Analyzing the Evolution of Social Groups in World of Warcraft®

Christian Thurau and Christian Bauckhage

***Abstract*— This paper investigates the evolution of social structures in the game WORLD OF WARCRAFT®. We analyze 192 million recordings of 18 million characters belonging to 1.4 million teams, spanning a period of 4 years. Using a recent matrix factorization method, we extract lower dimensional data embeddings. The embeddings provide intuitively interpretable categorizations and we find a tendency towards guilds comprised of casual gamers. To our knowledge, this is the first study considering such a vast amount of data for analyzing groups in MMORPGs.**

I. INTRODUCTION AND RELATED WORK

Massively Multiplayer Online Role Playing Games (MMORPGs) have become a considerable source of revenues for the computer game industry. This becomes apparat from considering the chart in Fig. 2 which displays the evolution of the number of subscribers of the most popular MMORPGs in western countries. As of this writing, the most profitable title by far is WORLD OF WARCRAFT®. It has attracted more than 12,000,000 active players who are willing to pay a monthly fee of about 15$ to play the game. In fact, the game is so economically important that there are several thousand people employed in the *gold-farming* industry (see Fig. 3).

In this paper, we present results of a comprehensive study of the evolution of group formation processes in WORLD OF WARCRAFT®. Given its immense fan-base, the game offers unique possibilities for the study of human interaction on a large scale. Since the game immerses its players in a closed world of considerable complexity, it allows for investigating realistic behavior patterns and social interaction in a semi controlled environment. Being able to characterize the evolution of player behaviors or player interactions will provide answers to obvious questions regarding possible reasons for the game's success. Our particular aims with the work reported here were to provide an interpretable categorization of different guilds of players, to analyze and visualize the development of guilds over time, and to compare the development of US and EU based guilds. In general, we expect such studies of in-game behaviors and in-game social networks to provide valuable insight for future game design.

In fact, the idea of in-game data mining has recently gained interest in academia and industry alike. Usually, the goal is to clone an individual player's behavior or to categorize it. Weber et al. [1] model game opponents using a data mining approach. They learn expert gameplay from vast amounts of game logs, leading to reasonable predictive models. Drachen et al. [2] construct models of players for the game TOMB RAIDER:UNDERWORLD®. From recordings of 1365 players they managed to extract 4 types of players by applying self-organizing maps. Their goal, too, was to assist game developers in the automation of play testing. In earlier work, starting with [3], [4] we used recordings of playing in QUAKE II® to reproduce (clone) individual human behavior. Applying machine learning techniques such as neural networks or mixtures of experts, we could reproduce reactive, tactical, or strategic behaviors.

Fig. 1: WORLD OF WARCRAFT® is a Massively Multiplayer Online Role-Playing Game played by millions of people worldwide. Each character in this picture is controlled by a human player. In order to master difficult quests, players organize themselves in teams, the so called guilds.

More recently, data mining has also been applied to analyze processes within MMORPGs. Ahmad et al. [5] target the task of gold-farmer detection in the MMORPG EVERQUEST II®. Interpreting the task as a binary classification problem, they trained and tested various classifiers using data from about 2 million characters. While their approach is sound, their results indicate that the problem of gold farmer detection is more difficult than expected. Ducheneaut et al. [6] investigate the structure of social networks in WORLD OF WARCRAFT® based on data collected from more than 300,000 characters. Their results show that social networks in MMORPGs are often sparse and that players experience a form of "collective solitude".

Next, we first describe the in-game mechanics of WORLD OF WARCRAFT® in more detail before we discuss the data that formed the empirical basis for our study. Then, we will discuss our findings and present our conclusion. Technical details regarding our data mining approach to analyze millions of in-game observations are provided in the Appendix.

Christian Thurau and Christian Bauckhage are with the Visual and Social Media Group at Fraunhofer IAIS, Sankt Augustin, Germany (e-mail: christian.thurau@iais.fraunhofer.de; christian.bauckhage@iais.fraunhofer.de).

978-1-4244-6297-1/10/$26.00 ©2010 IEEE

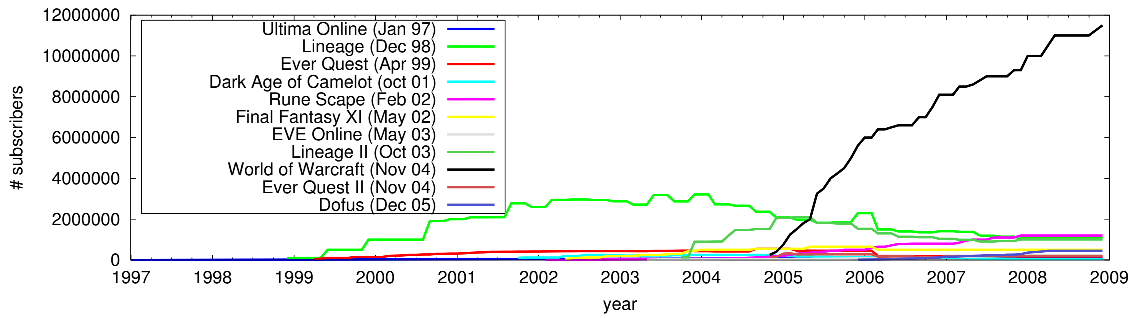


Fig. 2: Evolution of the number of active players of the most popular MMORPGs in western countries (according to `http://www.mmogchart.com`). WORLD OF WARCRAFT® was released in November 2004 and has since gained more than 12,000,000 players. This is more than 6 times as many players as any of the 10 next popular games can attract.

Fig. 3: In-game advertisement for a gold-farmer website. In WORLD OF WARCRAFT® , these are not permitted but constitute an abuse of in-game mechanics, e.g. by placing player corpses to form a web URL. Gold-farming is one of the most prominent fraudulent/illegal activities in WORLD OF WARCRAFT® .

II. WORLD OF WARCRAFT®

WORLD OF WARCRAFT® is a *Massively Multiplayer Online Role Playing Game (MMORPG)*. It is an open multiplayer game and takes place in a medieval fantasy world. It is played by millions of players world wide and is arguably the most successful and most popular game in video game history. To participate in the game, players are required to pay an monthly fee. Currently, around 12 million paying customers log in to WORLD OF WARCRAFT® every day (see Figure 1 for a few gameplay impressions). The gameworld is organized in a few hundred *realms*, i.e. a separated worlds each for a few thousand players who can only interact with characters from the same realm.

While playing, each player controls a single alter-ego character. The *strength* of this character increases with playing time and with successful completion of in-game tasks or quests. Strength is represented by the character's experience level, reaching from level 1 (a newly created character) to level 80 (the currently highest experience level which grants the most powerful abilities such as special attack moves for close combat or mighty spells for mage classes). The maximum experience level has just lately been raised to 80. In the initial release of WORLD OF WARCRAFT® , it was set to 60. With the first expansion package "Burning Crusade" (Jan. 2007) it was extended to level 70, with second expansion "Wrath of the Lich King" (Dec. 2008) it was further extended to level 80.

Due to the open character of the game world, the game lacks a properly defined goal. However, advancing a character's experience level from level 1 to level 80 is implicitly understood as the common goal among all players. Gathering treasures or acquiring better equipment is another commonly accepted goal in the game.

An important aspect of the game lies in the social interaction among players. While players could play the game on their own, a lot of the entertainment comes from the multiplayer experience. Certain quests can only be solved with the help of others. In fact, the most valuable in-game items can only be accessed by a group of more than ten level 80 characters. Therefore, the forming of teams, so called *guilds*, is an integral part of gameplay. Each character can only join a single guild. A player can of course leave a guild and join another one at any time. For an enjoyable game experiences, players usually try to find a guild that matches their own playing style (note again that not all players try to achieve the same goals, e.g. acquiring treasures).

III. DATA AND FEATURES

We gathered a vast amount of player/guild logs from `http://www.warcraftrealms.com`. The logs show for a certain number of dates the records of currently online players from European and United States WORLD OF WARCRAFT® realms. In addition to the player's name, level, class, and guild membership are recorded. In total, we gathered 192 million recordings of 18 million characters belonging to 1.4 million guilds. The recordings cover a period of 4 years, starting in 2005 (when WORLD OF WARCRAFT® was released) and ending in early 2009. The data we recorded (roughly) summarizes the social in-game activities of the players. that is to say, we know when players

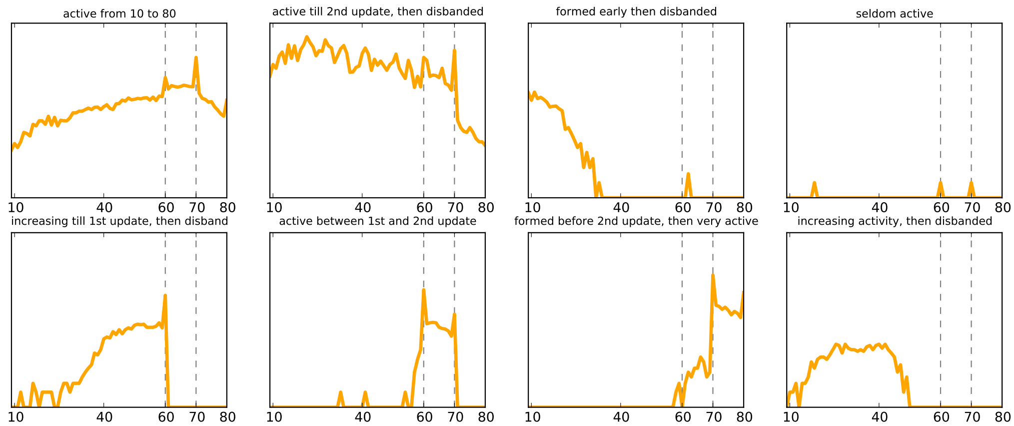


Fig. 4: Basis vectors resulting from the application of CH-NMF to the WORLD OF WARCRAFT® guild database. The x-axis denotes the level histogram bin, the y-axis denotes the number of observations for this bin. These guilds represent *archetypal* guilds and are automatically extracted from 1.4 million guilds.

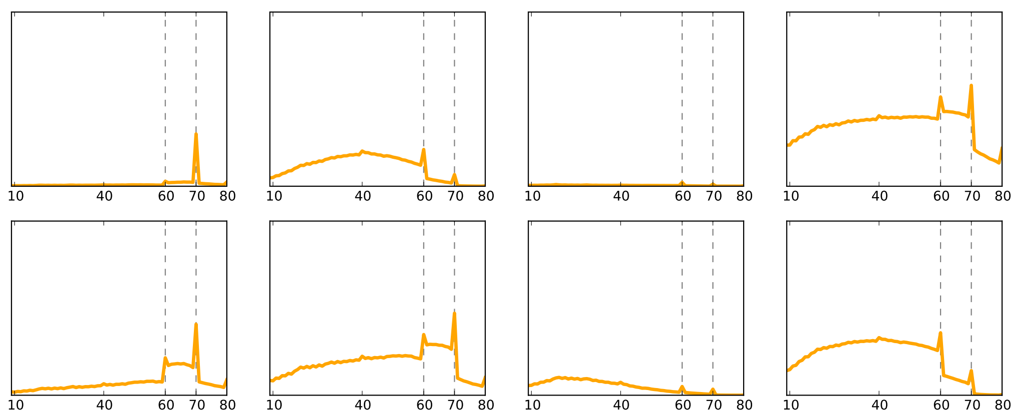


Fig. 5: Clusters resulting from the application of k-means clustering to the WORLD OF WARCRAFT® guild database. Cluster centroids correspond to average guilds. Overall, these centroids are not readily interpretable. In contrast to the CH-NMF results, k-means produces centroids of very similar characteristics, differences between guilds are not apparent.

left or joined a guild, how many players were with a guild at what time, and how experience levels were distributed among the members of a guild.

As mentioned before, the player's experience level provides a measure of the skill of a particular player. While a guild with more higher level players is more likely to be successful, a guild of only low level players is basically excluded from a large amount of the game content. The distribution of experience levels among guilds, i.e. the number of players of a certain level that are with a particular guild, therefore provides a feature that characterizes a guild in terms of game success.

The distribution can be efficiently approximated by means of building a histogram over experience levels of guild members (see Figure 4 and Figure 5 for examples). If we build these histograms over all observations of a specific period of time, they also summarize the temporal evolution of a guild. A brief example should clarify this: if a guild is newly formed by level 80 players, it does not contain any observations of level 10 players and the corresponding histogram bin would be empty. A guild which was formed by level 10 players should, over a longer period, also have observations of level 40, 60, 80 (and intermediate levels) players, as the guild members usually increase their level over time.

IV. RESULTS

Extracting meaningful information from a massive amount of data is clearly a non trivial task. Especially, if it is not entirely clear what to look out for. In these situations data mining resembles the proverbial search for a needle in a haystack. The results we present in the following were mostly obtained from applying Convex-Hull Non Negative Matrix Factorization (CH-NMF). This technique has recently been introduced as an efficient approach towards finding meaningful lower dimensional data embeddings by means of constrained matrix factorization. For the rather involved technical details, we refer the reader to the Appendix.

We applied CH-NMF to 1.4 million guild histograms that contain data covering a period of 4 years. We aim at the following goals: **(a)** an interpretable categorization, **(b)** analysis and visualization of the development of guilds over time, **(c)** a comparative study between US and EU WORLD OF WARCRAFT® guilds.

(a) Obtaining an interpretable categorization of guilds is certainly not a straight forward task. We applied CH-NMF to the guild histograms using 8 basis vectors. Note that we also tried different numbers of basis vectors. For a larger number visualization is getting more and more difficult, for a smaller number we could not capture the data variability at the desired resolution. We found that 8 basis vectors provide a convenient tradeoff between granularity and visualizability.

The 8 basis vectors can be seen in Figure 4. Following the definitions of CH-NMF, each basis vector resides on the convex hull of all guild histograms and thereby represents an *archetypal* guild. This makes the basis vectors easy to interpret as there is usually only one salient characteristic. In our case, the archetypal guilds are clearly distinguishable from each other. One can also recognize that a wide variety of guilds can be sufficiently formed by a convex combination of the archetypal guilds. Especially the raise of the level cap (from level 60 to 70 and level 70 to 80) that was introduced with the 1st and 2nd content update can be easily seen.

The interpretability of CH-NMF bases becomes even more obvious when compared to conventional clustering methods. Figure 5 shows 8 cluster centroids resulting from k-means clustering. K-means clusters s.t. common data regions are represented by their average. While this is often a reasonable way of analyzing data, it does not necessarily lead to interpretable results. In this case, the cluster centroids represent the data by (mostly) the same characteristic curve just differently scaled.

(b) Figure 7 and Figure 8 shows a projection of all guilds into the space spanned by the CH-NMF basis vectors (see also Figure 4) for the EU and US realms respectively. The projection is performed for different time spans (90 days, 180 days, 1 year, 2 years, 3 years, 4 years). It is important to note that the plots show an 8 dimensional space visualized in a 2D plane. While there are of course certain problematic aspects of this kind of visualization, it still preserves the main characteristics of the guild space.

The first thing to notice is that the total number of guilds increases considerably over time. Also, with more and more guilds to observe, a huge part of the guild space is densely covered. Interestingly, most guilds fall into the category of `seldom active` guilds (this could also indicate very small guilds) or are close to it. There is only a small number of guilds (still many thousands though) that completely fall into another category. One explanation could be the increase in the total number of guilds (this includes already disbanded guilds) over time which is shown in Figure 6. It can be seen the growth rate is (roughly) exponential. This indicates that the new formation of guilds is rather normal, as they are also abandoned often. For future work, it might be interesting to have a closer look at the life cycle of a guild.

(c) We compare US and EU realms and guilds by inspecting the distribution in guild space (Figure 7 and Figure 8) and also cumulative sum over the coefficients for each basis vectors. The cumulative sum over the coefficients serves as an indicator how much each basis vector contributes to the overall convex reconstruction of CH-NMF (how important is each basis vector for the complete guild space?). Figure 9 shows how the coefficient contributions change over time for EU and US realms respectively. Note that we computed the archetypal guilds (CH-NMF basis vectors) from the complete list of EU guilds. Interestingly, we could find similar archetypal guilds among the US guilds, however, we simply had to decide for one side for computing the embeddings.

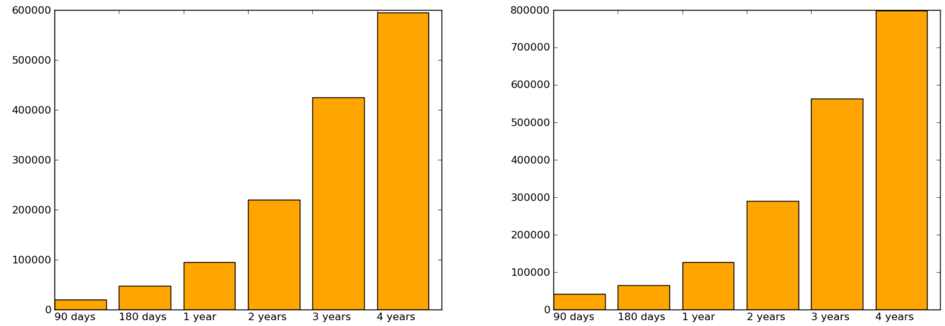
Considering both indicators, we could not find any significant difference in the average development of guilds from either the EU or the US.

V. CONCLUSION

This paper provides a first large-scale study to understand group formation processes in MMORPGs. To best of our knowledge, this is the first time that a vast amount of data on human behavior was used for analyzing or categorizing social behavior in MMORPGs. We were interested in identifying different types of guilds of players, in analyzing the evolution of guilds over time, and in comparing the development of US and EU based guilds. Applying Convex-hull NMF, we found the following archetypal guilds "active from 10 to 80", "active till 2nd update, then disbanded", "formed early then disbanded", "seldom active", "increasing till 1st update, then disband", "active between 1st and 2nd update", "formed before 2nd update, then very active", "increasing activity, then disbanded". Interestingly, we found a strong tendency towards more casual types of guilds. The vast majority of guilds in our study closely resemble archetypes representing guilds of non-professions ambitions. Our study of the temporal evolution of guilds revealed that from the release of the game onwards most players joined guilds that did never evolve towards professional competition. This was found for American and European guilds alike, and we were not able to discover cultural differences. For the design of future games, our conclusion at this point is that for commercial success it is of major importance to cater to the needs of the casual gamer.

APPENDIX

This appendix details our approach to analyzing massive amounts of in-game data. The method of *convex-hull non-*



(a) EU guild numbers

(b) US guild numbers

Fig. 6: Evolution of the total number of guilds over time in WORLD OF WARCRAFT® . Interestingly, for both, the US and the EU realms, the temporal development is surprisingly similar and follows a power law distribution.

negative matrix factorization was specifically developed to be able to mine data sets consisting of millions of observations in a timely manner. We shall motivate it and discuss its benefits for our purpose.

A. *Definitions and Notation*

In the following, vectors are denoted by bold lower case letters ($\boldsymbol{v}$) and their entries are denoted using subscripted lower case italics (v_k). $\mathbf{0}$ is the vector of all zeros and $\mathbf{1}$ is the vector of all ones. We write $\boldsymbol{v} \succeq \mathbf{0}$ if $v_k \geq 0$ for all k. The inner product of two vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ is written as $\boldsymbol{u}^T\boldsymbol{v}$. Consequently, $\mathbf{1}^T\boldsymbol{v}$ is a shorthand for $\sum_k v_k$.

Matrices are written using bold upper case letters ($\boldsymbol{M}$). If the columns of a matrix are known, we also write $\boldsymbol{M} = [\boldsymbol{m}_1\boldsymbol{m}_2 \ldots \boldsymbol{m}_n]$ where $\boldsymbol{m}_j \in \mathbb{R}^m$ is the jth column vector of $\boldsymbol{M}$.

We may identify a discrete set of vectors $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\}$ with the matrix $\boldsymbol{V} = [\boldsymbol{v}_1 \ldots \boldsymbol{v}_n]$ whose columns are given by the elements of the set. Moreover, we use $\|\boldsymbol{M}\|$ to denote the *Frobenius norm* of $\boldsymbol{M}$.

A set $\mathcal{S} \subset \mathbb{R}^m$ is *convex*, if every point on the line segment between any two points in $\mathcal{S}$ is also in $\mathcal{S}$. A vector $\boldsymbol{v} \in \mathbb{R}^m$ is a *convex combination* of vectors $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_l \in \mathbb{R}^m$, if $\boldsymbol{v} = \sum_i \lambda_i \boldsymbol{v}_i$ where $\lambda_i \geq 0$ and $\sum_i \lambda_i = 1$. Using matrix notation, we write convex combinations as $\boldsymbol{v} = \boldsymbol{V}\boldsymbol{\lambda}$ where $\boldsymbol{V} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_l]$ and $\boldsymbol{\lambda} \in \mathbb{R}^l$ such that $\mathbf{1}^T\boldsymbol{\lambda} = 1$ and $\boldsymbol{\lambda} \succeq \mathbf{0}$. An *extreme point* of a convex set $\mathcal{S}$ is any point $\boldsymbol{v} \in \mathcal{S}$ that is not a convex combination of other points in $\mathcal{S}$. The *convex hull* $\mathcal{C}$ of a set $\mathcal{S} \subset \mathbb{R}^m$ is the set of all convex combinations of points in $\mathcal{S}$. A *polytope* is the convex hull of finitely many points, i.e. it is the set $\mathcal{C}(\mathcal{S})$ for $|\mathcal{S}| < \infty$. The extreme points of a polytope are called *vertices*. We use $\mathcal{V}(\mathcal{S})$ to denote the set of all vertices of a polytope. Note that every point inside a polytope can be expressed as a convex combination of the points in $\mathcal{V}$.

B. *Clustering and NMF*

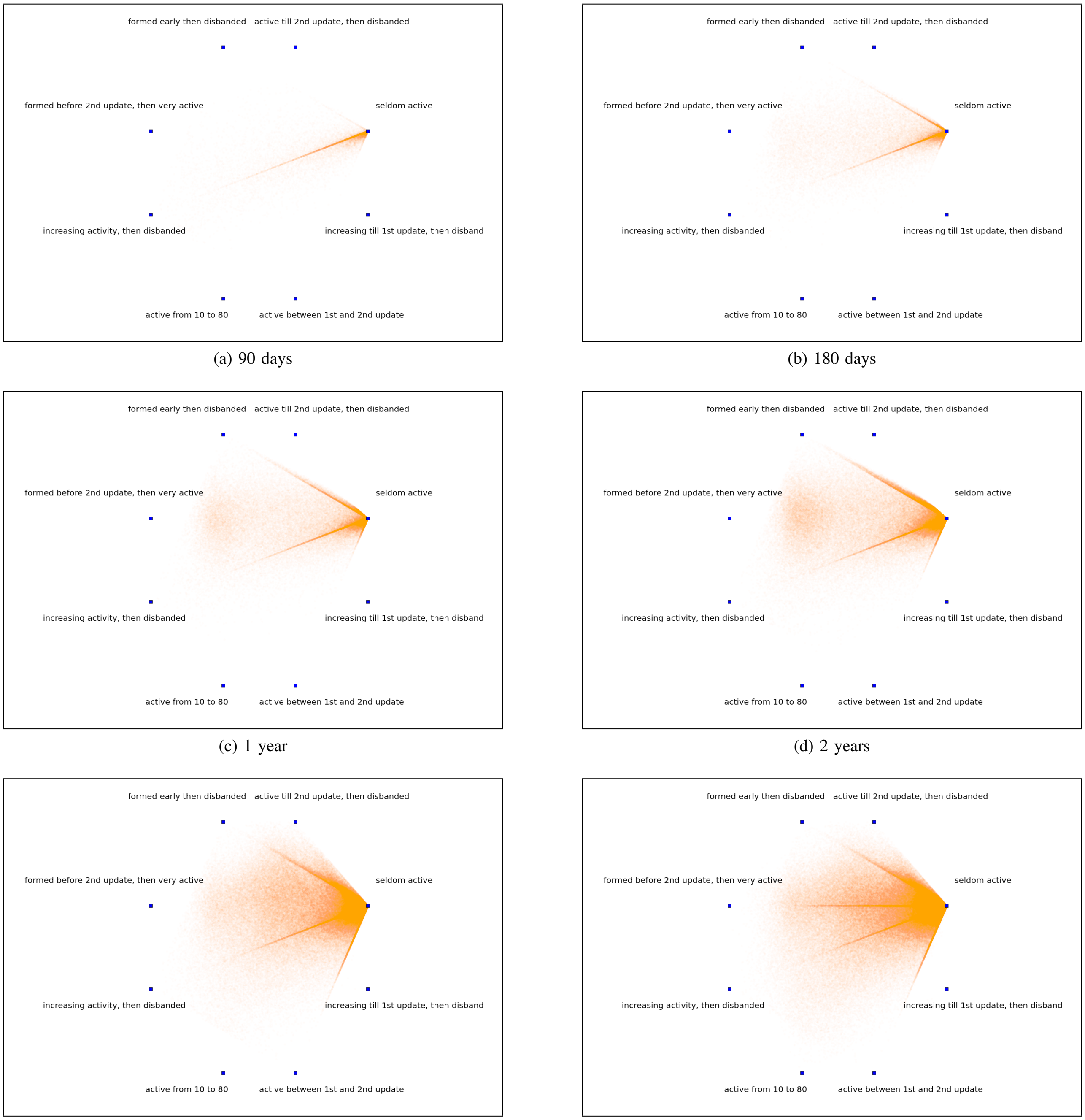
Matrix factorization is a fundamental step in many approaches to data mining, machine learning, and statistical pattern analysis. Recent work in machine learning has focused on matrix factorizations which obey particular constraints that are inherent to certain data and therefore should be accounted for in any analysis. In particular, non-negative matrix factorization (NMF) focuses on the analysis of data matrices whose elements are non-negative, a common occurrence in representations, for example, text or images data. Given a non-negative input matrix $\boldsymbol{V}$, NMF aims at determining two non-negative matrix factors $\boldsymbol{W}$ and $\boldsymbol{H}$ s.t.

$$\boldsymbol{V} \approx \boldsymbol{W}\boldsymbol{H}.$$

Convex non-negative matrix factorization (C-NMF) approaches additionally restrict the columns of $\boldsymbol{W}$ to be convex combinations of the data points gathered in $\boldsymbol{V}$, in order to enforce $\boldsymbol{W}$ to represent meaningful "cluster centroids". Thereby it contrasts agglomerative (single-linkage), mean-based (k-means), or mode seeking (mean-shift) clustering methods which rather search for representations by similarity. Meaningful centroids prove to be beneficial in applications such as text or genome mining, as well as image or social network analysis. Application of C-NMF on vast data is often not feasible and requires approximate methods such as Convex-Hull NMF.

C. *Convex-Hull NMF*

Convex-Hull NMF aims at a factorization that incorporates data points that reside on the data convex hull. The resulting representation has two interesting properties: first, the basis vectors are real data points and correspond to extremes rather than to averages. Second, any data point can be expressed as a convex combination of these basis elements. As convex combinations are nothing else but percentaged fractions of extremes (consider, for instance, a 70% chance of sunshine and a 30% chance of rain), they often offer intuitive interpretability [7].



(a) 90 days

(b) 180 days

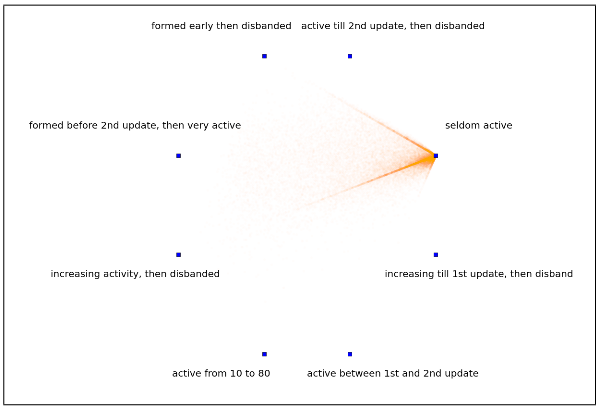
(c) 1 year

(d) 2 years

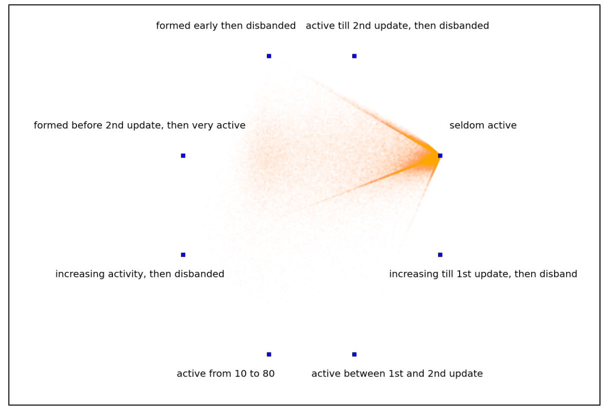
(e) 3 years

(f) 4 years

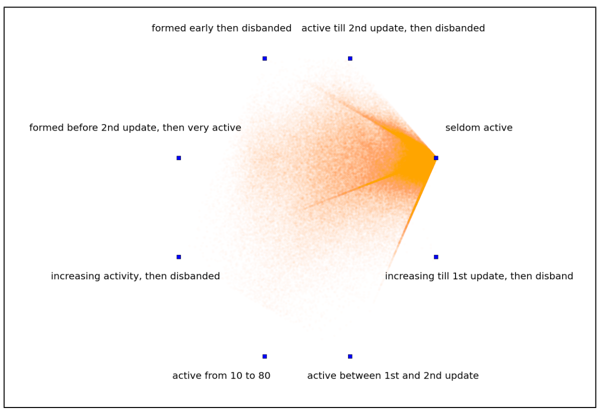
Fig. 7: Embeddings of European WORLD OF WARCRAFT® guilds and their development over time. Not only does the number of active and newly formed guilds increases dramatically, it can also be seen that more and more diverse guilds evolve. Especially in Figure 7c and Figure 7d, a clear separation between the very active and more casual guilds can be seen. Interestingly, the gap between these more professional and more casual guilds later fills continuously as can be seen in Figure 7e and Figure 7f.



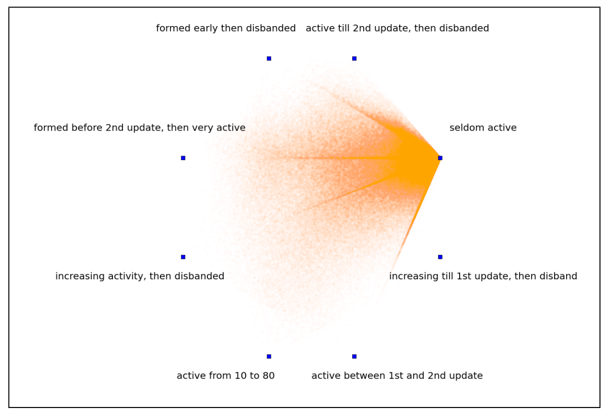
(a) 90 days



(b) 1 year

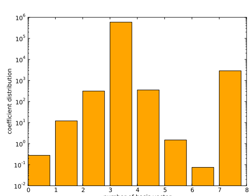


(c) 3 years

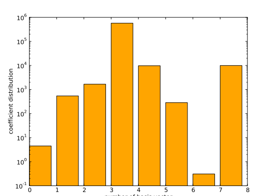


(d) 4 years

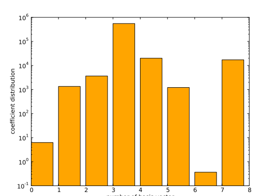
Fig. 8: Embeddings of US WORLD OF WARCRAFT® guilds and their development over time. Compared to the EU guilds the differences are negligible.



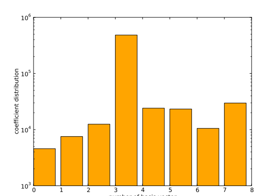
(a) EU: 90 days



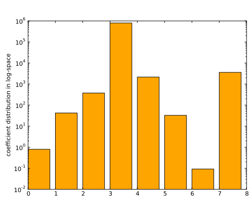
(b) EU: 1 year



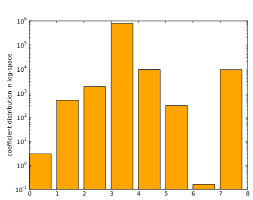
(c) EU: 2 years



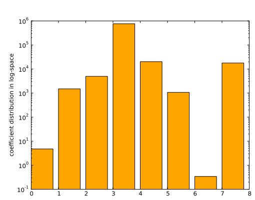
(d) EU: 4 years



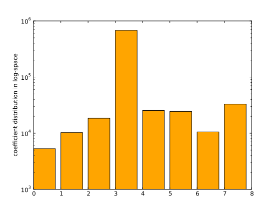
(e) US: 90 days



(f) US: 1 year



(g) US: 2 years



(h) US: 4 years

Fig. 9: Distribution of coefficient vectors over the 8 basis vectors (see Figure 4) for the EU and US Realms. Each bar represents how many guilds fall into the specified category or, more precisely, how much the corresponding basis vector is contributing to the overall reconstruction. Apparently, the overall distribution changes considerably over time.

We consider a factorization of the form

$$\boldsymbol{V} = \boldsymbol{V}\boldsymbol{G}\boldsymbol{H},$$

where $\boldsymbol{V} \in \mathbb{R}^{m\times n}, \boldsymbol{G} \in \mathbb{R}^{n\times k}, \boldsymbol{H} \in \mathbb{R}^{k\times n}$. If we restrict the columns of $\boldsymbol{G}$ and $\boldsymbol{H}$ to convexity, we obtain a factorization where each data point can be expressed as a *convex combination of convex combinations* of specific data points.

Our task is thus to solve the following constrained quadratic problem

$$\begin{aligned} \text{minimize} \quad & J = \left\|\boldsymbol{V} - \boldsymbol{V}\boldsymbol{G}\boldsymbol{H}\right\|^2 \\ \text{subject to} \quad & \mathbf{1}^T \boldsymbol{g}_i = 1,\ \boldsymbol{g}_i \succeq \mathbf{0} \\ & \mathbf{1}^T \boldsymbol{h}_j = 1, \boldsymbol{h}_j \succeq \mathbf{0} \end{aligned} \tag{1}$$

for $\boldsymbol{G}$ and $\boldsymbol{H}$. To facilitate our discussion, we set

$$\boldsymbol{X} = \boldsymbol{V}\boldsymbol{G}$$

so that $\boldsymbol{X} \in \mathbb{R}^{m\times k}$. Accordingly, due to the above properties of the coefficient matrix $\boldsymbol{G}$, the column vectors in $\boldsymbol{X}$ are convex combinations of columns in $\boldsymbol{V}$. Hence, the convex hull $\mathcal{C}(\boldsymbol{V})$ of $\boldsymbol{V}$ must contain $\boldsymbol{X}$. We could therefore achieve a perfect factorization of the data matrix by choosing the columns of $\boldsymbol{G}$ such that they would single out the vertices of $\mathcal{C}(\boldsymbol{V})$, i.e. such that they contain exactly one entry equal to 1 for each data point that is a vertex of the convex hull while all other entries were set to zero. Therefore, our goal becomes to solve Eq. (1) by finding k appropriate vertices of the convex hull. In other words, we aim at solving

$$\begin{aligned} \text{minimize} \quad & J = \left\|\boldsymbol{V} - \boldsymbol{X}\boldsymbol{H}^T\right\|^2 \\ \text{subject to} \quad & \boldsymbol{x}_i \in \mathcal{V}(\boldsymbol{V}),\ i = 1, \ldots, k. \end{aligned} \tag{2}$$

Solving problem (2) is not necessarily straight forward. Rather, it is known that the worst case complexity for computing the set of vertices $\mathcal{V}(\boldsymbol{V})$ of the convex hull of n data points in m dimensions is $\Theta(n^{\frac{m}{2}})$.

We therefore consider an approximate solution that subsamples the convex hull. Our approach exploits the fact that any data point on the convex hull of a linear lower dimensional projection of the data also resides on the convex hull in the original data dimension. Since $\boldsymbol{V}$ contains finitely many points and therefore forms a polytope in $\mathbb{R}^m$, we can resort to the main theorem of polytope theory which states that every image of a polytope P under an affine map $\pi : \boldsymbol{x} \rightarrow \boldsymbol{M}\boldsymbol{x} + \boldsymbol{t}$ is a polytope. In particular, every vertex of an affine image of P, i.e., every vertex of the convex hull of the image of P, corresponds to a vertex of P. For a proof of this important result, we refer the reader to [8].

Computing the vertices of the convex hulls of several 2D affine projections of the data therefore offers a way of subsampling $\mathcal{V}(\boldsymbol{V})$. Moreover, it is an efficient way of doing so since computing the extreme points of a set of 2D points can be done in $O(n \log n)$ time [9]. This allows us to approximate the convex hull of $\boldsymbol{X}$ as the union of points found on convex hulls of different 2D projections of the data. We project the data onto the $\frac{h(h-1)}{2}$ 2D subspaces spanned by pairwise combinations of the first h eigenvectors of the covariance matrix of $\boldsymbol{V}$ where h is chosen such that the first h eigenvectors account for 95% of the data variation. The mean and covariance matrix of $\boldsymbol{V}$ can be computed iteratively and the resulting matrices of size $m \times m$ and can be efficiently stored. Estimates of the resulting number of sampled points can be obtained from a result in [10] which states that the expected size of the convex hull of n Gaussian data points in the plane is $\Omega(\sqrt{\log n})$. For Gaussian data, we thus expect to sample $p = j\sqrt{\log n}$ points; for data that can be approximated using a mixture of q Gaussian, we expect to sample $p = jq\sqrt{\log(n/q)}$ data points. In both cases, the candidate set grows much slower than n.

Give a candidate set $\boldsymbol{S} \subset \mathcal{V}(\boldsymbol{V})$ containing p vertices of the data convex hull, we now select those $k < p$ vertices that yield the best reconstruction of the remaining points in $\boldsymbol{S}$. Since $\boldsymbol{S}$ is a $m \times p$ matrix, this, too, can be formulated as a constrained NMF optimization problem

$$\begin{aligned} \text{minimize} \quad & J_S = \left\|\boldsymbol{S} - \boldsymbol{S}\boldsymbol{I}\boldsymbol{J}\right\|^2 \\ \text{subject to} \quad & \mathbf{1}^T \boldsymbol{i}_i = 1,\ \boldsymbol{i}_i \succeq \mathbf{0} \\ & \mathbf{1}^T \boldsymbol{j}_j = 1,\ \boldsymbol{j}_j \succeq \mathbf{0} \end{aligned} \tag{3}$$

where $\boldsymbol{I} \in \mathbb{R}^{p\times k}$ and $\boldsymbol{J} \in \mathbb{R}^{k\times p}$. Since $p \ll n$, common quadratic programming routines solve (3) efficiently.

Once a suitable $\boldsymbol{I} \in \mathbb{R}^{p\times k}$ has been determined, the matrix $\boldsymbol{X}$ in Eq. (2) can be written as $\boldsymbol{X} = \boldsymbol{S}\boldsymbol{I}$ which guarantees that the problem in Eq. (2) is solely concerned with k data points on the convex hull of $\boldsymbol{V}$. We found that $\boldsymbol{I}$ usually results in unary representations. If this is not the case, we simply map $\boldsymbol{S}\boldsymbol{I}$ to their nearest neighboring data point in $\boldsymbol{S}$.

Given $\boldsymbol{X}$, the computation of the coefficients $\boldsymbol{H}$ may be done by minimizing $J_i = \|\boldsymbol{v}_i - \boldsymbol{X}\boldsymbol{h}_i\|^2, \mathbf{1}^T\boldsymbol{h}_i = 1, \boldsymbol{h}_i \succeq \mathbf{0}$ using common solvers.

References

[1] B. Weber and M. Mateas, "A data mining approach to strategy prediction," in *Proc. Int. Conf. on Computational Intelligence and Games (CIG)*, 2009.

[2] A. Drachen, A. Canossa, and G. N. Yannakakis, "Player modeling using self-organization in tomb raider: Underworld," in *Proc. Int. Conf. on Computational Intelligence and Games (CIG)*, 2009.

[3] C. Bauckhage, C. Thurau, and G. Sagerer, "Learning Human-like Opponent Behavior for Interactive Computer Games," in *Proc. Pattern Recognition Symposium of the German Association for Pattern Recognition (DAGM)*, 2003.

[4] C. Thurau, C. Bauckhage, and G. Sagerer, "Learning Human-Like Movement Behavior for Computer Games," in *Proc. International Conference on the Simulation of Adaptive Behavior (SAB)*, 2004.

[5] M. A. Ahmad, B. Keegan, J. Srivastava, D. Williams, and N. Contractor, "Mining for gold farmers: Automatic detection of deviant players in mmogs," in *Proc. Int. Conf. on Computational Science and Engineering*, 2009.

[6] N. Ducheneaut and N. Yee, "Collective solitude and social networks in world of warcraft," in *Social Networking Communities and E-dating Services: Concepts and Implications*, 2008.

[7] C. Thurau, K. Kersting, and C. Bauckhage, "Convex non-negative matrix factorization in the wild," in *Proc. Int. Conf. on Data Mining*, 2009.

[8] G. Ziegler, *Lectures on Polytopes*. Springer, 1995.

[9] M. de Berg, M. van Kreveld, M. Overmars, and O. Schwarzkopf, *Computational Geometry*. Springer, 2000.

[10] I. Hueter, "Limit Theorems for the Convex Hull of Random Points in Higher Dimensions," *Trans. of the American Mathematical Society*, vol. 351, no. 11, pp. 4337–4363, 1999.