Database Group Leipzig
Department of Computer Science

UNIVERSITÄT LEIPZIG

Research Report 2018/2019

https://dbs.uni-leipzig.de

Overview

Database Group in June 2019. r.t.l.: Prof. Dr. Erhard Rahm, Christopher Rost, Andre Valdestilhas, Jonas Kreusch, Daniel Baumgarten, Martin Grimmer, Dr. Eric Peukert, Dr. Christian Martin, Ziad Sehili, Elias Saalmann, Martin Franke, Florens Rohde, Alaa Ashour, Daniel Alaya, Victor Christen, Markus Reinisch, Georges Alkhouri, Moritz Wilke, Caroline Mösler, Prof. Dr. Andreas Thor, Matthias Täschner, Marcus Stelzer, Dr. Ying-Chi Lin, Markus Nentwig.

1 Staff

Prof. Dr. Rahm, Erhard	Professor
Hesse, Andrea	Secretary
Alkhouri, Georges	Research associate (BMWI)
Christen, Victor	Research associate
Franke, Martin	Research associate
Gladbach, Marcel (until Sept. 2018)	Research associate (BMBF)
Gomez, Kevin (since Apr. 2018)	Research associate
Grimmer, Martin	Research associate (BMBF)
Kricke, Matthias (until Sept. 2018)	Research associate (BMBF)
Dr. Lin, Ying-Chi	Research associate (DFG)
Dr. Martin, Christian (since Aug. 2019)	Postdoctoral Researcher
Nentwig, Markus (until Apr. 2019)	Research associate (DFG)
Obraczka, Daniel (since Oct. 2018)	Research associate
Pogany, Gergely (since Dec. 2019)	Research associate
Dr. Peukert, Eric	Postdoctoral Researcher (BMBF)
Rohde, Florens (since Oct. 2018)	Research associate (BMBF)
Rost, Christopher (since Mar. 2018)	Research associate
Dr. Rostami, Mohammad Ali (until Feb. 2019)	Research associate (BMBF)
Saeedi, Alieh	Ph. D. student
Schuchart, Jonathan (since Oct. 2019)	Research associate
Sehili, Ziad	Research associate
Täschner, Matthias (since Feb. 2019)	Research associate
Valdestilhas, Andre (Nov. 2018-Oct. 2019)	Ph.D. student
Wilke, Moritz (since Jan. 2018)	Research associate
Prof. Dr. Thor, Andreas (HfTL Leipzig)	Associated team member
Dr. Zschache, Johannes	Postdoctoral Researcher

2 Highlights

There have been several highlights in 2018 and 2019:

1. The kick-off of the AI center ScaDS.AI took place in November 2019 at Leipzig University. ScaDS.AI (Center for Scalable Data Analytics and Artificial Intelligence) Dresden/Leipzig is one of the six centers for Artificial Intelligence (AI) funded within the AI Strategy of the Federal Government. In ScaDS.AI, both Big Data and AI methods will be researched and transferred into scientific and economic applications, making ScaDS.AI also a research and transfer center for data science. The funding for ScaDS and its extension to ScaDS.AI will initially be provided by the Federal Ministry of Education and Research (BMBF) until 2022 with a volume of approx. 12 million Euros. Prof. Rahm is the project leader and coordinator of ScaDS.AI at the Univ. of Leipzig.

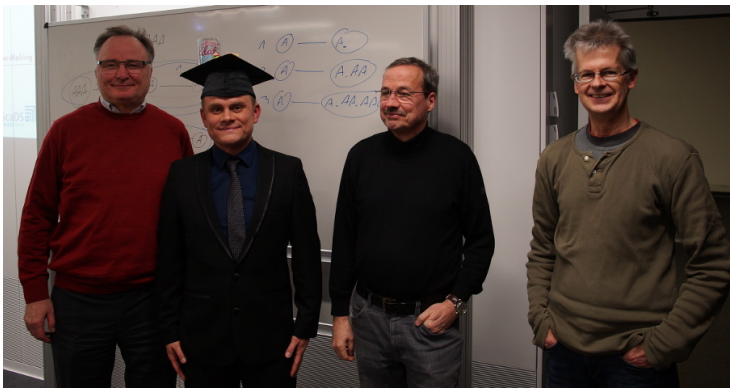
2. Prof. Rahm has initiated a new master degree course of study on Data Science. After the rectorate has approved the study program in 2019, it will start in April 2020

3. The following guests visited the database group: Peter Christen (ANU), Qing Wang (ANU), Daniel Ayala (Univ. de Sevilla). Prof. Rahm visited the Australian National Univ. (ANU) in March/April 2019

4. In May 2018, the traditional Zingst research seminar of the database group took place for the 16th time at the Leipzig University branch in Zingst / Baltic Sea.

Kickoff of ScaDS.AI (Nov. 2019), from left: Rector Prof. Dr. Schücking, Rector Prof. Dr. Müller-Steinhagen, State Minister Dr. Stange, Prof. Dr. Erhard Rahm, Prof. Dr. Wolfgang E. Nagel, Vice Rector Prof. Lenk

5. André Petermann successfully defended his Ph.D. thesis
6. The FAMER paper "Using Link Features for Entity Clustering in Knowledge Graphs" received the Best Research paper award of the Int. Extended Semantic Web Conference (ESWC) in 2018
7. The research prototypes GRADOOP and PRIMAT could be presented at the VLDB conferences 2018 (in Rio de Janeiro) and 2019 (in Los Angeles), respectively
8. From June 30th to July 6th 2018, the Big Data Center ScaDS (Competence Center for Scalable Data Services and Solutions) Dresden/Leipzig has hosted its second international summer school for Big Data and Machine Learning in Leipzig. About 70 participants attended it.
9. Prof. Rahm has been re-elected as member of the DFG review board for computer science and will serve in that capacity for the period 2020-2024. Moreover, he has been selected as a member of the expert committee of the National Research Data Initiative NFDI.
10. A significant number of new third-party funded projects, including (DE4L, VIP, TWIN, GRAMMY, could be secured in the period under review and will be described below. The projects BIGGR and EXPLOIDS have successfully been ended.



Dissertation defence of André Petermann (Jan. 2019).

Research Seminar in Zingst (May 2018).

Participants of the international Big Data summer school in Leipzig (July 2018).

3 Research Topics and Projects

ScaDS.AI - Center for Scalable Data Analytics and Artifical Intelligence

E. Peukert, C. Martin, D. Obraczka, J. Schuchart, M. Täschner, M. Wilke, E. Rahm

The „Competence Center for Scalable Data Services and Solutions Dresden/Leipzig (ScaDS)“ lead by Prof. Nagel from the TU Dresden and Prof. Rahm from the University of Leipzig started in 2014 as a nationwide competence center for Big Data in Germany. After a successful evaluation in 2018, the center was expanded in November 2019 to become one of the German centers for artificial intelligence (AI), which is funded as part of the federal government's AI strategy. This expanded center is called ScaDS.AI (Center for Scalable Data Analytics and Artificial Intelligence) Dresden/Leipzig. The project is funded by the Federal Ministry of Education and Research (FKZ: 01IS18026B) and is to be further strengthened by the Free State of Saxony with the establishment of 4 new AI professorships at both locations. In basic research on AI methods, the center strives to bridge the gap between the efficient use of mass data, advanced AI methods and knowledge management. In addition to new methods of machine learning and artificial intelligence, the focus is also on research topics on trust, protection of privacy, transparency, protection of minorities and traceability of AI-driven decisions.

The research is running at two locations, Dresden and Leipzig, by the partners Dresden University of Technology, Leipzig University, Max Planck Institute for Molecular Cell Biology and Genetics, Leibniz Institute for Ecological Spatial Planning, Helmholtz Center for Environmental Research, Leipzig and the Helmholtz Center Dresden Rossendorf.

ScaDS has become a success story at the University of Leipzig and its partner institutions. About 120 scientific publications were published and more than 200 keynotes and talks were presented by ScaDS members in the first there years (2014-2017). Together with application partners many challenges were solved which partly resulted in a number of Big Data Service that are offered by ScaDS. Through ScaDS a number of industry contacts were established and many collaborations were started which partly already lead to further research projects that run in close collaboration with ScaDS.AI.

The database group of Prof. Rahm is involved in several ScaDS projects, in particular related to graph analytics (Gradoop) and data integration including privacy-preserving record linkage. Within ScaDS.AI the database group will conduct research on graph-based data integration and analysis of dynamic graphs, privacy preserving data integration, machine learning for evolving graph data as well as privacy-preserving machine learning.

`http://www.scads.ai`

BIGGR – Big Graph Data Analysis Workflows

Ali M. Rostami, S. Dienst, M. Wilke, M. Täschner E. Peukert, E. Rahm

The analysis of big and network-structured data is a current trend in different fields like biological or social networks. This data, which can be interpreted as graphs, is central to many fields for extracting different information. Typical processes are data import, integration, transformation, analysis of corresponding graphs, and finally the visualization with the goal of identifying the relations and influences of data. However, the classical databases are not flexible enough and do not have a suitable support of analysis workflows and algorithms. Also, the modeled dependencies and parameters in the analysis can not be supported. On the other hand, the existing graph databases are very technical for the analysis big graphs by end users.

The goal of BIGGR was to develop a new software system for user-friendly and efficient analysis and visualization of big graphs. The project was successfully completed at the end of May 2019. The system developed in the process can be used without deep knowledge. More clearly, the graph analysis workflows can be defined and executed graphically from simple basic operators. Then, the user can see a graphical view of the results at the end. For this purpose, the KNIME Analytics platform and the Gradoop framework, both available as open source systems from both partners of this project, were specifically adapted, expanded and combined. In addition, the system can be easily extended with new operators, execution target systems and visualization techniques. Practical suitability was evaluated on the basis of different application cases. Being open-source makes the results widely usable for data analysts in Germany and worldwide. The corresponding plug-in for the KNIME Analytics Platform is available via this link: `https://www.knime.com/biggr`.

Interactive tool for visual exploration of big graph data (SAB project)

C. Rost, K. Gomez, E. Rahm

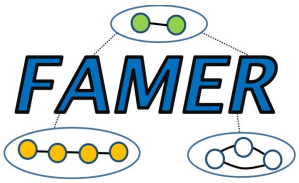
The analysis of highly connected data as graphs becomes more and more important in many different domains. Prominent examples are social networks, e.g., Facebook and Twitter, as well as information networks like the World Wide Web or biological networks. One important similarity of these domain specific data is their inherent graph structure which makes them eligible for analytics using graph algorithms. Besides that, the datasets share two more similarities: they are huge in size, making it hard or even impossible to process them on a single machine and they are heterogeneous in terms of the objects they represent and their attached data. With the objective of analyzing these large-scale, heterogeneous graphs, we continue developing a framework called "Gradoop" (Graph Analytics on Hadoop). Gradoop is built around the so called Extended Property Graph Model (EPGM) which supports not only single but also collections of heterogeneous graphs and includes a wide range of combinable operators. These operators allow the definition of complex analytical programs as they take single graphs or graph collections as input and result in either of those. Gradoop is build on top of the distributed dataflow framework Apache Flink, and makes use of the provided APIs to implement the EPGM and its operators. The system is publicly available (www.gradoop.com) and gets code contributions from other institutes and companies. A demo application showing the usage and resultset of the graph grouping and pattern matching operator, has been implemented and the corresponding articles published.

In 2018, a two-year cooperation has started with the industry partner TIQ Solutions GmbH, Leipzig. The project "Development of an interactive tool for the visual analysis of very large graph data", funded by the Sächsische Aufbau Bank (SAB), has focused on the interactive analysis of networked data for business intelligence. The researchers and associate students of our institute are continuously developing the Gradoop system by improving performance and features, as well as providing an extensive interface to use all developed operators. The project partner focused on a explorative user interface and the visualization of the analytical results, created by Gradoop. The API of the system has been extensively expanded as the basis for the interaction of the frontend and backend.

In our ongoing work we focus on the processing and analysis of temporal graphs, i.e., graphs that change over time, with continuous addition and removal of vertices and edges, as well as frequent changes of their attributes. We already started with the development of several extensions to the graph data model as well as the operators. Graph snapshot retrieval, i.e., accessing the past state of a graph, the search and discovery for temporal patterns with a well defined chronological order and comparative analysis that compare the evolution of properties are typical examples of analyses that focus on the additional time-domain of the graph.

Distributed Large-Scale Graph Data Integration

A. Saeedi, D. Obraczka, E. Peukert, E. Rahm



Nowadays knowledge graphs lay the foundation for a diverse range of applications, such as e.g. Question Answering. These knowledge graphs usually are constructed from multiple sources across different domains. This necessitates high-quality data integration tools and due to the vast number of data needed in modern application these integration tools need to be scalable as well. To address this need we provide the framework FAMER (FAst Matching and Entity Resolution). FAMER supports the matching and clustering of entities from many data sources and utilizes Apache Flink for a distributed execution. It supports different blocking and matching strategies to first determine a so-called similarity graph. Furthermore, several clustering schemes can be applied to determine clusters of matching entities from similarity graphs.
Knowledge graphs usually contain multiple entity types. This can be beneficial in the matching process since the neighborhood of an entity can be used to aid in the matching process. Furthermore, previous matching decisions of certain entity types can be used to determine duplicates of another entity type. Creating a framework that extends FAMER to match graphs with multiple entity types and using neighborhood information is an ongoing effort.

KOBRA: learning-based deduplication of customer data

G.Alkhouri, E. Peukert, E. Rahm

For businesses, well-maintained customer and partner data is often the most valuable asset they own. A key challenge to achieve a high qulity of this data is to identify and eliminate duplicates. The Uniserv GmbH from Pforzheim, is offering a sophisticatged rule-based system to find such duplicates in customer data. A key goal in the joint project KOBRA (Configuration of Business Rules for Users of Duplicate Detection Systems) is to simplify the configuration of this tool by applying learning-based methods. The project should help Data Stewards, Data (Quality) Analysts and Citizen Data Scientists to perform data preparation and tool configuration to a large extent automatically via easy-to-understand and agile self-service tools. In doing so, task and company-specific rules will be adapted to the specific problem by adding positive and negative samples given by the users. This, in turn shall be achieved by a combination of different machine learning techniques with training data selection, historization, reinforcement learning, and a simulation environment. Within KOBRA we successfully developed a learning-based approach for the automatic configuration of business rules in duplicate detection systems. Users of a matching system can now adapt and configure matching rules with given examples. The developed processes were evaluated and integrated in a prototype solution, which will be developed into a marketable product after the project has ended in 2020.

DE4L - Data Economy for advanced Logistics

M. Schneider, E. Peukert, E. Rahm

The DE4L project is pursuing the development of an intelligent ecosystem as part of a platform for data exchange for logistics service companies. This is to avoid high congestion on delivery vehicles, costs due to incorrect delivery and repeated delivery and pickup attempts. The so-called "last mile" of the supply chain, meaning the exact delivery and collection of parcels at the front door, offers a great deal of potential for increasing efficiency. With the platform DE4L strengthens the cooperation of the service companies and promotes the digitization of the information.

DE4L is a BMBF-funded cooperation (FKZ: 01MD19008D) with different Partners from the Logistics domain, the Fraunhofer IML and the Data-Science-Center ScaDS.AI. We are striving to build an innovative Blockchain/Distributed Ledger-based trading platform for sensor data from logistics. For this purpose sensors are developed and different types of data are recorded. Moreover, we build methods and techniques for privacy-preserving trading, machine learning and data exchange, especially of sensor and business partner data.

The project started in August 2019 and is running for three years.

`https://de4l.io/en/about-de4l/`.

TWIN - Transformation of complex product development processes into knowledge-based services for additive manufacturing

E. Peukert, E. Rahm

In TWIN, development processes of laser-based generative manufacturing (metal and plastic) and additive manufacturing processes are to be digitilized. For this purpose, TWIN is developing a digital product service system (with a digital twin as the core object) with the participation of the entire value chain of industrial additive manufacturing. Particular emphasis is placed on support for using machine learning processes.

The University of Leipzig is concerned with two main areas in the project: (1) integration and storage of heterogeneous sensor and process data as well as (2) the subsequent analysis and modeling. Data integration and analysis is implemented as an iterative process, i.e. the digital twin in this project progresses gradually and is expanded to include data sources and models.

The project is funded by the BMWi (FKZ: 02K18D055) and started in October 2019. It will run for three years.

GRAMMY - InteGRAtive analysis of tuMor, Microenvironment, immunitY and patient expectation for personalized response prediction in Gastric Cancer

G. Pogany, C. Martin, E. Rahm

Gastric cancer (CG) is a complex disease, the fifth most common malignant tumor in the world and the third leading cause of death from cancer. CG is very heterogeneous and affects twice as many men as women. Chemotherapy combined with surgery represents the standard of care for stage II to III CG, but the efficacy of such treatments is still limited for many patients. It is therefore imperative to develop an innovative approach aimed at identifying new predictive markers, including those deduced from taking into account the impact of the psychosocial and cultural environment of each patient. We defend the idea that the style of communication, the degree of acceptance of the treatment by the patient, as well as the doctor-patient interaction, can influence the response to treatment, with in particular differences in compliance. The integration of different levels of information, biological and psychosocial, is very promising, although it is particularly difficult, to identify the links between the specific biological characteristics of the disease, the patient's perception and the prognosis. The consortium consists of an number of European partners from Italy (lead), Greece and France.

The "GRAMMY" project is funded by the European ERA PerMed call (Antragsnummer-SAB: 100394103) and will run for 3 years until 2022/23. The database group is responsible for the data integration and is also supporting the analyses of heterogeneous medical data sources of the project.

Determining annotations in the life sciences

V. Christen, Y.-C. Lin, E. Rahm

The automatic annotation of real world objects with concepts from an ontology is an active field of research in the life sciences, e.g. to support a better data integration and analysis of electronic health records or clinical data. In our previous work, we investigated the annotation of medical forms typically used in clinical studies, e.g., forms asking for eligibility criteria (e.g. specific disease symptoms). Often there are many heterogeneous forms for similar topics impeding the integration of study results. To overcome such issues, it is a crucial aim to annotate medical forms with standardized vocabularies such as the Unified Medical Language System (UMLS). Therefore, we developed novel methods to (semi-) automatically annotate medical forms. However, automatic matching of form questions (items) is a complex task since questions are written in free text, use different synonyms for the same semantics and can cover several different medical concepts. Our annotation workflow includes several preprocessing steps, different linguistic match approaches and a novel group-based strategy to select the most promising concepts for annotating a question in the medical form.

We investigate the usage of neural networks for generating vector representations of textual information. Hence, a mention and a concept can be represented as vectors that are used as input for further neural networks. Initially, we reimplemented a CNN-based annotation approach and observed that the results are not completely reproducible. We implement an approach for generating embeddings using CNN-based autoencoders. The realized network also considers that the embeddings of two synonyms of one concept are similar in the vector space. The quality of embeddings was evaluated using feed-forward networks.

Furthermore, we transformed the annotation problem into a link prediction problem. Therefore, we used the name entities from documents as well as concepts as vertices and the relationships from an ontology as well as the concurrences of the named entities as edges. We investigate node embedding approaches that utilize structural as well as local information for generating node embeddings. The resulting node representations are used for link prediction.

ELISA - Evolution of Semantic Annotations

Y.-C. Lin, V. Christen, E. Rahm

Annotating documents or datasets using concepts of biomedical ontologies has become increasingly important. Such ontology-based semantic annotations can improve the interoperability and the quality of data integration in health care practice and biomedical research. For instance, PubMed, the search engine for MEDLINE database, uses MeSH (Medical Subject Headings) terms to retrieve more relevant results. The use of the hierarchy information within the ontologies can further expand the potential matches. Furthermore, annotating data across multiple disconnected databases using concepts from same ontologies enables data integration.

With the development of the medical knowledge, the ontologies are changing continuously. On the other hand, the documents to be annotated, such as medical forms, can also be revised into different versions or adapted into different languages. The ELISA (Evolution of Semantic Annotations) project aims to investigate the impacts of such changes, in both ontologies and the documents, on the semantic annotations. The project is a cooperation with the Luxembourg Institute of Science and Technology (LIST), the University of Paris-Sud.

We designed and implemented a (semi-)automatic approach to insure the validation of the semantic annotations when the underlying ontology is evolving. The maintenance framework considers rules that exploit the morphosyntactic form of terms denoting attribute values, such as split or merge. Secondly, it also includes further background knowledge such as additional biomedical terminologies to determine the correct update of the annotation. Finally, the framework adapts the new annotation

using Semantic Change Patterns that regards the lexical and semantic similarities of the terms.

To investigate the evolution of the documents being annotated, we target the cross-lingual annotation of the medical forms used in epidemiological studies. Many of these study forms are originated from English and in our case the forms are in German. We examine two main strategies to annotate such forms using relevant ontologies from UMLS: (1) using all available German ontologies to annotate the German forms and (2) integrating machine translators to translate German forms and annotate the translated corpus using English ontologies. The results show that using German ontologies only produces very restricted results, whereas translation achieves better annotation quality and is able to retain almost 70% of the annotations.

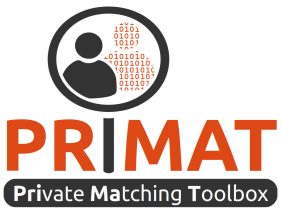
DAAD project: Advancing data integration: Privacy and semantics for record linkage

E. Rahm, Z. Sehili, V. Christen, A. Groß

This project is a joint research DAAD project with the Australian National University and the University Leipzig. The project focuses on privacy-preserving record linkage and advanced matching techniques. A general issue in record linkage is the classification of matching and non matching record pairs. To reduce the amount of manual labelling to generate suitable training data, we developed an active learning approach that does not require any prior knowledge about true matches and that is independent of the learning method used. Our approach successively identifies new training examples based on an informativeness measure for similarity vectors by considering their relationship to already classified vectors and the uncertainty in the similarity vector space covered by the current training set. Experiments on several data sets show that even for a small labelling effort our approach achieves comparable results to fully supervised approaches and it can outperform previous active learning approaches for entity resolution.

Privacy-preserving Record Linkage

Z. Sehili, M. Franke, M. Gladbach, F. Rohde, E. Rahm



Record linkage aims at linking records that refer to the same real-world entity, such as persons. Typically, there is a lack of global identifiers, therefore the linkage can only be achieved by comparing available quasi-identifiers, such as name, address or date of birth. However, in many cases, data owners are only willing or allowed to provide their data for such data integration if there is sufficient protection of sensitive information to ensure the privacy of persons, such as patients or customers. Privacy-preserving Record Linkage (PPRL) addresses this problem by providing techniques to securely encode and match records. By combining data from different sources data analysis and research can be improved significantly. The linkage of person-related records is based on encoded quasi-identifiers while the data needed for analysis, e.g., health data, is excluded from the linkage.

PPRL is confronted with several challenges needing to be solved to ensure its practical applicability. In particular, a high degree of privacy has to be ensured by suitable encoding of sensitive data and organizational structures, such as the use of a trusted linkage unit. PPRL must achieve a high linkage quality by avoiding false or missing matches. Furthermore, a high efficiency with fast linkage time and scalability to large data volumes are needed. A main problem for performance is the inherent quadratic complexity of the linkage problem when every record of the first source is compared with every record of the second source. For better efficiency, the number of comparisons can be reduced by adopting blocking or filtering approaches. Furthermore, the matching can be performed in parallel on multiple processing nodes.

In recent years, we have focused our research on improving the performance and scalability of PPRL workflows. As the records are represented as bit vectors (using Bloom filter for encoding), we used metric space similarity measures for filtering. In particular, the pivot-based approach for metric spaces utilizing the triangle inequality to reduce the search space showed significant improvement of performance compared to previous filter techniques. One data source is indexed by determining some records as pivots and assigning the leftover records to them. We can save many similarity computations by comparing the records of the second source with only the pivots first and exclude most records as possible matches.

In various experiments it became apparent that high linkage quality is often not achievable with current PPRL methods. Even for fine-tuned encoding and linkage parameters we experienced that precision drops quite drastically, in particular for datasets containing millions of records or many families or households. By analyzing the linkage result, we found that a single record is often linked to many records of the other source as they agree on certain attributes. For instance, all members of a family will very likely share their last name as well as their address. If we assume, that each source does not contain any duplicates or is deduplicated before the linkage process, we can utilize post-processing methods after the matching. Post-processing methods ensure a one-to-one match mapping, such that one record of the first source is matched to at maximum one record of the second source. Therefore, the similarity graph (match mapping) is analyzed and links are removed by certain heuristics. We investigated three well-known algorithms from graph theory for this problem, namely the Hungarian algorithm for calculation a maximum weight matching, the Gale-Shapley algorithm for solving the stable marriage problem as well as a heuristic best match strategy. We found that all methods can significantly improve linkage quality in terms of precision and thus f-measure. However, in most cases the best match selection strategy performed best closely followed by the Gale-Shapley algorithm which achieves better recall but a slightly lower f-measure due to lower precision values.

Additionally, we analyzed methods for multiparty PPRL considering applications, where more than two data owners are involved in a linkage process. The intuitive way is to link the sources sequentially such that in each iteration a new source is linked with the other already matched sources. The first challenge to solve is to keep the linkage process scalable for several large sources. To this end we devised a new method to index records from different source in a unique heterogeneous and dynamic pivot-based metric space. Hence, for each new source and in particular for each new records the algorithm checks if new pivots should be created to better distribute records on the pivots and reduce the number of similarity computations. The second challenge is to find subset matches, i.e., sets of matching records that are not in all sources, but in subsets of them. for this second challenge we investigated several method to post-process the result of the linkage process. some methods a run during the linkage after each iteration like the Hungarian algorithm or max-both. Other methods, however, take the final linkage result of all the sources as a similarity graph and try to delete from each connected component as much edges as possible to have clusters that contain at most one records from each source.

One application area for PPRL is medical research, since the investigation of many scientific questions is only possible by merging distributed patient data where privacy and data protection are essential requirements as medical information is very sensitive personal data. We therefore contribute our experience to the SMITH consortium within the Medical Informatics Initiative.

Furthermore we conducted an evaluation of the record linkage facilities of the so-called Mainzelliste, an open-source software for identity and pseudonym management of patients. The software is used by a growing number of medical joint research projects but has never been evaluated with respect to its linkage quality and runtime performance. Our results show that the tool achieves excellent linkage quality but has unacceptable runtimes for larger registries. Therefore we introduced established blocking methods for plain text matching and developed new blocking methods on encoded records that drastically improve the runtime while having none or only minor effect on the linkage quality. Our improvements were integrated into the official source code repository so that the users, e.g. the medical community, can benefit from it.

Finally, we developed a PPRL toolbox named PRIMAT (**Pri**vate **Ma**tching **T**oolbox) that includes our previously developed methods for fast and scalable PPRL based on the use of blocking, metric space filtering and parallel matching. Additionally, we included many state-of-the-art PPRL methods for Bloom-filter-based encoding. PRIMAT facilitates the adoption of PPRL in real-world applications by providing different components for the definition, execution and evaluation of tailored PPRL workflows. In the future, we plan to extend PRIMAT to support incremental linkage as well as multi-party PPRL. Besides, we plan to investigate the privacy and security properties of current PPRL encoding techniques.

VIP – Visual Product Matching

M. Wilke, E. Peukert, E. Rahm

A very useful application of record linkage techniques is the growing field of e-commerce. Linking product offers from different vendors allows to compare these and to gain valuable insight into the market. Unfortunately data from the web is very heterogeneous and not easy to integrate. E.g. an equal product can be described with a very differing level of detail, different attributes, description text and so on. Worse yet the description of two non-equal products can be identical (often when its verbose).

Recording linkages approaches as of today are largely based on columnar and textual data. However, online product data typically also consists of images and in some domains (e.g. fashion) this visual information is much more reliable and relevant for the user. The goal of the VIP project is to explore whether the additional information provided by the images can be used to improve the results of existing matching systems. To achieve this, a variety of image similarity metrics from computer vision and deep learning shall be investigated. Furthermore it is to examine how the image matching approaches can be integrated into current record linkage systems to allow the matching of heterogeneous, multi-modal data.

The VIP project started in October 2019, it is a joint project with the company Web Data Solutions. It is funded by the Sächsische Aufbaubank (SAB)

Exploids - Explicit Privacy-Preserving Host Intrusion Detection System

M. Grimmer, E. Peukert, E. Rahm

The research project Explicit Privacy-Preserving Host Intrusion Detection System (EXPLOIDS, https://www.exploids.de) aims to increase the security of virtual machines in data centers and cloud environments. The idea is to monitor a Linux guest system for attacks as well as to ensure that any traces of an attack are detected for legal clarification at a later date and to protect data privacy. Prof. Rahm's group is researching anomaly detection techniques for this application, where only knowledge about the normal behavior of the processes is brought in. This so-called one-class classification is thus different from the binary or multi-class classifications that are more widely used in the literature. With such methods, it is possible to detect previous unknown attacks. We started with graph-based approaches in combination with existing methods to increase recognition rates and reduce false alarm rates. By taking the inherent structure of the underlying data and its meta data into account, it is possible to gain more insights compared to other known methods. Later we also applied modern methods of machine learning to further improve the results. The project was funded by the BMBF (Förderkennzeichen: 16KIS0523) and ran from July 2016 until December 2019.

The first steps in the project analyzed known algorithms that analyze the sequences of the occurring system calls. They were evaluated on the latest publicly available data set (the ADFA-LD). We then developed algorithms that interpret the underlying sequence of the processes to be

monitored as a graph. Thus it was possible to calculate a probability for each change of state of the processes and to determine an anomaly score with the help of this probability.

These algorithms and most other scientific work in the field deal only with the sequences of the system calls without considering their diverse metadata, e.g. the executing processes, threads, users, timestamps, parameters and return values. Since this presumably helpful information is not available in existing evaluation datasets, we started to record a new dataset.The newly determined dataset for testing and evaluating host-based intrusion detection systems is called the Leipzig Intrusion Detection - Data Set (LID-DS)[1]. It was presented at the IT Security Congress of the Federal Office for Information Security (BSI).

With the LID-DS we have been able to develop better algorithms that use both the sequences of the system calls and their metadata. First experiments show that just taking the thread information into account leads to a significant improvement in false alarm rates by about a factor of 10 with the same detection rate. Our ongoing work aims at further improvements, especially by using the parameters and return values in combination with more advanced machine learning techniques such as autoencoders, CNNs or RNNs.

4 Publications and Theses

Book editorships

[1] Michael Böhlen, R. Pichler, Norman May, Erhard Rahm, S. Wu, and K. Hose. *Advances in Database Technology - Proc. 21th International Conference on Extending Database Technology EDBT, Vienna, Austria*. Mar. 2018.

[2] Torsten Grust, Felix Naumann, Alexander Böhm, Wolfgang Lehner, Theo Härder, Erhard Rahm, Andreas Heuer, Meike Klettke, and Holger Meyer, eds. *Datenbanksysteme für Business, Technologie und Web (BTW 2019), 18. Fachtagung des GI-Fachbereichs „Datenbanken und Informationssysteme" (DBIS), 4.-8. März 2019, Rostock, Germany, Proceedings*. Vol. P-289. LNI. Gesellschaft für Informatik, Bonn, 2019. ISBN: 978-3-88579-683-1. URL: https://dl.gi.de/handle/20.500.12116/21526.

Journal Publications

[1] Silvio Domingos Cardoso, Chantal Reynaud-Delaître, Marcos Da Silveira, Ying-Chi Lin, Anika Gross, Erhard Rahm, and Cédric Pruski. "Evolving semantic annotations through multiple versions of controlled medical terminologies." In: *Health and Technology* 8 (2018), pp. 361–376.

[2] Martin Franke, Marcel Gladbach, Ziad Sehili, Florens Rohde, and Erhard Rahm. "ScaDS research on scalable privacy-preserving record linkage." In: *Datenbank-Spektrum* 19.1 (2019), pp. 31–40.

[3] Johannes Frey, Kay Müller, Sebastian Hellmann, Erhard Rahm, and Maria-Esther Vidal. "Evaluation of metadata representations in RDF stores." In: *Semantic Web* 10.2 (2019), pp. 205–229. DOI: 10.3233/SW-180307. URL: https://doi.org/10.3233/SW-180307.

[4] Kevin Gómez, Matthias Täschner, M. Ali Rostami, Christopher Rost, and Erhard Rahm. "Graph Sampling with Distributed In-Memory Dataflow Systems." In: *CoRR* abs/1910.04493 (2019). arXiv: 1910.04493. URL: http://arxiv.org/abs/1910.04493.

[5] René Jäkel, Eric Peukert, Wolfgang E Nagel, and Erhard Rahm. "ScaDS Dresden/Leipzig–A competence center for collaborative big data research." In: *it-Information Technology* 60.5-6 (2018), pp. 327–333.

[1] https://www.exploids.de/lid-ds/

[6] Martin Junghanns, Max Kießling, Niklas Teichmann, Kevin Gómez, André Petermann, and Erhard Rahm. "Declarative and distributed graph analytics with GRADOOP." In: *PVLDB* 11.12 (2018), pp. 2006–2009. DOI: 10.14778/3229863.3236246. URL: http://www.vldb.org/pvldb/vol11/p2006-junghanns.pdf.

[7] Erhard Rahm and Theo Härder. "Editorial." In: *Datenbank-Spektrum* 19.1 (2019), pp. 1–3. DOI: 10.1007/s13222-019-00310-1. URL: https://doi.org/10.1007/s13222-019-00310-1.

[8] Erhard Rahm, Wolfgang E. Nagel, Eric Peukert, René Jäkel, Fabian Gärtner, Peter F. Stadler, Daniel Wiegreffe, Dirk Zeckzer, and Wolfgang Lehner. "Big Data Competence Center ScaDS Dresden/Leipzig: Overview and selected research activities." In: *Datenbank-Spektrum* 19.1 (2019), pp. 5–16. DOI: 10.1007/s13222-018-00303-6. URL: https://doi.org/10.1007/s13222-018-00303-6.

[9] Christopher Rost, Andreas Thor, and Erhard Rahm. "Analyzing Temporal Graphs with Gradoop." In: *Datenbank-Spektrum* 19.3 (2019), pp. 199–208. DOI: 10.1007/s13222-019-00325-8. URL: https://doi.org/10.1007/s13222-019-00325-8.

[10] M Ali Rostami, Matthias Kricke, Eric Peukert, Stefan Kühne, Moritz Wilke, Steffen Dienst, and Erhard Rahm. "BIGGR: Bringing GRADOOP to applications." In: *Datenbank-Spektrum* 19.1 (2019), pp. 51–60.

[11] Alieh Saeedi, Markus Nentwig, Eric Peukert, and Erhard Rahm. "Scalable matching and clustering of entities with FAMER." In: *Complex Systems Informatics and Modeling Quarterly* 16 (2018), pp. 61–83.

[12] Dinusha Vatsalan, Peter Christen, and Erhard Rahm. "Incremental Clustering Techniques for Multi-Party Privacy-Preserving Record Linkage." In: *CoRR* abs/1911.12930 (2019). arXiv: 1911.12930. URL: http://arxiv.org/abs/1911.12930.

[13] Alfred Winter, Sebastian Stäubert, Danny Ammon, Stephan Aiche, Oya Beyan, Verena Bischoff, Philipp Daumke, Stefan Decker, Gert Funkat, Jan E Gewehr, et al. "Smart medical information technology for healthcare (SMITH)." In: *Methods of information in medicine* 57.1 (2018).

Book Chapters and Conference/Workshop Publications

[1] Giacomo Bergami, André Petermann, and Danilo Montesi. "THoSP: an algorithm for nesting property graphs." In: *Proceedings of the 1st ACM SIGMOD Joint International Workshop on Graph Data Management Experiences & Systems (GRADES) and Network Data Analytics (NDA)*. 2018, pp. 1–10.

[2] Silvio Domingos Cardoso, Marcos Da Silveira, Ying-Chi Lin, Victor Christen, Erhard Rahm, Chantal Reynaud-Delaître, and Cédric Pruski. "Combining Semantic and Lexical Measures to Evaluate Medical Terms Similarity." In: *International Conference on Data Integration in the Life Sciences*. Springer. 2018, pp. 17–32.

[3] Victor Christen, Peter Christen, and Erhard Rahm. "Informativeness-Based Active Learning for Entity Resolution." In: *DINA Workshop*. 2019.

[4] Victor Christen, Ying-Chi Lin, Anika Groß, Silvio Domingos Cardoso, Cédric Pruski, Marcos Da Silveira, and Erhard Rahm. "A Learning-Based Approach to Combine Medical Annotation Results." In: *DILS*. 2018.

[5] Martin Franke, Ziad Sehili, Marcel Gladbach, and Erhard Rahm. "Post-processing methods for high quality privacy-preserving record linkage." In: *Data Privacy Management, Cryptocurrencies and Blockchain Technology*. Springer, 2018, pp. 263–278.

[6] Martin Franke, Ziad Sehili, and Erhard Rahm. "Parallel Privacy-preserving Record Linkage using LSH-based Blocking." In: *IoTBDS*. 2018, pp. 195–203.

[7] Martin Franke, Ziad Sehili, and Erhard Rahm. "PRIMAT: A toolbox for fast privacy-preserving matching." In: vol. 12. 12. 2019, pp. 1826–1829.

[8] Johannes Frey, Marvin Hofer, Daniel Obraczka, Jens Lehmann, and Sebastian Hellmann. "DBpedia FlexiFusion the Best of Wikipedia ¿ Wikidata ¿ Your Data." In: *The Semantic Web – ISWC 2019*. Ed. by Chiara Ghidini, Olaf Hartig, Maria Maleshkova, Vojtěch Svátek, Isabel Cruz, Aidan Hogan, Jie Song, Maxime Lefrançois, and Fabien Gandon. Cham: Springer International Publishing, 2019, pp. 96–112. ISBN: 978-3-030-30796-7.

[9] Kleanthi Georgala, Daniel Obraczka, and Axel-Cyrille Ngonga Ngomo. "Dynamic planning for link discovery." In: *European Semantic Web Conference*. Springer. 2018, pp. 240–255.

[10] Marcel Gladbach, Ziad Sehili, Thomas Kudrass, Peter Christen, and Erhard Rahm. "Distributed privacy-preserving record linkage using pivot-based filter techniques." In: *ICDEW*. IEEE. 2018, pp. 33–38.

[11] M. Grimmer, J. Hoffmann, and M. M. Röhling. "Exploids: Host-basierte Angriffserkennung auf Linux-VMs." In: *LINUX-MAGAZIN 03/2018*, pp. 74–77. URL: https://www.linux-magazin.de/ausgaben/2018/03/intrusion-detection/.

[12] M. Grimmer, M. M. Röhling, D. Kreusel, and S. Ganz. "A Modern and Sophisticated Host Based Intrusion Detection Data Set." In: *IT-Sicherheit als Voraussetzung für eine erfolgreiche Digitalisierung*. 2019, pp. 135–145. ISBN: 978-3-922746-82-9.

[13] M. Grimmer, M. M. Röhling, M. Kricke, B. Franczyk, and E. Rahm. "Intrusion Detection on System Call Graphs." In: *Sicherheit in vernetzten Systemen*. 2018, G1–G18. ISBN: 9783746086378.

[14] Matthias Kricke, Eric Peukert, and Erhard Rahm. "Graph Data Transformations in Gradoop." In: *Datenbanksysteme für Business, Technologie und Web (BTW 2019), 18. Fachtagung des GI-Fachbereichs „Datenbanken und Informationssysteme" (DBIS), 4.-8. März 2019, Rostock, Germany, Proceedings*. Ed. by Torsten Grust, Felix Naumann, Alexander Böhm, Wolfgang Lehner, Theo Härder, Erhard Rahm, Andreas Heuer, Meike Klettke, and Holger Meyer. Vol. P-289. LNI. Gesellschaft für Informatik, Bonn, 2019, pp. 193–202. DOI: 10.18420/btw2019-12. URL: https://doi.org/10.18420/btw2019-12.

[15] Markus Nentwig and Erhard Rahm. "Incremental Clustering on Linked Data." In: *2018 IEEE International Conference on Data Mining Workshops, ICDM Workshops, Singapore, Singapore, November 17-20, 2018*. Ed. by Hanghang Tong, Zhenhui Jessie Li, Feida Zhu, and Jeffrey Yu. IEEE, 2018, pp. 531–538. DOI: 10.1109/ICDMW.2018.00084. URL: https://doi.org/10.1109/ICDMW.2018.00084.

[16] Daniel Obraczka and Axel-Cyrille Ngonga Ngomo. "Dragon: Decision Tree Learning for Link Discovery." In: *Web Engineering*. Ed. by Maxim Bakaev, Flavius Frasincar, and In-Young Ko. Cham: Springer International Publishing, 2019, pp. 441–456. ISBN: 978-3-030-19274-7.

[17] Daniel Obraczka, Alieh Saeedi, and Erhard Rahm. "Knowledge Graph Completion with FAMER." In: *Proceedings of the 1st International Workshop on Challenges and Experiences from Data Integration to Knowledge Graphs co-located with the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD 2019)*. Ed. by Donatella Firmani, Valter Crescenzi, Andrea De Angelis, Xin Luna Dong, Maurizio Mazzei, Paolo Merialdo, and Divesh Srivastava. Vol. 2512. CEUR Workshop Proceedings. 2019. URL: http://ceur-ws.org/Vol-2512/paper1.pdf.

[18] Erhard Rahm and Eric Peukert. "Holistic Schema Matching." In: *Encyclopedia of Big Data Technologies*. Ed. by Sherif Sakr and Albert Y. Zomaya. Springer, 2019. DOI: 10.1007/978-3-319-63962-8_12-1. URL: https://doi.org/10.1007/978-3-319-63962-8%5C_12-1.

[19] Erhard Rahm and Eric Peukert. "Large Scale Entity Resolution." In: *Encyclopedia of Big Data Technologies*. Ed. by Sherif Sakr and Albert Y. Zomaya. Springer, 2019. DOI: 10.1007/978-3-319-63962-8_4-1. URL: https://doi.org/10.1007/978-3-319-63962-8%5C_4-1.

[20] Erhard Rahm and Eric Peukert. "Large-Scale Schema Matching." In: *Encyclopedia of Big Data Technologies*. Ed. by Sherif Sakr and Albert Y. Zomaya. Springer, 2019. DOI: 10.1007/978-3-319-63962-8_330-1. URL: https://doi.org/10.1007/978-3-319-63962-8%5C_330-1.

[21] Martin Max Röhling, Martin Grimmer, Dennis Kreußel, Jorn Hoffmann, and Bogdan Franczyk. "Standardized container virtualization approach for collecting host intrusion detection data." In: *2019 Federated Conference on Computer Science and Information Systems (FedCSIS)*. IEEE. 2019, pp. 459–463.

[22] Christopher Rost, Andreas Thor, and Erhard Rahm. "Temporal Graph Analysis using Gradoop." In: *Datenbanksysteme für Business, Technologie und Web (BTW 2019), 18. Fachtagung des GI-Fachbereichs „Datenbanken und Informationssysteme" (DBIS), 4.-8. März 2019, Rostock, Germany, Workshopband*. 2019, pp. 109–118. DOI: 10.18420/btw2019-ws-11. URL: https://doi.org/10.18420/btw2019-ws-11.

[23] M Ali Rostami, Alieh Saeedi, Eric Peukert, and Erhard Rahm. "Interactive Visualization of Large Similarity Graphs and Entity Resolution Clusters." In: *EDBT*. 2018, pp. 690–693.

[24] M. Ali Rostami, Eric Peukert, Moritz Wilke, and Erhard Rahm. "Big graph analysis by visually created workflows." In: *BTW 2019*. Ed. by Torsten Grust, Felix Naumann, Alexander Böhm, Wolfgang Lehner, Theo Härder, Erhard Rahm, Andreas Heuer, Meike Klettke, and Holger Meyer. Gesellschaft für Informatik, Bonn, 2019, pp. 559–563. DOI: 10.18420/btw2019-45.

[25] Alieh Saeedi, Eric Peukert, and Erhard Rahm. "Using Link Features for Entity Clustering in Knowledge Graphs." In: *Proc. ESWC 2018 (Best research paper award)*. Springer International Publishing, 2018, pp. 576–59.

[26] Andre Valdestilhas, Tommaso Soru, Markus Nentwig, Edgard Marx, Muhammad Saleem, and Axel-Cyrille Ngonga Ngomo. "Where is my URI?" In: *European Semantic Web Conference*. Springer. 2018, pp. 671–681.

Bachelor, Master and Ph.D. Theses

[1] Hans Angermann. "Entwicklung eines Werkzeug zur graphischen Datenmodellierung (ERM, RM)." B.Sc. Leipzig University, 2019.

[2] Daniel Baumgarten. "Distributed Graph Layouting Algorithm with GRADOOP." M.Sc. Leipzig University, 2019.

[3] Jan Buchholz. "Datenerfassung für die Untersuchung von Primaten." B.Sc. Leipzig University, 2019.

[4] Nico Duldhardt. "Tree-based Learning Methods for Match Classification with Apache Flink." B.Sc. Leipzig University, 2018.

[5] Katja Englert. "Evaluation and Comparison Big Data ETL Frameworks." M.Sc. Leipzig University, 2018.

[6] Simon Ganz. "Ein moderner Host Intrusion Detection Datensatz." M.Sc. Leipzig University, 2019.

[7] Johannes Geisler. "Erstellung und Evaluierung eines geeigneten Modells zur Betrugserkennung in Echtzeit mit Apache Spark." M.Sc. Leipzig University, 2018.

[8] David Geistert. "Visualisierung von Annotation-Mappings für klinische Formulare." B.Sc. Leipzig University, 2018.

[9] Kevin Gomez. "Speicheroptimierung beim verteilten Frequent Subgraph Mining mittels (Multi-)Pattern Matching." M.Sc. Leipzig University, 2018.

[10] Stephan Kemper. "Erweiterte Graph-Gruppierung mit Gradoop." M.Sc. Leipzig University, 2018.

[11] Alexander Kern. "Declarative Information Fusion for Graph Collections." M.Sc. Leipzig University, 2018.

[12] Julius Kluge. "Evaluierung von Workflowmanagementsystemen am Beispiel der Annotierung von medizinischen Fragebögen." B.Sc. Leipzig University, 2018.

[13] Michael Koch. "Integration von EAGLE in FAMER." B.Sc. Leipzig University, 2019.

[14] Maria Kömmpel. "Efficient Storage and Retrieval of Temporal Graph Data." M.Sc. Leipzig University, 2019.

[15] Dennis Kreußel. "Simulation and analysis of system call traces for adversial anomaly detection." B.Sc. Leipzig University, 2019.

[16] Steve Lehmann. "Evaluierungstool für eine interaktive Validierung von Record-Links." B.Sc. Leipzig University, 2019.

[17] Caroline Mösler. "Visualisierungskonzept für PPRL-Workflows." B.Sc. Leipzig University, 2018.

[18] André Petermann. "On Pattern Mining in Graph Data to Support Decision-Making." Dr. rer. nat. Leipzig University, 2019.

[19] Elias Saalmann. "An Evaluation of the EPGM based on Apache Flink's Table API." M.Sc. Leipzig University, 2019.

[20] Anna Schröder. "Realisierung eines NF2-Prototypen zum Einsatz in der Lehre." B.Sc. Leipzig University, 2019.

[21] Marcus Stelzer. "Realisierung eines Frameworks zur forensischen Gutachtenerstellung und Auswertung von Browserverläufen." B.Sc. Leipzig University, 2019.

[22] Matthias Täschner. "Interaktive visuelle Exploration großer Graphen." M.Sc. Leipzig University, 2019.

[23] Yannik Völker. "Entwicklung eines dezentralen sozialen Netzwerks." B.Sc. Leipzig University, 2018.

[24] Lukas Werner. "Verteilte Berechnung exakter Perzentile von Fließkommazahlen." M.Sc. Leipzig University, 2018.

5 Talks

[1] Victor Christen. *Informativeness-Based Active Learning for Entity Resolution.* DINA Workshop, ECML, Würzburg. Sept. 2019.

[2] Victor Christen. *Learning-Based Approach to Combine Medical Annotation Results*. DILS,Hannover. Nov. 2018.

[3] Steffen Dienst and Stefan Kühne. *Analysis of Large Graph Data with Gradoop and KNIME.* Big-Data.AI Summit, Bitkom, Berlin. Feb. 2018.

[4] Martin Franke. *Parallel PPRL using LSH-based Blocking*. IoTBDS, Funchal, Madeira. Mar. 2018.

[5] Martin Franke. *Post-processing Methods for High Quality PPRL*. DPM Workshop, ESORICS, Barcelona. Sept. 2018.

[6] Marcel Gladbach. *Distributed privacy-preserving record linkage using pivot-based filter techniques*. ICDE Workshops, Paris, France. 2018.

[7] Marcel Gladbach. *Privacy-Preserving Record Linkage*. ScaDS summer school, Leipzig. July 2018.

[8] Kevin Gomez. *Keynote: Scalable Graph Analytics*. GRIDKA19, Karlsruhe. Aug. 2019.

[9] Martin Grimmer. *A Modern and Sophisticated Host Based Intrusion Detection Data Set*. 16. Deutscher IT-Sicherheitskongress (BSI), Bonn. Mai 2019.

[10] Martin Grimmer. *Intrusion Detection on System Call Graphs*. 25. DFN-Konferenz: Sicherheit in vernetzten Systemen, Hamburg. Feb. 2018.

[11] Markus Nentwig. *Incremental Clustering on Linked Data*. ICDM Workshops, Singapore. 2018.

[12] Daniel Obraczka. *Knowledge Graph Completion with FAMER*. DI2KG, Anchorage, Alaska. Aug. 2019.

[13] Eric Peukert. *Gradoop Overview*. ScaDS summer school, Dresden. Aug. 2019.

[14] Eric Peukert. *Graph Data Transformations in Gradoop*. BTW, Rostock, Germany. Mar. 2019.

[15] Eric Peukert. *Graph-based data integration and analysis at ScaDS*. Blackbee Techtalk, Web Data Solutions GmbH, Leipzig. Jan. 2018.

[16] Eric Peukert. *Graph-basierte Datenintegration*. Big Data Konferenz INNOVATIVE, Frankfurt. June 2018.

[17] Erhard Rahm. *Competence Center for Scalable Data Services and Solutions (ScaDS) Dresden/Leipzig Phase 2*. Kickoff ScaDS II, Dresden. Jan. 2019.

[18] Erhard Rahm. *Data Science research and education at ScaDS / Univ. Leipzig*. Data Science workshop, Fakultätentag Informatik, Osnabrück. Nov. 2018.

[19] Erhard Rahm. *Educating Data Scientists in the context of ScaDS Dresden/Leipzig*. GI Präsidiumsarbeitskreis, Berlin. Jan. 2018.

[20] Erhard Rahm. *FAst Multi-source Entity Resolution system (FAMER)*. Leipziger Semantik Web Tag, Leipzig. June 2018.

[21] Erhard Rahm. *Fast-ML: Laufzeitoptimierung der Mainzelliste für Privacy-Preserving Record Linkage*. TMF, Berlin. Sept. 2019.

[22] Erhard Rahm. *From ScaDS to ScaDS.AI*. Kickoff ScaDS.AI, Leipzig. Nov. 2019.

[23] Erhard Rahm. *Introduction into summer school and ScaDS overview*. ScaDS summer school, Leipzig. July 2018.

[24] Erhard Rahm. *Introduction into workshop and ScaDS overview*. Workshop Big Data and AI in Business (BiDiB), Leipzig. Sept. 2019.

[25] Erhard Rahm. *Research in ScaDS II*. All-Hands-Meeting, Berlin. Nov. 2018.

[26] Erhard Rahm. *Scalable Graph Data Management and Analytics With GRADOOP*. CS colloquia at ANU Canberra and UNSW Sydney. Apr. 2019.

[27] Erhard Rahm. *Standortübergreifendes PPRL von Patientendaten mit zentraler Linkage-Einheit*. GMDS, Dortmund. Sept. 2019.

[28] Erhard Rahm and Eric Peukert. *Entity Resolution for Large-Scale Data*. EDBT Summerschool, Lyon. Sept. 2019.

[29] Christopher Rost. *Evolution Analysis of Evolving Graphs with Gradoop.* LEG Workshop, ECML, Würzburg. Sept. 2019.

[30] Christopher Rost. *Temporal Graph Analysis using Gradoop.* BigDS Workshop, BTW, Rostock. März 2019.

[31] Alieh Saeedi. *FAMER - FAst Multi-source Entity Resolution*. ScaDS summer school, Leipzig. July 2018.

[32] Alieh Saeedi. *Using Link Features for Entity Clustering in Knowledge Graphs*. ESWC (Best paper award), Heraklion, Crete, Greece. 2018.

[33] Moritz Wilke. *Gradoop and KNIME for patent analysis*. Big Data AI-Summit – with BIGGR-Project and Bosch. Feb. 2018.