,8].

⁴We do not consider executable goals that were already executed because they are not useful for constructing a threat. We prefer to use goals that are in the opponent's agenda. For example, if agent *son* already got a Nintendo, threaten this goal would be useless. Therefore, a threat is stronger when the threatened goal is in the opponent's agenda.

$x, y \in \{a, p, c, agd\}$ and $x \neq y$. Finally, let $State(go_i) = z$ be a function that returns the state of a given goal; for $z \in \{1, 2, 3, 4\}$ where 1 means that the goal is active, 2 pursuable, 3 chosen and 4 that it is in the agenda;

- $\mathcal{T}hs$ stores the threats constructed by the agent. The definition of a threat is given in Section 4.

Definition 3.2. (Compound structures) These store characteristics of the basic structures.

- $\mathcal{O}p_{det} = \{(op_i, \delta)\}$ such that $op_i \in \mathcal{O}p$ and δ is the execution credibility level of threats the proponent has from the point of view of opponent op_i . Hereafter, we denote that $\delta \in [0, 1]$ such that δ is a real from the given interval. Let $Level_Exec_{th}(op_i) = \delta$ be a function that returns the execution credibility level for a given opponent agent;

- $\mathcal{G}O_{det} = \{(go_i, \delta, op_j)\}$ such that $go_i \in \mathcal{G}O$ is a goal of opponent $op_j \in \mathcal{O}p$ whose importance is given by δ . Let $Importance(go_i, op_j) = \delta$ be a function that returns the importance of a given goal. The opponent is taken into account as an opponent goal may be the same for more than one opponent, but the importance may be different for each case. Finally, let $Op_Goals(op_j) = \{go_i, \dots, go_k\}$ be a function that returns all the goals of a given opponent;

- $\mathcal{T}hs_{det} = \{(th_i, st_i)\}$ such that $th_j \in \mathcal{T}hs$ is a threat whose strength value is st_i .

4. Construction of threats

A threat is constructed based on two goals:

1. An outsourced goal of the proponent: This kind of goal needs the opponent involvement in order to be achieved. For example, the goal of *mom* is that *son* does his homework, for this goal to be achieved is necessary that *son* executes the required action. Considering the BBGP stages defined in Section 2 the state of this goal is executable.

Definition 4.1. (Outsourced goal) An outsourced goal g_i is an expression of the form $g_i(op_k, g'_i)$, such that, $op_k \in \mathcal{O}p$ and g'_i is an action that op_k has to execute. Let $first(g_i) = op_k$ and $second(g_i) = g'_i$ be the functions that return each component of g_i .

2. The goal of an opponent: It is a goal that the proponent knows its opponent wants to achieve. For example, *mom* knows that *son* wants a Nintendo. Besides knowing the goal of his opponent, the proponent has to know the state of that goal and its importance.

The construction of a threat begins when (i) an outsourced goal g_i passes all the goal processing stages and becomes executable and, (ii) after a failed first attempt of proponent agent to make his opponent to do the requested action g'_i . The process of construction of a threat is the following:

1. Function Op_goals returns the set of all goals the proponent knows an opponent op_j wants to achieve. Let $\mathcal{S}_{go} = Op_goals(op_j)$ be the returned set.
2. If $\mathcal{S}_{go} \neq \emptyset$
 - (a) For each $go_j \in \mathcal{S}_{go}$:
 - i. Generate a threat rule rth_k , which links the proponent and the opponent goals on which is based a threat. This is an expression of the form $\neg g'_i \rightarrow \neg go_j$.

ii. Construct a threat and save it in $\mathcal{T}hs$.

Threats in $\mathcal{T}hs$ are called candidates. After the strength calculation, the strongest one is sent to his opponent to try to persuade him.

Following, we present the definition of a threat, which is based on the definition given in [3], with some modifications that consider the agent architecture proposed in Section 3.

Definition 4.2. (Threat) A threat is a triple $th = \langle rth_k, g'_i, go_j \rangle$, where:

- rth_k is a threat rule,
- $g'_i = \text{second}(g_i)$, such that $g_i \in \mathcal{G}_e$,
- $rth_k \cup \{\neg g'_i\} \vdash \neg go_j$ such that $go_j \in \mathcal{GO}$.

Let's call rth_k and g'_i the support of the threat and go_j its conclusion.

Example 4.1. Let us define the mental state of agent *mom*:

$\mathcal{G}_e = g_1$ where $g_1 = \text{make}(\text{son}, \text{do}(\text{homework}))$ is an outsourced goal,
 $\mathcal{GO}_p = \{go_2\}$ where $go_2 = \text{have}(\text{nintendo})$, $\mathcal{GO}_c = \{go_1\}$ where $go_1 = \text{go}(\text{party})$,
 $\mathcal{GO}_{agd} = \{go_3\}$ where $go_3 = \text{go}(\text{skating})$, $\mathcal{GO}_{det} = \{(go_1, 0.8, \text{son}), (go_2, 0.5, \text{son}), (go_3, 0.3, \text{son})\}$,
 $\mathcal{OP} = \{\text{son}\}$, $\mathcal{OP}_{det} = \{(\text{son}, 1)\}$, $\mathcal{T}hs = \{\}$, $\mathcal{T}hs_{det} = \{\}$

Let us suppose that agent *son* rejected to do action $g'_1 = \text{do}(\text{homework})$. Therefore, *mom* begins the process of construction of candidate threats:

1. $\mathcal{S}_{go} = \mathcal{OP_goals}(\text{son}) = \{go_1, go_2, go_3\}$
2. $\mathcal{S}_{go} \neq \emptyset$, then
 - (a) For go_1 , generate $rth_1 = \neg g'_1 \rightarrow \neg go_1$ and construct $th_1 = \langle (rth_1, g'_1, go_1) \rangle$
 - (b) For go_2 , generate $rth_2 = \neg g'_1 \rightarrow \neg go_2$ and construct $th_2 = \langle (rth_2, g'_1, go_2) \rangle$
 - (c) For go_3 , generate $rth_3 = \neg g'_1 \rightarrow \neg go_3$ and construct $th_3 = \langle (rth_3, g'_1, go_3) \rangle$

Therefore, $\mathcal{T}hs = \{th_1, th_2, th_3\}$

5. Strength calculation

The strength of a threat is mainly based on the “value” that the threatened goal has for the opponent. Besides, the credibility the proponent has in the face of his opponent(s) regarding his ability to execute his threats is an aspect that also influences the strength calculation. The strength calculation is done after the agent generates all the candidate threats and it is done for all candidate threats.

The strength calculation of a threat depends on:

1. The goal of the opponent go_i (or threatened goal): two aspects are considered:

- *The importance of goal go_i :* like in some related works ([3,4,11]), we will take into account the importance of the threatened goal.
- *The state of goal go_i :* Let's recall that we use 1 for denoting that a goal is active, 2 for denoting that it is pursuable, 3 for denoting that it is chosen and 4 when the goal is executable but it has not been executed yet and hence it is in the agenda of the opponent.

2. Execution credibility level: It is also important that the proponent agent be able to execute its threats, from the point of view of his opponent. This value (represented in the proponent) reflects what the proponent believes the opponent thinks about his execution level and can be different for each opponent.

Considering these aspects, the formalization of our proposal is defined as follows.

Definition 5.1. (Basic strength of threats) The basic strength of a threat depends on the importance and the state of the threatened goal. Let $th = \langle rth_k, g'_i, go_j \rangle$ be a threat, the basic strength of th is obtained applying:

$$ST_{basic}(th) = \min \left(\frac{State(go_j)}{num_states}, Importance(go_j) \right) \quad (1)$$

where $num_states = 4$ is the number of goal states.

A direct consequence of the above definition is that the value of the basic strength of a threat is a real value between 0 and 1. Formally:

Property 5.1. Let $th = \langle rth_k, g'_i, go_j \rangle$ be a threat. Since the value of the importance of $go_j \in [0, 1]$ and the normalization of the states value of go_j is also between 0 and 1: $ST_{basic}(th) \in [0, 1]$, where 0 represents the minimum value and 1 represents the maximum value the basic strength can have.

When the proponent agent constructs a set of threats only for one opponent, the value of the basic strength is enough to choose the threat that will be sent. Nevertheless, a more exact value can be obtained if the execution credibility level is also considered. This aspect is even more important when the proponent agent generates threats for more than one opponent as it will let him know which opponent may be weaker when faced with one of his threats.

Definition 5.2. (Combined strength of threats) The combined strength of a threat depends on the basic strength of the threat and the execution credibility level of the proponent. Let $th = \langle rth_k, g'_i, go_j \rangle$ be a threat and $op_n \in \mathcal{O}p$ the opponent whose threatened goal is go_j . The combined strength of th is obtained applying:

$$ST_{comb}(th) = ST_{basic}(th) \times Level_Exec_{th}(op_n) \quad (2)$$

Property 5.2. The maximum value of the combined strength of a threat is at most the value of its basic strength: $ST_{comb}(th) \in [0, ST_{basic}(th)]$.

Example 5.1. Let us continue with example 4.1:

$$State(go_1) = 3, Importance(go_1) = 0.8 \quad State(go_3) = 3, Importance(go_3) = 0.3$$

$$State(go_2) = 4, Importance(go_2) = 0.5$$

$Level_Exec_{th}(son) = 1$ is the execution credibility level of *mom* from the point of view of *son*.

Applying equation 1, the basic strengths of threats in $\mathcal{T}hs$ are $ST_{basic}(th_1) = 0.75$, $ST_{basic}(th_2) = 0.5$ and $ST_{basic}(th_3) = 0.3$. Since agent *mom* generated threats only for one opponent, she can choose the strongest one without calculating the combined strengths, even more when in this case the values of the combined strengths are the same of the basic strengths.

Finally, this information must be added to structure $\mathcal{T}hs_{det}$. Let's recall that the second component of each element of $\mathcal{T}hs_{det}$ is the value of the strength.

Therefore, $\mathcal{T}hs_{det} = \{(th_1, 0.75), (th_2, 0.5), (th_3, 0.3)\}$ and *mom* would send th_1 because it is the strongest threat.

Example 5.2. Let us suppose that *mom* has another goal: $make(x, 'clean(house)')$ and has two opponents: *son* and *daughter*. Let us also suppose that *mom* has generated three threats for *son* and three threats for *daughter*. The threats for *son* have the following basic strengths: $ST_{basic}(th_5) = 0.85$, $ST_{basic}(th_6) = 0.6$ and $ST_{basic}(th_7) = 0.4$ and the threats of *daughter* have the following ones: $ST_{basic}(th_8) = 0.9$, $ST_{basic}(th_9) = 0.45$ and $ST_{basic}(th_{10}) = 0.4$, with $Level_Exec_{th}(daughter) = 0.8$.

Taking into consideration only the basic strengths, the strongest threat is $th_8 = 0.9$. This means that *mom* could send this threat to *daughter* as it seems that it would be more effective than sending a threat to *son*. However, the execution credibility level of *mom* (from the point of view of *daughter*) is lower than the execution credibility level for *son*. Thus, the combined strengths (equation 2) have the same values of the basic strengths for *son*, but different values for *daughter*:

<i>son</i>	<i>daughter</i>
$ST_{comb}(th_5) = 0.85 \times 1 = 0.85$	$ST_{comb}(th_8) = 0.9 \times 0.8 = 0.72$
$ST_{comb}(th_6) = 0.6 \times 1 = 0.6$	$ST_{comb}(th_9) = 0.45 \times 0.8 = 0.36$
$ST_{comb}(th_7) = 0.4 \times 1 = 0.4$	$ST_{comb}(th_{10}) = 0.4 \times 0.8 = 0.32$

Therefore, the best option for *mom* is to send threat th_5 to her opponent *son*.

As in the previous example, these values must be added to $\mathcal{T}h_{det}$. Notice that the strength value of a threat can be obtained using either the basic strength equation or the combined one.

6. Related works

Kraus et al. [10] present a set of axioms for threats generation. In these axioms, when the rule body is satisfied, a candidate threat is generated. For selecting an action the proponent will be able to execute, the time and the opponent's preference value are considered. In our proposal the opponent's preferences are also taken into account, but as part of the goal processing. Thus, when a goal state is chosen it means that it is most preferable with relation to other goals. With respect to the strength of threats, the authors claim that a threat is the strongest rhetorical argument (compared to rewards and appeals), however a calculation model is not defined.

Servapali et al. [1] propose a model where the rhetorical strength of threats varies during the negotiation depending on the environmental conditions. For calculating the strength value of threats, it is taken into account a set of world states an agent can be carried to by using a certain threat. The intensity of the strength depends on the desirability of each of these states. For a fair calculation, an average over all possible states is used. The criteria and the way of calculation are completely different from our proposal and threats generation is not studied.

In the work of Amgoud and Prade [3], a formal definition of threats and an evaluation system are presented. For the evaluation of strength of threats, the certainty of beliefs that are used for the generation of the threat and the importance of the goal of the opponent are considered. The same authors have other later articles about rhetorical arguments

([4,11]). In these works, the calculation of strength of threats is done always by taking into account the two criteria previously mentioned. For our proposal, we made a further analysis of the components of a threat and defined new criteria for calculating the strength of threats. Another difference is in relation to the threat rule, meanwhile in these works it is part of the knowledge base since the beginning, in our work it is constructed from an own goal of the proponent agent and the goals of the opponent, giving more flexibility for the generation of threats.

7. Conclusions and future work

This work makes a further analysis of the components of a threat and considers new criteria for the calculation of the strength of threats. Using these criteria, two forms for calculating the strength of a threat were proposed: the basic strength and the combined strength calculation. These two different ways of calculus is one of the advantages of our proposal as, depending on the need of the situation, the proponent can use either the basic or the combined equation.

It will be interesting to do this kind of analysis for other rhetorical arguments (like rewards and appealings). This is object of future work.

We also presented a process for threats construction and an agent architecture based on the BBGP model of Castelfranchi and Paglieri. We believe that our proposed process gives more flexibility for the generation of threats as the threat rules are generated dynamically from the set goals of the opponent, which can be updated depending on the new information the agent receives.

Finally, we will work on the experience-based calculation, which is a calculation after the proponent receives an answer to his threat. We think that it could be worthwhile in the study of strength calculation and it could be used in the opponent modeling.

References

- [1] Sarvapali D Ramchurn, Nicholas R Jennings, and Carles Sierra. Persuasive negotiation for autonomous agents: A rhetorical approach. *IJCAI Workshop on Computational Models of Natural Argument*, pages 9-17, 2003.
- [2] Katia P Sycara. Persuasive argumentation in negotiation. *Theory and decision*, 28(3):203-242, 1990.
- [3] Leila Amgoud and Henri Prade. Threat, reward and explanatory arguments: generation and evaluation. In *Proceedings of the ECAI Workshop on Computational Models of Natural Argument*, pages 73-76, 2004.
- [4] Leila Amgoud and Henri Prade. Formal handling of threats and rewards in a negotiation dialogue. In *Argumentation in Multi-Agent Systems*, pages 88103. Springer, 2006.
- [5] Cristiano Castelfranchi and Fabio Paglieri. The role of beliefs in goal dynamics: Prolegomena to a constructive theory of intentions. *Synthese*, 155(2):237-263, 2007.
- [6] Michael Bratman. Intention, plans, and practical reason. 1987.
- [7] Christos Hadjinikolis, Yiannis Siantos, Sanjay Modgil, Elizabeth Black, and Peter McBurney. Opponent Modelling in Persuasion Dialogues. In *IJCAI*. 2013.
- [8] Tjitze Rienstra, Matthias Thimm, and Nir Oren. Opponent Models with Uncertainty for Strategic Argumentation. In *IJCAI*. 2013.
- [9] Anthony Hunter. Base Logics in Argumentation. In *COMMA*, pages 275–286. 2010.
- [10] Sarit Kraus, Katia Sycara, and Amir Evenchik. Reaching agreements through argumentation: a logical model and implementation. *Artificial Intelligence*, 104(1):169, 1998.
- [11] Leila Amgoud and Henri Prade. Handling threats, rewards, and explanatory arguments in a unified setting. *International journal of intelligent systems*, 20(12):1195-1218, 2005.

A Heuristic Strategy for Persuasion Dialogues

Josh MURPHY, Elizabeth BLACK and Michael LUCK

Department of Informatics, King's College London, UK

Abstract Argument-based persuasion dialogues provide an effective mechanism for agents to communicate their beliefs, and their reasons for those beliefs, in order to convince another agent of some topic argument. In such dialogues, the persuader has strategic considerations, and must decide which of its known arguments should be asserted, and the order in which they should be asserted. Recent works consider mechanisms for determining an optimal strategy for persuading the responder. However, computing such strategies is expensive, swiftly becoming impractical as the number of arguments increases. In response, we present a strategy that uses heuristic information of the domain arguments and can be computed with high numbers of arguments. Our results show that not only is the heuristic strategy fast to compute, it also performs significantly better than a random strategy.

Keywords. Argument-based dialogues, dialogue strategies, persuasion dialogues

1. Introduction

Argument-based dialogues are a useful mechanism for agent co-ordination, particularly in the domains of human-machine interaction and agreement technologies [6]. In this paper, we focus on a simple type of persuasion dialogue (where one agent presents arguments to another with the aim of convincing it to accept some argument that is the topic of the dialogue) and consider the problem of how the persuader can determine which arguments to present during the dialogue, *i.e.*, what dialogue *strategy* it should employ.

The development of methods for generating agent dialogue strategies is an active area of research [9]. So far, work on this problem has shown that computing an optimal strategy for two-party dialogues is computationally expensive, and becomes intractable as the number of arguments in the dialogue domain increases. Black *et al.* [1] consider the same simple persuasion dialogue setting that we focus on here, modelling it as a planning problem so that a planner can be used to generate an optimal strategy for the persuader, while Hadoux *et al.* [4] and Rienstra *et al.* [8] each support richer models of argument dialogue, generating optimal strategies using Mixed Observability Markov Decision Problems (MOMDPs) and a variant of the minimax algorithm respectively. While each of these approaches [1,4,8] determines an optimal strategy for the persuader, none have been shown to scale to domains with more than 10 arguments.

This work was partially supported by the UK Engineering and Physical Sciences Research Council, grant ref. EP/M01892X/1.

The key contribution of this paper is a heuristic strategy for persuasion that can easily scale to domains with 50 arguments (with computation time of less than 1 second). Although this heuristic strategy is not optimal, it gives a reasonable chance of successful persuasion and significantly outperforms a strategy that randomly selects arguments. Our heuristic strategy does not require the persuading agent to have any knowledge of the persuadee, relying only on arguments the persuader knows may exist in the domain, and uses a measure of distance from the topic argument to estimate the likelihood that any argument would (if asserted) affect the persuadee's perception of the topic acceptability.

The remainder of the paper is structured as follows. Section 2 provides the preliminary background on argumentation and argument dialogues, in particular the two-player simple persuasion dialogue we use as a testbed for our strategy. Section 3 introduces the heuristic used to estimate the likelihood of each argument to persuade the responder, and in Section 4 the strategy is formally defined. Section 5 details the experimental set-up, and Section 6 presents the results. Section 7 concludes with a discussion.

2. Argumentation and simple persuasion dialogues

Dung-style *argumentation frameworks* [3] are comprised of two key elements: arguments and attacks (the directed relationship between the arguments representing conflict).

Definition 1. An *argument framework* is a tuple $AF = \langle A, R \rangle$, s.t. A is a set of arguments, and $R \subseteq A \times A$, is a set of attacks where $\langle x, y \rangle \in R$ is an attack, x to y .

Given an argument framework, we can determine which *extensions* (sets of arguments) are rational for an agent to consider acceptable. While different extensions are based on different intuitions, a desirable property for a set of acceptable arguments is often that of *admissibility*. An argument is admissible with respect to a set of arguments S if all of its attackers are attacked by some argument in S , and no argument in S attacks an argument in S . For the rest of this paper, we consider an argument to be *acceptable* to an agent (w.r.t. an argumentation framework) if it is part of all maximal admissible sets. These criteria for acceptability are known as the *preferred sceptical semantics* (as in [3]).

Definition 2. We define a function, $\text{Acc}(AF)$, to return the set of acceptable arguments under the preferred sceptical semantics of the given argumentation framework AF .

To investigate the effectiveness of the heuristic strategy we apply it to a persuasion dialogue (adapted from [1]) that has two participating agents: a *persuader* and a *responder*. The persuader's goal is to convince the responder of the dialogue topic (an argument). The responder replies truthfully as to whether it finds the topic acceptable given its (private) beliefs and the arguments asserted by the persuader. Agents engage in a dialogue under an argument framework — the *global knowledge* (all possible arguments in the domain, and the attacks between them) — from which their own personal knowledge is a subset.

Definition 3. A *simple persuasion dialogue scenario*, under global knowledge $AF_G = \langle A_G, R_G \rangle$, is a tuple $\langle AF_P, AF_R, t \rangle$, such that:

- $AF_P = \langle A_P, R_P \rangle$, where $A_P \subseteq A_G$ and $R_P = R_G \cap (A_P \times A_P)$, is the persuader's initial knowledge base,

- $AF_R = \langle A_R, R_R \rangle$, where $A_R \subseteq A_G$ and $R_R = R_G \cap (A_R \times A_R)$, is the responder's initial knowledge base, and
- $t \in A_P$, is the dialogue topic.

During the dialogue, the persuader and responder take turns to make utterances to one another; the persuader may assert arguments or choose to terminate the dialogue, while the responder makes a *yes* or *no* move, indicating whether it finds the topic acceptable. A *well-formed simple persuasion dialogue* is one in which the persuader only asserts arguments from its knowledge base and the responder replies truthfully, and that terminates once either the responder is convinced or the persuader chooses to give up.

Definition 4. A *well-formed simple persuasion dialogue* of a simple persuasion dialogue scenario $\langle AF_P, AF_R, t \rangle$ under global knowledge $\langle A_G, R_G \rangle$, is a sequence of moves $[M_0^P, M_0^R, \dots, M_n^P, M_n^R]$, such that:

- $\forall i$ such that $0 < i < n$, $M_i^P \in A_P$,
- $M_n^P \in A_P \cup \{\text{terminate}\}$,
- $\forall i$ such that $0 < i < n$, $M_i^R = \text{no}$ and $t \notin \text{Acc}(\langle A_R \cup \{M_0^P, \dots, M_i^P\}, R_G \rangle)$,
- $M_n^R \in \{\text{yes}, \text{no}\}$, and
- $M_n^R = \text{yes}$ iff $t \in \text{Acc}(\langle A_H \cup \{M_0^P, \dots, M_n^P\}, R_G \rangle)$.

A dialogue is *terminated* iff either $M_n^P = \text{terminate}$ or $M_n^R = \text{yes}$. A terminated dialogue is said to be *successful* iff $M_n^R = \text{yes}$, and *unsuccessful* otherwise.

Over the course of a well-formed simple persuasion dialogue, the responder has no strategic concerns, as it must reply honestly if it finds the topic acceptable. However, each turn of the persuader requires a decision as to whether an argument should be asserted, and if so, which arguments in its knowledge base should be asserted. Previous work [1] has applied automated planning techniques to find an optimal strategy for the persuader to apply, but does not scale well beyond 8 domain arguments. In Section 4 we present a heuristic strategy, and show that this can easily scale to domains with up to 50 arguments. First, however, we give the intuition on which this heuristic strategy relies.

3. Evaluating the influence of arguments

We consider the *local* topological properties of argument graphs to estimate how beneficial an argument would be if asserted. The estimate is based on the intuition that arguments topologically closer to the topic are more likely to affect its acceptability. We estimate the likelihood that an argument affects the acceptability of the topic and whether the argument defends or attacks (perhaps indirectly) the topic. Note that argument acceptability not only depends on the attackers of the argument, but on the acceptability of the attackers. Thus, we are interested in *argument paths* terminating in the topic argument.

Definition 5. An *argument path*, in an argument graph $AF = \langle A, R \rangle$ with topic t , is a list of arguments $p = [a_0, a_1, \dots, a_k]$, such that:

- $a_0 = t$,
- $\forall i$ such that $1 \leq i < k$, $\langle a_{i+1}, a_i \rangle \in R$,
- $\forall i, j$ such that $0 \leq i, j \leq k$, $a_i = a_j$ iff $i = j$ (arguments are distinct).

The *depth* of an argument a in an argument path $p = [a_0, a_1, \dots, a_i]$ is given by the function: $\text{depth}(a, p) = x$ where $a = a_x$.

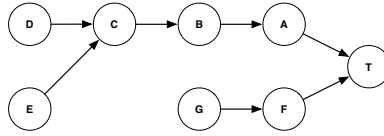


Figure 1. An example argument graph with 8 arguments.

Example 1. Consider the example argumentation framework in Figure 1 and the topic being T . Valid argument paths include $[T, F, G]$, $[T, A, B]$, and $[T, A, B, C]$; invalid argument paths include $[A, B, C]$ (the first argument is not the topic), and $[T, A, F]$ (there is no such path in the argumentation framework).

The distance of an argument from the topic argument provides an estimate of how likely it is that asserting the argument will affect the acceptability of the topic. The intuition behind this is as follows: for an argument to affect the topic through a particular argument path, all preceding arguments on that path must be present; furthermore, any arguments that precede the argument in question and support the topic cannot be defeated by an acceptable argument from another path. The more arguments that proceed the argument on a particular path, the more chance that one of these conditions may not hold, thus the more likely it is that the argument will not affect the topic through that path.

Example 2. Consider the example argumentation framework in Figure 1. The persuader wishes to convince the responder (whose arguments are unknown) that the topic T is acceptable. Consider that the persuader chooses to assert the argument G ; in order for this to have a chance of changing the responder's perception of the acceptability of the topic, the responder must know F . Consider instead that the persuader chooses to assert the argument D (which is twice as far away from the topic as G); for this to have a chance of changing the responder's perception of the acceptability of T , not only must the responder know A , B and C , but it must also be that the responder cannot know E .

To obtain an estimate of how likely each argument is to affect the acceptability of the topic, we must consider all argument paths in the argument graph that start with the topic. The number of possible argument paths grows exponentially as the size of argumentation framework increases, so we consider only argument paths up to a specified depth.

Definition 6. The **complete set of argument paths** with depth d of an argumentation framework AF and topic argument t , is a set of argument paths $C_{AF,t}^d$ where $C_{AF,t}^d = \{[t, a_1, \dots, a_x] \mid [t, a_1, \dots, a_x] \text{ is an argument path in } AF, x \leq d, \text{ and } \nexists [t, a_1, \dots, a_x, \dots, a_y] \text{ such that } [t, a_1, \dots, a_x, \dots, a_y] \text{ is an argument path in } AF \text{ and } x < y \leq d\}$

An argument at an even depth in a path will be a *supporting argument* of the topic, and its presence in an agent's knowledge *increases* the likelihood that it finds the topic acceptable (the argument is either the topic argument itself, or an argument that attacks an argument that attacks an opposing argument). Similarly, an argument at an odd depth will be an *opposing argument*, and its presence *decreases* the likelihood that it finds the topic to be acceptable (the argument is an attacker of a supporting argument).

With respect to a particular argument path, the magnitude of an argument's *value* is an estimation of the likelihood that the argument will affect the acceptability of the topic, and the sign indicates whether it is likely to make the topic acceptable or unacceptable.

Definition 7. The *value* of an argument a with depth $d = \text{depth}(a, p)$ w.r.t. an argument path $p = [a_0, a_1, \dots, a_i]$ is given by the function:

$$\text{value}(a, p) = \begin{cases} 0 & \text{if } a \notin \{a_0, \dots, a_i\} \\ 1/2^d & \text{if } a \in \{a_0, \dots, a_i\} \text{ and } d \bmod 2 = 0 \\ -1/2^d & \text{if } a \in \{a_0, \dots, a_i\} \text{ and } d \bmod 2 = 1 \end{cases}$$

To determine the *estimated utility* of an argument, we sum the values of that argument with respect to each argument path to the topic.

Definition 8. The *estimated utility* of an argument A in an argumentation framework AF with topic t to a depth d , is a real number given by the function eu such that:

$$\text{eu}(A, C_{AF,t}^d) = \sum_{p \in C_{AF,t}^d} \text{value}(a, p).$$

4. Heuristic strategy

A persuader using the heuristic strategy will not give up trying to convince the responder until it has run out of arguments to assert (known as an *exhaustive persuader* [2]). It uses estimated utility to determine which argument to assert, choosing one not yet asserted.

Definition 9. Consider a persuader with a knowledge base $AF_P = \langle A_P, R_P \rangle$ participating in a dialogue $D = [M_0^P, M_0^R, \dots, M_n^P, M_n^R]$, under a global knowledge $AF_G = \langle A_G, R_G \rangle$. The **heuristic strategy** for a depth d is given by the function hStrategy_d such that:

- if $A_P - \{M_0^P, \dots, M_n^P\} = \emptyset$ then $\text{hStrategy}_d(D) = \text{terminate}$, otherwise
- $\text{hStrategy}_d(D) = M$ where $M \in \{A \in A_P - \{M_0^P, \dots, M_n^P\} \mid \forall B \in A_P - \{M_0^P, \dots, M_n^P\}, \text{eu}(A, C_{AF_G,t}^d) \geq \text{eu}(B, C_{AF_G,t}^d)\}$

Note that a persuader using the heuristic strategy can only assert arguments from its knowledge base, but uses global knowledge to determine which argument to assert. Similar to the virtual argument approach taken by Rienstra *et al.* [8], we assume that the persuader can only assert arguments it is aware of, but is aware of the potential existence of all arguments in the domain, even those that it cannot itself assert. Other works that determine strategies for argument dialogues make similar assumptions and further assume that the persuader has a model of its opponent's knowledge [1] or behaviour [4].

5. Implementation

To evaluate our heuristic strategy we generate random simple persuasion dialogue scenarios, in which the persuader selects which arguments to assert. As a benchmark for evaluation, we use a random strategy, by which a persuader will assert its unasserted arguments at random until the responder is persuaded or there are no unasserted arguments.

To generate a random simple persuasion dialogue scenario, an argument graph representing the global knowledge must be selected. In our experiments, we randomly generate two types of argument graph: tree-like and grid-like (full details of their generation are available at github.com/joshlmurphy). This allows us to generate a large number of dialogue scenarios on which to run experiments. Except where noted, we use

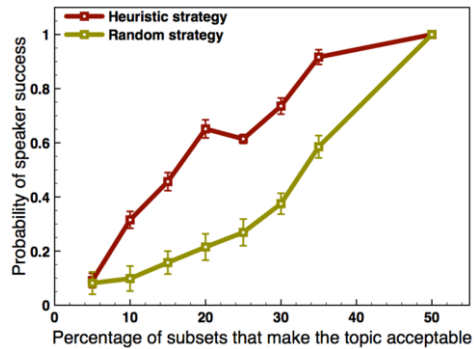


Figure 2. Percentage success rate of strategies. Error bars indicate standard error.

sparse, fully-connected, tree-like graphs which rarely contain cycles. These properties are based loosely on argument frameworks transcribed from BBC Radio 4's Moral Maze program, in which experts aim to persuade a panel of an opinion [7].

Once the global knowledge has been generated, arguments are evenly distributed into the persuader's responder's knowledge bases at random. The topic argument of the dialogue is then selected randomly from the persuader's knowledge base so that the topic is initially known by the persuader, but not by the responder. For our experiments the heuristic strategy considers argument paths up to depth 5; initial testing showed this allowed for a strong success rate while remaining fast to compute. We leave an analysis of how depth affects success strategy and computation time for future work.

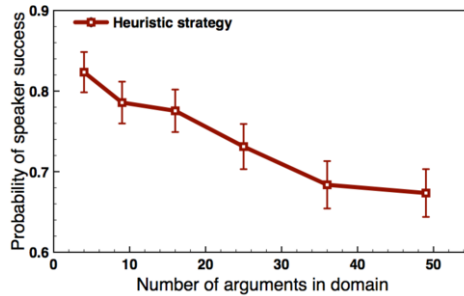
The implementation for the generation and testing of simple persuasion dialogues was done in Java, and run on a standard PC (1.86 GHz dual-core processor, 2GB RAM). We used libraries from Tweety [10] to determine whether the argument topic was acceptable under the preferred sceptical semantics for a given argument graph.

6. Results

The heuristic strategy has a high success rate It is desirable for a dialogue strategy to have a high success rate in achieving an agent's dialogue goals no matter what the agents know. For simple persuasion dialogues, this means that the persuader's strategy should have a high probability of persuading the responder of the topic argument. Both the heuristic and random strategies were run on domains with 8 arguments, with different rates of argument subsets making the topic acceptable. The probability of persuader success for the strategies was determined by running many simulations of dialogues, each with a different randomly generated argumentation framework, and recording the percentage of argument subsets that make the topic acceptable in the argumentation framework, as well as whether the persuader is successful when using the heuristic or random strategy. The results are shown in Figure 2. We observe a similar trend for both strategies: as the proportion of argument subsets of the global knowledge that make the topic acceptable increases, so does the likelihood that the strategy is successful. The results show that the heuristic strategy is more likely to be successful than the random strategy.

Table 1. Time to compute heuristic strategy (seconds). *Args* is the number of arguments in the domain.

Args	10	20	30	40	50
Time	<0.1	0.21	0.37	0.56	0.77

**Figure 3.** The heuristic strategy remains successful with increasing numbers of arguments.

The heuristic strategy is fast to compute To determine the computational cost of generating the heuristic dialogue strategy, we measure the time taken to compute the estimated utility of each argument that is assigned to the persuader in a randomly generated dialogue scenario. The results are shown in Table 1, giving the average time for 1,000 random dialogue scenarios. For domains with fewer than 10 arguments the generation of the strategy took less than 0.1 seconds. At 11 arguments, the increase in time is noticeable, and appears to be somewhat linear, allowing computation of the heuristic strategy in less than a second for as many as 50 arguments in the domain. The results show that the heuristic strategy is efficiently scalable for domains with large numbers of arguments.

The heuristic strategy succeeds with many arguments As can be seen from the results in Figure 2, the chance of successfully convincing the responder depends heavily on the particular argument graph that determines the global knowledge. The more subsets of arguments from the global knowledge that determine the topic to be acceptable, the more chance of reaching a point in the dialogue where such a set of arguments is available to the responder, causing it to terminate the dialogue successfully. To investigate how the performance of the heuristic strategy scales with the number of arguments we needed to generate global knowledge argument graphs in such a way that the proportion of argument subsets that determine the topic to be acceptable remains near constant as the size of the graphs increases. Thus, here we used partial grids, which allowed us to keep the average percentage of subsets of the global knowledge that make the topic acceptable within the range 28%–33% for all argument graphs we experimented with. We observe in Figure 3 that there is a slight decrease in the success rate of the heuristic strategy as the number of arguments increases because, as the argument graph grows, so does its complexity, and these complexities are ignored by the heuristic strategy. The decrease in success can be considered a necessary sacrifice for a computationally tractable strategy.

7. Discussion

In this paper we have presented and evaluated a heuristic strategy that can be used in persuasion dialogues. Our results show that this heuristic strategy is fast to compute,

even for domains with a large number of arguments, which is not the case for existing approaches that generate optimal strategies [1,4,8].

In future work, we intend to investigate the performance of the heuristic strategy in more complex scenarios, specifically persuasion dialogues involving more than two participants, each of which may assert arguments with the aim of convincing the others. We expect that existing approaches for determining optimal strategies [1,4,8] would be intractable here, since the probabilistic information about the opponent used determines the state space that must be searched to find an optimal solution and so as the number of opponents increases, the number of possible states to consider increases exponentially.

Argument strategies that use heuristic information have also been investigated in different types of dialogue. Kontranis *et al.* evaluate a set of heuristic-style strategies that agents use in a dialogue-type scenario, in which participants vote on the attacks between globally known arguments, with the goal to reach a consensus [5]. In comparison, the heuristic strategy we present is based on a typical dialogue game in which agents assert arguments, rather than the focus of communication being on attack relations. Wardeh *et al.* investigate PADUA, a dialogue protocol allowing agents to classify objects based on evidence from previous examples of object classification [11]. Depending on whether the opponent is agreeable or not, the persuader can select the appropriate heuristic strategy in order to increase their success rate in deciding upon their desired classification. However, Wardeh *et al.* do not investigate the scalability of their proposed strategies.

References

- [1] E. Black, A. Coles, and S. Bernardini. Automated planning of simple persuasion dialogues. In N. Bulling, L. van der Torre, S. Villata, W. Jamroga, and W. Vasconcelos, editors, *Computational Logic in Multi-Agent Systems*, volume 8624 of *LNCS*, pages 87–104. Springer, 2014.
- [2] E. Black and A. Hunter. Reasons and options for updating an opponent model in persuasion dialogues. In E. Black, S. Modgil, and N. Oren, editors, *Theory and Applications of Formal Argumentation*, volume 9524 of *LNCS*, pages 21–39. Springer, 2015.
- [3] P. Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. In *Artificial Intelligence*, volume 77, pages 321–357. Elsevier, 1995.
- [4] E. Hadoux, A. Beynier, N. Maudet, P. Weng, and A. Hunter. Optimization of probabilistic argumentation with Markov decision models. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence*, pages 2004–2010. AAAI Press, 2015.
- [5] D. Kontarinis, E. Bonzon, N. Maudet, and P. Moraitis. Empirical evaluation of strategies for multiparty argumentative debates. In N. Bulling, L. van der Torre, S. Villata, W. Jamroga, and W. Vasconcelos, editors, *Computational Logic in Multi-Agent Systems*, volume 8624 of *LNCS*, pages 105–122. Springer, 2014.
- [6] S. Modgil, F. Toni, F. Bex, I. Bratko, C. Chesñevar, W. Dvořák, M. Falappa, et al. The added value of argumentation. In S. Ossowski, editor, *Agreement Technologies*, pages 357–403. Springer, 2013.
- [7] C. Reed. Argument corpora. Technical report, University of Dundee Technical Report, Available online at www.arg.dundee.ac.uk/corpora, 2013.
- [8] T. Rienstra, M. Thimm, and N. Oren. Opponent models with uncertainty for strategic argumentation. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence*, pages 332–338, 2013.
- [9] M. Thimm. Strategic argumentation in multi-agent systems. *Künstliche Intelligenz*, 28:159–168, 2014.
- [10] M. Thimm. Tweak - A comprehensive collection of Java libraries for logical aspects of artificial intelligence and knowledge representation. In *Proceedings of the Fourteenth International Conference on Principles of Knowledge Representation and Reasoning*, 2014.
- [11] M. Wardeh, T. Bench-Capon, and F. Coenen. PADUA protocol: Strategies and tactics. In K. Mellouli, editor, *Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, volume 4724 of *LNCS*, pages 465–476. Springer, 2007.

Rethinking the Rationality Postulates for Argumentation-Based Inference

Henry PRAKKEN

Department of Information and Computing Sciences, Utrecht University and Faculty of Law, University of Groningen, The Netherlands

Abstract. Much research on structured argumentation aims to satisfy the rationality postulates of direct and indirect consistency and strict (deductive) closure. However, examples like the lottery paradox indicate that it is sometimes rational to accept sets of propositions that are indirectly inconsistent or not deductively closed. This paper proposes a variant of the *ASPIC*⁺ framework that violates indirect consistency and full strict closure but satisfies direct consistency and restricted forms of strict closure and indirect consistency.

Keywords. Rational acceptance, Rationality postulates, Lottery paradox

1. Introduction

Much current work on structured argumentation (e.g. [6,9,2]) concerns the so-called rationality postulates of [1]. The idea is that argument extensions [4] should be closed under subarguments and that the sets of conclusions of all arguments in an extension should be directly consistent (no formulas that negate each other should be in the set), closed under strict (deductive) inference and indirectly consistent (the strict closure should be directly consistent). Most work on these postulates simply assumes that they should be satisfied, but examples like the lottery paradox [7] suggest that it may sometimes be rational to jointly accept indirectly inconsistent propositions or not to accept deductive consequences of acceptable propositions.

Imagine a fair lottery with one million tickets and just one prize. If the principle is accepted that it is rational to accept a proposition if its truth is highly probable, then for each ticket T_i it is rational to accept that T_i will not win while at the same time it is rational to accept that exactly one ticket will win. If we also accept that everything that deductively follows from a set of rationally acceptable propositions, then we have two rationally acceptable propositions that contradict each other: we can join all individual propositions $\neg T_i$ into a big conjunction $\neg T_1 \wedge \dots \wedge \neg T_{1,000,000}$ with one million conjuncts, which contradicts the certain fact that exactly one ticket will win.

The problem does not only arise in precisely defined probabilistic settings (cf. [11]). First, non-statistical examples of the lottery paradox can easily be imagined. For example, for each arbitrary part of a complex machine we can rationally accept that it will not malfunction but at the same time we know that some part will at some point in time malfunction. Moreover, the problem arises in any model of 'fallible' rational acceptance. Rational acceptance is usually fallible, either because one starts from uncertain premises or

because one applies defeasible inferences. Now whenever a deductive inference is made from at least two ‘fallible’ pieces of information, the deductive inference can be said to aggregate the degrees of fallibility of the individual elements to which it is applied. This in turn means that the deductive inference may be weaker than either of these elements, so that a successful attack on the deductive inference does not necessarily imply a successful attack on one of the fallible elements to which it was applied.

In discussions of the lottery paradox several positions have been defended. For example, Pollock [10] argued that sets of rationally acceptable propositions should always be deductively closed. Moreover, in the lottery paradox he argued that for no ticket is it rational to accept that it will not win. However, this position is not quite self-evident: if propositions cannot be accepted even if their truth is highly probable, then many propositions that seem clearly acceptable would not formally come out as such. Others (including Kyburg [7]) reject the conjunction principle for rational acceptance, motivated by the fact that according to probability theory a conjunction of two highly probable propositions need not be highly probable. However, this also has its issues, since people often conjoin their beliefs, and regarding this as always irrational seems too strong. Therefore intermediate positions have also been considered. For example, Makinson [8] argues that (in the lottery example) conjunctions $\neg T_i \wedge \dots \neg T_j$ are rationally acceptable for up to a particular (not too large) number of conjuncts. And [3] argue that examples like the lottery paradox are exceptional cases where strict closure fails since their underlying probability structure is uniform: no particular event is typical and randomness prevails. In this paper we want to explore whether such an intermediate position can be formalised in an argumentation setting. In doing so, we will make two assumptions.

First, problems like these do not arise when rational acceptance is seen as a matter of degree. In epistemology there is a debate whether rational acceptance is always a matter of degree or whether it makes sense to speak of full (though still possibly defeasible) acceptance [5]. Taking a stance in this debate goes beyond the scope of this paper but since the notion of full acceptance is in epistemology often defended, it makes sense to explore its consequences in an argumentation setting. This holds the more since most formal and computational models of argument model non-gradual notions of full acceptance.

Second, Pollock [10] also argued that what can be rationally accepted in the lottery paradox is that it is *highly probable* that it will win. At first sight, this approach would seem attractive, until one realises that if it is applied to the lottery example, it should be applied to many other examples of defeasible reasoning, since many of those arguably have an underlying probabilistic justification. So why require in the lottery example that the probability of a statement is expressed in the object language while not requiring this for, for instance, ‘If P then usually Q ’ and ‘ P ’ defeasibly imply ‘ Q ’? Accordingly, in this paper we will make a second assumption that is often adopted in formal and computational models of argument, namely, that the probability of statements is not expressed in the logical object language of a system but in its metalanguage, in the nonmonotonicity of its consequence notion. Just as the assumption that full acceptance is possible, this assumption is debatable, but both assumptions are widely adopted, which justifies this paper’s aim to explore their logical consequences.

Summarising, the purpose of this paper is to formally investigate the relevance of examples like the lottery paradox for models of argumentation that model non-gradual notions of full acceptance and that express the probability of statements in the metalanguage in the nonmonotonicity of their consequence notion. In particular, we will explore

how the intermediate position can be formalised that conclusions of deductive inferences from fallibly acceptable propositions can but need not be rationally acceptable. We will argue that under the adopted assumptions the rationality postulate of direct consistency should be retained but that the postulates of indirect consistency and strict closure have to be weakened in general (although they may apply in special cases). We will carry out the investigations in terms of the $ASPIC^+$ framework, motivated by its generality: as shown earlier [12,9] it can be instantiated in many different ways and some of these ways capture other models of structured argumentation as special cases.

2. The $ASPIC^+$ framework

$ASPIC^+$ generates abstract argumentation frameworks in the sense of [4]. Formally, an **abstract argumentation framework** (AF) is a pair $(\mathcal{A}, \mathcal{D})$, where \mathcal{A} is a set of *arguments* and $\mathcal{D} \subseteq \mathcal{A} \times \mathcal{A}$ is a binary relation of *defeat*. We say that A *strictly defeats* B if A defeats B while B does not defeat A . A semantics for AFs returns sets of arguments called *extensions*, which are subsets of \mathcal{A} with particular properties:

Definition 1 Let $(\mathcal{A}, \mathcal{D})$ be an AF. For any $X \in \mathcal{A}$, X is *acceptable* w.r.t. some $S \subseteq \mathcal{A}$ iff $\forall Y$ s.t. $(Y, X) \in \mathcal{D}$ implies $\exists Z \in S$ s.t. $(Z, Y) \in \mathcal{D}$. Let $S \subseteq \mathcal{A}$ be *conflict free*, i.e., there are no A, B in S such that $(A, B) \in \mathcal{D}$. Then S is: an *admissible* set iff $X \in S$ implies X is acceptable w.r.t. S ; a *complete* extension iff $X \in S$ whenever X is acceptable w.r.t. S ; a *preferred* extension iff it is a set inclusion maximal admissible set; the *grounded* extension iff it is the set inclusion minimal complete extension; a *stable* extension iff it is conflict-free and $\forall Y \notin S, \exists X \in S$ s.t. $(X, Y) \in \mathcal{D}$.

For $T \in \{\text{complete, preferred, grounded, stable}\}$, X is *sceptically* or *credulously* justified under the T semantics if X belongs to all, respectively at least one, T extension.

We next summarise $ASPIC^+$ as defined in [9]. It defines the notion of an abstract *argumentation system* as a structure consisting of a logical language \mathcal{L} with negation, two sets \mathcal{R}_s and \mathcal{R}_d of strict and defeasible inference rules, and a naming convention n in \mathcal{L} for defeasible rules in order to talk about the applicability of defeasible rules in \mathcal{L} .

Definition 2 [Argumentation systems] An *argumentation system* is a triple $AS = (\mathcal{L}, \mathcal{R}, n)$ where:

- \mathcal{L} is a logical language with a unary negation connective \neg .
- $\mathcal{R} = \mathcal{R}_s \cup \mathcal{R}_d$ is a set of strict (\mathcal{R}_s) and defeasible (\mathcal{R}_d) inference rules of the form $\varphi_1, \dots, \varphi_n \rightarrow \varphi$ and $\varphi_1, \dots, \varphi_n \Rightarrow \varphi$ respectively (where φ_i, φ are meta-variables ranging over wff in \mathcal{L}), such that $\mathcal{R}_s \cap \mathcal{R}_d = \emptyset$.
- n is a partial function from \mathcal{R}_d to \mathcal{L} .

We write $\psi = -\varphi$ just in case $\psi = \neg\varphi$ or $\varphi = \neg\psi$. Note that $-$ is not a connective in \mathcal{L} but a function symbol in the metalanguage of \mathcal{L} .

$ASPIC^+$ leaves the choice of inference rules free. If desired, the strict rules can be based on a given deductive logic L by letting $\varphi_1, \dots, \varphi_n \rightarrow \varphi \in \mathcal{R}_s$ iff $\varphi_1, \dots, \varphi_n \vdash_L \varphi$. However, for simplicity this paper's examples will not encode full logics in \mathcal{R}_s .

Example 1 An example argumentation system is with $\mathcal{L} = \{p, \neg p, q, \neg q, r, \neg r, s, \neg s, t, \neg t, r_1, r_2, \neg r_1, \neg r_2\}$, $\mathcal{R}_s = \{p, r \rightarrow s; \neg s \rightarrow \neg r_1\}$, $\mathcal{R}_d = \{q \Rightarrow r; t \Rightarrow \neg s\}$ where $n(q \Rightarrow r) = r_1$ and $n(t \Rightarrow \neg s) = r_2$.

Definition 3 [Consistency] For any $S \subseteq \mathcal{L}$, let the *closure of S under strict rules*, denoted $Cl_{\mathcal{R}_s}(S)$, be the smallest set containing S and the consequent of any strict rule in \mathcal{R}_s whose antecedents are in $Cl_{\mathcal{R}_s}(S)$. Then a set $S \subseteq \mathcal{L}$ is *directly consistent* iff $\nexists \psi, \varphi \in S$ such that $\psi = \neg\varphi$, and *indirectly consistent* iff $Cl_{\mathcal{R}_s}(S)$ is directly consistent.

Example 2 In our example argumentation system, an example of a directly inconsistent set is $\{p, \neg p\}$ and an example of an indirectly inconsistent set is $\{p, r, \neg s\}$.

Definition 4 [Knowledge bases] A *knowledge base* in an $AS = (\mathcal{L}, \mathcal{R}, n)$ is a set $\mathcal{K} \subseteq \mathcal{L}$ consisting of two disjoint subsets \mathcal{K}_n (the *axioms*) and \mathcal{K}_p (the *ordinary premises*).

Arguments can be constructed from knowledge bases by applying inference rules. In what follows, for a given argument the function Prem returns all its premises, Conc returns its conclusion, Sub returns all its sub-arguments and DefRules and TopRule return, respectively, all defeasible rules and the last rule applied in the argument.

Definition 5 [Arguments] An *argument* A on the basis of a knowledge base \mathcal{K} in an argumentation system $(\mathcal{L}, \mathcal{R}, n)$ is:

1. φ if $\varphi \in \mathcal{K}$ with: $\text{Prem}(A) = \{\varphi\}$; $\text{Conc}(A) = \varphi$; $\text{Sub}(A) = \{\varphi\}$; $\text{DefRules}(A) = \emptyset$; $\text{TopRule}(A) = \text{undefined}$.
2. $A_1, \dots, A_n \rightarrow \psi$ if A_1, \dots, A_n are arguments such that $\text{Conc}(A_1), \dots, \text{Conc}(A_n) \rightarrow \psi \in \mathcal{R}_s$.
 $\text{Prem}(A) = \text{Prem}(A_1) \cup \dots \cup \text{Prem}(A_n)$; $\text{Conc}(A) = \psi$; $\text{Sub}(A) = \text{Sub}(A_1) \cup \dots \cup \text{Sub}(A_n) \cup \{A\}$; $\text{DefRules}(A) = \text{DefRules}(A_1) \cup \dots \cup \text{DefRules}(A_n)$; $\text{TopRule}(A) = \text{Conc}(A_1), \dots, \text{Conc}(A_n) \rightarrow \psi$.
3. $A_1, \dots, A_n \Rightarrow \psi$ if A_1, \dots, A_n are arguments such that $\text{Conc}(A_1), \dots, \text{Conc}(A_n) \Rightarrow \psi \in \mathcal{R}_d$.
 $\text{Prem}(A)$, $\text{Conc}(A)$ and $\text{Sub}(A)$ are defined as in (2) while $\text{DefRules}(A) = \text{DefRules}(A_1) \cup \dots \cup \text{DefRules}(A_n) \cup \{\text{Conc}(A_1), \dots, \text{Conc}(A_n) \Rightarrow \psi\}$ and $\text{TopRule}(A) = \text{Conc}(A_1), \dots, \text{Conc}(A_n) \Rightarrow \psi$.

For any argument A , $\text{Prem}_n(A) = \text{Prem}(A) \cap \mathcal{K}_n$ and $\text{Prem}_p(A) = \text{Prem}(A) \cap \mathcal{K}_p$. An argument A is *infallible* if $\text{DefRules}(A) = \emptyset$ and $\text{Prem}(A) \subseteq \mathcal{K}_n$; otherwise it is *fallible*. For any set S of arguments, $\text{Conc}(S) = \{\varphi \mid \varphi = \text{Conc}(A) \text{ for some } A \in S\}$. We write $S \vdash \varphi$ if there exists a strict argument for φ with all premises taken from S .

Example 3 If our example argumentation system is combined with a knowledge base with $\mathcal{K}_n = \{p\}$ and $\mathcal{K}_p = \{q, t\}$, then the following arguments can be constructed, of which only A_1 is infallible:

$$\begin{array}{lll}
 A_1 = & p & A_4 = A_2 \Rightarrow r & A_7 = A_5 \rightarrow \neg r_1 \\
 A_2 = & q & A_5 = A_3 \Rightarrow \neg s & \\
 A_3 = & t & A_6 = A_1, A_4 \rightarrow s &
 \end{array}$$

Arguments can be attacked in three ways: on an application of a defeasible rule, on the conclusion of such an application or on an ordinary premise.

Definition 6 [Attack] An argument A attacks an argument B iff A undercuts or rebuts or undermines B , where:

- A undercuts B (on B') iff $\text{Conc}(A) = -n(r)$ and $B' \in \text{Sub}(B)$ such that B' 's top rule r is defeasible.
- A rebuts B (on B') iff $\text{Conc}(A) = -\varphi$ for some $B' \in \text{Sub}(B)$ of the form $B'_1, \dots, B'_n \Rightarrow \varphi$.
- A undermines B (on φ) iff $\text{Conc}(A) = -\varphi$ for some $\varphi \in \text{Prem}(B) \cap \mathcal{K}_p$.

Example 4 In our running example A_6 rebuts A_5 and A_7 on A_5 . Note that A_5 does not rebut A_6 since A_6 has a strict top rule. Furthermore, A_7 undercuts A_4 and A_6 on A_4 .

Argumentation systems plus knowledge bases induce structured argumentation frameworks.

Definition 7 [Structured Argumentation Frameworks] Let AT be an argumentation theory (AS, \mathcal{K}) . A structured argumentation framework (SAF) defined by AT , is a triple $\langle \mathcal{A}, \mathcal{C}, \preceq \rangle$ where \mathcal{A} is the set of all finite arguments constructed from \mathcal{K} in AS , \preceq is an ordering on \mathcal{A} , and $(X, Y) \in \mathcal{C}$ iff X attacks Y . A *c-structured argumentation framework* (c-SAF) is defined likewise except that \mathcal{A} is the set of all finite arguments constructed from \mathcal{K} with indirectly consistent set of premises.

The notion of *defeat* can then be defined as follows ($A \prec B$ is defined as usual as $A \preceq B$ and $B \not\preceq A$ and $A \approx B$ as $A \preceq B$ and $B \preceq A$).

Definition 8 [Defeat] A defeats B iff either A undercuts B ; or A rebuts or undermines B on B' and $A \not\prec B'$.

Example 5 In our running example A_6 defeats A_5 unless $A_6 \prec A_5$. Furthermore, regardless of the argument ordering, A_7 defeats A_4 (and thus A_6).

Abstract argumentation frameworks are then generated from (c-)SAFs as follows:

Definition 9 [Argumentation frameworks] An abstract argumentation framework (AF) corresponding to a (c-)SAF $= \langle \mathcal{A}, \mathcal{C}, \preceq \rangle$ is a pair (\mathcal{A}, D) such that D is the defeat relation on \mathcal{A} determined by (c-)SAF.

A nonmonotonic consequence notion can then be defined as follows. Let $T \in \{\text{complete, preferred, grounded, stable}\}$ and let \mathcal{L} be from the AT defining $(c) - SAF$. A wff $\varphi \in \mathcal{L}$ is *sceptically T-justified* in $(c-)SAF$ if φ is the conclusion of a sceptically T -justified argument, and *credulously T-justified* in $(c-)SAF$ if φ is not sceptically T -justified and is the conclusion of a credulously T -justified argument.

[9] prove that for so-called ‘well-defined’ argumentation theories with so-called ‘reasonable’ argument orderings the extensions induced by Definition 9 satisfy all four rationality postulates of the rationality postulates of [1]. These and some related notions are defined as follows.

Definition 10 [Well defined (c-)SAFs] Let $AT = (AS, \mathcal{K})$ be an argumentation theory, where $AS = (\mathcal{L}, \mathcal{R}, n)$. We say that AT is:

- *closed under contraposition* iff for all $S \subseteq \mathcal{L}$, all $\varphi \in \mathcal{L}$ and all $\psi \in S$: if $S \vdash \varphi$, then $S \setminus \{\psi\} \cup \{\varphi'\} \vdash \psi'$ for all φ' such that $\varphi' = \neg\varphi$ and all ψ' such that $\psi' = \neg\psi$.
- *closed under transposition* iff if $\varphi_1, \dots, \varphi_n \rightarrow \psi \in \mathcal{R}_s$, then for $i = 1 \dots n$, $\varphi_1, \varphi_{i-1}, \psi', \varphi_{i+1}, \dots, \varphi_n \rightarrow \varphi'_i \in \mathcal{R}_s$ for all φ'_i such that $\varphi'_i = \neg\varphi_i$ and all ψ' such that $\psi' = \neg\psi$.
- *axiom consistent* iff \mathcal{K}_n is indirectly consistent.

If a (c-)SAF is defined by an AT that is axiom consistent and closed under contraposition or transposition, then the SAF is said to be *well defined*.

Henceforth, any (c-)SAF is assumed to be well defined.

Example 6 The argumentation theory in our running example is axiom consistent since $\{p\}$ is indirectly consistent. It can be made closed under contraposition or transposition by adding $p, \neg s \rightarrow \neg r$ and $r, \neg s \rightarrow \neg p$ and $r_1 \rightarrow s$ to \mathcal{R}_s .

We now define strict continuations of arguments slightly differently than in [9].¹

Definition 11 [Strict continuations] The set of *strict continuations* of any set of arguments from \mathcal{A} is the smallest set satisfying the following conditions:

1. Any argument A is a strict continuation of $\{A\}$.
2. If A_1, \dots, A_n and S_1, \dots, S_n are sets of arguments such that all A_i are a strict continuation of S_i and all of B_1, \dots, B_n are infallible arguments, then $A_1, \dots, A_n, B_1, \dots, B_n \rightarrow \varphi$ is a strict continuation of $S_1 \cup \dots \cup S_n$.

Example 7 In our running example all arguments are strict continuations of themselves while A_6 is a strict continuation of $\{A_4\}$ and A_7 is a strict continuation of A_5 .

Definition 12 [Reasonable Argument Orderings] An argument ordering \preceq is *reasonable* iff:

1. i) $\forall A, B$, if A is infallible and B is fallible, then $B \prec A$;
 ii) $\forall A, B$, if B is infallible then $B \not\prec A$;
 iii) $\forall A, A', B$ such that A' is a strict continuation of $\{A\}$, if $A \not\prec B$ then $A' \not\prec B$, and if $B \not\prec A$ then $B \not\prec A'$ (i.e., applying strict rules to a set of arguments of which at most one is fallible does not weaken, resp. strengthen, arguments).
2. Let $\{C_1, \dots, C_n\}$ be a finite subset of \mathcal{A} , and for $i = 1 \dots n$, let $C^{+\setminus i}$ be some strict continuation of $\{C_1, \dots, C_{i-1}, C_{i+1}, \dots, C_n\}$. Then it is not the case that: $\forall i, C^{+\setminus i} \prec C_i$.

Example 8 In our running example, Conditions 1(i,ii) make that $A_1 \not\prec A_1$ and $A_i \prec A_1$ for all i such that $1 < i \leq 7$. Suppose we further have $A_5 \not\prec A_6$. Then by 1(iii) we also have $A_7 \not\prec A_6$. Suppose we also have $A_2 \not\prec A_7$; then by 1(ii) we also have $A_2 \not\prec A_5$. To illustrate Condition (2), let us temporarily move p from \mathcal{K}_n to \mathcal{K}_p and suppose \mathcal{R}_s is closed under transposition. Then the following new arguments can be constructed:

$$A_8 = A_1, A_5 \rightarrow \neg r \quad A_9 = A_4, A_5 \rightarrow \neg p$$

¹The new definition is arguably simpler but does not affect the proofs of [9].

Note that A_6 strictly continues $\{A_1, A_4\}$, A_8 strictly continues $\{A_1, A_5\}$ and A_9 strictly continues $\{A_4, A_5\}$. Then we cannot have all of $A_6 \prec A_5$ and $A_8 \prec A_4$ and $A_9 \prec A_1$.

Finally, in some proofs below the notion of a maximum fallible subargument is used. The following definition improves the one of [9], which does not satisfy Lemma 11 below.

Definition 13 [Maximal fallible subarguments] For any argument A , the set $M(A)$ of *maximal fallible subarguments* of A is inductively defined as:

1. If $A \in \mathcal{K}_n$, then $M(A) = \emptyset$;
2. If $A \in \mathcal{K}_p$ or A has a defeasible top rule, then $M(A) = \{A\}$;
3. otherwise, i.e., if A is of the form $A_1, \dots, A_n \rightarrow \varphi$, then $M(A) = M(A_1) \cup \dots \cup M(A_n)$.

Example 9 In our running example we have that $M(A_1) = \emptyset$, $M(A_2) = \{A_2\}$, $M(A_3) = \{A_3\}$, $M(A_4) = M(A_6) = \{A_4\}$, $M(A_5) = M(A_7) = \{A_5\}$.

3. Changing the $ASPIC^+$ framework

We now reconsider the rationality postulates of [1] in light of our discussion in Section 1 and then propose a modified version of $ASPIC^+$. Our proposal applies to both sceptical and credulous justification (cf. Definition 1), since an extension can be seen as a set of arguments that a rational agent could accept. We will discuss the rationality postulates as applying to single extensions, but note that if they are satisfied for single extensions, they are easily provable for the intersection of all extensions (cf. [1,9]).

We first discuss the consistency and strict-closure postulates². Direct consistency is not put into question by the lottery paradox or similar examples: it seems plainly irrational to simultaneously accept two propositions that negate each other. However, for strict closure and indirect consistency things are different. As discussed in Section 1, if a deductive inference is applied to at least two fallible subarguments, then it *aggregates* the ‘amounts’ of fallibility of its subarguments. This in turn means that the argument applying the deductive inference may be less preferred than either of these subarguments, so a successful attack on it does not imply a successful attack on one of these subarguments. Note that this line of reasoning does not apply to cases where a deductive inference is applied to at most one fallible element: then the amount of fallibility of the new argument is exactly the same as the amount of fallibility of the single fallible argument to which the deductive inference is applied. So we want to weaken the demand of strict closure to those subsets of an extension that contain at most one fallible argument. Combined with the wish to retain direct consistency, this implies a wish to restrict indirect consistency in the same way as strict closure.

We next discuss the changes in $ASPIC^+$. Consider the following modelling of the lottery paradox. Let \mathcal{L} be a propositional language built from the set of atoms $\{T_i \mid 1 \leq i \leq 1,000,000\}$. Then let X denote a well-formed formula $X_1 \vee \dots \vee X_{1,000,000}$ where \vee is exclusive or and where each X_i is of one of the following forms:

²For reasons of space, we do not formally list the postulates of [1] and leave the formulation of the new postulates implicit in the formal results of Section 4.

- If $i = 1$ then $X_i = T_1 \wedge \neg T_2 \wedge \dots \wedge \neg T_n$
- If $i = n$ then $X_i = \neg T_1 \wedge \neg T_2 \wedge \dots \wedge \neg T_{n-1} \wedge T_n$
- Otherwise $X_i = \neg T_1 \wedge \dots \wedge \neg T_{i-1} \wedge T_i \wedge \neg T_{n+1} \wedge \dots \wedge \neg T_n$

Next we choose $\mathcal{K}_p = \{\neg T_i \mid 1 \leq i \leq 1,000,000\}$, $\mathcal{K}_n = \{X\}$, \mathcal{R}_s as consisting of all propositionally valid inferences from finite sets and $\mathcal{R}_d = \emptyset$.

We want to formalise an account of the paradox in which for each individual ticket the statement that it will not win is sceptically justified, in which the statement that exactly one ticket will win is sceptically justified and in which the justification status of conjunctions of statements that a ticket will not win depends on the size of the conjuncts. In this section we only discuss the first two demands; the last one will be discussed in Section 5. Our analysis does not depend on the choice of semantics. The following arguments are relevant for any i such that $1 \leq i \leq 1,000,000$.

$$\neg T_i \quad \text{and} \quad \neg T_1, \dots, \neg T_{i-1}, \neg T_{i+1}, \dots, \neg T_{1,000,000}, X \rightarrow T_i \text{ (call it } A_i)$$

This requires for all i that $A_i \prec \neg T_i$, to prevent A_i from defeating $\neg T_i$. This in turn requires that Condition (2) of Definition 12 of reasonable argument orderings is dropped, since it excludes such an argument ordering. On the other hand, Condition (1) of Definition 12 can be retained. In particular, Condition (1.iii) captures that applying a strict rule to the conclusion of a single argument A to obtain an argument A' does not change the ‘preferredness’ of A' compared to A . This is reasonable in general, since A and A' have exactly the same set of fallible elements (ordinary premises and/or defeasible inferences).

Finally, we need to allow rebutting attacks on strict-rule applications applied to at least two fallible subarguments, since otherwise A_i is not defeated and both A_i and $\neg T_i$ are justified, which violates direct consistency. However, such rebuttals should not be allowed on strict rules applied to just one fallible argument, since then strict closure and indirect consistency do for preferred and stable semantics not even hold for strict inferences from at most one fallible subargument. A counterexample is $\mathcal{R}_d = \mathcal{K}_n = \emptyset$, $\mathcal{R}_s = \{b \rightarrow \neg m, m \rightarrow \neg b\}$ and $\mathcal{K}_p = \{b, m\}$. Then $\{b, m\}$ is an admissible set [1].

Based on this analysis, *ASPIC*⁺ is now adapted as follows. First, the definition of rebutting attack in Definition 6 is replaced with the following definition.³

Definition 14 [Semi-restricted rebut] *A* rebuts argument *B* (on *B'*) iff for some $B' \in \text{Sub}(B)$ it holds that $\text{Conc}(A) = \neg\varphi$ and either:

1. B' is of the form $B_1, \dots, B_n \Rightarrow \varphi$; or
2. B' is of the form $B_1, \dots, B_n \rightarrow \varphi$ and $n \geq 2$ and at least two of B_1, \dots, B_n are fallible.

Example 10 In our running example A_5 does still not rebut A_6 since A_6 applies its strict top rule to just one fallible subargument. However, if p is moved from \mathcal{K}_n to \mathcal{K}_p , then A_5 does rebut A_6 .

Definition 8 of defeat then directly applies to the modified framework. Finally, argument orderings are from now on assumed to be *weakly reasonable* in that they satisfy Condition (1) of Definition 12.

³[1,2] investigate similar notions of rebutting attack. However, they allow rebuttals on strict rules applied to only one fallible argument and do not investigate weakened versions of the rationality postulates.

4. The new rationality postulates verified

We now verify that the changed $ASPIC^+$ framework satisfies [1]’s postulates of closure under subarguments and direct consistency plus the new postulates of ‘restricted’ strict closure and ‘restricted’ indirect consistency. The results and proofs are based on those of [9] but reformulated or adapted when needed. For ease of comparison the original numbering of [9] is retained. In fact, for c-SAFs the results can only be proven under the assumption that an argument’s premises joined with \mathcal{K}_n is consistent. Accordingly, the notion of a c-SAF is redefined as follows:

Definition 15 [c-Structured Argumentation Frameworks redefined] Let $AT = (AS, \mathcal{K})$ be an *argumentation theory*. A *c-structured argumentation framework* (c-SAF) defined by AT , is a triple $\langle \mathcal{A}, \mathcal{C}, \preceq \rangle$ where \mathcal{A} is the set of all finite arguments constructed from \mathcal{K} in AS such that for all $A \in \mathcal{A}$ it holds that $\text{Prem}(A) \cup \mathcal{K}_n$ is indirectly consistent, \preceq is an ordering on \mathcal{A} , and $(X, Y) \in \mathcal{C}$ iff X attacks Y .

Well-defined structured argumentation frameworks for $ASPIC^+$ with semi-restricted rebut and a weakly reasonable argument ordering are below denoted with $(c-)SAF^{sw}$, where $c- SAF^{sw}$ ’s are defined as in Definition 15.

Lemma 11 For any argument A : $\text{Conc}(M(A)) \cup \text{Prem}_n(A) \vdash \text{Conc}(A)$.

PROOF. By induction on the structure of arguments. The result is obvious if $A \in \mathcal{K}$ or $\text{TopRule}(A) \in \mathcal{R}_d$. If $\text{TopRule}(A) \in \mathcal{R}_s$, then by the induction hypothesis $\text{Conc}(A_i) \in \text{Cl}_{\mathcal{R}_s}(\text{Conc}(M(A_i)) \cup \text{Prem}_n(A_i))$ for all A_i ($1 \leq i \leq n$). Since $\text{Prem}_n(A) = \text{Prem}_n(A_1) \cup \dots \cup \text{Prem}_n(A_n)$, the result follows. QED

Proposition 8 For any argument A and fallible argument B that have contradictory conclusions: (1) A defeats B ; or (2) some strict continuation $A+$ of A defeats B .

PROOF. If B has no strict top rule or a top rule applied to at least two fallible arguments, then clearly A defeats B . Otherwise, consider first systems closed under contraposition (Def. 10). By Lemma 11 it holds that $\text{Conc}(M(B)) \cup \text{Prem}_n(B) \vdash \text{Conc}(B)$. By contraposition, and since $\text{Conc}(A)$ and $\text{Conc}(B)$ contradict each other and $M(B) = \{B'\}$, we have that $\text{Prem}_n(B) \cup \text{Conc}(A) \vdash \varphi$ for some φ such that $\varphi = \neg \text{Conc}(B')$. Hence, one can construct a strict continuation $A+$ of A that concludes φ . Since by construction of $M(B)$ either B' is an ordinary premise or ends with a defeasible inference, $A+$ either undermines or rebuts B' . But then $A+$ also undermines or rebuts B .

For systems closed under transposition the existence of argument $A+$ is proven by straightforward generalisation of Lemma 6 of [1]. Then the proof is completed as above. In the case of c-SAFs, it must also be shown that $\text{Prem}(A+) \cup \mathcal{K}_n$ is indirectly consistent, which follows given $\text{Prem}(A+) \subseteq \text{Prem}(A) \cup \text{Prem}_n(B)$ and $\text{Prem}_n(B) \subseteq \mathcal{K}_n$, and $\text{Prem}(A) \cup \mathcal{K}_n$ is indirectly consistent by assumption.

2) Since $A+$ is a strict extension of A and B is a strict extension of B' and $A \not\leq B$, we have $A+ \not\leq B'$ by Condition (1c) of Definition 12, so $A+$ defeats B' and B . QED

Lemma 37 Let $(\mathcal{A}, \mathcal{C}, \preceq)$ be a $(c-)SAF^{sw}$. Let $A \in \mathcal{A}$ be a strict continuation of $S = \{A_1, \dots, A_n\} \subseteq \mathcal{A}$ such that at most one member of S is fallible, and for $i = 1 \dots n$, A_i is acceptable w.r.t. an admissible set $E \subseteq \mathcal{A}$. Then A is acceptable w.r.t. E .

PROOF. Let B be any argument defeating A . By Def. 6 of attack and Def. 14 of semi-restricted rebut, B attacks A by undercutting or rebutting on defeasible rules in A or undermining on an ordinary premise in A . Hence, by definition of strict continuations (Def. 11), it must be that B attacks A iff B attacks A_i for the unique fallible $A_i \in \{A_1, \dots, A_n\}$. Either:

- 1) B undercuts A_i , and so by Def. 8, B defeats A_i , or:
- 2) B does not undercut A_i . Suppose $B \prec A'_i$. This contradicts B defeats A . Hence, B defeats A_i .

We have shown that if B defeats A then B defeats some $A_i \in S$. By assumption of A_i acceptable w.r.t. E and E being admissible, $\exists C \in E$ s.t. C defeats B . Hence, A is acceptable w.r.t. E . QED

Proposition 9 Let $(\mathcal{A}, \mathcal{C}, \preceq)$ be a $c - SAF^{sw}$. If A_1, \dots, A_n are acceptable w.r.t. some admissible set $E \subseteq \mathcal{A}$ and at most one of A_1, \dots, A_n is fallible, then $\bigcup_{i=1}^n \text{Prem}(A_i) \cup \mathcal{K}_n$ is indirectly consistent.

PROOF. Suppose for contradiction otherwise and let S be any minimally indirectly inconsistent subset of $\bigcup_{i=1}^n \text{Prem}(A_i)$. Then for all $\varphi \in S$, $S \setminus \{\varphi\} \vdash \varphi'$ for all φ' such that $\varphi' = -\varphi$ and $S \setminus \{\varphi\}$ is indirectly consistent. Since at most one of A_1, \dots, A_n is fallible, we thus have for some A_i the set of ordinary premises $S = \{\varphi_1, \dots, \varphi_m\} \subseteq \text{Prem}(A_i)$ (that must be non-empty given that \mathcal{K}_n is indirectly consistent by assumption of axiom consistency (Def. 10)), that S is consistent but $S \cup \mathcal{K}_n$ is inconsistent. But this contradicts the fact that $\text{Prem}(A_i) \cup \mathcal{K}_n$ is indirectly consistent. QED

Theorem 12 [Sub-argument Closure] Let $\Delta = (\mathcal{A}, \mathcal{C}, \preceq)$ be a $(c-)SAF$ and E a complete extension of Δ . Then for all $A \in E$: if $A' \in \text{Sub}(A)$ then $A' \in E$.

PROOF. As in [9]. QED

Theorem 13 [Restricted closure under Strict Rules] Let $\Delta = (\mathcal{A}, \mathcal{C}, \preceq)$ be a $(c-)SAF^{sw}$ and E a complete extension of Δ and let $S \subseteq E$ be such that at most one element of S is fallible. Then $\text{Conc}(S) = Cl_{R_s}(\text{Conc}(S))$.

PROOF. It suffices to show that any strict continuation X of S is in E . By Lemma 37, any such X is acceptable w.r.t. E . By Proposition 10 of [9], $E \cup \{X\}$ is conflict free. Hence, since E is complete, $X \in E$. Note that if Δ is a $c-SAF$, then Proposition 9 guarantees that $\text{Prem}(X) \cup \mathcal{K}_n$ is indirectly consistent. QED

Theorem 14 [Direct Consistency] Let $\Delta = (\mathcal{A}, \mathcal{C}, \preceq)$ be a $(c-)SAF^{sw}$ and E a complete extension of Δ . Then $\{\text{Conc}(A) \mid A \in E\}$ is directly consistent.

PROOF. We show that if $A, B \in E$, $\text{Conc}(A) = -\text{Conc}(B)$, a contradiction results.

1. A is infallible, and: **1.1** if B is infallible, then this contradicts the assumption that \mathcal{K}_n is consistent. **1.2** if B is fallible, and **1.2.1** B is an ordinary premise or has a defeasible top rule or has a strict top rule applied to at least two fallible subarguments, then A defeats B contradicting E is conflict free, or **1.2.2** B has a strict top rule applied to at most one fallible subargument (see **3** below).

2. A is fallible, and: **2.1** if B is infallible then either **2.1.1** A is an ordinary premise or has

a defeasible top rule or has a strict top rule applied to at least two fallible subarguments, in which case B defeats A , contradicting E is conflict free, or **2.1.2** A has a strict top rule applied to at most one fallible subargument (see **3** below); **2.2** if B is fallible and **2.2.1** B is an ordinary premise or has a defeasible top rule or has a strict top rule applied to at least two fallible subarguments, then either A defeats B or B defeats A , contradicting E is conflict free, or **2.2.2** B has a strict top rule applied to at most one fallible subargument (see **3** below).

3. Each of **1.2.2**, **2.1.2** and **2.2.2** describes the case where $X, Y \in E$, $\text{Conc}(X) = -\text{Conc}(Y)$, Y is fallible and has a strict top rule applied to at most one fallible subargument. In the case that Δ is a $c\text{-SAF}$, since $X, Y \in E$, then X, Y are acceptable w.r.t. E , and so by Proposition 9, $\text{Prem}(A) \cup \text{Prem}(B) \cup \mathcal{K}_n$ is indirectly consistent. By Proposition 8 there is a strict continuation $X+$ of X that defeats Y . By Lemma 37 $X+$ is acceptable w.r.t. E , and by Proposition 10 of [9], $E \cup \{X+\}$ is conflict free, contradicting $X+$ defeats Y . QED

Then Theorem 15 follows from Theorems 13 and 14.

Theorem 15 [Restricted Indirect Consistency] Let $\Delta = (\mathcal{A}, \mathcal{C}, \preceq)$ be a $(c-)\text{SAF}^{sw}$ and E a complete extension of Δ and let $S \subseteq E$ be such that at most one element of S is fallible. Then $\text{Conc}(S)$ is indirectly consistent.

5. Conclusion

We first verify that the new variant of ASPIC^+ is a middle ground between the extremes of Pollock and Kyburg in that whether a deductive consequence of multiple rationally acceptable propositions is also rationally acceptable depends on the specific example. The crucial element here is the argument ordering. Recall the modelling in Section 3 of the lottery paradox and assume that arguments have a numerical fallibility degree f , being the number of ordinary premises that they use. Next we define a ‘bandwidth’ for strict argument preference, by letting for any pair of fallible arguments A and B , $A \prec B$ iff $f(A) - f(B) > n$ for some natural number n . More sophisticated argument orderings may be possible but this one suffices to illustrate our point. Now if, for example, $n = 600,000$ and adopting preferred semantics for illustration, then all arguments for conjunctions $\neg T_i \wedge \dots \wedge \neg T_j$ with fewer than 200,000 conjuncts strictly defeat their rebutting counterarguments and are thus in all preferred extensions, the arguments for conjunctions between 200,000 and 800,000 conjuncts defeat and are defeated by their rebutting counterarguments so are in some but not all preferred extensions, while the arguments with more than 800,000 conjuncts are strictly defeated by their rebutting counterarguments so are not in any preferred extension.

We next conclude. In this paper we presented an argumentation-based notion of fallible rational acceptance according to which one can sometimes rationally accept sets of propositions that are indirectly inconsistent or not strictly closed. We proposed new rationality postulates capturing this idea and proposed a variant of ASPIC^+ that satisfies the new postulates while not satisfying their original versions. While we illustrated these ideas with a purely probabilistic example, the basic intuition is more general, being that an argument formed by strictly extending more than one fallible subargument has more

fallibility than each of the combined arguments alone. Therefore, the relevance of this paper is not confined to discussions of the lottery paradox but extends to any application of argumentation in which arguments can have multiple fallible elements.

Our approach captures the intermediate position that deductive inferences from multiple fallibly acceptable propositions can but need not be acceptable. The argumentation approach here provided a fresh logical perspective compared to other logical approaches. First, the truth-preserving nature of deductive inference rules is respected by allowing their application inside arguments as strict rules. A key observation here is that preservation of truth does not imply preservation of rational acceptability, since truth and rational acceptability are different things. A virtue of an argumentation approach is that it can naturally model this distinction, since the strict-closure postulate does not capture preservation of truth but preservation of rational acceptability. Second, argumentation can make a natural distinction between cases where strict closure and indirect consistency do and do not hold, since if an argument that applies a deductive inference to fallible subarguments is not rebutted on this inference or if none of its rebuttals are strong enough to defeat it, then this argument can still be acceptable. The notion of an argument ordering is crucial here, since it can make fine-grained distinctions between cases where applications of deductive inferences are and are not strong enough to survive attack.

Having said so, it remains to be investigated how argument orderings can be defined in principled ways. For example, can they help in modelling argumentation-based counterparts of [8]’s “lossy” inference rules, or [3]’s “big-step probabilities” (their attempt to distinguish between cases with and without uniform underlying probability structures)? Such investigations could shed further light on the relation between argumentation-based and other logical modellings of reasoning with uncertain information.

References

- [1] M. Caminada and L. Amgoud. On the evaluation of argumentation formalisms. *Artificial Intelligence*, 171:286–310, 2007.
- [2] M. Caminada, S. Modgil, and N. Oren. Preferences and unrestricted rebut. In S. Parsons, N. Oren, C. Reed, and F. Cerutti, editors, *Computational Models of Argument. Proceedings of COMMA 2014*, pages 209–220. IOS Press, Amsterdam etc, 2014.
- [3] D. Dubois, H. Fargier, and H. Prade. Ordinal and probabilistic representations of acceptance. *Journal of Artificial Intelligence Research*, 22:23–56, 2004.
- [4] P.M. Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming, and n -person games. *Artificial Intelligence*, 77:321–357, 1995.
- [5] R. Foley. Beliefs, degrees of belief, and the Lockean thesis. In F. Huber and C. Schmidt-Petri, editors, *Degrees of belief*, volume 342 of *Synthese Library*, pages 37–47. Springer, 2009.
- [6] N. Gorgiannis and A. Hunter. Instantiating abstract argumentation with classical-logic arguments: postulates and properties. *Artificial Intelligence*, 175:1479–1497, 2011.
- [7] H. Kyburg. *Probability and the Logic of Rational Belief*. Wesleyan U. P. Middletown, CT, 1961.
- [8] D. Makinson. Logical questions behind the lottery and preface paradoxes: lossy rules for uncertain inference. *Synthese*, 186:511–529, 2012.
- [9] S. Modgil and H. Prakken. A general account of argumentation with preferences. *Artificial Intelligence*, 195:361–397, 2013.
- [10] J.L. Pollock. Defeasible reasoning. In J. Adler and L. Rips, editors, *Reasoning: Studies of Human Inference and its Foundations*, pages 451–470. Cambridge, Cambridge University Press, 2007b.
- [11] D.L. Poole. The effect of knowledge on belief: Conditioning, specificity and the lottery paradox in default reasoning. *Artificial Intelligence*, 49:281–307, 1991.
- [12] H. Prakken. An abstract framework for argumentation with structured arguments. *Argument and Computation*, 1:93–124, 2010.

Assessing Weight of Opinion by Aggregating Coalitions of Arguments

Pavithra RAJENDRAN^a, Danushka BOLLEGALA^a and Simon PARSONS^b

^a*Department of Computer Science, University of Liverpool*

^b*Department of Informatics, King's College London*

Abstract. Argument mining promises to be able to extract information from unstructured text that can help us to understand that text. This paper suggests a novel way to use such information once it has been extracted. Attack and support relations between arguments from a set of test texts are identified, the strength of the arguments is computed based on the relations, and arguments are grouped into coalitions. The resulting set of arguments is then used to predict the weight of opinion in new text, by identifying arguments whose weight has been computed, and aggregating these weights. Our approach is evaluated on a corpus of hotel reviews, and compared with an existing method of predicting the sentiment of reviews.

Keywords. argument mining, bipolar argumentation, coalitions of arguments

1. Introduction

Argumentation mining is an emerging field that focuses on the identification and extraction of arguments from natural language texts. The aim of such work is to pinpoint what opinions are expressed for and against some point of view. These arguments can then be used in understanding the text, perhaps to highlight the key issues raised, or to summarise the overall view expressed in the text. In this paper we contribute to the growing literature around argument mining, studying the use of arguments that have been extracted. In particular, we are interested in investigating how techniques from computational argumentation can be used to process the results of argument mining, establishing what can be done to gain insights about the texts from which the arguments were mined.

In this paper the texts we process are online reviews. We take a set of arguments that are hand-extracted from reviews, and, based on ideas from bipolar argumentation [1], extract coalitions of arguments that relate to the products (hotels) that are reviewed. We then evaluate different approaches for aggregating these coalitions, and assess whether the result of the aggregation can be used to predict the weight of opinion about the products, as expressed by the star rating of the reviews. Note that we are not interested in reviews *per se* — for a set of reviews, the overall star rating is probably the best guide to the weight of opinion. However, precisely because of these star ratings, reviews are a very convenient dataset to refine our approach before analysing more general texts.

Any work on argument mining will be dependent on the precise definition of “argument” that is used in that work. This term has several definitions. A typical definition is the combination of a set of premises and the conclusion that these premises lead to. Argu-

Table 1. Statements present in a review annotated as argument or not using our definitions.

Statements	Sentiment	Aspect	Argument	Type
My mother and I stayed at the Warwick for 2 nights in November.	objective	none	no	-
The hotel itself was ok, fairly clean and decent location.	positive	yes	yes	support
The front desk staff, however are not helpful and pass the buck so as not to have to deal with a problem.	negative	yes	yes	attack

ments of this form are difficult to extract from the unstructured text present in online reviews, forums, blogs etc and methods to accurately extract them are under-development. Wyner et al. [2], for example, describes work extracting such arguments using a set of argumentation schemes. Garcia-Villalba and Saint-Dizier [3] also show how this approach can help in generating arguments using evaluative expressions such as “*well located hotel*” and also evaluate such statements using rhetorical relations for argument extraction.

Rather than focussing on extracting structured arguments, we use knowledge of product attributes to extract statements that can be considered arguments for or against a product. We deal with what we call *aspects*. In our terminology an aspect is an entity relating to a product or service about which a review writer expresses an opinion. Both aspects and the relation between aspect properties and opinions about the product or service are highly domain dependent. For example, “*battery is small and lightweight*” in an electronics review is a positive statement about the battery aspect while “*rooms are small*” is a negative statement about the room aspect in a hotel review.

Statements about aspects are then classified as arguments for or against a product:

Argument A statement that is either a supporting argument or an attacking argument.

Supporting argument A statement that has positive polarity and can be considered to support the product by supporting an aspect of the product or the product itself.

Attacking argument A statement that has a negative polarity and can be considered to attack the product by attacking an aspect of the product or the product itself.

Examples of arguments can be found in Table 1. Note that we group related aspects into *aspect categories*, and consider that there is some level of equivalence between statements about aspects in the same category. Given a set of arguments of this form, this paper examines whether they can be used to establish the overall opinion in a way that agrees with the review writers.

2. Background

Dung’s abstract argumentation framework [4] provides a framework for analysing a set of arguments with attack relations between them. Bipolar argumentation [5] extends this by introducing the notion of support as an independent interaction among arguments:

Definition 1. An abstract bipolar argumentation framework is a 3-tuple $\langle \mathcal{A}, \mathcal{S}, \mathcal{R} \rangle$ where \mathcal{A} is a set of arguments such that \mathcal{S} represents the support relation and \mathcal{R} represents the attack relation between the arguments.

A bipolar argumentation framework can be represented as a bipolar interaction graph in which arguments are nodes and support and attack relations are edges. Cayrol & Lagasquie-Schiex [1] further proposed the structuring of a bipolar argumentation framework into coalitions of arguments.

Definition 2. A coalition of arguments is a set of arguments supporting each other directly or indirectly where conflicts occur among such coalitions. These coalitions of arguments satisfy the following properties:

1. There is no direct attack among pairs of arguments belonging to the same coalition.
2. Any pair of arguments in a coalition will have a direct or indirect support relation between them.
3. If an argument in coalition A attacks an argument in coalition B, then A attacks B.

Since a set of reviews of a given product will contain multiple arguments for and against different aspects of that product, we consider such a set of reviews as a coalition of arguments.

3. Data preparation

3.1. Dataset

We used an existing dataset, the ArguAna corpus [6], which contains manually annotated hotel reviews from TripAdvisor.com. The corpus contains each review, identified with a review id, the author name, the local sentiment of each statement (positive or negative) in the review, and the aspects present in the statement. Each review has the star rating provided by the reviewer. Several existing classifiers are available for automatically identifying sentiment, but since sentiment data was already available, we used it. We manually collected the aspects present in each review, and each statement that contained any of the aspects and was labelled as positive or negative was considered to be an argument. Every statement was extracted from each review for a given hotel, and the arguments were collected together regardless of whether or not they belonged to the same review.

3.2. Automatic identification: Support/Attack

The ArguAna corpus does not contain relations between arguments present in the reviews. To extract this information we used the Takelab STS¹ System. There are three types of relations that we wanted to identify between pairs of arguments — support (arguments about aspects in the same aspect category with same sentiment), attack (arguments about aspects in the same aspect category with opposite sentiment) and unknown (arguments about aspects in different aspect categories).

These definitions are based on inferring whether two statements support/attack in different ways but target the same conclusion. The aspects of the product or service are grouped into different categories based on their common properties. For instance, in a hotel review, the aspects *staff* and *manager* belong to the same aspect category.

To detect relations, we took a sample set of arguments, paired them according to the above definitions and manually annotated the relations. We then used Takelab STS to obtain the semantic similarity scores for each pair of arguments. TakeLab STS accepts

¹<http://takelab.fer.hr/sts/>

two statements as input and produces a semantic similarity score ranging from 0 (lowest semantic similarity) to 5 (highest semantic similarity). In our experiments, the maximum similarity score was 3. To avoid errors, we set a minimum similarity score of 1.0 as a threshold below which we consider that there is no relation between statements (and above which we considered a relation to hold), which gave a macro-averaged F1-score of 0.18 for automatically predicting the manually annotated relations. While relation prediction was not perfect, it was sufficient for our purposes.

3.3. Coalitions of arguments in reviews

Arguments present in reviews, according to our definition, relate to aspects. Considering the properties of the support and attack relations with respect to aspects, we noticed that the support relations naturally fall into coalitions, where each argument within a coalition relates to the same aspect and all the arguments support each other directly or indirectly. This gives rise to several questions such as what kind of coalitions of arguments are formed in a single review, and in a set of randomly selected reviews. We were not able to find coalitions of arguments in a single review, since it seems that in our dataset each review contains at most one statement about each aspect. For the remainder of the paper we study coalitions of arguments across sets of **Low** reviews (reviews with 1 star and 2 star rating) and sets of **High** reviews (4 star and 5 star ratings).

4. Aggregating natural language arguments

We are interested in interpreting a set of reviews for a particular hotel. Across all the reviews of that hotel, a number of aspects will have been mentioned by the reviewers. We consider all the comments about a specific aspect as being an argument for or against the hotel, and we will aggregate these arguments to get the overall opinion about the hotel.

4.1. Arguments for aspects

The first step in the process is to identify attack and support relations between arguments. This is done, as described above, using TakeLab STS. The second step is to compute the weight of each argument. There are several methods that we could use to compute the weight of an argument on the basis of the arguments that support and attack it, and from these possibilities we picked the intrinsic generic gradual valuation method proposed by Cayrol and Lagasque-Schiex [7] which takes into account arguments that support and attack the argument in question:

Definition 3. For every argument $a \in \mathcal{A}$ with a set of supporters $\mathcal{B} = \{b_1, b_2, \dots, b_n\}$, and attackers $\mathcal{C} = \{c_1, c_2, \dots, c_m\}$, the gradual valuation function $g : \mathbb{R}^2 \rightarrow \mathbb{R}$ is defined as:

$$v(a) = g_{\text{agg}}(h_{\text{agg}}^{\text{sup}}, h_{\text{agg}}^{\text{att}}) = \left(\frac{1}{h_{\text{agg}}^{\text{att}} + 1} - \frac{1}{h_{\text{agg}}^{\text{sup}} + 1} \right); h_{\text{agg}}^{\text{sup}}(\mathcal{A}) = \sum_{i=1}^n v(b_i), h_{\text{agg}}^{\text{att}}(\mathcal{A}) = \sum_{i=1}^m v(c_i) \quad (1)$$

We assume the initial strength value of each argument satisfies $v(b_i) = 1$ for $i = 1, \dots, n$, and $v(c_j) = 1$ for $j = 1, \dots, m$, irrespective of whether they are supporting or attacking.

The previous step gives us a value for each individual argument. Before combining arguments to summarise reviews, we structure arguments into coalitions. We do this exactly following Definition 2. This gives us a set of coalitions, each of which supports or attacks an aspect. Each coalition is a set of arguments, and each argument has a weight.

4.2. Aggregating coalitions

We consider the opinion about each aspect category of a hotel to be an argument about the hotel. The strength of opinion about the hotel is then a combination of the strengths of opinions about the aspect categories (which depend on the arguments in the coalitions that relate to the aspects). We could establish the opinion about the hotel by combining all the aspect categories and all the arguments for each aspect category, but it isn't clear that we want to include either all the arguments that bear on each aspect category or all the aspect categories that relate to each hotel. We infer this on the basis of the work of Wachsmuth et al. [6] who studied the patterns of positive and negative statements in the same corpus that we use and showed that the most negative reviews contain most of the negative statements and the most positive reviews contain most of the positive statements. This suggests that we should only consider a subset of the arguments present when assessing hotels. To do this, we divided the arguments into two categories, **Low** and **High**, using the ground-truth data provided by the star ratings of the reviews in which each of the arguments were present. Arguments in **Low** reviews were rated **Low**, those in **High** reviews were rated **High**. We then considered four different ways in which to choose the arguments that should be taken into account:

ArgAll All arguments, regardless of the coalition they belong to, are taken into account, and we consider arguments for all aspect categories when rating a hotel.

AttSupCoal All attacking arguments from coalitions of arguments in **Low** rated reviews, and all supporting arguments in coalitions of arguments in **High** rated reviews are taken into account. Again we consider arguments for all aspect categories.

AttSupArg This is a refinement of *AttSupCoal* in which we only consider the arguments relating to the aspect category attacked by the strongest attacking coalition, and the arguments relating to the aspect category supported by the strongest supporting coalition when rating a hotel.

AttSupBoth A hybrid version of *AttSupCoal* and *AttSupArg*. It initially picks all attacking arguments from coalitions of arguments in **Low** rated reviews, and all supporting arguments in coalitions of arguments in **High** rated reviews just like *AttSupCoal*. However, these arguments are then filtered by only including arguments for those aspects (rather than aspect categories) that are present in the review being rated. The strengths of the resulting sub-coalitions are computed, and the strongest attacking and supporting sub-coalitions are used to rate the hotel (echoing *AttSupArg*).

These four approaches all identify a set of arguments to take into account. We then experimented with two ways of using these sets of arguments, both taking all the arguments in the set into account, an approach we call f_{agg} , and just taking the strongest arguments in a set into account, an approach we call f_{max} . Again, the idea of focusing on the strongest arguments comes from [6].

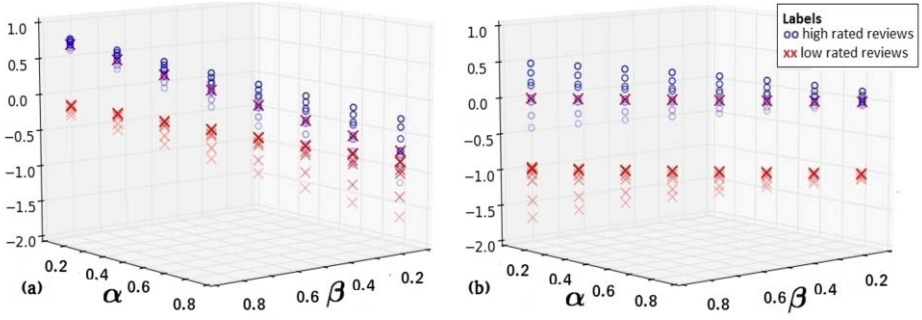


Figure 1. Scores of each review for a single hotel. A red cross denotes a review that belongs to **LOW** and a blue circle denotes a review that belongs to **High**. (a) Scores vs α and β aggregating using *ArgAll* (b) Scores vs α and β aggregating using *AttSupBoth*. Both use f_{max} .

4.3. Argument aggregation value function

For each hotel review we consider, we now have a set of arguments for and against it, selected using the methods described above. Having already used Eq. 1 to compute the strength of each argument, it is natural to use Eq. 1 to combine the strengths of the arguments for and against each hotel. However, such a combination does not distinguish well between **Low** and **High** rated reviews. As a result, we introduce a generalisation of Eq. 1 in which supporting and attacking arguments are weighted differently. In particular we considered the overall strength of an argument to be a function of the strength of the coalition of supporting arguments (SCV: supporting coalition value) and the coalition of attacking arguments (ACV: attacking coalition value):

$$f(h^{sup}(SCV), h^{att}(ACV)) = \left(\frac{1}{\beta h^{att}(ACV) + 1} - \frac{1}{\alpha h^{sup}(SCV) + 1} \right) \quad (2)$$

where, $\alpha, \beta, \alpha + \beta = 1$ provide a simple way of weighting the support and attack components differently. Exactly which arguments are included in ACV and SCV depends on the choice from $\{AllArg, AttSupArg, AttSupCoal, ArgSupBoth\}$ and $\{f_{max}, f_{agg}\}$.

To establish the optimum values of α and β to use with Eq. 2, we computed results for values of α and β across $[0, 1]$. For each pair of values, we followed a process analogous to 10-fold cross-validation, training on 90% of the reviews and testing on 10%, and averaging results across 10 repetitions². We did this for 14 different random hotel datasets, each of which contains an average of 25 individual reviews. We performed the experiment for both balanced and unbalanced sets of reviews, recognising that this gave us three different categories of hotel that we were attempting to categorise — hotels with a majority of low rated reviews (unbalanced), hotels with a majority of high rated reviews (unbalanced) and hotels with a balanced set of low rated and high rated reviews (balanced). We repeated the experiment for *ArgAll* and *AttSupBoth* with f_{max} . Figure 1 shows the results, scores for a set of reviews belonging to a particular hotel. Each **Low**-rated review is represented by a red cross and each **High**-rated review is represented by a blue circle. The score of each review is computed using varying values of α and β , and

²In this work “training” is going through the process of extracting arguments from reviews, weighting the arguments and identifying coalitions. “Testing” is then using these arguments to rate new reviews.

Table 2. Results for prediction of reviews. The numbers are the percentage of reviews correctly predicted into category **Low** and **High**. The highest value on each line is highlighted. Test data was reviews from 14 different hotels. There were 217 **Low** reviews and 148 **High** reviews. Because of the imbalance between **Low** and **High**, we report results for hotels where the majority of reviews were **Low**, where the majority of reviews were **High**, and where the number of reviews were approximately equal, as well as the overall results.

	Category	<i>AttSupBoth</i>		<i>AttSupArg</i>		<i>AttSupCoal</i>		<i>AllArg</i>	
		f_{agg}	f_{max}	f_{agg}	f_{max}	f_{agg}	f_{max}	f_{agg}	f_{max}
Majority Low reviews	Low	96	97	88	92	74	90	80	68
	High	37	50	22	35	31	22	16	16
Balanced reviews	Low	90	93	85	90	76	87	85	72
	High	35	35	33	45	54	26	23	23
Majority High reviews	Low	84	92	88	92	52	88	80	64
	High	23	40	38	38	76	25	28	28
Overall	Low	93	96	86	90	72	87	80	70
	High	36	46	31	39	54	24	20	20

Table 3. Comparison with ArguAna. Conditions as in Table 2.

	Category	Majority High	Balanced	Majority Low	Overall
<i>AttSupBoth</i> , f_{max}	Low	97	93	92	96
	High	50	35	40	46
ArguAna	Low	99	93	100	97
	High	29	21	30	28

from the figures it is evident that, (a) for *ArgAll* aggregation there is no clear separation between **Low** and **High** reviews for any value of α and β whereas (b) for *AttSupBoth* aggregation, there is a clear gap between the scores of low rated and high rated reviews that seems to widen for particular values of α and β . This suggests that our approach, along with aggregation of arguments based on Eq. 2, *AttSupBoth* and f_{max} can weigh up arguments in a review in a way that broadly agrees with the writer of the review.

4.4. Evaluation

Having established the potential of our approach, we carried out a more detailed evaluation. First we examined the relative performance of the four methods for picking which coalitions to take into account (*ArgAll*, *AttSupCoal*, *AttSupArg*, *AttSupBoth*) and the two methods for selecting arguments to aggregate (f_{agg} and f_{max}). We ran the same 10-fold cross-validation exercise as before, set $\alpha = 0.75$ and $\beta = 0.25$, and evaluated the methods by predicting whether reviews for 14 randomly selected hotels were **High** or **Low**. This was a set of 217 **Low** reviews and 148 **High** reviews. The results are given in Table 2 which reports the percentage of reviews that were correctly predicted. We ran two-tailed t-tests on each pair of comparable results — that is every pair of results on the same line of the table. All differences in value are significant at the 0.05 level except those between the predictions made by *AttSupArg*/ f_{agg} and *AttSupCoal*/ f_{max} for the **LOW** category.

The results suggest that the combination of *AttSupBoth* and f_{max} is the best predictor across the different segments, though it is outperformed by *AttSupCoal* and f_{agg} in terms of the prediction of **High** reviews. We interpret this as evidence that focusing on the most strongly held relevant opinion (as Wachsmuth et al. [6] suggest) is the key to good

prediction, but that the best way to pick the relevant opinions (the ones from which the strongest are selected) varies depending on whether the review is positive or negative. To come back to our original question — whether we can combine arguments to reach a view that matches the opinion of the review writers — the results suggest we can do this well for **High**, and with some accuracy for **Low** reviews, though with the latter there is considerable room for improvement. The dataset we used was developed to test a sentiment classification tool called *ArguAna* [8]. We compared our approach (*AttSupBoth*, f_{max}) with *ArguAna*. The results are given in Table 3. Considering the Overall results, a two-tailed t-test tells us that our approach is significantly better in predicting **High** reviews and not significantly worse in predicting **Low** reviews. In fact, in all categories, our approach does much better in predicting **High** reviews.

5. Conclusion

This paper considered the task of weighing up the arguments in a text to determine the overall opinion being expressed. We proposed a method that starts from a set of arguments extracted from online reviews. This involves identifying support and attack relations between these arguments, computing the weight of the arguments, and identifying coalitions of arguments. Having established a training set of such arguments, we showed that the arguments, weights and coalitions could be used to evaluate new reviews in a way that can distinguish between two broad classes of positive and negative reviews. Our approach compares well with an existing approach to sentiment analysis of reviews, outperforming the existing approach in identifying positive reviews and doing no worse on negative reviews. Note that the overall aim of this work is not to predict the sentiment of reviews. We concentrated on reviews here because reviews come with star ratings that provide a form of ground truth data about the opinion of the review writer. Our aim is to be able to summarise the opinion expressed in general texts.

Acknowledgement: PR is supported by a scholarship from the University of Liverpool.

References

- [1] C. Cayrol and M.-C. Lagasquie-Schiex. Coalitions of arguments: A tool for handling bipolar argumentation frameworks. *Int. J. Intell. Syst.*, 25(1):83–109, 2010.
- [2] A. Wyner, J. Schneider, K. Atkinson, and T. J. M. Bench-Capon. Semi-automated argumentative analysis of online product reviews. In B. Verheij, S. Szeider, and S. Woltran, editors, *COMMA'12*, volume 245 of *Frontiers in Artificial Intelligence and Applications*, pages 43–50. IOS Press, 2012.
- [3] M. P. Garcia-Villalba and P. Saint-Dizier. A framework to extract arguments in opinion texts. *IJCI*, 6(3):62–87, 2012.
- [4] P. M. Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n -person games. *Artif. Intell.*, 77:321–357, 1995.
- [5] L. Amgoud, C. Cayrol, and M.-C. Lagasquie-Schiex. On the bipolarity in argumentation frameworks. In J. P. Delgrande and T. Schaub, editors, *NMR'04*, pages 1–9, 2004.
- [6] H. Wachsmuth, M. Trenkmann, B. Stein, G. Engels, and T. Palakarska. A review corpus for argumentation analysis. In *ICCLITP'14*, pages 115–127, April 2014.
- [7] C. Cayrol and M.-C. Lagasquie-Schiex. Gradual valuation for bipolar argumentation frameworks. In L. Godo, editor, *ECSQARU'05*, volume 3571 of *LNCS*, pages 366–377. Springer, 2005.
- [8] H. Wachsmuth, M. Trenkmann, B. Stein, and G. Engels. Modeling review argumentation for robust sentiment analysis. In *ICCL'14*, pages 553–564, 2014.

Perfection in Abstract Argumentation¹

Christof SPANRING

*University of Liverpool, Department of Computer Science, Agent ART Group
TU Wien, Faculty of Informatics, Database and Artificial Intelligence Group*

Abstract. It is a well-known fact that stable semantics might not provide any extensions for some given abstract argumentation framework. Arguably such frameworks might be considered futile, at least with respect to stable semantics. We propagate σ -perfection stating that for a given argumentation graph all induced subgraphs provide σ -extensions. We discuss perfection and conditions for popular abstract argumentation semantics and possibly infinite frameworks.

Keywords. argumentation, semantics, foundations, existence, perfection

Introduction

Abstract argumentation uses arguments and a two-valued attack relation as atomic structure, and semantics to assign acceptance states to sets of arguments. In his seminal paper Dung in 1995 [1] already gave conditions for semantics to provide extensions but also examples of meaningful argumentation systems without stable extensions. Subsequently various semantics have been introduced not least to circumvent the problem of vanishing extension sets. In this work we elaborate on structural extension existence conditions. To this end we draw inspiration from kernel-perfection [2]. Given semantics σ , an argumentation framework is σ -perfect if every induced subframework provides σ -extensions. To flesh out σ -perfection in abstract argumentation we advance on known results and present novel approaches particularly for semi-stable and stage semantics.

Non-interference, contaminating frameworks and crash have been popularized as properties of argumentation semantics [3]. For various reasons these properties do not match our intuitions. When thinking about abstract argumentation semantics intuitively we want to be able to evaluate independent components of some framework independently from each other. We introduce this property as *well-definedness*. We elaborate on issues with the other properties in the Background section and use the term *collapse* from [4] to refer to our intuitive concept of crash (vanishing extension sets).

The remaining parts of this paper are organized as follows:

- In Section 1 we introduce all necessary background definitions and discuss the issue of well-definedness and collapse vs. non-interference and crash.
- In Section 2 we introduce perfection and present a fine collection of related results. This culminates in a rather sophisticated tool for stage semantics.
- In Section 3 we wrap up, relate to the literature, present a conjecture and discuss other possible future research directions.

¹This research has been supported by the Austrian Science Fund (FWF) through projects I1102 and I2854.

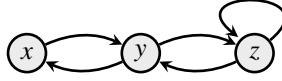


Figure 1. A simple AF as discussed in Example 1. AFs frequently are visualized as graphs where nodes reflect arguments and directed edges reflect attacks between arguments.

1. Argumentation and Fairness

Let us first introduce common definitions and basic framework operations.

Definition 1. An *argumentation framework (AF)* is an ordered pair $F = (A, R)$ where A is an arbitrary set of *arguments* and $R \subseteq A \times A$ is called the *attack relation*. For $(a, b) \in R$ we say that a attacks b . Furthermore, for $S \subseteq A$ and $a \in A$ we say that a attacks S (or S attacks a) if for some $b \in S$ we have a attacks b (or b attacks a). We use the term *defense* to denote some argument(s) attacking all attackers of some (other) argument(s). Finally, for $S \subseteq A$ we call $S^+ = S \cup \{a \in A \mid S \text{ attacks } a\}$ the range of S in F .

For a given AF $F = (B, S)$ use $A_F = B$ and $R_F = S$ to denote its arguments and attacks respectively. For given AFs F, G with $A_F \cap A_G = \emptyset$ we use the disjunct union $F \uplus G = (A_F \cup A_G, R_F \cup R_G)$. For given AF F and argument set $X \subseteq A_F$ we use the restriction operator $F|_X = (X, X \times X \cap R_F)$.

Investigating some arbitrary AF we consider sets of arguments, and investigate whether these sets appear to be justified under some principles, also called argumentation semantics. For a comprehensive introduction into argumentation semantics see [3]. Additional to semantics discussed in [1] we consider semi-stable and stage semantics [5,6].

Definition 2. A *semantics* is a mapping from AFs to sets of arguments, where for any AF F and semantics σ we have $\sigma(F) \subseteq \wp(A_F)$. The members of $\sigma(F)$ are then called σ -*extensions* of F . By stating properties a specific extension has to fulfill, we will now define the semantics of interest for this work.

A set $S \subseteq A_F$ is called *conflict-free (cf)*, $S \in cf(F)$ if no member attacks any other member. $S \in cf(F)$ is called *admissible (ad)*, $S \in ad(F)$ if it defends itself against attacks from the outside. An extension $S \subseteq A_F$ is called

- *complete (co)*, $S \in co(F)$ if $S \in cf(F)$ and S contains all arguments defended by S ,
- *grounded (gr)*, $S \in gr(F)$ if $S = \bigcap co(F)$,
- *naive (na)*, $S \in na(F)$ if $S \in cf(F)$ and there is no $S' \in cf(F)$ with $S \subset S'$,
- *preferred (pr)*, $S \in pr(F)$ if $S \in ad(F)$ and there is no $S' \in ad(F)$ with $S \subset S'$,
- *stage (sg)*, $S \in sg(F)$ if $S \in cf(F)$ and there is no $S' \in cf(F)$ with $S^+ \subset S'^+$,
- *semi-stable (ss)*, $S \in ss(F)$ if $S \in ad(F)$ and there is no $S' \in ad(F)$ with $S^+ \subset S'^+$,
- *stable (sb)*, $S \in sb(F)$ if $S \in cf(F)$ and $S^+ = A_F$.

Example 1. Consider the AF $F = (\{x, y, z\}, \{(x, y), (y, x), (y, z), (z, y), (z, z)\})$ as depicted in Figure 1. Here the arguments could for instance refer to sentences such as x :(everything is finite), y :(infinity is real), z :(reality is finite infinity). We have $cf(F) = ad(F) = co(F) = \{\emptyset, \{x\}, \{y\}\}$, $gr(F) = \{\emptyset\}$, $na(F) = pr(F) = \{\{x\}, \{y\}\}$, $sg(F) = ss(F) = sb(F) = \{\{y\}\}$. Observe that these equality relations do not hold for arbitrary AFs. However for any AF F it holds that $sb(F) \subseteq sg(F) \subseteq na(F) \subseteq cf(F)$ and $sb(F) \subseteq ss(F) \subseteq pr(F) \subseteq ad(F) \subseteq cf(F)$.

In opposition to the traditional semantics properties of crash-resistance and non-interference we will use a different word to denote a formally different meaning.

Definition 3 (Collapse). A semantics σ is said to *collapse* for some AF F if $\sigma(F) = \emptyset$.

We now give intuitive properties for semantics with the main principle of fairness in mind. There should be acceptable arguments for some frameworks. Arguments should be treated equally. We should be able to evaluate components of the union of disjunct AFs independently from each other.

Definition 4 (Fairness). An argumentation semantics σ is called

1. *basic* if there is some AF F and argument set $S \neq \emptyset$ such that $S \in \sigma(F)$;
2. *language independent* [3] if isomorphic AFs produce isomorphic extension sets;
3. *well-defined* if it evaluates separate components separately, for AFs F, G, H with $H = F \uplus G$ we have $\sigma(H) = \{S \cup T \mid S \in \sigma(F), T \in \sigma(G)\}$;
4. *fair* if it is basic, language independent and well-defined.

All semantics under consideration are fair semantics. We even go a bit further and state that only fair semantics are of use for abstract argumentation. For the purpose of reference we give a formal definition of non-interference and crash-resistance and follow up by showing equivalence of collapse with crash and interference for fair semantics.

Definition 5. A semantics σ is *non-interfering* if for AFs F, G, H with $H = F \uplus G$ we have $\sigma(F) = \{S \cap A_F \mid S \in \sigma(H)\}$. A semantics σ is *crash-resistant* if there is no AF F such that for all disjunct AFs G we have $\sigma(F \uplus G) = \sigma(F)$, otherwise it *crashes* at F .²

Lemma 1. A given fair semantics σ collapses for some AF F if and only if it violates crash-resistance and non-interference.

Proof. Assume $\sigma(F) = \emptyset$ for some AF F . By well-definedness for any disjoint AF G we get $\sigma(F \uplus G) = \{S \cup T \mid S \in \emptyset, T \in \sigma(G)\} = \emptyset$, i.e. σ crashes at F and (in case $\sigma(G) \neq \emptyset$, granted σ is basic language-independent) also violates the non-interference property.

Now assume σ does not collapse for any AF and consider some arbitrary syntactically disjoint AFs F and G , and $H = F \uplus G$. Since σ does not collapse we have $\sigma(F) \neq \emptyset$ and $\sigma(G) \neq \emptyset$. By well-definedness we then get $\sigma(H) = \{S \cup T \mid S \in \sigma(F), T \in \sigma(G)\}$ and hence non-interference. With σ being basic wlog. there is some AF F with $S \in \sigma(F)$ and $S \neq \emptyset$. By definition of semantics and disjointness we get $S \cap \bigcup \sigma(G) = \emptyset$. With $\sigma(G) \neq \emptyset$ there is $T \in \sigma(G)$ and hence with $S \cup T \notin \sigma(G)$ no AF G can crash σ . \square

Regarding erratic behaviour of non-interference and crash-resistance we resume by letting go of well-definedness for the brief moment of the following example. Then, e.g. non-interference does not literally prevent interference anymore. Since we firmly believe that all reasonable semantics are fair, the main benefit of collapse over interference, contamination and crash though is a substantially less complicated characterization.

Example 2. Consider a semantics σ such that for some AFs $F, G, H = F \uplus G$ we have $\sigma(F) = \{S_i \mid i \in \mathbb{N}\}$, $\sigma(G) = \{T_i \mid i \in \mathbb{N}\}$ and $\sigma(H) = \{S_1 \cup T_i, T_1 \cup S_i \mid i \in \mathbb{N}\}$. For all we know σ might be basic, language-independent, non-interfering and not crashing. However it is not well-defined and shows strong preference for the extensions S_1 and T_1 .

²Traditionally crash-resistance is defined via contamination, which we consider redundant.

For the next section of this paper we will characterize AFs that do not collapse for some semantics. To this end we will make use of various framework or graph classes. The remainder of this section is dedicated to introducing those.

Definition 6. An AF F is called *finite* if $|A_F| < \infty$, it is called *infinite* if it is not finite. It is called *finitary* if each argument has only finitely many attackers.

Definition 7. Given some AF F . It is called

1. *bipartite* if there is partition $B \cap C = \emptyset$, $A_F = B \cup C$ such that for each $(x, y) \in R_F$ we have either $x \in B$ and $y \in C$ or $y \in B$ and $x \in C$;
2. *symmetric* if for any $(x, y) \in R_F$ also $(y, x) \in R_F$;
3. *loop-free* if there is no $a \in A_F$ such that $(a, a) \in R_F$;
4. *well-founded* if there exists no infinite sequence $a_0, a_1 \dots$ such that $(a_{i+1}, a_i) \in R_F$ for all i .

Fact 1. It is well known [1,7,8] that

1. for bipartite AFs semantics *pr*, *sg*, *ss*, *sb* coincide,
2. for symmetric AFs every *cf* and *ad* sets (and thus *na* and *pr*, *sg* and *ss* semantics) coincide,
3. for symmetric loop-free AFs *na*, *pr*, *sg*, *ss*, *sb* coincide,
4. for well-founded AFs *gr*, *co*, *na*, *pr*, *sg*, *ss*, *sb* coincide.

2. Perfection in Abstract Argumentation

This section is the name-giving section of this paper. We start by introducing the core definition.

Definition 8. Given some semantics σ an AF F is called σ -perfect if for any induced sub-AF F' ($F' = F|_X$ for some $X \subseteq A_F$) we have $\sigma(F) \neq \emptyset$.

The following theorem might be considered basic knowledge of abstract argumentation. The mere reason we provide proof is to highlight that Zorn's Lemma is not needed here after all.

Theorem 1. For $\sigma \in \{cf, ad, co, gr\}$ every AF is σ -perfect.

Proof. First the empty set always is conflict-free and admissible and is thus an extension for *cf* and *ad*. Further every AF has a grounded extension, e.g. constructed via characteristic function:³ starting with the empty set. At each induction step we select all arguments defended (and not attacked) by the before collected arguments. At limit steps we collect all arguments collected up to this limit step. For any AF F the (limited) set of arguments A_F witnesses that at some cardinality this procedure stops as eventually it will not be able to gather any more arguments. Finally since the grounded extension always is a complete extension every AF provides a complete extension. \square

³The characteristic function takes a set of arguments as input and gives all defended and not attacked arguments as output. It is used in [1] to characterize grounded semantics.

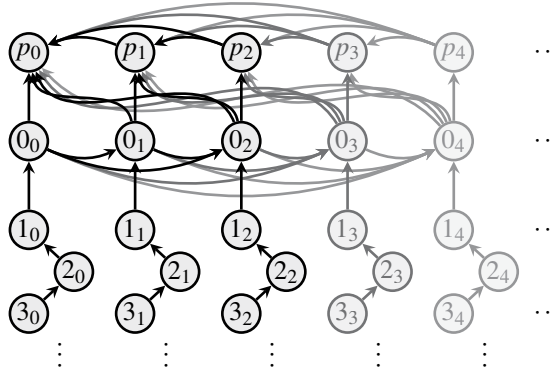


Figure 2. A cycle-free AF without stage, semi-stable or stable extensions, cf. Example 3.

Equivalence of existence of naive and preferred extensions to the Axiom of Choice is shown in [9]. For the remainder of this paper we assume ZFC and hence for instance Zorn's Lemma but do not discuss theoretical foundations thereof anymore.

Theorem 2. *Every AF is na-perfect and pr-perfect.*

We now focus on the remaining semantics *sg*, *ss* and *sb* and proceed by giving a cycle-free example of collapse.

Example 3. Consider the AF F as depicted in Figure 2. First observe that for the sequence of maximal admissible sets $S_i = \{0_i, 2_i, 4_i, \dots\} \cup \{1_j, 3_j, 5_j, \dots \mid j \neq i\}$ we have $S_i^+ \subset S_j^+$ for all $i < j$. Further observe that the p_i as well as the 0_i are pairwise in conflict and thus any conflict-free set S contains at most one of each, wlog. $p_i, 0_j \in S$. But now $S^+ \subset S_{\max(i,j)+1}^+$ and hence F collapses for *sg*, *ss* and *sb*.

It should be noted that the AF from Example 3 is cycle-free, which is why we do not overly discuss this graph-property in this paper. Now recall Fact 1 regarding basic AF classes and deduce the following.

Theorem 3. *For $\sigma \in \{sg, ss, sb\}$ the following hold:*

- bipartite AFs are σ -perfect,
- symmetric loop-free AFs are σ -perfect, and
- well-founded AFs are σ -perfect.

To see that neither symmetric nor loop-free AFs are σ -perfect on their own for $\sigma \in \{sg, ss, sb\}$ (and hence round out Theorem 3) we present the following two examples.

Example 4. Consider the symmetric AF F as illustrated in Figure 3(a). We have as only *pr* and *na* extensions $S = \{q_i \mid i \in \mathbb{N}\}$ and for $n \in \mathbb{N}$ the sets $S_n = (S \cup \{p_n\}) \setminus \{q_n\}$, where for $i < j$ we have $S^+ \subset S_i^+ \subset S_j^+$. So in effect for any *pr* or *na* extension there is another one of larger range and thus *sg*, *ss* and *sb* collapse.

Example 5. Consider the AF F as illustrated in Figure 3(b). The only preferred extensions are $S_q = \{q_i \mid i \in \mathbb{N}\}$ and for each $n \in \mathbb{N}$ the sets $S_n = \{q_i, p_n, s_j \mid i < n, j \geq n\}$. Here p_n

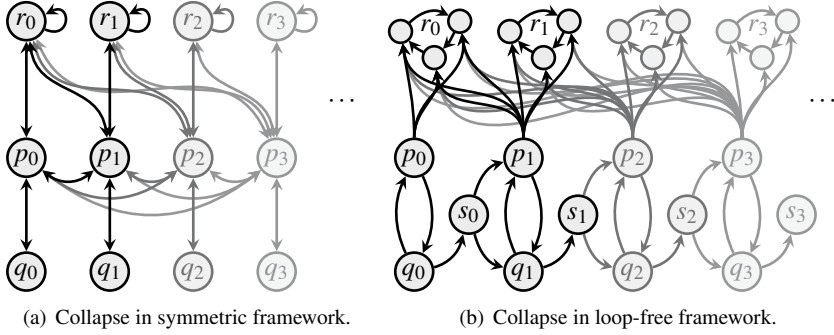


Figure 3. AFs without semi-stable or stage extensions, cf. Examples 4 and 5.

defends s_n , and accepting s_n for admissibility reasons means that we will accept each s_j for $j > n$. Again for $i < j$ we have $S_q^+ \subset S_i^+ \subset S_j^+$, and hence the collapse of semi-stable semantics. It can be shown that F collapses also for stage semantics, see [4] for a proof for a similar example.

As σ -perfection is inspired by kernel-perfection from graph theory and for any AF F the digraph $D = (A_F, \{(b, a) \mid (a, b) \in R_F\})$ has the set S as a kernel if and only if $S \in sb(F)$ we continue by importing the following two theorems.

Theorem 4 (Imported and transformed from [10]). *An AF F is sb-perfect if every induced sub-AF provides a non-empty admissible set. A finitary AF F is sb-perfect if and only if every finite induced sub-AF provides a sb extension.*

Theorem 5 (Imported and transformed from [11]). *Some given finite AF F is sb-perfect if every cycle of odd length is symmetrical.*

With this we close the case on stable semantics and move on to stage and semi-stable semantics. We start with the remark that *sb*-perfection of course implies *ss*- and *sg*-perfection and a last import.

Theorem 6 (Imported and adjusted from [12]). *Finitary AFs are sg- and ss-perfect.*

Upon our quest of searching for extensions of the given perfection-conditions for semi-stable semantics we might consider cases where the conditions are violated only marginally, for instance by one argument. The following example witnesses that this approach is of no help in the case of finitary planar⁴ loop-free AFs.

Example 6. Consider the AF $F = (A, R)$ as illustrated in Figure 4. Observe that only z_0 violates the finitary condition here and that this AF is planar and loop-free.

We have as only preferred extensions the set $S_x = \{\bar{z}_0\} \cup \{x_i \mid i \in \mathbb{N}\}$ and for each $n \in \mathbb{N}$ the sets $S_n = \{x_i, y_j, \bar{z}_j \mid j \leq n, i > n\}$. Again for $i < j$ we have $S_x^+ \subset S_i^+ \subset S_j^+$ and hence semi-stable semantics collapses. For stage semantics on the other hand, the set

⁴In this paper we do not give a formal definition of an AF being planar. Informally planar AFs can be sketched on a plane without crossing attack lines.

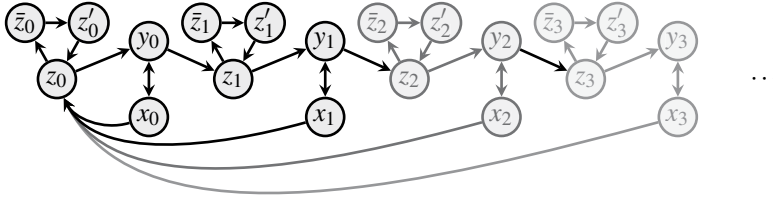


Figure 4. Loop-free planar AF with all but one finitary arguments and *ss*-collapse, cf. Example 6.

$S_y = \{y_i, \bar{z}_i \mid i \in \mathbb{N}\}$ is maximal in range, as only $z_0 \notin S_y^+$. But attacking z_0 means including z'_0 or x_j for some j and thus one of $\bar{z}_0, z_{j+1}, \bar{z}_{j+1}$ or z'_{j+1} drops out of range.

We now turn to stage semantics and start straightforward with a powerful result. We will then give example applications of this characterizing theorem.

Theorem 7 (Stage Perfection Characterization). *Given some AF $F = (A, R)$ where there is a finite set $Y \subseteq A$ such that the restriction $F|_{A \setminus Y}$ is *sg*-perfect. Then also F is *sg*-perfect.*

Proof. We use induction on the size of Y where the base case is given by assumption. We hence assume $Y = \{x\}$ as induction step. Observe that for every naive extension $S \in na(F)$ we can distinguish three cases:

1. $x \in S$ (x is a member of S),
2. $x \in S^+ \setminus S$ (S attacks x),
3. $x \in A \setminus S^+$ (due to maximality then however x attacks S).

For a contradiction assume $\sigma(F) = \emptyset$, yet for every proper induced sub-AF $F' = F|_{A \setminus Y}$ for $x \in Y \subseteq A$ we have $\sigma(F') \neq \emptyset$. This means that there is an unbounded range-chain $(S_i)_{i \in \mathbb{N}}$ of $S_i \in na(F)$ such that for $i < j$ we have $S_i^+ \subset S_j^+$. As this range-chain clearly can not be finite there is an infinite amount of S_i that can be filed under one and the same of above three cases. We proceed by considering each of these cases separately.

Case (1), wlog. $x \in S_i$ for all i : Then for each i we have $x^+ \subseteq S_i^+$ and hence $(S_i \setminus \{x\})_i$ is an unbounded naive range-chain of $F|_{A \setminus \{x, a, b \mid (a, x), (x, b) \in R\}}$ already.⁵

Case (2), wlog. S_i attacks x for all i : Then $x \in S_i^+$ for all i and hence $(S_i)_i$ is an unbounded range-chain of $F|_{A \setminus \{x\}}$ already.

Case (3), wlog. $x \notin S_i^+$: Then clearly x is also not member of the chain-range $\bigcup_{i \in \mathbb{N}} S_i^+$ and thus $(S_i)_i$ is an unbounded range-chain for $F|_{A \setminus \{x\}}$ already again. \square

The full power of Theorem 7 comes into play when considering classes of AFs we already know to be *sg*-perfect. We can immediately extend these classes and do so with the following corollaries. The first is dual to and thus proof of a conjecture from [4], i.e. *sg* collapses only if there are infinitely many arguments with infinitely many attackers. Recall that finitary AFs are *sg*-perfect.

Corollary 1. *AFs where most arguments have only finitely many attackers are *sg*-perfect.*

For the following recall that in symmetric AFs *cf* and *ad* and thus *sg* and *ss* coincide, and that symmetric loop-free AFs (see Theorem 3) are *sg*-perfect.

Corollary 2. *Symmetric AFs with finitely many self-attacking arguments are *sg/ss*-perfect.*

⁵In case of semi-stable this case is the reason the theorem fails, as $S_i \setminus \{x\}$ might not be admissible.

3. Discussion

In a way this paper is a collection of subtle details. In Section 1, Lemma 1 and Example 2 we critically discuss non-interference, contamination and crash-resistance. We proclaim (Definitions 3 and 4) well-definedness, fair semantics and collapse instead. In Section 2 we introduce and raise awareness for σ -perfection. Naturally such an intuitive property provides several results almost for free, or as corollaries from e.g. [1,7,8,9,12]. Still, especially for semi-stable and stage semantics we advance on known results and collapsing examples, proof a conjecture from [4] and elaborate on the surprisingly profound resistance of stage semantics against collapse (Theorem 7). With this we get by themselves already very powerful results (e.g. Corollaries 1 and 2) seemingly for free.

As obvious future research questions there are several other semantics out in the wild to be considered. Further results from graph theory on kernel-perfection can deliver additional immediate results for *sb*-perfection (and thus *ss*- and *sg*-perfection). It might also prove rather useful to consider classes of finitely generated infinite argumentation frameworks. Finally, also other syntactical AF-properties might be of interest in terms of σ -perfection. For instance, above results, the dynamics of chain-ranges and range-chains [12] and observations on the density of attacks in *sg*-collapsing AFs [4] let us propose this closing conjecture.

Conjecture 1. *Planar AFs are sg-perfect.*

References

- [1] Phan Minh Dung. On the Acceptability of Arguments and its Fundamental Role in Nonmonotonic Reasoning, Logic Programming and n-Person Games. *Artif. Intell.*, 77(2):321–358, 1995.
- [2] Hortensia Galeana-Sánchez and Victor Neumann-Lara. On kernels and semikernels of digraphs. *Discrete Mathematics*, 48(1):67–76, 1984.
- [3] Pietro Baroni, Martin Caminada, and Massimiliano Giacomin. An introduction to argumentation semantics. *Knowledge Eng. Review*, 26(4):365–410, 2011.
- [4] Christof Spanring. Hunt for the Collapse of Semantics in Infinite Abstract Argumentation Frameworks. In *ICCSW*, volume 49 of *OASICS*, pages 70–77. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, 2015.
- [5] Bart Verheij. DefLog: on the Logical Interpretation of Prima Facie Justified Assumptions. *J. Log. Comput.*, 13(3):319–346, 2003.
- [6] Martin Caminada, Walter A. Carnielli, and Paul E. Dunne. Semi-stable semantics. *J. Log. Comput.*, 22(5):1207–1254, 2012.
- [7] Sylvie Coste-Marquis, Caroline Devred, and Pierre Marquis. Symmetric Argumentation Frameworks. In *ECSQARU*, volume 3571 of *Lecture Notes in Computer Science*, pages 317–328. Springer, 2005.
- [8] Paul E. Dunne. Computational properties of argument systems satisfying graph-theoretic constraints. *Artif. Intell.*, 171(10-15):701–729, 2007.
- [9] Christof Spanring. Axiom of Choice, Maximal Independent Sets, Argumentation and Dialogue Games. In *ICCSW*, volume 43 of *OASICS*, pages 91–98. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, 2014.
- [10] Pierre Duchet and Henry Meyniel. Kernels in directed graphs: a poison game. *Discrete Mathematics*, 115(1-3):273–276, 1993.
- [11] Moses Richardson. On weakly ordered systems. *Bulletin of the American Mathematical Society*, 52(2):113–116, 1946.
- [12] Ringo Baumann and Christof Spanring. Infinite Argumentation Frameworks - On the Existence and Uniqueness of Extensions. In *Advances in Knowledge Representation, Logic Programming, and Abstract Argumentation*, volume 9060 of *Lecture Notes in Computer Science*, pages 281–295. Springer, 2015.

Gödel Fuzzy Argumentation Frameworks

Jiachao WU^{a,1}, Hengfei LI^b, Nir OREN^c, Timothy J. NORMAN^d

^a *Department of Mathematics, Shandong Normal University, China*

^b *School of Computer Science and Technology, Shandong Jianzhu University, China*

^c *Department of Computing Science, University of Aberdeen, UK*

^d *School of Electronics and Computer Science, University of Southampton, UK*

Abstract. In this paper we combine fuzzy set theory and argumentation to facilitate the use of fuzzy arguments and attacks. Unlike many existing approaches, our work does not require the use of any parameters, bringing it closer to Dung's work in spirit. We begin by introducing Fuzzy Argumentation Frameworks, and specialise them using the Gödel t-norm. We then examine this framework's properties and show that the standard Dung extensions are obtained, though the stable semantics coincide with the preferred. Finally, we examine the relationship between our framework and Dung's original system, as well as the existing fuzzy frameworks, describing where they overlap and differ.

Keywords. Fuzzy argumentation, abstract argumentation, semantics

1. Introduction

Following on from Dung's seminal paper [5], a variety of abstract argumentation frameworks have been proposed. These extensions to Dung's original work seek to identify a subset of arguments which is considered justified under a variety of inter-argument interactions, including support [2,12]; attacks which are joint [10] or recursive [3]; and preferences over arguments [1]. The properties assigned to arguments and argument interactions in such systems are typically binary (e.g., an attack is, or is not present), or qualitative (e.g., one argument is preferred to another). Such approaches can be contrasted with work on weighted argument frameworks [6], probabilistic argument frameworks [9,13] and multi-valued or fuzzy frameworks [4,11,7,8], where quantitative properties are considered.

Unlike qualitative approaches, which identify a justified set of arguments according to some semantics, quantitative approaches (with few exceptions) provide a justified set of arguments together with some additional information. For example, weighted argumentation frameworks determine justified arguments with respect to some inconsistency budget; probabilistic argumentation frameworks compute the likelihood that some set of arguments is justified, and fuzzy frame-

¹Corresponding Author: Dept. Mathematics, Shandong Normal University, Jinan 250014, China; E-mail: wujiachao1981@hotmail.com

works compute acceptability with regards to some parameter. Our goal within this paper is to more closely align fuzzy argumentation frameworks with classical argumentation approaches, unequivocally identifying a justified set of arguments with no reference to parameters. We do so by considering the notion of *sufficient attacks* and *weakening defends*, which we use to determine when one argument is sufficiently strong to defeat another. Before doing so, we first justify the importance of fuzziness in argumentation, contrasting it with uncertainty as found in probabilistic argumentation frameworks. Sections 3 and 4 provides our main contribution, formalising our framework and examining its properties. In Section 5 we compare our approach with existing work, before concluding.

2. Why Fuzziness in Argumentation?

To understand why fuzzy reasoning is necessary, we consider the following example [4] that considers whether a batch of tomatoes should be eaten:

B: The tomatoes are rotten.

C: The tomatoes can be eaten.

B attacks *C*: If a tomato is rotten, it should not be eaten.

Within a standard argumentation formalism, argument *B* would be justified: one could not conclude that the tomatoes can be eaten. Argument *B* may, however, be partially true — a tomato may have mold on one side, but the remaining half could be consumed. Similarly, some people may have differing judgments of how rotten the fruit is, and one should be able to aggregate these judgments to make a final decision. Probabilistic frameworks could, conceivably, capture the latter case, but would not help us in the former — such frameworks are designed to deal with uncertainty rather than fuzziness. When treated as fuzzy sets, arguments *B* and *C* could potentially both be considered justified — in situations where the tomatoes are only very slightly rotten, they can still be eaten.

Graduation or strength of this type is captured by associating a fuzziness degree to each argument. Different fuzziness degrees then result in different outcomes. For example, giving *B* a degree of 0.8 (i.e., most of the tomatoes are rotten), together with a belief² that most of the tomatoes can be eaten (e.g., associating 0.9 to *C*), it is clear that the two arguments are in conflict. On the other hand, giving *B* a degree of 0.1 while maintaining *C* at 0.9 should result in the two arguments being justified together. In the first instance, we may view the attack from *B* on *C* as *sufficient* to cause them to be judged in conflict, while in the second case, the attack can be tolerated by the system. We refer to such a situation as a *tolerable* attack.

Consider an additional argument and attack:

A: The tomatoes are stored well.

A attacks *B*: If tomatoes are stored well, they will not go rotten.

²We will utilise the term "degree of belief" interchangeably with "degree of fuzziness". Such degrees, rather than representing uncertainty, capture the belief in the level of fuzziness of the concept under consideration.

We may assign a degree of belief to the attack here such as 0.9, as we know (for example) that in most cases this relationship holds. Now if the tomatoes are stored well (e.g., assigning A a high fuzziness degree), then we should expect that most will be edible — a high degree of belief in A defends a high degree of belief in C by weakening the degree of belief in B . Similarly, a low degree of belief in A should weaken the defense it provides to C from an attack by B ³.

Within this paper we will formalise these concepts in order to construct a framework whose outputs are similar to standard argumentation frameworks: given a set of arguments, attacks between arguments, and an appropriate semantics, we will identify sets of justified arguments.

3. Fuzzy Argumentation

Our work builds on both fuzzy set theory [15] and abstract argumentation [5], in the spirit of de Costa Pereira et al. [11]⁴. We begin this section by providing an overview of fuzzy set theory and abstract argumentation.

3.1. Fuzzy set theory

Let X be a nonempty set. A fuzzy set (X, S) is determined by its membership function $S: X \rightarrow [0, 1]$, such that for each $x \in X$ the value $S(x)$ is interpreted as the grade of membership of x within X . Given some constant set X , we may denote a fuzzy set (X, S) as S for convenience.

A fuzzy set S is contained in another fuzzy set S' , if $\forall x \in X, S(x) \leq S'(x)$, which is denoted by $S \subseteq S'$.

The set $\{x \in X \mid S(x) > 0\}$ is called the *support* of (X, S) and the set $\{x \in X \mid S(x) = 1\}$ is called its *kernel*, or core.

A fuzzy set S is called a *fuzzy point* if its support is a single point $x \in X$, and is denoted by $(x, S(x))$. A fuzzy point $(x, S(x))$ is contained in a fuzzy set S if it is a subset of S .

3.2. Abstract argumentation frameworks

An abstract argumentation framework (AF) [5] contains a set of arguments and an attack relation:

Definition 1. An AF is a pair $(Args, \mathcal{R})$ where $Args$ is a set of arguments and $\mathcal{R} \subseteq Args \times Args$ is a set of attacks. An argument A attacks an argument B iff $(A, B) \in \mathcal{R}$.

Dung defines a number key concepts and various types of *extension* or ways to interpret an argument graph. In this paper, we build upon the following:

Defends A set $S \subseteq Args$ defends⁵ an argument $A \in Args$, if for every $B \in Args$ such that $(B, A) \in \mathcal{R}$, there is some $C \in S$ such that $(C, B) \in \mathcal{R}$.

³There are some clear similarities between this principle and reinstatement.

⁴A detailed comparison with this work is provided in Section 5.

⁵Dung introduced this concept as *acceptability* [5].

Conflict-free A set $S \in \text{Args}$ is conflict-free if there are no arguments $A, B \in S$ such that $(A, B) \in \mathcal{R}$.

Admissibility A conflict-free set S is admissible if it defends each argument in S .

Characteristic function The characteristic function of an AF $(\text{Args}, \mathcal{R})$ is a function $F: 2^{\text{Args}} \rightarrow 2^{\text{Args}}$, where $\forall S \subseteq \text{Args}$, $F(S) = \{A: S \text{ defends } A\}$.

Grounded extension The grounded extension is the least fixed point of F .

Complete extension A conflict-free set, S , is complete if $S = F(S)$.

Preferred extension A preferred extension is a maximal admissible set.

Stable extension A stable extension is a conflict-free set, S , that attacks each argument in $\text{Args} \setminus S$.

3.3. Fuzzy Argumentation Frameworks

Existing fuzzy argumentation models (such as [7,11]) consider either fuzzy arguments or fuzzy attacks between arguments. In this work we create a system with both fuzzy arguments and attacks. Furthermore, unlike work such as [7,9,6], our work takes an objective view to fuzzy extensions, not requiring a budget-like parameter to be specified. We begin by describing our approach, and then analyze its properties in Section 4. We refer to argumentation frameworks within our approach as *fuzzy argumentation frameworks*, abbreviated FAF.

Definition 2. (Fuzzy Argumentation Framework) A fuzzy argumentation framework is a tuple (\mathcal{A}, ρ) where $\mathcal{A}: \text{Args} \rightarrow [0, 1]$ and $\rho: \text{Args} \times \text{Args} \rightarrow [0, 1]$ are total functions. We refer to \mathcal{A} as a fuzzy set of arguments, and ρ as a fuzzy set of attacks, while Args is a set of crisp arguments.

A valid fuzzy argument can be encoded by the tuple (A, a) where $A \in \text{Args}$ and $a \in [0, 1]$, subject to the constraint that $a \leq \mathcal{A}(A)$. Similarly, a valid fuzzy attack can be written as $((A, B), \rho_{AB})$ if $\rho_{AB} \leq \rho(A, B)$.

It is important to differentiate between the value of a within (A, a) and $\mathcal{A}(A)$. The function \mathcal{A} identifies the *maximum* degree of belief associated with every argument that the system can permit. Therefore, any degree of belief smaller than this can also be accepted by the system. A similar argument applies to ρ . Given that our goal is to provide a means to select arguments from the FAF with some associated maximum degree (i.e., upper bound), any selected argument with a lesser degree of belief will also be accepted. It should also be noted that attacks within FAFs are between arguments; i.e., Args , rather than fuzzy arguments, \mathcal{A} . This is because attacks are determined by the relations between arguments, rather than the degree of belief in those arguments⁶.

Since ρ is a total function, we assume that if $\rho((A, B))$ is not specified, then $\rho((A, B)) = 0$. Returning to the rotten tomato example, where A = “The tomatoes are stored well”, and B = “The tomatoes are rotten”.

⁶Additionally, if the domain of ρ was $\mathcal{A} \times \mathcal{A}$, the system could be represented as a standard Dung argument system with infinite arguments of the form $\text{Args}' = \{(A, a) : A \in \text{Args}, a \in [0, 1]\}$, together with attacks between these arguments based on their different strengths.

Example 1. Assume that $(\mathcal{A}, \rho) = (\{(A, 0.7), (B, 0.8)\}, \{((A, B), 0.9)\})$. Here $A(A) = 0.7$, and we could accept that “The tomatoes are stored well” with a degree of belief 0.6 (that is, $(A, 0.6)$), but doing so with a degree of belief of 0.9 (i.e., $(A, 0.9)$) would be counter-intuitive.

To capture the above intuitions, we introduce the concepts of *sufficient attacks* and *weakening defends*.

3.4. Sufficient Attacks and Weakening Defends

Given Example 1, we may identify two types of attacks: tolerable and sufficient. A tolerable attack is one for which the target of the attack may be included within an extension without considering reinstatement (i.e., the attack is too “weak” to succeed in some sense), while a sufficient attack has sufficient strength to cause its target to be excluded from the extension. These terms are taken from Da Costa Pereira *et al.* [11], and they argue that these two types of attacks can be distinguished through the following principle:

“Suppose an argument A attacks an argument B . If we strongly believe A , then we hardly believe B , and if we strongly believe the negation of A , we should believe B strongly. Additionally, the belief of B should be no more than the belief of the negation of A .”

Formally, given a fuzzy argument (A, a) attacking another fuzzy argument (B, b) requires that the degree of belief b in B be no more than the value of $\neg(A, a)$ ⁷. One simple assignment for the strength of belief of $\neg(A, a) = 1 - a$, which therefore requires that $b \leq 1 - a$.

We can extend this idea to frameworks containing both fuzzy arguments and fuzzy attacks. Suppose that A attacks B with degree ρ_{AB} . Following Janssen *et al.* [7], the degree of belief associated with B given such an attack is based on the composition of the degree of belief in A (before considering B), together with the degree of belief placed in the attack itself. The degree of belief placed in B should, therefore, be no more than the negation of the composition of these two factors. In other words, if (A, a) (fuzzily) attacks another argument (B, b) with an attack of degree ρ_{AB} , then the following inequality must be satisfied.

$$b \leq 1 - a \star \rho_{AB} \quad (1)$$

Here, \star is a composition operator, and the question immediately arises as to what desirable properties are for such an operator.

Following [7], we believe it is reasonable for \star to satisfy the following conditions.

1. If, for some $(A, a) \in \mathcal{A}$ and $((A, B), \rho_{AB}) \in \rho$, $a = 1$ and $\rho_{AB} = x$, or $\rho_{AB} = 1$ and $a = x$, the value of the composition should be x ; i.e., $x \star 1 = 1 \star x = x$;

⁷In an ASPIC-like system, one could interpret this as the degree of belief b being no more than the contrary of (A, a) .

2. If, for some $(A, a) \in \mathcal{A}$ and $((A, B), \rho_{AB}) \in \rho$, $a = 0$ or $\rho_{AB} = 0$ (A is selected out or ρ_{AB} disappears), then the composition should be 0; and
3. Operator \star should be monotone on both sides.

These conditions mean that \star is a non-commutative t-norm; the simplest such operator is the Gödel t-norm: $a \star \rho_{AB} = \min\{a, \rho_{AB}\}$ ⁸. Substituting this operator into Equation 1 yields:

$$\min\{a, \rho_{AB}\} + b \leq 1 \quad (2)$$

We refer to a fuzzy argumentation framework using the Gödel t-norm as a Gödel Fuzzy Argumentation Framework (or GFAPF).

Note that if $\rho_{AB} = 1$, Equation 2 reduces to $a + b \leq 1$, which is that used by the system of Da Costa Pereira *et al.* Furthermore, if the degree of belief in all attacks is 0 or 1, the model reduces to one in which all attacks are crisp. In such cases, our system is consistent with Da Costa Pereira *et al.*'s method for distinguishing between tolerable and sufficient attacks.

We are now in a position to formalise tolerable and sufficient attacks for GFAPFs.

Definition 3. *Given two arguments, (A, a) and (B, b) as well as an attack $((A, B), \rho_{AB})$, if Equation 2 is satisfied, then the attack is tolerable, otherwise it is sufficient.*

Example 2. *Returning to the rotten tomato example, assume that $(A, 0.1)$, $(B, 0.8)$ and $((A, B), 0.9)$. In this situation, the attack is tolerable. However, if instead we have that $(A, 0.7)$, then the attack becomes sufficient.*

As mentioned above, a tolerable attack has no influence on (B, b) . However, a sufficient attack *weakens* the attacked argument.

Definition 4. *Given an attack $((A, B), \rho_{AB})$ from (A, a) to (B, b) within a GFAPF (\mathcal{A}, ρ) , (A, a) weakens (B, b) to (B, b') by the attack $((A, B), \rho_{AB})$, thus:*

$$b' = \min\{1 - \min\{a, \rho_{AB}\}, b\}$$

Note that this definition captures both tolerable and sufficient attacks, with the latter resulting in $b' = b$.

Example 3. *Returning to Example 2, if the degree of belief of A is 0.1, $(B, 0.8)$ is weakened to $(B, 0.8)$ by the attack $((A, B), 0.9)$. However, if we have $(A, 0.7)$, $(B, 0.8)$ is weakened to $(B, 0.3)$ by this attack. Given the attack $((A, B), 0.6)$, $(A, 0.7)$ weakens $(B, 0.8)$ to $(B, 0.4)$.*

⁸We concentrate on the Gödel t-norm in this work, but other operators, such as the product t-norm could also be utilized; an investigation of the properties of such operators is an avenue for future research.

For convenience, we may say that (A, a) weakens (B, b) to (B, b') without referring to the degree of belief in the attack. In this case, we mean that b' is minimal — A weakens B by the maximal value of the argument A (i.e., a) or the attack (ρ_{AB}) .

Since tolerable attacks do not change the degree of belief in the attacked argument, such attacks are ignored when computing a conflict-free set of fuzzy arguments.

Definition 5. *Given a GFAP (\mathcal{A}, ρ) , a fuzzy set of arguments $S \subseteq \mathcal{A}$ is conflict-free (abbreviated Cf) if all attacks between the fuzzy arguments in S are tolerable.*

The conflict-freeness of a set is, therefore, determined by considering the maximum degree of belief of the fuzzy attacks between arguments.

Example 4. *Consider the GFAP $(\{(A, 0.7), (B, 0.8)\}, ((A, B), 0.9))$. Both fuzzy sets $\{(A, 0.7), (B, 0.3)\}$ and $\{(A, 0.4), (B, 0.5)\}$ are conflict-free. In contrast, neither $\{(A, 1), (B, 0)\}$ nor $\{(A, 0.7), (B, 0.8)\}$ are conflict-free. This is because $\{(A, 1), (B, 0)\}$ is not a fuzzy subset of \mathcal{A} , and the attack $((A, B), 0.9)$ with $\{(A, 0.7), (B, 0.8)\}$ is sufficient.*

Having defined tolerable and sufficient attacks and introduced the concept of weakening, we are now in a position to define how a fuzzy set provides a weakening defense of a fuzzy argument.

Definition 6. *Given a GFAP (\mathcal{A}, ρ) , a fuzzy set $S \subset \mathcal{A}$ weakening defends a fuzzy argument $(C, c) \in \mathcal{A}$ if for any $(B, b) \in \mathcal{A}$ there is some $(A, a) \in S$ such that (A, a) weakens (B, b) to (B, b') and (B, b') tolerably attacks (C, c) .*

Theorem 1. *Given a GFAP (\mathcal{A}, ρ) , a set $S \subset \mathcal{A}$ weakening defends $(C, c) \in \mathcal{A}$, iff $\forall (B, b) \in \mathcal{A}$,*

$$\min_{A \in \text{Args}} \{1 - \min\{S(A), \rho((A, B))\}, b, \rho((B, C))\} + c \leq 1. \quad (3)$$

Proof. (\Leftarrow) Suppose Equation 3 is satisfied. For a finite set Args , there will be some $A \in \text{Args}$ such that

$$\min\{\min\{1 - \min\{S(A), \rho((A, B))\}, b\}, \rho((B, C))\} + c \leq 1,$$

which means $(A, S(A)) \in S$ weakens (B, b) to (B, b') , where, by Definition 6, $b' = \min\{1 - \min\{S(A), \rho((A, B))\}, b\}$, and (B, b') does not sufficiently attack (C, c) ; i.e. S weakening defends (C, c) .

(\Rightarrow) Suppose S weakening defends (C, c) . Then, for any $(B, b) \in \mathcal{A}$, there is some $(A, a) \in S$ such that (A, a) weakens (B, b) to (B, b') , and (B, b') tolerably attacks (C, c) ; i.e., $\min\{\min\{1 - \min\{a, \rho((A, B))\}, b\}, \rho((B, C))\} + c \leq 1$. Because $a \leq S(A)$, we have

$$\min\{\min\{1 - \min\{S(A), \rho((A, B))\}, b\}, \rho((B, C))\} + c \leq 1,$$

which immediately reduces to Equation 3. \square

Example 5. Suppose we have a GFAF $(\{(A, 0.7), (B, 0.8), (C, 0.9)\}, \{((A, B), 0.9), ((B, C), 0.7)\})$. The fuzzy argument $(A, 0.6)$ weakening defends $(C, 0.6)$, but $(A, 0.8)$ does not weakening defend $(C, 0.9)$.

Using Theorem 1, we may extend Definition 6 to utilise sets.

Definition 7. Suppose $S \subset \mathcal{A}$ and $(B, b) \in \mathcal{A}$ for GFAF (\mathcal{A}, ρ) . The set S weakens (B, b) to (B, b') , such that

$$b' = \min_{A \in \text{Args}} \{1 - \min\{S(A), \rho(A, B)\}\}$$

In other words, (B, b) is weakened by every argument in S , and the minimum value due to attacks from S is b' .

The following proposition follows naturally from this definition.

Lemma 1. If S weakens (A, a) to (A, a') , then S tolerably attacks (A, a') .

4. Semantics of GFAFs

In this section we define various argumentation semantics within GFAFs, namely the grounded, complete, preferred and stable extensions. Following this, we examine the relationships between them. In defining these semantics, we utilise the concepts of an admissible set and the characteristic function of a GFAF.

Definition 8. A conflict-free set of fuzzy arguments, $S \in \mathcal{A}$, in a GFAF (\mathcal{A}, ρ) is admissible (abbreviated AE), if S weakening defends each element in S .

Example 6. Consider again our fuzzy argumentation framework with arguments A , B and C : $(\{(A, 0.7), (B, 0.8), (C, 0.9)\}, \{((A, B), 0.9), ((B, C), 0.7)\})$. Here, both $\{(A, 0.6), (B, 0.3), (C, 0.6)\}$ and $\{(A, 0), (B, 0), (C, 0)\}$ (the empty set) are admissible sets of fuzzy arguments. In contrast, $\{(A, 0.4), (B, 0.2), (C, 0.6)\}$ is not an admissible set, because $(A, 0.4)$ is not strong enough to defend $(C, 0.6)$; i.e. $(A, 0.4)$ can only weaken $(B, 0.8)$ to $(B, 0.6)$, which still sufficiently attacks $(C, 0.6)$. Similarly, $\{(A, 0.4), (B, 0.4), (C, 0.4)\}$ is not an admissible set, because $(B, 0.4)$ is sufficiently attacked by $(A, 0.7)$, which is not weakened by any other fuzzy argument in this set.

Definition 9. The characteristic function of a GFAF (\mathcal{A}, ρ) is a function \mathcal{F} from the set of all the subsets of \mathcal{A} to itself, such that $\forall S \subseteq \mathcal{A}$, $\mathcal{F}(S) = \{(A, a) : S \text{ weakening defends } (A, a)\}$.

From this definition, \mathcal{F} is monotonic with respect to set inclusion; i.e., if $S_1 \subset S_2$, then $\mathcal{F}(S_1) \subset \mathcal{F}(S_2)$.

Given our formulation of fuzzy argumentation frameworks and the definitions presented, the definitions of different semantics for FAFs follow those for Dung argumentation frameworks.

Definition 10. The grounded extension (GE) is the least fixed point of the characteristic function \mathcal{F} .

Definition 11. A conflict-free set S is a complete extension (CE) if it contains all the fuzzy arguments in \mathcal{A} that S weakening defends; i.e., $\mathcal{F}(CE) = CE$.

Example 7. Consider the GFAF

$$(\{(A, 0.7), (B, 0.8), (C, 0.9)\}, \{((A, B), 0.9), ((B, C), 0.7)\})$$

The sets of fuzzy arguments $\{(A, 0.6), (B, 0.3), (C, 0.6)\}$ and $\{(A, 0), (B, 0), (C, 0)\}$ are both admissible, but neither is complete. The reason for this is that the empty set defends $(A, 0.7)$, which is not within either set. In this case, there is a single complete extension: $\{(A, 0.7), (B, 0.3), (C, 0.7)\}$.

Definition 12. An admissible extension is a preferred extension (PE) if it is maximal.

A preferred extension E is a maximal self-defended conflict-free set of fuzzy arguments. Unlike Dung-like systems, conflict here arises due to changes in the degree of belief placed in arguments, rather than simply from the presence of arguments.

Example 8. Consider the GFAF $(\{(A, 0.7), (B, 0.8), (C, 0.9), (D, 0.7)\}, \{((A, B), 0.9), ((B, C), 0.7), ((C, D), 0.8), ((D, C), 0.8)\})$. Here, $\{(A, 0.7), (B, 0.3), (C, 0.5), (D, 0.4)\}$ is complete but not preferred, since both the complete extension $\{(A, 0.7), (B, 0.3), (C, 0.5), (D, 0.5)\}$ and $\{(A, 0.7), (B, 0.3), (C, 0.6), (D, 0.4)\}$, which are preferred, strictly contains it.

Definition 13. A conflict-free extension E is stable (abbreviated SE) if it sufficiently attacks every elements in \mathcal{A} not in E .

The stable extensions E is both the maximal conflict-free set and the minimal set that can attack all other arguments.

Example 9. Suppose a GFAF is given as $(\{(A, 1)\}, \{((A, A), 1)\})$. Then the extension $\{(A, 0.5)\}$ is preferred and stable.

Next, we consider the relationship between the different extensions. These relationships are identical to those found in Dung frameworks, with the exception that an extension is preferred if and only if it is stable (i.e., preferred and stable extensions coincide).

Theorem 2.

$$PE = SE \Rightarrow CE \Rightarrow AE \Rightarrow Cf, \quad GE \Rightarrow CE.$$

The converse is not valid.

Proof. (sketch) The examples above show that the converse of the implications are invalid.

$AE \Rightarrow Cf$ and $GE \Rightarrow CE$ are trivially valid.

$CE \Rightarrow AE$: From the definition of the characteristic function, CE weakening defends each element in CE . Thus, it is admissible.

$PE \Rightarrow CE$: Consider $\mathcal{F}(PE)$. From the definition of \mathcal{F} , $\mathcal{F}(PE)$ contains all the fuzzy arguments that are weakening defended by PE . Since PE is admissible, $PE \subset \mathcal{F}(PE)$, that is $\mathcal{F}(PE)$ is also admissible. Additionally, because PE is maximal in all the admissible extensions, $\mathcal{F}(PE)$ is not a superset of PE ; i.e., $PE = \mathcal{F}(PE)$. Thus, PE is complete.

$SE \Rightarrow PE$: Consider an argument (A, a) which sufficiently attacks SE . Such an argument is not in SE , as SE is conflict-free. By Lemma 1, SE weakens (A, a) to (A, a') , such that (A, a') is not sufficiently attacked by SE , which means (A, a') does not sufficiently attack SE . Thus, SE weakening defends any argument in SE ; i.e., SE is admissible.

Obviously, SE is maximal. Therefore, SE is preferred.

$PE \Rightarrow SE$: Suppose PE weakens every argument (A, a) , which is sufficiently attacked by PE , to $\mathcal{A}'(A)$, i.e.

$$\mathcal{A}'(A) = \min_{B \in \text{Args}} \{1 - \min\{PE(B), \rho(B, A)\}, \mathcal{A}(A)\}.$$

Obviously, any elements not in \mathcal{A}' , are sufficiently attacked by PE .

Additionally, it is not difficult to show that \mathcal{A}' is just $\mathcal{F}(PE)$, with \mathcal{F} the characteristic function, by Definitions 6 and 9. For PE is CE , we have $\mathcal{A}' = PE$. This means PE sufficiently attacks all the other arguments not in PE ; i.e., PE is stable. \square

5. Discussion

In this section we explore the relationships between GFAs and Dung argument systems, between GFAs and the system of Da Costa Pereira *et al.* [11], and between GFAs and Janssen's approach *et al.* [7]. We note that a variety of other fuzzy approaches to argumentation have been proposed (e.g., [14], which considers a fuzzy instantiated argumentation system), but these are not closely related to our work, and therefore omitted due to space constraints. Additionally, we state only our main results, with proofs provided in a technical report⁹.

5.1. Dung Argument Frameworks

A DAF can be viewed as a *crisp* GFAF where the degree of belief of arguments and attacks is 1 or 0 (1 for those attacks present in the DAF, and 0 for absent attacks).

Theorem 3. *A conflict-free; stable; or admissible extension within a DAF is also a conflict-free; stable; or admissible extension within the corresponding crisp GFAF.*

⁹<http://homepages.abdn.ac.uk/n.oren/pages/CS2016-01.pdf>

As mentioned above, the stable and preferred extensions coincide within GFAFs. A preferred, but non-stable extension within a DAF will not form a preferred extension within a GFAF.

5.2. Da Costa Pereira et al.

The model proposed by Da Costa Pereira *et al.* [11] can be seen as a FAF with crisp attacks, which utilises the following (convergent) function to provide a fuzzy labelling for arguments [11, Definition 12, page 5]:

$$\alpha_{t+1}(A) = \frac{1}{2}\alpha_t(A) + \frac{1}{2} \min\{\mathcal{A}(A), 1 - \max_{B: B \rightarrow A} \alpha_t(B)\}, \quad (4)$$

Theorem 4. *Given a fuzzy set E , defined as $E(A) = \alpha(A), \forall A \in \text{Args}$, E is a preferred/stable extension of the GFAF obtained using crisp attacks and fuzzy arguments with degree of belief obtained using α .*

5.3. Janssen et al.

We consider a restricted form of Janssen et al.'s framework [7] (referred to as JAFs) where: the truth lattice \mathcal{L} is binary; the tnorm \wedge is the Gödel t-norm, meaning that the implication can be defined by the residual (i.e., for any $a, b, c \in [0, 1]$, $a \leq b \rightsquigarrow c$ iff $\min\{a, b\} \leq c$); and $\neg a = 1 - a$.

Even with these restrictions, the argumentation frameworks differ. In our work, a GFAF contains a given fuzzy subset of *Args* with some associated upper bound on degree of belief of the arguments, and a fuzzy set of attacks between arguments. Within a JAF, the sets of arguments are crisp, and the extensions are fuzzy subsets of a crisp set. Attacks in GFAFs are based on Dung's notion of attack, while in a JAF, they are from a fuzzy set of arguments to an argument or another fuzzy set of arguments. Given this, basic concepts such as conflict-freeness differ between JAFs and GFAFs, meaning that extensions typically also differ. However, JAFs and GFAFs coincide in the following situation (based on Proposition 2 of [7]).

Theorem 5. *Let (Args, ρ) be a JAF and a GFAF, with $\mathcal{A} = \text{Args}$. A fuzzy set S is stable in GFAF, iff it is conflict-free in GFAF and 1-stable in JAF using the Gödel t-norm.*

6. Conclusions

In this paper we introduced Gödel Fuzzy Argumentation Frameworks, for which extensions are defined in a manner that is consistent with those defined for Dung's abstract argument frameworks, while utilising fuzzy arguments and attacks. Using the notions of sufficient attacks and weakening defends enables us to rigorously model reasoning over arguments and attacks that have degrees of belief associated with them. When restricted to crisp arguments and attacks, the extensions

obtained are similar to those of a Dung system. The main thrust of our future research is to further investigate the properties of GFAs, such as in the context of the semi-stable semantics, and to examine the properties of other t-norms.

Acknowledgements

This work is supported by the Excellent Young Scholars Research Fund of Shandong Normal University. This research was sponsored by the U.S. Army Research Laboratory and the U.K. Ministry of Defence and was accomplished under Agreement Number W911NF-06-3-0001.

References

- [1] L. Amgoud and C. Cayrol. A reasoning model based on the production of acceptable arguments. *Annals of Mathematics and Artificial Intelligence*, 34(1-3):197–215, 2002.
- [2] L. Amgoud, C. Cayrol, and M.-C. Lagasque-Schiek. On the bipolarity in argumentation frameworks. In *Proceedings of the 10th International Workshop on Non-monotonic Reasoning*, pages 1–9, 2004.
- [3] P. Baroni, F. Cerutti, M. Giacomin, and G. Guida. AFRA: Argumentation framework with recursive attacks. *International Journal of Approximate Reasoning*, 55(1):19–37, 2011.
- [4] P. Dondio. Multi-value and probabilistic argumentation frameworks. In *Proceedings of the 5th International Conference on Computational Models of Argument*, pages 253–260, 2014.
- [5] P. M. Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77:321–357, 1995.
- [6] P. Dunne, A. Hunter, P. McBurney, S. Parsons, and M. Wooldridge. Weighted argument systems: Basic definitions, algorithms, and complexity results. *Artificial Intelligence*, 175:457–486, 2011.
- [7] J. Janssen, M. De Cock, and D. Vermeir. Fuzzy argumentation frameworks. In *Proceedings of the 12th Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 513–520, 2008.
- [8] S. Kaci and C. Labreuche. Argumentation framework with fuzzy preference relations. In *Computational Intelligence for Knowledge-Based Systems Design*, pages 554–563. Springer, 2010.
- [9] H. Li, N. Oren, and T. J. Norman. Probabilistic argumentation frameworks. In *Theories and Applications of Formal Argumentation*, pages 1–16. Springer, 2012.
- [10] S. H. Nielsen and S. Parsons. A generalization of Dung’s abstract framework for argumentation: Arguing with sets of attacking arguments. In *Proceedings of the Third International Workshop on Argumentation in Multi-Agent Systems*, pages 7–19, 2006.
- [11] C. Da Costa Pereira, A. Tettamanzi, and S. Villata. Changing ones mind: Erase or rewind? In *Proceedings of the 22nd International Joint Conference Artificial Intelligence*, pages 164–171, 2011.
- [12] S. Polberg and N. Oren. Revisiting support in abstract argumentation systems. In *Proceedings of the 5th International Conference on Computational Models of Argument*, pages 369–376, 2014.
- [13] T. Rienstra. Towards a probabilistic Dung-style argumentation system. In *Proceedings of the 1st International Conference on Agreement Technologies*, pages 138–152, 2012.
- [14] Nouredine Tamani and Madalina Croitoru. *Fuzzy Argumentation System for Decision Support*, pages 77–86. Springer International Publishing, Cham, 2014.
- [15] L.A. Zadeh. Fuzzy sets. *Information and Control*, 8:338–353, 1965.

Demonstrations

This page intentionally left blank

DALEK: A Tool for Dialectical Explanations in Inconsistent Knowledge Bases

Abdallah ARIOUA^a, Madalina CROITORU^b, Patrice BUCHE^a

^a IATE, INRA, INRIA Graphik/LIRMM University Montpellier 2, France

^b INRIA Graphik/LIRMM University Montpellier 2, France

Abstract. In this paper we present a prototype of a framework called DALEK (**Di**ALectical **Ex**planation in **K**nowledge-bases). This framework implements dialectical approaches to explain query answers in inconsistent knowledge bases. The motivation behind the prototype is as follows: given an inconsistent knowledge base represented within *Datalog \pm* , a semantics for handling inconsistency and a query Q , the goal is to explain why Q is accepted or not accepted under such semantics. The explanation takes a dialogical form (cf. [1,3]).

Keywords. Applications of Argumentation, Explanation and Argumentation Dialogues, *Datalog \pm* .

1. DALEK Framework: Explain!

DALEK engages a **User** and the **Reasoner** in a dialogue about the entailment of any boolean conjunctive query in *Datalog \pm* knowledge bases. The dialogue could be of argumentative or explanatory nature. In DALEK the **User** can shift between dialogue types (i.e. dialectical shifts). The framework is general enough to carry out a standalone argumentation dialogue as well as a standalone explanatory dialogue. DALEK also implements commitments and understanding stores.¹

When the **User** interacts with the GUI, the latter communicates with *the dialogue manager* which possesses the *configuration structure* and the *stores*. Then, the dialogue manager, at its turn, communicates with *the semantics structure* through the sub-module “Syntax and semantics handler” and with *the dialogue planner* through the sub-module “Utterance dispatcher”. Next, *the dialogue planner* and *the semantics structure* communicate directly with *the logical model* that uses the *Datalog \pm* GRAAL library [2] to query *the knowledge base*. Hereafter we detail each module of Figure 1.

Configuration structure. This module specifies: (1) the set of allowed locutions with their legal replies, (2) the parameters of the protocol, e.g. unique-move, multiple-move, the participants, etc. and (3) the parameters of the planner.

Dialogue manager. This is the referee between the **User** and the **Reasoner** (i.e. dialogue planner), it dispatches their utterances through the sub-module “Utter-

¹See <http://www.lirmm.fr/~arioua/dkb/#rulesdalek> for more details.

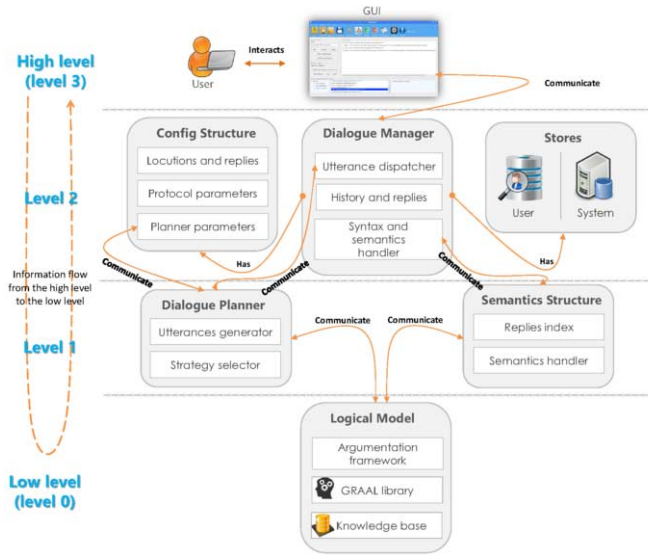


Figure 1. The DALEK's architecture. Each layer is composed of modules and each module is composed of sub-modules.

ance dispatcher” after ensuring their legality. To verify the legality the dialogue manager communicates with the module semantics structure through the sub-module “Syntax and semantics handler” that makes use of the stores. The syntactical verification ensures the legality of any advanced utterance with respect to : (1) legality of the utterance itself, and (2) legality of the reply within the dialogue. The semantics verification ensures, among other things, the legality of the utterances with respect to the content. It checks whether the advanced utterance holds a legal content and it replies with a legal content.

Semantics structure. This structure implements an *operational semantics* of the dialogue. It associates with each reply a procedure that should be called by the dialogue manager to check the legality of the reply.

Dialogue planner. This module receives the utterances from the **User** through the dialogue manager and plans the next utterance. The planner in its current state tries to answer **User**’s utterances as they come.

2. Acknowledgments

The authors acknowledge the support of QUALINCA (ANR-12-0012) and DUR-DUR (ANR-13-ALID-0002) grants.

References

- [1] A. Arioua and M. Croitoru. Formalizing explanatory dialogues. In *Proceedings of SUM’15*, pages 282–297. Springer, 2015.
- [2] J.-F. Baget, M. Leclère, M.-L. Mugnier, S. Rocher, and C. Sipieter. Graal: A toolkit for query answering with existential rules. In *Proceedings of RuleML’15*, pages 328–344. Springer, 2015.
- [3] H. Prakken. Coherence and flexibility in dialogue games for argumentation. *Journal of logic and computation*, 15(6):1009–1040, 2005.

ConArg: A Tool for Classical and Weighted Argumentation

STEFANO BISTARELLI^a, FABIO ROSSI^a, FRANCESCO SANTINI^a

^a*Dipartimento di Matematica e Informatica, University of Perugia, Italy*
[bista,rossi,francesco.santini]@dmi.unipg.it

Abstract. ConArg is a tool for solving different problems related to extension-based semantics: e.g., enumeration of extensions, sceptical and credulous acceptance of arguments. We have extended it in order to deal with *Weighted Abstract Argumentation Frameworks*, where each attack is associated with a strength score. Classical notions of defence and conflict-freeness have been redefined with the purpose to have different (weighted) degrees of their relaxation. The ultimate aim is to let an agent choose between a higher internal consistency or a stronger defence.

Keywords. Abstract Argumentation Frameworks, Extension-based semantics, Weighted attacks, Weighted defence, Inconsistency tolerance.

ConArg¹ [5,2] is an Argumentation-related reasoner based on *Gecode*², which is an open, free, and efficient C++ library where to develop constraint-based applications. ConArg is able to find all the classical extensions on a given *Abstract Argumentation Framework* (AAF) [7]: conflict-free, admissible, complete, stable, grounded, preferred, semi-stable, and ideal extensions. In addition, it can check the credulous or sceptical acceptance of a given argument. The tool is offered to users as a stand-alone command-line executable, or through a Web-interface that can be found at the official site of ConArg.

Besides classical unweighted problems [7], ConArg has been extended to also deal with *Weighted Abstract Argumentation Frameworks* (WAAFs) [4]. This is accomplished *i)* by allowing an internal conflict *inside* the extensions satisfying a given semantics, and *ii)* by relaxing defence taking into account the difference between the two weights of attacks (aggregated per attacker) and defence. Hence, two parameters influence new semantics: α is the amount of internal conflict that can be tolerated, while γ represents how much defence can be relaxed. The result is the definition of α^γ -semantics (e.g., α^γ -admissible). The strictest (not relaxed) level of defence corresponds to w -defence [3]: an extension $\mathcal{B} \subseteq \mathcal{A}_{rgs}$ defends an argument $b \in \mathcal{A}_{rgs}$ from $a \in \mathcal{A}_{rgs}$, if the sum of all the attack weights from \mathcal{B} to a is stronger than the sum of all the attacks from a to $\mathcal{B} \cup \{b\}$.

For instance, looking at Fig. 1, \mathcal{B} is w -defended (or 0-defended) from the attacks of a ($2 + 6 = 8 \geq 7$), while \mathcal{B} is not w -defended from f ($5 + 2 = 7 < 8 = 5 + 3$): \mathcal{B} is only 1-defended (i.e., the difference between attack and defence, $8 - 7 = 1$). \mathcal{B} is 2-conflict-free, since it encompasses two attacks with weight 1 each (between b and e , and e and c). To summarise, \mathcal{B} is 2^1 -admissible ($\alpha = 2$ and $\gamma = 1$): we tolerate an internal conflict of 2, and that the defence is weaker (by 1) than the aggregated weight of attacks (from f).

¹<http://www.dmi.unipg.it/conarg/>.

²<http://www.gecode.org>.

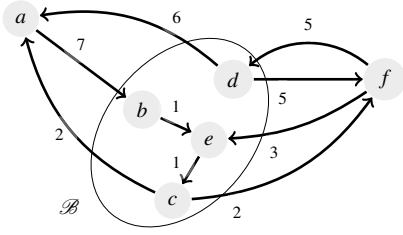


Figure 1. \mathcal{B} is w -defended from a , but only 1-defended from f . \mathcal{B} is also 2-conflict-free.

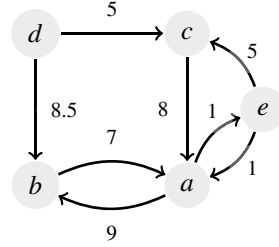


Figure 2. $\{a, d\}$ is 0^3 -admissible, $\{a, d, e\}$ is 2^0 -admissible.

ConArg can import (W)AAFs with a format as, e.g., $arg(a)$, $arg(b)$, $att(a, b)$. If $att(a, b) : -6$, then it means that the attack from a to b is associated with a weight of 6.

Parameters α and γ mutually influence each other: allowing a small conflict may lead to have one more argument inside an extension, which consequently may be more strongly defended by exploiting the attacks of this additional argument, or more weakly, in case such additional argument receives attacks from external arguments. Figure 2 is presented to show how internal and defence relaxations are strictly linked together: the set $\{a, d\}$ is 0^3 -admissible, since a is attacked by c with weight of 8, but only a counter-attack with weight 5 is present from d to c (hence, the difference to be tolerated is $8 - 5 = 3$). However, if an internal inconsistency of 2 can be tolerated (inconsistency is ubiquitous in every-day life [1]), the set $\{a, d, e\}$ is 2^0 -admissible: by allowing a small internal conflict, the defence against b and c becomes stronger (no defence-relaxation is needed to defend them). Therefore, we provide a means to an agent to decide between $\{a, d\}$ or $\{a, d, e\}$, satisfying either the first (with a higher internal consistency) or the second semantics (with a stronger defence).

In the future we will study two-criteria (α and γ) decision-making procedures to help an agent choose between internal or defence relaxations (as in the example in Fig. 2). We will also extend weighted relaxations to coalitions of arguments [6].

References

- [1] L. E. Bertossi, A. Hunter, and T. Schaub, editors. *Inconsistency Tolerance [result from a Dagstuhl seminar]*, volume 3300 of *Lecture Notes in Computer Science*. Springer, 2005.
- [2] S. Bistarelli, F. Rossi, and F. Santini. Benchmarking hard problems in random abstract afs: The stable semantics. In *Computational Models of Argument - Proceedings of COMMA 2014*, volume 266, pages 153–160. IOS Press, 2014.
- [3] S. Bistarelli, F. Rossi, and F. Santini. A collective defence against grouped attacks for weighted abstract argumentation frameworks. In *Proceedings of the Twenty-Ninth International Florida Artificial Intelligence Research Society Conference, FLAIRS 2016*, pages 638–643. AAAI Press, 2016.
- [4] S. Bistarelli and F. Santini. A common computational framework for semiring-based argumentation systems. In Helder Coelho, Rudi Studer, and Michael Wooldridge, editors, *ECAI*, volume 215 of *Frontiers in Artificial Intelligence and Applications*, pages 131–136. IOS Press, 2010.
- [5] S. Bistarelli and F. Santini. Conarg: A constraint-based computational framework for argumentation systems. In *IEEE 23rd International Conference on Tools with Artificial Intelligence, ICTAI 2011*, pages 605–612. IEEE, 2011.
- [6] S. Bistarelli and F. Santini. Coalitions of arguments: An approach with constraint programming. *Fundam. Inform.*, 124(4):383–401, 2013.
- [7] P. M. Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77(2):321–357, 1995.

Efficient and Off-The-Shelf Solver: jArgSemSAT

Federico CERUTTI^a, Mauro VALLATI^b, Massimiliano GIACOMIN^c

^aCardiff University, School of Computer Science & Informatics, UK

^bSchool of Computing and Engineering, University of Huddersfield, UK

^cDepartment of Information Engineering, University of Brescia, Italy

Abstract. jArgSemSAT is a Java re-implementation of ArgSemSAT—a SAT-based solver for abstract argumentation problems—that can be easily integrated in existing argumentation systems (1) as an off-the-shelf, standalone, library; (2) as a Tweety compatible library; and (3) as a fast and robust web service freely available on the Web. Despite being written in Java, jArgSemSAT is very efficient.

Keywords. Dung’s *AF*, semantics, solver

Introduction

Dung’s argumentation framework (*AF*) consists of a set of arguments and an *attack* relation between them [3]. Different *argumentation semantics* introduce in a declarative way the criteria to determine which arguments emerge as “justified” from the conflict, by identifying a number of *extensions*, i.e. sets of arguments that can “survive the conflict together”. In [3] three “traditional” semantics are introduced, namely *grounded*, *stable*, and *preferred* semantics, as well as the auxiliary notion of *complete* extension.

ArgSemSAT scored second at the first International Competition on Computational Models of Argumentation ICCMA2015 [6] and was ranked first or second in each track associated to the highest in complexity problems for stable and preferred semantics—except one due to an implementation bug discovered after the competition.¹

Building on top of the success of ArgSemSAT, we re-coded it in Java designing jArgSemSAT [2], for being easily integrated within existing argumentation systems, such as Dung-O-Matic [4], the Tweety libraries [5], and ArgTech [1].

1. jArgSemSAT

jArgSemSAT is a mature application that now exists in four different versions:

1. Stand-alone application compatible with the Probo interface [6];
2. Dung-O-Matic (DoM) [4] compatible library: this ensures compatibility for works already using DoM;

¹Details in <http://goo.gl/sRFaSi>

3. Tweety [5] compatible library: we proudly support the Tweety project whose aim is to provide a general framework for implementing and testing knowledge representation formalisms;
4. ArgTech [1] compatible web-service: we created a Tomcat web-service exporting jArgSemSAT with ArgTech-compatible RESTful interfaces.

jArgSemSAT is freely (MIT licence) available on SourceForge² and as Maven projects directly accessible from the central repository. It is composed by two `jar` files and a `war` file.

`jArgSemSAT-VERSION.jar` provides both the stand-alone application compatible with the **Probo** interface and the DoM compatible library: we chose not to distribute the library without the **Probo** interface to facilitate future experiments also from different research groups and to improve the awareness in the community of the ICCMA competition.

`jArgSemSAT-Tweety-VERSION.jar` is a self-contained, Tweety-compatible, library: it includes `jArgSemSAT-VERSION.jar` and provides a Tweety-compatible interface.

`jArgSemSATWeb-VERSION.war` is a self-contained Tomcat web-service archive compatible with ArgTech specifications. This web-service is also freely available at <http://cicero.cs.cf.ac.uk/jArgSemSATWeb/restapi/argtech/> (best effort SLA.) Its source code is also freely available.

2. Conclusions

jArgSemSAT is an efficient off-the-shelf solver for abstract argumentation problems and in [2] we proved that it is only slightly less efficient than its ancestor ArgSemSAT, which is written in C++.

To give an hint of jArgSemSAT performance, in a re-run of ICCMA 2015, it made the podium in regard to credulous acceptance w.r.t. preferred semantics; enumeration of preferred extensions; skeptical acceptance w.r.t. stable semantics; and enumeration of stable extensions.

References

- [1] Floris Bex, John Lawrence, Mark Snaith, and Chris Reed. Implementing the argument web. *Communications of the ACM*, 56(10):66, oct 2013.
- [2] Federico Cerutti, Mauro Vallati, and Massimiliano Giacomin. jargsemsat: An efficient off-the-shelf solver for abstract argumentation frameworks. In *KR'16. AAI*, 2016.
- [3] Phan Minh Dung. On the Acceptability of Arguments and Its Fundamental Role in Nonmonotonic Reasoning, Logic Programming, and n-Person Games. *Artificial Intelligence*, 77(2):321–357, 1995.
- [4] Mark Snaith, Joseph Devereux, John Lawrence, and Chris Reed. Pipelining argumentation technologies. In *COMMA 2010*, pages 447–453, 2010.
- [5] Matthias Thimm. Tweety - a comprehensive collection of java libraries for logical aspects of artificial intelligence and knowledge representation. In *KR'14*, pages 528–537, July 2014.
- [6] Matthias Thimm, Serena Villata, Federico Cerutti, Nir Oren, Hannes Strass, and Mauro Vallati. Summary Report of The First International Competition on Computational Models of Argumentation. *AI Magazine*, 2016.

²<https://sourceforge.net/projects/jargsemsat/>

Generating Structured Argumentation Frameworks: AFBenchGen2

Federico CERUTTI^a, Massimiliano GIACOMIN^b, Mauro VALLATI^c

^aCardiff University, School of Computer Science & Informatics, UK

^bDepartment of Information Engineering, University of Brescia, Italy

^cSchool of Computing and Engineering, University of Huddersfield, UK

Abstract. In this paper we describe AFBenchGen2, which allows to randomised argumentation frameworks for testing purposes with a large variety of structures.

Keywords. argumentation frameworks, benchmarks, algorithm evaluation

Introduction

Dung’s abstract argumentation framework (*AF*) [4] provides a fundamental reference in computational argumentation. An *AF* consists of a set of arguments and an *attack* relation between them. The concept of *extension* plays a key role in this simple setting, where an *extension* is a set of arguments which can “survive the conflict together.” Different notions of extensions and of the requirements they should satisfy correspond to alternative *argumentation semantics*.

In previous research we introduced AFBenchGen [3], allowing for the generation of challenging *AF*s based on the Erdős-Rényi model [5]. However, as [2] discussed, different structures can give rise to interesting different results w.r.t. performance for existing solvers of decision and enumeration problems on Dung’s *AF*s. In this paper we present AFBenchGen2, the first open-source, configurable system for generating *AF*s with a variety of structures.

1. AFBenchGen2

Differently from its predecessor, AFBenchGen2¹ is written in Java and can create *AF*s with a configurable number of arguments, and of type: (1) Erdős-Rényi [5]; (2) Watts-Strogatz [8]; (3) Barabasi-Albert [1].

Erdős-Rényi Erdős-Rényi graphs [5] are generated by randomly selecting attacks between arguments. AFBenchGen2 allows the selection of the probability of attacks via the parameter `-ER_probAttacks` (between 0 and 1).

¹<https://sourceforge.net/projects/afbenchgen/>

Watts-Strogatz Watts and Strogatz [8] show that many biological, technological and social networks are neither completely regular nor completely random, but something in the between. These systems can be highly clustered, like regular lattices, yet have small characteristic path lengths, like random graphs, and they are named *small-world* networks by analogy with the small-world phenomenon.

AFBenchGen2 generates a ring of n arguments where each argument is connected to its k nearest neighbors in the ring: k can be specified via the parameter `-WS_baseDegree` and it must satisfy $n \gg k \gg \log(n) \gg 1$ to ensure a connected graph. Then AFBenchGen2 considers each argument and *rewires* each of its edges toward the not yet processed arguments with randomly chosen arguments with a probability β that can be specified with the parameter `-WS_beta` (between 0 and 1).

Barabasi-Albert As discussed in [1], a common property of many large networks is that the node connectivities follow a scale-free power-law distribution. Therefore, generating a Barabasi-Albert graph requires to iteratively connect a given number of new nodes and to prefer sites that are already well connected. In order to resemble online discussions, we chose to tune AFBenchGen2 to add a single new argument at every iteration: however, this can be made configurable.

Both Watts-Strogatz and Barabasi-Albert would result in undirected graph (or, directed graph with no cycles); we therefore added an additional parameter `-BA_WS_probCycles` (between 0 and 1) that describes the probability of an argument to be in at least one cycle. AFBenchGen2 will therefore add extra attacks accordingly.

2. Conclusions

In the last years, thank also to the ICCMA15 [7] there has been an increased attention in the community towards implementations and experimental analysis. However, benchmarks and raw data are as important as papers and systems code: in certain disciplines the majority of published findings cannot be reproduced [6]. Making AFBenchGen2 freely available and open source goes in the direction of reducing such a risk for the argumentation in AI community.

References

- [1] Albert-Laszlo Barabasi and Reka Albert. Emergence of scaling in random networks. *Science*, 286(5439):11, oct 1999.
- [2] Stefano Bistarelli, Fabio Rossi, and Francesco Santini. Enumerating Extensions on Random Abstract-AFs with ArgTools, Aspartix, ConArg2, and Dung-O-Matic. In *CLIMA'14*, pages 70–86. 2014.
- [3] Federico Cerutti, Massimiliano Giacomin, and Mauro Vallati. Generating Challenging Benchmark AFs. In *COMMA'14*, pages 457–458, 2014.
- [4] Phan Minh Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77(2):321–357, 1995.
- [5] Paul Erdős and Alfréd Rényi. On random graphs. I. *Publ. Math. Debrecen*, 6:290–297, 1959.
- [6] John P A Ioannidis. Why most published research findings are false. *PLoS medicine*, 2(8):e124, 2005.
- [7] Matthias Thimm, Serena Villata, Federico Cerutti, Nir Oren, Hannes Strass, and Mauro Vallati. Summary Report of The First International Competition on Computational Models of Argumentation. *AI Magazine*, 2016.
- [8] Duncan J. Watts and Stephen H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–2, jun 1998.

A System for Supporting the Detection of Deceptive Reviews Using Argument Mining

Oana COCARASCU^{a,1}, Francesca TONI^a

^a*Department of Computing, Imperial College London, UK*

Abstract. The unstoppable rise of social networks and the web is facing a serious challenge: identifying the truthfulness of online opinions and reviews. We propose a system to identify two new argumentative features that a trained classifier can use to help determine whether a review is deceptive.

Keywords. Deception, Argument mining, Abstract argumentation

1. Introduction

Nowadays the decision of purchasing a specific item (e.g. product or service) is mainly done based on online reviews. However, the truthfulness of these reviews is not guaranteed and content communities and review websites are susceptible to deception. The standard approach to detecting deceptive reviews is to use a classifier trained on a dataset consisting of truthful and deceptive reviews (e.g. see [2]), typically relying on syntactic features extracted from reviews (e.g. frequencies of part-of-speech tags). We propose a system, to process unstructured text from reviews and determine whether they are deceptive, using a trained classifier that relies on standard features as well as two novel *argumentative features*. These result from measuring the strength of arguments, for the goodness and badness of the items being reviewed, within abstract argumentation frameworks (AFs) [1] mined from reviews about the items; here the strength of arguments is measured using a variant of the method in [3]. Within our system, the AFs used to determine the novel argumentative features are also used to explain classifications (as deceptive or truthful) of user-given reviews.

2. Argumentative features for detecting deception

We illustrate our system with a simple example. For hotel H, consider reviews:

r_1 : ‘It had nice rooms but terrible food.’

r_2 : ‘The service was amazing and we loved the room. They don’t offer free Wi-Fi.’

First, each review is mapped (with a tokenizer) to one or more arguments, each contained in a sentence. Sentences containing *but/although* are split into different arguments

¹Corresponding Author: Oana Cocarascu, Department of Computing, Imperial College London, United Kingdom; E-mail: oana.cocarascu11@imperial.ac.uk.

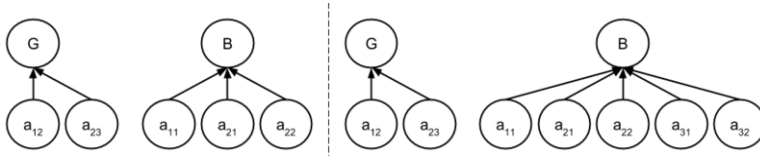


Figure 1. AFs extracted from reviews r_1, r_2 (left) and r_1, r_2, r_3 (right)

since generally the phrases before and after these separators express different sentiments. Second, we assign a polarity (positive +, negative -) to each argument extracted from reviews. For example, arguments (with their polarity) extracted from r_1, r_2 are

a_{11} : *nice rooms* (+) a_{12} : *terrible food* (-)
 a_{21} : *service was amazing* (+) a_{22} : *loved the room* (+)
 a_{23} : *they don't offer free Wi-Fi* (-)

Third, we construct an AF whose arguments include, in addition to the arguments extracted from reviews, two special arguments: G (for 'good') and B (for 'bad'), and use the polarity of arguments to determine the attack relationship. The AF obtained for reviews r_1 and r_2 , for example, is shown graphically on the left of Figure 1.

Finally, to help determine whether a review r about an item is deceptive, given a set of reviews R for the item including r , we compute two argumentative features for r , representing the impact of r on how good/bad a product is with respect to R . These new features are obtained from measuring (using a variant of [3]) the strength of G and B in the AFs obtained from R and from $R \setminus \{r\}$. For example, consider the problem of determining whether the following review for hotel H is deceptive:

r_3 : 'The staff was super friendly and helpful and the location was fantastic.'
 with respect to $R = \{r_1, r_2, r_3\}$. The arguments extracted from r_3 are:

a_{31} : *staff was super friendly and helpful* (+) a_{32} : *location was fantastic* (+)

We construct a new AF, shown graphically on the right of Figure 1. Then, the two argumentative features for r_3 are the absolute difference between the measure of how good/bad hotel H is deemed to be given all reviews R and how good/bad (respectively) it is deemed to be given $R \setminus \{r_3\}$, namely the absolute difference between the strength of G/B (respectively) in the AF on the right of Figure 1 and on the left of Figure 1. The use of these two new features (one for 'good' and one for 'bad'), in addition to standard features, allows to obtain a classifier with accuracy of up to 85%, using 5-fold cross-validation for training on the dataset in [2]. Our system uses the trained classifier to predict deceptive reviews and offers explanations in terms of the AFs used to determine the new features for predictions of deception or truthfulness of user-given reviews.

References

- [1] Dung, P.M.: On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence* 77(2), 321 – 357 (1995)
- [2] Ott, M., Cardie, C., Hancock, J.T.: Negative deceptive opinion spam. In: *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (2013)
- [3] Rago, A., Toni, F., Aurisicchio, M., Baroni, P.: Discontinuity-free decision support with quantitative argumentation debates. In: *Principles of Knowledge Representation and Reasoning: Proceedings of the Fifteenth International Conference*. pp. 63–73 (2016)

DIAMOND 3.0 – A Native C++ Implementation of DIAMOND

Stefan ELLMAUTHALER^a and Hannes STRASS^a

^a*Intelligent Systems Group, Computer Science Institute, Leipzig University, Germany*¹

Abstract. We present a reimplementaion of the DIAMOND system for computing with abstract dialectical frameworks. The original DIAMOND was a script-based tool that called an external ASP solver. This reimplementaion uses the clingo library in a native C++ environment and thus avoids communication overhead.

Keywords. abstract argumentation, abstract dialectical frameworks, DIAMOND, answer set programming, clingo, potassco

1. Motivation

Abstract dialectical frameworks (ADFs, [2,1]) are a logic-based formalism for representing knowledge about arguments (statements) and relationships between them. They can be represented by directed graphs, where each node denotes an argument and an edge from a to b encodes that a has an influence on b . The precise nature of this influence (and the influences of other parent nodes of b) are expressed using an acceptance condition, a Boolean function on the parents of b . The semantics of ADFs indicate (according to different justification standards) what statements can be collectively justified [1].

To compute ADF semantics, researchers quickly came to capitalise on the advantages of answer set programming (ASP, [8]). The first ADF system, ADFsys, was presented at COMMA 2012 [6]. With the introduction of additional semantics, also a new system was presented: DIAMOND [1,5].

The DIAMOND system uses ASP encodings of ADF semantics to solve reasoning problems. (For example deciding whether an ADF has a (non-trivial) interpretation for a specific semantics, or a statement is sceptically/credulously accepted/rejected by an ADF according to a semantics.) As an ASP back-end, DIAMOND uses tools from the Potsdam answer set solving collection, Potassco [7], more specifically the solver clingo.

Since ADFs are a generalisation of Dung's abstract argumentation frameworks (AFs, [4]), DIAMOND can also be used to compute semantics for AFs. In this function, we submitted DIAMOND 2.0.1 to the First International Competition on Computational Models of Argumentation, ICCMA 2015 (<http://argumentationcompetition.org>). From the competition results, we could observe that DIAMOND 2.0.1 gave an unexpectedly large amount of wrong answers to decision problems. We traced these problems back to technical problems in DIAMOND's communication with its solver back-end. In this paper, we announce the next version of DIAMOND, version 3.0.x, that aims at overcoming those problems with a change in architecture.

¹email: {ellmauthaler,strass}@informatik.uni-leipzig.de

2. Main Changes

The “old” DIAMOND (versions up to 2.0.x) basically consisted of a python script that took user commands and an instance filename to combine the right ASP encodings to give to a solver. It used python’s interface to operating system pipes and command line calls to communicate with its solver back-end. However, python makes no assertions about data loss in pipes; and indeed, for large ADF instances we observed differences between manual ASP solver calls and the python wrapper script.

The DIAMOND system was completely re-written in C++. The new DIAMOND now uses the clingo library and calls the solver clingo internally by creating a solver object and directly communicating with it within the program instead of invoking the operating system.

The ASP encodings themselves (actually computing the semantics) have not changed, but are now compiled into the program at compile time. In principle, this allows for a standalone executable.

The new version of DIAMOND is available at sourceforge at <https://sourceforge.net/p/diamond-adf/code/ci/cdevelop/tree/>.

3. Conclusion

For future work, we want to do an extensive experimental evaluation of DIAMOND 3.0.x. In particular, we want to compare DIAMOND 2.0.2 and DIAMOND 3.0.x using the probos software [3].

References

- [1] Gerhard Brewka, Stefan Ellmauthaler, Hannes Strass, Johannes Peter Wallner, and Stefan Woltran. Abstract Dialectical Frameworks Revisited. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence (IJCAI)*, pages 803–809. IJCAI/AAAI, August 2013.
- [2] Gerhard Brewka and Stefan Woltran. Abstract Dialectical Frameworks. In *Proceedings of the Twelfth International Conference on the Principles of Knowledge Representation and Reasoning (KR)*, pages 102–111, 2010.
- [3] Federico Cerutti, Nir Oren, Hannes Strass, Matthias Thimm, and Mauro Vallati. A benchmark framework for a computational argumentation competition. In *Proceedings of the Fifth International Conference on Computational Models of Argument (COMMA)*, volume 266 of *FAIA*, pages 459–460. IOS Press, September 2014.
- [4] Phan Minh Dung. On the Acceptability of Arguments and its Fundamental Role in Nonmonotonic Reasoning, Logic Programming and n-Person Games. *Artificial Intelligence*, 77(2):321–358, 1995.
- [5] Stefan Ellmauthaler and Hannes Strass. The DIAMOND System for Computing with Abstract Dialectical Frameworks. In *Proceedings of the Fifth International Conference on Computational Models of Argument (COMMA)*, volume 266 of *FAIA*, pages 233–240. IOS Press, September 2014.
- [6] Stefan Ellmauthaler and Johannes P. Wallner. Evaluating Abstract Dialectical Frameworks with ASP. In *Proceedings of the Fourth International Conference on Computational Models of Argument (COMMA)*, volume 245 of *FAIA*, pages 505–506. IOS Press, 2012.
- [7] M. Gebser, R. Kaminski, B. Kaufmann, M. Ostrowski, T. Schaub, and M. Schneider. Potassco: The Potsdam Answer Set Solving Collection. *AI Communications*, 24(2):105–124, 2011. Available at <http://potassco.sourceforge.net>.
- [8] Martin Gebser, Roland Kaminski, Benjamin Kaufmann, and Torsten Schaub. Answer set solving in practice. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 6(3):1–238, 2012.

GrappaVis - A System for Advanced Graph-Based Argumentation

Georg HEISSENBERGER^a Stefan WOLTRAN^b

^a Institute of Information Systems 184/2, TU Wien, Austria

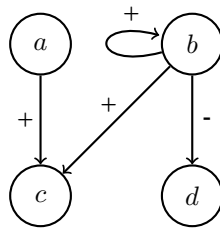
^b Institute of Information Systems 184/2, TU Wien, Austria

Abstract. We present a new system for specifying and evaluating frameworks in the recently proposed argumentation formalism of GRAPPA.

Keywords. Abstract Argumentation, Systems, Answer-Set Programming

GRAPPA (G**R**aph-based Argument Processing with Patterns of Acceptance) [3] is new formalism for abstract argumentation that allows for specifying arbitrary relations between the arguments without giving up the popularity of graphical models in argumentation. Frameworks in GRAPPA are edge-labelled graphs where each argument (i.e. each node) comes with an acceptance condition that is evaluated with respect to the labels of active links (i.e. links to accepted nodes). Rather than specifying acceptance conditions in a logic, like e.g. in abstract dialectical frameworks (ADFs) [2], GRAPPA offers a powerful language of acceptance patterns which make the design of the frameworks more intuitive.

The following graph shows a GRAPPA instance over arguments $S = \{a, b, c, d\}$ and labels $\{+, -\}$.



For simplicity, let's assume all nodes have the same acceptance condition stating that a node becomes accepted if all positive links (+) are active (i.e. all respective parents must be accepted) and no negative link (-) is active. In GRAPPA, this is simply expressed by assigning the term $(\#_t(+) - \#(+)) = 0 \wedge (\#(-) = 0)$ to each node of S . For details on the language of acceptance pattern, we refer to [3].

The definition of the semantics for GRAPPA only requires a minor modification of the operator underlying the ADF semantics [2], taking into account the new type of acceptance conditions. For the example above, we obtain two so-called models, namely $\{a, b, c, \neg d\}$ and $\{a, \neg b, \neg c, d\}$; the grounded interpretation is $\{a\}$. In fact, not only ADF semantics are covered by GRAPPA, it is also a generalization of Dung-style argumentation frameworks (those are expressed by labelling all

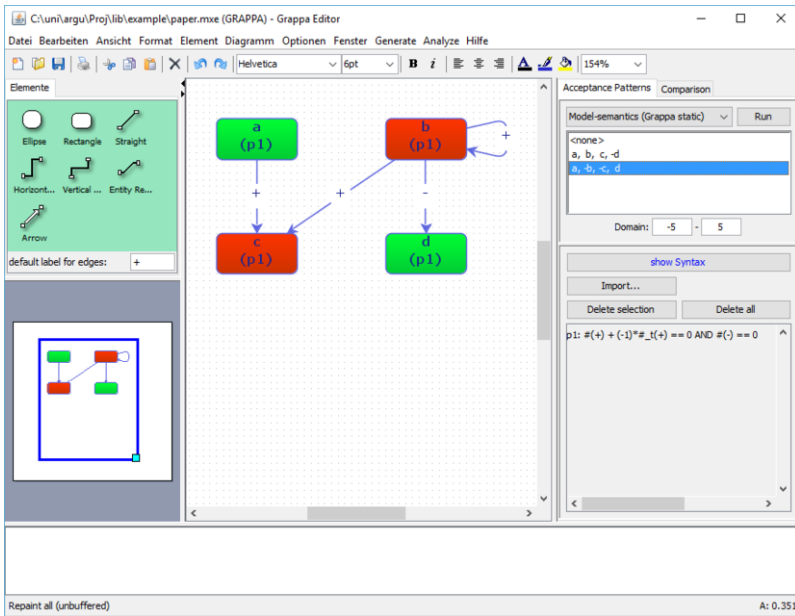


Figure 1. Screenshot of GrappaVis

edges with a single label, say “-”, and assign to each node the acceptance pattern $\#(-) = 0$.

GrappaVis is a system for specifying and evaluating GRAPPA instances under different semantics. The system is implemented in Java and thus works on different platforms. It consists of a graphical frontend (see Figure 1 for a screenshot) based on the JGraphX library which makes it easy to specify and manage GRAPPA instances. Acceptance patterns can be directly assigned to the nodes of the framework. The actual evaluation of the GRAPPA framework is delegated to Answer-Set Programming systems (for an overview on this programming paradigm, see [1]). To this end, novel encodings have been developed; for details see the first author’s master thesis [4]. The resulting extensions are processed again by the system and shown via highlighting of nodes. The system is available under

<http://dbai.tuwien.ac.at/proj/adf/grappavis/>

Acknowledgements This research has been supported by the Austrian Science Fund (FWF) through projects Y698, I1102 and I2854.

References

- [1] G. Brewka, T. Eiter, and M. Truszczynski. Answer set programming at a glance. *Commun. ACM*, 54(12):92–103, 2011.
- [2] G. Brewka, S. Ellmauthaler, H. Strass, J. P. Wallner, and S. Woltran. Abstract Dialectical Frameworks Revisited. In *Proc. IJCAI 2013*, pages 803–809. AAAI Press / IJCAI, 2013.
- [3] G. Brewka and S. Woltran. GRAPPA: A semantical framework for graph-based argument processing. In *Proc. ECAI 2014*, volume 263 of *FAIA*, pages 153–158. IOS Press, 2014.
- [4] G. Heissenberger. A system for advanced graphical argumentation formalisms. Master’s thesis, TU Wien, 2016. Available at the GrappaVis webpage.

The ARGTEACH Web-Platform

Claudia SCHULZ^a and Dragos DUMITRACHE^a

^a *Department of Computing, Imperial College London, UK*

Abstract. ARGTEACH is a web-platform for teaching and learning the labelling semantics of abstract argumentation frameworks. The user's task is to find the complete labellings of an abstract argumentation framework, supported by ARGTEACH in the form of hints about sensible next labelling steps as well as error checking of partial and total labellings. The ARGTEACH web-platform considerably improves and extends an earlier java application, both with respect to functionality and design.

Keywords. abstract argumentation, labelling semantics, teaching and learning

The ARGTEACH web-platform¹ assists the user in learning how to find complete labellings of abstract argumentation frameworks (AFs) [1]. Pre- or user-defined AFs are visualised as directed graphs, whose argument nodes can be labelled as *in* (green), *out* (red), or *undec* (white) by clicking on the respective argument. ARGTEACH supports the user in the form of *labelling hints* as well as *error checking*, which have been considerably improved compared to the earlier ARGTEACH java application² [2].

Labelling Hints If the user is at any point unsure which argument to label next, ARGTEACH provides hints about sensible next labelling steps, i.e. what labels may be assigned to which unlabelled arguments based on the labels assigned already. The hints make use of the three conditions satisfied by a complete labelling, namely for each argument *A* it holds that

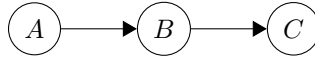
- *A* is labelled *in* if and only if all attackers of *A* are labelled *out*;
- *A* is labelled *out* if and only if some attacker of *A* is labelled *in*;
- *A* is labelled *undec* if and only if no attacker of *A* is labelled *in* and some attacker of *A* is labelled *undec*.

In contrast to the earlier java application, the ARGTEACH web-platform makes use of both the “if” and the “only if” direction of these conditions, thus generating hints not only about arguments attacked by an already labelled argument but also about arguments attacking an already labelled argument.

Consequently, *manifold evidence* for assigning a certain label to an argument can be generated. For example, if arguments *A* and *C* in the below AF are labelled *in*, and argument *B* is unlabelled, then ARGTEACH provides a two-fold hint: “Argument *B* should be *out* because it has an *in* labelled attacker. Furthermore, argument *B* should be *out* because it attacks an *in* labelled argument.”

¹<http://www-argteach.doc.ic.ac.uk/>

² <http://www.doc.ic.ac.uk/~cis11/ArgTeach/argTeach.html>



On the other hand, a hint may comprise *contradictory evidence*, that is evidence for assigning different labels to the same argument, indicating a mistake in the already labelled arguments. For example, if the user labelled argument *A* as *out* and *C* as *in*, the hint created by ARGTEACH is contradictory: “Argument *B* should be *in* because all of its attackers are labelled *out*. On the other hand, argument *B* should be *out* because it attacks an *in* labelled argument.” This shows that the hints merely teach the user how to apply the three conditions of complete labellings step by step to unlabelled arguments; they do not rectify mistakes in already labelled arguments.

Error Checking Once all arguments are labelled, the user can check whether the total labelling is indeed a complete labelling. ARGTEACH points out every argument, if any, which violates one of the three conditions of complete labellings, again checking not only the “if” direction as done in the earlier java application, but also the “only if” direction. In contrast to the hints, only one violation is pointed out for an argument, even if more violations exist. On request, ARGTEACH also indicates whether or not the label of a violating argument should be changed in order to obtain a complete labelling.

For example, if all three arguments in the above AF are labelled *in*, ARGTEACH indicates a violation for every argument: “Argument *A* is labelled *in* but attacks an argument that is labelled *in*”; “Argument *B* is labelled *in* but is attacked by an argument that is not labelled *out*”; and “Argument *C* is labelled *in* but is attacked by an argument that is not labelled *out*”. On request ARGTEACH suggests to only change the label of argument *B*, whereas for the violations of arguments *A* and *C* it advises “Better try to fix another error”.

ARGTEACH also provides error checking for partial labellings, applying the same algorithm as for total labellings. If no explicit violations are found, ARGTEACH checks if the current labelling can lead to a complete labelling.

Additional Functionality Additional features of the ARGTEACH web-platform are the option to check if a found complete labelling is a grounded, preferred, semi-stable, or stable labelling, as well as the user login, which saves previously found complete labellings of an AF.

Acknowledgements

Special thanks go to K. Cyraś and O. Cocarascu for exhaustively testing ARGTEACH and providing feedback about its design and functionality.

References

- [1] Pietro Baroni, Martin Caminada, and Massimiliano Giacomin. An Introduction to Argumentation Semantics. *The Knowledge Engineering Review*, 26(04):365–410, 2011.
- [2] Jeremie Dauphin and Claudia Schulz. ArgTeach - A Learning Tool for Argumentation Theory. In *ICTAI'14*, pages 776–783, 2014.

Gorgias-B: Argumentation in Practice

Nikolaos I. SPANOUDAKIS ^{a,1}, Antonis C. KAKAS ^b and Pavlos MORAITIS ^c

^a *Applied Mathematics and Computers Lab., Technical University of Crete, Greece*

^b *Department of Computer Science, University of Cyprus, Cyprus*

^c *Laboratory of Informatics Paris Descartes (LIPADE), France*

Abstract. *Gorgias-B* is a new tool that supports a methodology for the development of real life applications. It can be used by non-argumentation experts generating and testing automatically the target argumentation theory in *Gorgias*.

1. Gorgias-B: Supporting applications of preference-based argumentation

Argumentation technology is well suited for implementing decision making mechanisms under conflicting, incomplete and contextual knowledge. It allows choosing preferred options (e.g. actions) among a list of possible (usually conflicting) alternatives under some decision policy of an application. *Gorgias* is a system based on preference-based argumentation that has been used during the past ten years by different users for developing a variety of real life applications (see <http://www.amcl.tuc.gr/gorgiasb/Apps.html>). Based on the study of these applications we have developed a new tool, *Gorgias-B*, to support the development of applications of argumentation under *Gorgias*, following a general software methodology. *Gorgias-B* guides the developer to structure his/her knowledge at several levels. The first level serves for enumerating the possible decisions and arguments that can support these options under some conditions, while each higher level serves for resolving conflicts at the previous level by taking into account default or contextual knowledge.

Figure 1 illustrates the development of a seller agent using the *Gorgias-B* tool. In the first screen (bottom) the user has defined two conflicting options, i.e. to sell products to agents at a high price or to sell them at a low price. Options appear on the left and then defined as complementary (or conflicting) on the right. After defining the various options the user can press the button "Add arguments for options" opening the dialogue in the next screen (second from bottom-up in the figure). This shows two arguments, a default one for selling high and another for selling low (*when* the buyer agent accepts to pay cash). By following the button "Resolve conflicts" a new dialogue (third from bottom-up in the figure) appears. Here, we can select possible scenarios (produced by combining the contexts of the arguments with conflicting options) of the previous level and select the winning option in the specific scenario. In the specific screen, "sell high" is preferred as a default policy for the seller. However, in the case of a regular customer, selling low is preferred over selling high. In this more specific context, i.e. of [*pay cash, regular customer*], we still have a conflict as both options can be selected. In the fourth

¹Corresponding Author: Nikolaos I. Spanoudakis, Technical University of Crete Campus, 73100, Chania, Greece; E-mail: nikos@amcl.tuc.gr.

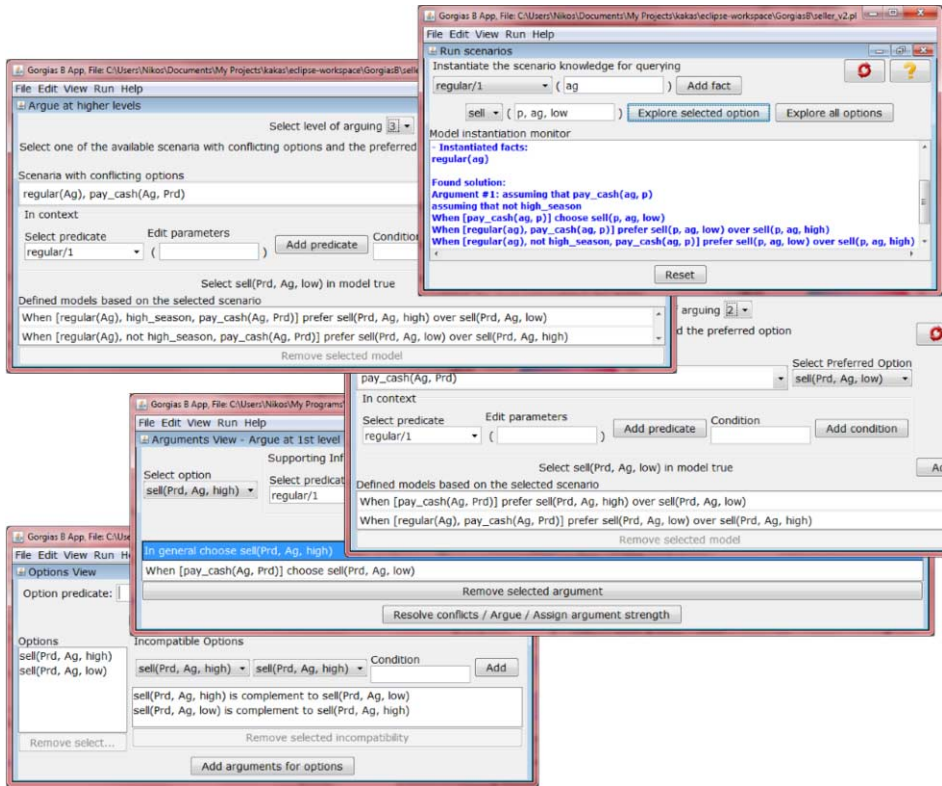


Figure 1. Gorgias-B screenshots

screenshot we see that we can resolve this conflict at a next level preferring the option low when it is not high season and high otherwise. Due to the complementarity of these two refinements of the context there is no need for further resolution at a next level (no scenarios appear if the user selects that level).

During this process the tool generates automatically an argumentation theory that captures the high-level specification entered by the developer within the argumentation framework of *Gorgias* [1]. This theory can be executed through the *Gorgias-B* tool (see last screen from bottom-up) by specifying scenarios of interest and asking which options are credulously or sceptically entailed in the scenarios. The tool returns these together with the admissible arguments that support them. *Gorgias-B* allows also to specify some predicates as *abducible*, and the tool can find scenario conditions under which an option or a conclusion will be entailed. In the last screenshot, we see that to support a low selling scenario, the assumptions of payment in cash and that it is not high season must hold.

Gorgias-B is freely available from <http://gorgiasb.tuc.gr>. Apart from application development it can also be used to demonstrate argumentation and how it supports defeasible reasoning.

References

- [1] Antonis C. Kakas and Pavlos Moraitis: Argumentation based decision making for autonomous agents. *2nd Int. Joint Conf. on Autonomous Agents and Multiagent Systems, (AAMAS)*, (2003), 883–890.

The RationalGRL Toolset for Goal Models and Argument Diagrams

Marc VAN ZEE^{a,1}, Diana MAROSIN^b, Floris BEX^c, Sepideh, GHANAVATI^d

^a Computer Science and Communication, University of Luxembourg, Luxembourg

^b Luxembourg Institute of Science and Technology, Luxembourg

^c Information and Computing Sciences, Utrecht University, The Netherlands

^d ICIS, Radboud University, The Netherlands

Problem Statement. The Goal-oriented Requirements Language (GRL) aims at modeling high-level business and system goals, subgoals and tasks and analyzing the alternative ways of achieving these goals and subgoals. However, GRL models are only the end product of a modeling process, and they do not provide any insight on how the models were created. For instance, they do not show what reasons were used to choose certain elements in the model and to reject the others and what evidence was given as the basis of this reasoning. There are, thus, several questions that are not answered in GRL: Why is a goal created? Why are some goals evaluated positively and some negatively? Do we have any evidence for the fact that performing a certain task contributes to a goal?

Overview of the Framework The main components of the RationalGRL framework are shown in Figure 1. The four main parts of the framework, Argumentation, Translation, Goal Modeling, and Update, are numbered and depicted in **bold**. For each component, the technology used to implement it is marked in a filled rectangle.

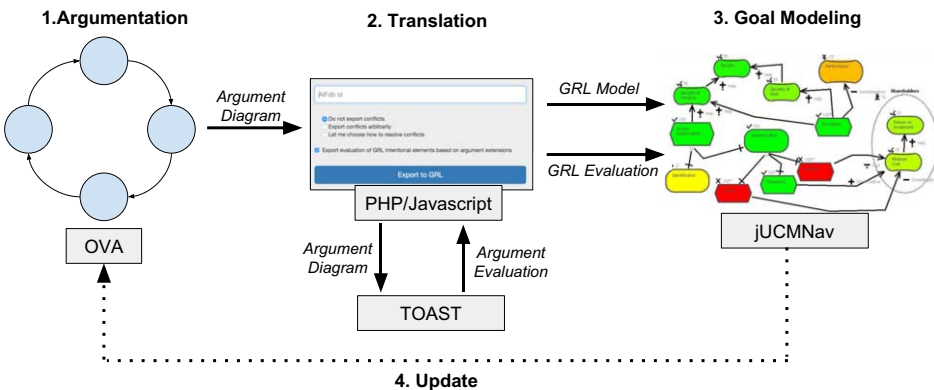


Figure 1. Overview of the RationalGRL Framework

¹Corresponding Author: marcvanzee@gmail.com.

In *Step 1 - Argumentation*, stakeholders discuss the requirements of their organization. In this process, stakeholders put forward arguments for or against certain elements of the model (e.g. goals, tasks,...). Arguments about why certain tasks can contribute to the fulfillment of goals and an evidence to support a claim are also part of this process. Furthermore, stakeholders can challenge claims by forming counterarguments. The complete set of claims, arguments and counterarguments can be represented in an argument diagram. We have implemented a formal argumentation theory for goal-based reasoning about evidence [4,3] into the browser-based argument diagramming tool OVA² [1].

In *Step 2 - Translation*, the argument diagram is translated to a goal model, in our case GRL. The process consists of two translations/mappings: the first generates the GRL elements and relationships, and the second mapping generates the satisfaction values of the GRL elements from the acceptability status of underlying arguments. We implemented the translation tool in PHP³. The tool requests an argument from the Argument Web, and then generates a GRL model using mapping rules. This is exported in XML format, which can be imported in jUCMNav, an Eclipse based tool for GRL modeling⁴. The tool then requests argument evaluations from TOAST, a tool for evaluating the Dung semantics⁵, and uses this to set the evaluation of the GRL elements.

In *Step 3 - Goal Modeling*, the goal model that is generated by the Translation process is evaluated by the stakeholders. These models can be used as a discussion means to investigate whether the goals in the model are in line with the original requirements of the stakeholders. This allows a better rationalization of the goal modeling process, with a clear traceability from the goals of the organization to the arguments and evidence that were used in the discussions.

Step 4 - Update involves translating GRL models with its analysis back into an argument diagram. This falls outside the scope of the current paper.

Future work. It would be interesting to explore the effect of different argumentation semantics on goal models. Moreover, we would like to add the *Update* step of our framework in order to automatically translate goal models to argument diagrams. For this, we see the recent proposal by Mirel and Villata as a useful starting point [2].

References

- [1] Floris Bex, John Lawrence, Mark Snaith, and Chris Reed. Implementing the argument web. *Communications of the ACM*, 56(10):66–73, 2013.
- [2] Isabelle Mirbel and Serena Villata. An Argumentation-based Support System for Requirements Reconciliation. In *Proceedings of COMMA 2014*, pages 467–468, 2014.
- [3] Marc van Zee, Floris Bex, and Sepideh Ghanavati. Rationalization of Goal Models in GRL using Formal Argumentation. In *Proceedings of RE:Next! track at RE 2015*, 2015.
- [4] Marc van Zee and Sepideh Ghanavati. Capturing Evidence and Rationales with Requirements Engineering and Argumentation-Based Techniques. In *Proc. of the 26th Benelux Conf. on Artificial Intelligence (BNAIC)*, 2014.

²<http://ova.arg-tech.org/>

³<http://www.marcvanze.nl/RationalGRL>

⁴<http://jucmnav.softwareengineering.ca/ucm/bin/view/ProjetSEG/WebHome>

⁵<http://toast.arg-tech.org>

Subject Index

AATS	71	bipolar argumentation	
abstract argumentation	83, 107, 151, 191, 199, 207, 231, 243, 255, 447, 469, 471, 473, 475	frameworks	41
abstract argumentation		bipolar models	151
frameworks	167, 275, 287, 463	bipolar networks	311
abstract dialectical		c-semiring	159
frameworks	127, 471	case-based reasoning	243
acceptability semantics	231	causal reasoning	127
algorithm configuration	199	character of speakers	299
algorithm evaluation	467	clingo	471
answer set programming	471, 473	coalitions of arguments	431
applications and structured		collaborative argumentation	33
argumentation and Datalog \pm	61	collaborative work	33
applications of argumentation	461	community of agents	339
argument accrual	327	computational complexity	275
argument-based dialogues	411	computational models of argument	5
argument evaluation	327	computational persuasion	5
Argument Interchange Format		computer science	33
(AIF)	119, 371	corpus resources	371
argument mining	21, 379, 431, 469	cumulative arguments	327
argument visualisation	371	Datalog \pm	461
argumentation	119, 139, 219, 311, 359, 391, 439	deception	469
argumentation-based dialogue	339	decision support	53
argumentation dynamics	391	defeasible logic	359
argumentation framework		degrees of acceptance	319
configuration	199	DGDL	351
argumentation frameworks	159, 467	DGEP	351
argumentation schemes	53, 71, 379	dialog-based approach	33
argumentation semantics	83, 159	dialog games	33
argumentation strategies	5	dialogical argumentation	5
arguments strength	403	dialogue game	351
Arvina	351	dialogue strategies	411
ASPIC $^+$	359	DIAMOND	471
ASWO	119	directed graph spectrum	167, 287
attack relation assignment	263	discussion games	179
automated analysis	53	dispute mediation	351
balancing arguments	327	Dung's <i>AF</i>	465
belief revision	391	dynamic programming on tree	
benchmarks	467	decompositions	107
bipolar argumentation	191, 231, 431	ethos attack	299
		ethos support	299
		existence	439
		explanation	243

explanation and argumentation		proof procedures	179
dialogues	461	propagation	139
extension-based semantics	167, 255, 275, 463	quantitative analysis	151
fixed-parameter tractability	107	ranking-based semantics	139
formal argumentation	127	rational acceptance	419
formal dialectical systems	351	rationality postulates	419
foundations	439	recursive interactions	191, 231
fuzzy argumentation	447	regular properties	263
group polarization	41	rhetoric	119
inconsistency tolerance	463	rhetorical arguments	403
information input	311	second-order argument	339
intermediate semantics	83	second-order belief	339
judgment aggregation	179	semantics	439, 447, 465
knowledge acquisition	4	semilattice	263
labelling semantics	475	sentiment analysis	299
lottery paradox	419	SIOC	119
machine learning	4, 219	social media	21, 119
model selection	53	social web	119
natural language processing	299, 379	solver	465
natural language understanding	4	solvers for argumentation	
numerical argumentation	319	problems	207
numerical methods	319	strong equivalence	83
online argumentation	33	structured argumentation	327
parliamentary debates	299	supervised classification	
perfection	439	approaches	21
Perron-Frobenius theory	287	systems	473
persuasion dialogues	5, 411	teaching and learning	475
persuasive arguments	5	threats	403
persuasive arguments theory	41	value-based argumentation	
persuasive negotiation	403	frameworks	41
portfolios methods for		values	71
argumentation	207	verifiability	83
potassco	471	weighted argumentation	
practical reasoning	71	frameworks	159
preferences	53	weighted attacks	463
probabilistic argumentation	5	weighted defence	463

Author Index

Allwood, J.	3	Gordon, T.F.	v, 327
Arioua, A.	461	Gottifredi, S.	231
Atkinson, K.	71	Governatori, G.	359
Baroni, P.	v	Hecher, M.	107
Baumann, R.	83	Heissenberger, G.	473
Baurmann, M.	33	Hosseini, S.A.	339
Bench-Capon, T.	71	Hunter, A.	5
Betz, G.	33	Janier, M.	351
Bex, F.	95, 479	Kakas, A.C.	477
Bistarelli, S.	463	Keppens, J.	53
Black, E.	411	Konieczny, S.	139
Bliem, B.	107	Krauthoff, T.	33
Blount, T.	119	Lagasquie-Schiex, M.-C.	191
Bochman, A.	127	Lam, H.-P.	359
Bollegala, D.	431	Lawrence, J.	351, 371, 379
Bonzon, E.	139	Li, H.	447
Booth, R.	179	Linsbichler, T.	83
Bosc, T.	21	Luck, M.	411
Buche, P.	461	Mailly, J.-G.	255
Budán, M.C.D.	151	Marosin, D.	479
Budzynska, K.	299, 351, 371	Maudet, N.	139
Bundo, S.	159	Mauve, M.	33
Butterworth, J.	167	McBurney, P.	53
Cabrio, E.	21	Millard, D.E.	119
Caminada, M.	179	Modgil, S.	339
Cayrol, C.	191	Moens, M.-F.	4
Cerutti, F.	199, 207, 465, 467	Moguillansky, M.O.	391
Cocarascu, O.	219, 469	Moraitis, P.	477
Cohen, A.	191, 231	Morveli-Espinoza, M.	403
Croitoru, M.	61, 461	Murphy, J.	411
Čyras, K.	243	Norman, T.J.	447
Delobelle, J.	139	Oren, N.	447
Doutre, S.	255	Parsons, S.	431
Dumitrache, D.	475	Possebom, A.T.	403
Dung, P.M.	263	Prakken, H.	419
Dunne, P.E.	167, 275, 287	Proietti, C.	41
Duthie, R.	299, 371	Rajendran, P.	431
Ellmauthaler, S.	471	Reed, C.	299, 351, 371, 379
Gabbay, D.M.	311, 319	Renooij, S.	95
Gabbay, M.	311	Riveret, R.	359
García, A.J.	231	Rodrigues, O.	319, 339
Ghanavati, S.	479	Rossi, F.	463
Giacomin, M.	199, 207, 465, 467	Santini, F.	463

Sassoon, I.	53	Tacla, C.A.	403
Satoh, K.	243	Toni, F.	219, 243, 469
Scheffler, T.	v	Vallati, M.	199, 207, 465, 467
Schulz, C.	475	van Zee, M.	479
Simari, G.I.	151	Villata, S.	21
Simari, G.R.	151, 231, 391	Walton, D.	327
Snaith, M.	351	Weal, M.J.	119
Spanoudakis, N.I.	477	Woltran, S.	83, 107, 473
Spanring, C.	439	Wu, J.	447
Stede, M.	v	Yamaguchi, K.	159
Strass, H.	471	Yun, B.	61