Metric Indexing for Efficient Data Access in the Internet of Things

Christian Beecks*‡, Alexander Grass‡ and Shreekantha Devasya‡
*University of Münster, Germany
*christian.beecks@uni-muenster.de
‡Fraunhofer Institute for Applied Information Technology FIT, Germany
‡{christian.beecks,alexander.grass,shreekantha.devasya}@fit.fraunhofer.de

***Abstract*—Data are a central phenomenon in our digital information age. They impact the way we live, work, and play and provide unprecedented opportunities to simplify our daily life and behavior. They implicate enormous potential and impact society, economy, and science. Due to the advancement of cyber-physical systems and Internet of Things technologies, it is expected that the majority of real-time data will be generated from devices interconnected within the Internet of Things by the year 2025. In this paper, we tackle the problem of managing Internet of Things data in an efficient way. To this end, we introduce the metric approach for storing and querying Internet of Things data and investigate the ability of pivot-based tables for indexing and searching this type of data. Along with the introduction of two real-world, large-scale Internet of Things datasets from the EU projects COMPOSITION and MONSOON (under grant no. 723145 and 723650), we show that the metric approach facilitates efficient data access in the Internet of Things.**

***Index Terms*—Internet of Things, Big Data Management, Metric Indexing**

I. Introduction

Data are a central phenomenon in our digital information age. They impact the way we live, work, and play and provide unprecedented opportunities to simplify our daily life and behavior. They implicate enormous potential and impact society, economy, and science.

Digitizing data and algorithmizing their inherent information are among the great challenges of our time. The continuous data evolution is going to reach a volume of 163 zettabytes by the year 2025, that is ten times the volume of data generated in the year 2016 [1]. Not only the *volume* of data is supposed to increase, but also the *velocity*. Due to advanced and efficient technologies of embedded systems and Internet of Things, we are supposed to interact with digital devices several thousand times per day, generating a high *variety* of data in a massive scale. In addition, intelligent algorithmic analyses of historical and live data are predicted to be applied on a data volume of more than 5 zettabytes by the year 2025 [1].

The problem of managing and analyzing big data becomes even more obvious in the industrial sector. Nearly all industry segments are accelerating the adoption of cyber-physical systems and Internet of Things technologies in order to improve any kind of processes. Reinsel et al. [1] expect that the majority of real-time data will be generated from devices interconnected within the Internet of Things by the year 2025. While gaining insight into data acquired from the Internet of Things and managing the discovered knowledge in a structured way is a non-trivial issue that has been addressed differently in the literature, cf. Section II, we are focusing on managing Internet of Things data in an efficient way. Since this a core operation for nearly any algorithmic analysis, we believe that efficient access to Internet of Things data is of crucial importance for any further data-driven investigation.

In this paper, we propose the *metric approach* [2], [3] for storing and querying Internet of Things data. To this end, we investigate the family of *pivot-based tables* [4], which are considered to be the most fundamental metric access methods, and study their abilities of indexing and thereupon searching Internet of Things data. Along with our methodological investigation, we introduce two real-world, large-scale Internet of Things datasets acquired within the scope of two different EU projects COMPOSITION and MONSOON. These datasets will provide a novel standard for researchers and practitioners in the field of Internet of Things and will help to study and advance methods for big data management and analytics. To sum up, our contributions are two-fold:

- We introduce and investigate the metric approach to data from the Internet of Things.
- We propose and make available two real-world, large-scale Internet of Things datasets.

The paper is structured as follows. In Section II we outline related work, before we introduce the principles and approaches of metric indexing in Section III. The data sets are described in Section IV. The methodology and results of our performance study are given in Section V. We finally conclude this paper with an outlook on future research directions in Section VI.

II. Related Work

The automotive, consumer, health, and manufacturing sectors are among the most promising application segments of the Internet of Things. A general overview of enabling technologies, protocols, and applications is given in the work of Al-Fuqaha et al. [5], while Díaz et al. [6] provide an additional survey of platforms and infrastructures when integrating the Internet of Things with cloud-based solutions. An overview about the state of the art in data mining methods for the Internet of Things including classification, clustering, as well as association and time series analysis is given, for instance,

in the works of Tsai et al. [7] and Chen et al. [8]. A more data-centric overview about techniques and methods for the Internet of Things including data stream processing, data storage models, complex event processing, and searching is given in the work of Quin et al. [9]. The work of Abu-Elkheir et al. [10] highlights different data management solutions and design primitives. While the work of Fathy et al. [11] provides a holistic overview of the state of the art on indexing, discovery and ranking of Internet of Things data, the work of Kardeby et al. [12] is more focused on indexing methods for the Internet of Things.

Although recent research activities are tackling the challenge of efficient big data management and access in the Internet of Things, none of the approaches included in the aforementioned surveys are utilizing metric access methods. To what extent the metric approach facilitates efficient indexing and searching of Internet of Things data is investigated in this paper. For this purpose, we summarize the basic principles behind metric indexing in the next section.

III. Metric Indexing

Fundamental to all metric approaches is a metric space that abstracts from a concrete application domain. A metric space $(\mathbb{U}, \delta)$ comprises a data space $\mathbb{U}$ and a distance function $\delta : \mathbb{U} \times \mathbb{U} \rightarrow \mathbb{R}^+$ which satisfies the metric postulates for all elements $x, y, z \in \mathbb{U}$:

$$\begin{aligned} \delta(x,y) = 0 \Leftrightarrow x = y \quad & \text{(identity)} \\ \delta(x,y) \geq 0 \quad & \text{(non-negativity)} \\ \delta(x,y) = \delta(y,x) \quad & \text{(symmetry)} \\ \delta(x,y) \leq \delta(x,z) + \delta(z,y) \quad & \text{(triangle inequality)} \end{aligned}$$

Among these metric postulates, the triangle inequality is then frequently used to derive lower bounds of the exact distances between a query $q \in \mathbb{U}$ and a data object $o \in \mathbb{U}$ via a finite set of pivot elements $\mathbb{P} \subseteq \mathbb{U}$ as follows:

$$\delta_{\triangle}(q,o) := \max_{p \in \mathbb{P}} |\delta(q,p) - \delta(p,o)| \leq \delta(q,o).$$

Following the triangle lower bound and indexing the exact distances $\delta(o,p)$ of all data objects $o \in \mathbb{U}$ to the pivots elements $p \in \mathbb{P}$ gives us the first principle of metric indexing, namely *pivoting* [13]. The idea of precomputing and storing distances prior the query processing leads to the concept of a *pivot table*, which was originally introduced as *AESA* [14] respectively *LAESA* [15]. These approaches belong to the family of *pivot-based tables* [4], while the utilization of further metric principles, such as *ball partitioning* and *generalized hyperplane partitioning* [13], lead to more complex and often hierarchically organized structures. Overviews about fundamental metric principles and indexing techniques can be found for instance in the works of Hetland [13] and Chen et al. [4].

IV. Internet of Things Datasets

In this section, we introduce two real-world, large-scale Internet of Things datasets acquired within the scope of different EU projects COMPOSITION and MONSOON. The datasets and their characteristics are briefly described below.

A. COMPOSITION Dataset

The first IoT dataset is collected within the EU project COMPOSITION[1]. The goal of this project is to develop an integrated information management system which optimizes internal production processes by exploiting existing data, knowledge and tools to increase productivity and dynamically adapt to changing market requirements.

This IoT dataset is based on a production process of pacemakers. Within this production process, we collected time-annotated sensor measurements in combination with additional process information over a period of more than seven years. More specifically, we gathered the sensor measurements based on five minute intervals about the set and observed temperature as well as the power consumption of individual fans inside a reflow oven. The resulting dataset comprises 619,909 data records with 57 real-valued attributes.

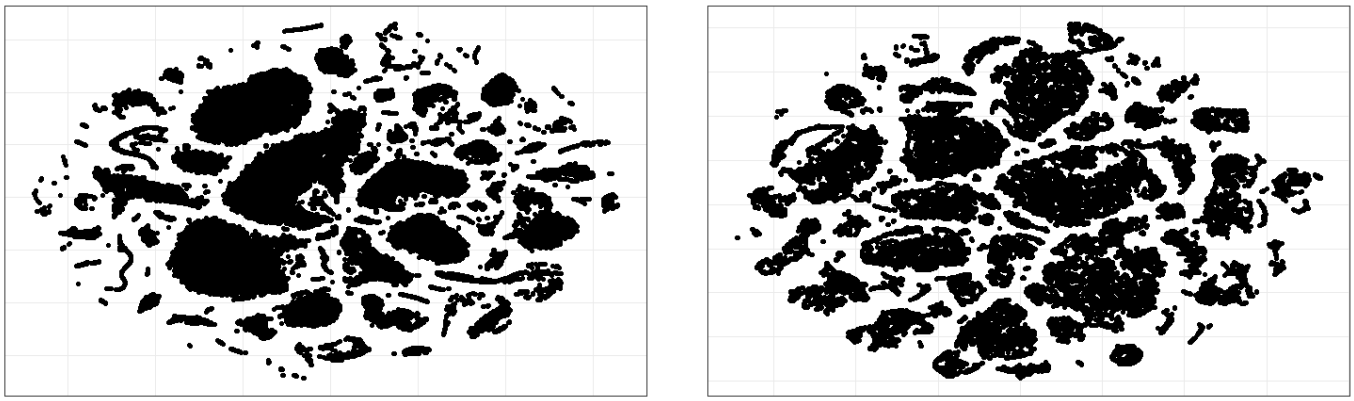
B. MONSOON Dataset

The second IoT dataset is collected within the EU project MONSOON[2]. The goal of this project is to establish a data-driven methodology to support identification and exploitation of optimization potentials by applying model-based predictive controls so as to perform plant and site-wide optimization of production processes.

The IoT dataset is based on a production process of coffee capsules, where the production is performed by an injection molding method. That is, the coffee capsules are produced by first heating the raw material and injecting it into a mold. After the holding pressure phase and the cooling phase the mold is opened again and the coffee capsules are extracted. The core of this process is the injection molding machine. We gathered sensor measurements about the machine's internal states such as temperature and pressure values as well as timings about the different phases within each production cycle. The resulting dataset comprises 357,383 data records with 16 real-valued attributes.

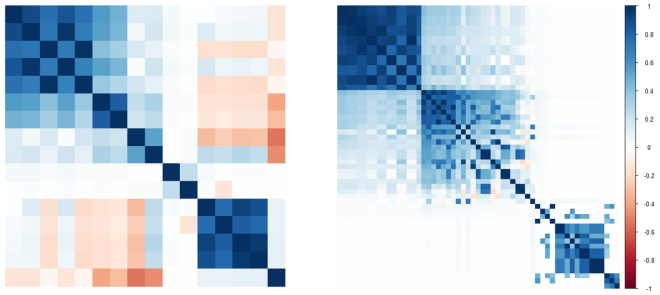
In order to give a first hint into the structure of both datasets, we sampled and visualized each dataset by means of the t-Distributed Stochastic Neighbor Embedding (t-SNE) [16] in Figure 1. This dimensionality reduction technique allows us to visualize high-dimensional data in a low-dimensional space by preserving the data's inherent structure. As can be seen in the figure, both datasets are showing a naturally inherent structure comprising several clusters and outliers. While the clusters arise due to the different production procedures and parameters of the machines, the outliers might indicate potential misbehavior of the machines and thus anomalies in

[1] www.composition-project.eu
[2] www.spire2030.eu/monsoon



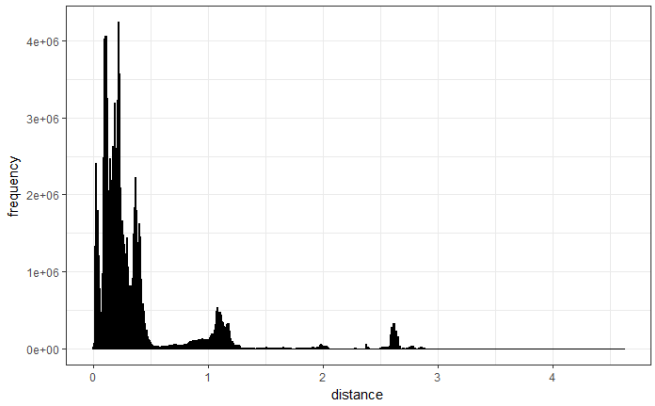
(a) COMPOSITION (b) MONSOON

Fig. 1. The introduced Internet of Things datasets at a glance.

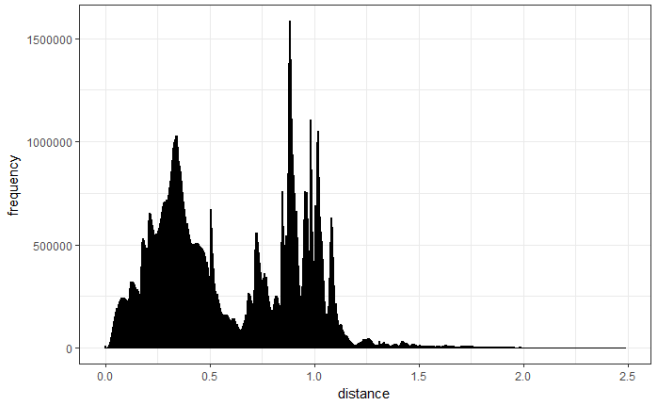


(a) COMPOSITION (b) MONSOON

Fig. 2. Correlation matrices of both datasets. Attributes are sorted according to the first principal component order.



(a) COMPOSITION



(b) MONSOON

Fig. 3. Distance distributions of both datasets.

production. Although these concrete structures are not the topic of our analysis, they imply a first hint on the indexability of the datasets. Since the data seems to be naturally grouped in both datasets, we expect a positive impact on the indexability. Before we investigate the indexability in Section V, we first investigate the correlation among the datasets' attributes. For this purpose, we evaluated the *Pearson correlation coefficient* between all pairs of attributes in order to determine their pairwise linear correlation.

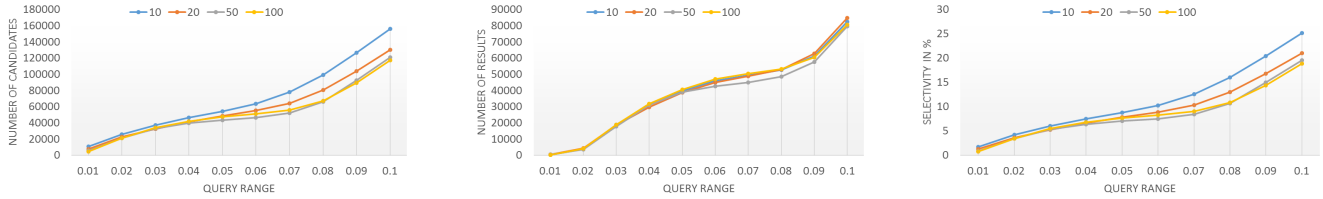
The results are visualized in Figure 2, where negative and positive correlations are indicated by reddish and bluish colors, respectively. The attributes are ordered according to the first principal component order. As can be seen in the figure, both datasets contain positively and negatively correlated attributes that indicate a grouping of attributes and partial redundancy of the sensor data measurements. As there are no completely dependent attributes, we include all attributes in our performance evaluation, whose results are given in the next section.

V. Experimental Results

In this section, we investigate the indexability of the proposed Internet of Things datasets by means of the metric approach and discuss the results of our empirical performance evaluation.

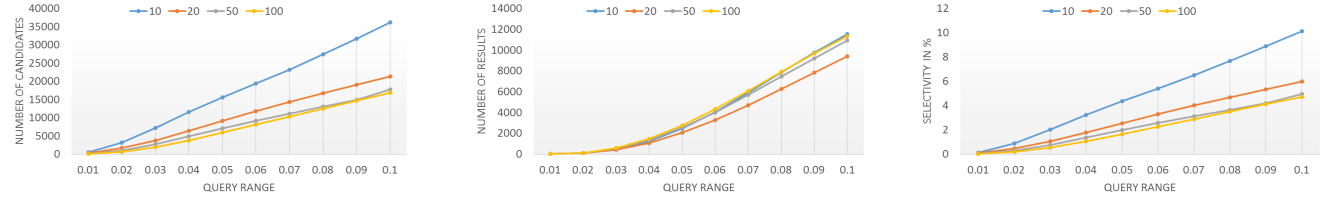
To this end, we first quantify the indexability of both datasets based on their distance distributions and the corresponding *intrinsic dimensionality* [17]:

$$\rho(\mathbb{U}, \delta) = \frac{\mathrm{E}[\delta(\mathbb{U}, \mathbb{U})]}{2 \cdot \mathrm{Var}[\delta(\mathbb{U}, \mathbb{U})]},$$



(a) (b) (c)

Fig. 4. Results for the COMPOSITION dataset: (a) number of candidates, (b) number of results, and (c) selectivity.



(a) (b) (c)

Fig. 5. Results for the MONSOON dataset: (a) number of candidates, (b) number of results, and (c) selectivity.

where $\mathrm{E}[\delta(\mathbb{U},\mathbb{U})]$ denotes the expected distance and $\mathrm{Var}[\delta(\mathbb{U},\mathbb{U})]$ denotes the variance of the distance within the data space $\mathbb{U}$. The intrinsic dimensionality $\rho \in \mathbb{R}$ reflects the indexability of a data distribution within a metric space $(\mathbb{U},\delta)$ by means of its distance distribution, cf. Figure 3. The lower the intrinsic dimensionality the better the indexability, and vice versa. Regarding the proposed IoT datasets, we measured the following intrinsic dimensionalities, both based on the Euclidean distance L_2:

- COMPOSITION: $\rho(\text{IoT-1}, \mathrm{L}_2) = 0.306$
- MONSOON: $\rho(\text{IoT-2}, \mathrm{L}_2) = 1.560$

The low intrinsic dimensionality of both datasets is also reflected in their distance distributions, which are shown in Figure 3. The distance distribution of the COMPOSITION dataset is more compact in comparison to the one of the MONSOON dataset. This compactness leads to a smaller intrinsic dimensionality and thus a potentially higher indexability. Whether this observation holds empirically when processing range queries, is investigated in the next series of experiments.

In order to study the indexability of both datasets with respect to range query processing, we make use of the metric approach *LAESA*. For this purpose, we randomly selected between 10 and 100 pivot elements and indexed both datasets with the Euclidean distance correspondingly. We have decided for a simple pivot-based table since we are aiming at investigating the general indexability of both Internet of Things datasets, not the highest overall performance of the underlying metric approach. The results presented in this paper are averaged over a query workload of 100 randomly chosen queries. All methods are implemented in the programming language *R* and are evaluated on a single-core 2.6 GHz machine with 16 GB of main memory, without parallelization.

The results in terms of *number of candidates* necessary to compute the final results of a range query, the *number of results*, and the *selectivity* of a query are shown as a function of the query range in Figure 4 and Figure 5 for the COMPOSITION and MONSOON datasets, respectively. As can be seen in the figures, both datasets show the same tendency. Increasing the query range yiels a larger number of candidates and thus an increase in the result size. Therefore the selectivity increases. In addition, by increasing the number of pivot elements from 10 to 100, the number of candidates, and thus the selectivity, decreases. Regarding the indexability, we observe that both datasets show an extraordinary high performance. Processing selective range queries with less than 1,000 retrieved data objects is achieved with a selectivity of smaller than 2% in both datasets. We thus conclude that the metric approach is a suitable indexing model for both IoT datasets.

VI. Conclusions and Future Work

In this paper, we have investigated the indexability of Internet of Things data by means of metric access methods. For this purpose, we have introduced two real-world, large-scale Internet of Things datasets acquired within the scope of two different EU projects COMPOSITION and MONSOON. We have briefly analyzed the datasets' inherent structures and benchmarked their indexability potential. Although both datasets are of high-dimensionality, we were able to index the IoT data and process range queries efficiently by a simple pivot-based table. The results of our performance evaluation indicate the high potential of the metric approach for indexing and searching Internet of Things data efficiently at large scale.

As future work, we plan to investigate more advanced metric access methods as well as ptolemaic access methods [18] in order to improve the efficiency of query processing even further. In addition, we aim to investigate adaptive and thus more expressive distance-based similarity models [19] such as the Signature Quadratic Form Distance [20] and the Signature

Matching Distance [21] in combination with different multi-step query processing algorithms [22], [23].

ACKNOWLEDGMENTS

The projects underlying this paper have received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 723145 (COMPOSITION) and No 723650 (MONSOON). This paper reflects only the authors' views and the commission is not responsible for any use that may be made of the information it contains.

REFERENCES

[1] D. Reinsel, J. Gantz, and J. Rydning, "Data age 2025: The evolution of data to life-critical," *Don't Focus on Big Data*, 2017.

[2] H. Samet, *Foundations of multidimensional and metric data structures*. Morgan Kaufmann, 2006.

[3] P. Zezula, G. Amato, V. Dohnal, and M. Batko, *Similarity search: the metric space approach*. Springer Science & Business Media, 2006, vol. 32.

[4] L. Chen, Y. Gao, B. Zheng, C. S. Jensen, H. Yang, and K. Yang, "Pivot-based metric indexing," *PVLDB*, vol. 10, no. 10, pp. 1058–1069, 2017.

[5] A. I. Al-Fuqaha, M. Guizani, M. Mohammadi, M. Aledhari, and M. Ayyash, "Internet of things: A survey on enabling technologies, protocols, and applications," *IEEE Communications Surveys and Tutorials*, vol. 17, no. 4, pp. 2347–2376, 2015.

[6] M. Díaz, C. Martín, and B. Rubio, "State-of-the-art, challenges, and open issues in the integration of internet of things and cloud computing," *J. Network and Computer Applications*, vol. 67, pp. 99–117, 2016.

[7] C. Tsai, C. Lai, M. Chiang, and L. T. Yang, "Data mining for internet of things: A survey," *IEEE Communications Surveys and Tutorials*, vol. 16, no. 1, pp. 77–97, 2014.

[8] F. Chen, P. Deng, J. Wan, D. Zhang, A. V. Vasilakos, and X. Rong, "Data mining for the internet of things: Literature review and challenges," *IJDSN*, vol. 11, pp. 431 047:1–431 047:14, 2015.

[9] Y. Qin, Q. Z. Sheng, N. J. Falkner, S. Dustdar, H. Wang, and A. V. Vasilakos, "When things matter: A survey on data-centric internet of things," *Journal of Network and Computer Applications*, vol. 64, pp. 137 – 153, 2016.

[10] M. Abu-Elkheir, M. Hayajneh, and N. A. Ali, "Data management for the internet of things: Design primitives and solution," *Sensors*, vol. 13, no. 11, pp. 15 582–15 612, 2013.

[11] Y. Fathy, P. Barnaghi, and R. Tafazolli, "Large-scale indexing, discovery, and ranking for the internet of things (iot)," *ACM Computing Surveys (CSUR)*, vol. 51, no. 2, p. 29, 2018.

[12] V. Kardeby, U. Jennehag, and M. Gidlund, "Surveying indexing methods for the internet of things," in *Internet of Things. IoT Infrastructures - Second International Summit, IoT 360° 2015*, ser. Lecture Notes of the Institute for Comp. Sc., Social Informatics and Telecommunications Engineering, vol. 170, 2015, pp. 282–291.

[13] M. L. Hetland, *The Basic Principles of Metric Indexing*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 199–232.

[14] E. V. Ruiz, "An algorithm for finding nearest neighbours in (approximately) constant average time," *Pattern Recognition Letters*, vol. 4, no. 3, pp. 145–157, 1986.

[15] M. L. Micó, J. Oncina, and E. Vidal, "A new version of the nearest-neighbour approximating and eliminating search algorithm (aesa) with linear preprocessing time and memory requirements," *Pattern Recognition Letters*, vol. 15, no. 1, pp. 9–17, 1994.

[16] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[17] E. Chávez, G. Navarro, R. A. Baeza-Yates, and J. L. Marroquín, "Searching in metric spaces," *ACM Comput. Surv.*, vol. 33, no. 3, pp. 273–321, 2001.

[18] M. L. Hetland, T. Skopal, J. Lokoc, and C. Beecks, "Ptolemaic access methods: Challenging the reign of the metric space model," *Inf. Syst.*, vol. 38, no. 7, pp. 989–1006, 2013.

[19] C. Beecks and T. Seidl, "On stability of adaptive similarity measures for content-based image retrieval," in *MMM*, ser. Lecture Notes in Computer Science, vol. 7131. Springer, 2012, pp. 346–357.

[20] C. Beecks, M. S. Uysal, and T. Seidl, "Signature quadratic form distance," in *CIVR*. ACM, 2010, pp. 438–445.

[21] C. Beecks, S. Kirchhoff, and T. Seidl, "Signature matching distance for content-based image retrieval," in *ICMR*. ACM, 2013, pp. 41–48.

[22] C. Beecks and A. Graß, "Multi-step threshold algorithm for efficient feature-based query processing in large-scale multimedia databases," in *BigData*. IEEE, 2016, pp. 596–605.

[23] C. Beecks and M. Berrendorf, "Optimal k-nearest-neighbor query processing via multiple lower bound approximations," in *BigData*. IEEE, 2018, accepted.