UMR IRISA

Activity Report 2021

Team DRUID

Declarative & Reliable management of Uncertain user-generated Interlinked Data

D7 – Data and Knowledge Management

1 Team composition

Researchers and faculty

Tristan Allard, Associate Professor, ISTIC Rennes 1 - Rennes (until 15 September 2021)
Jean-Christophe Dubois, Associate Professor, IUT Lannion - Lannion
Mickaël Foursov, Associate Professor, ISTIC Rennes 1 - Rennes
David Gross-Amblard, Professor, ISTIC Univ. Rennes 1 - Rennes
Mouloud Kharoune, Associate Professor, IUT Lannion - Lannion (until March 2021, retired)
Yolande Le Gall, Associate Professor, IUT Lannion - Lannion
Arnaud Martin, Professor, IUT Lannion, Univ. Rennes 1 - Lannion
Zoltan Miklos, Associate Professor, ESIR Rennes 1 - Rennes, **head of the team**
Virginie Sans, Associate Professor, ISTIC Rennes 1 - Rennes

ATER

Constance Thierry, IUT Lannion (2021/22)

Engineer

Javier Rojas Balderrama (until 15 September 2021)

PhD students

Louis Béziaud, UQÀM/CominLabs Profile, co-advised with Sébastien Gambs (UQÀM), (until 15 September 2021)
Ian Jeantet, ANR EPIQUE, (until January 2021)
Gauthier Lyan, CIFRE Keolis, co-advised with Jean-Marc Jezequel (Diverse), (until September 2021)
Maria Massri, CIFRE OrangeLabs
Fancois Mentec, CIFRE ALTEN
Rituraj Singh, ANR HEADWORK, co-advised with Loic Helouet (Sumo), (until Mars 2021)
Constance Thierry, CD22/ANR HEADWORK, (until Decembre 2021)
Zuowei Zhang, CSC grant, (until January 2022)

Administrative assistant

Sophie Maupile (until December 2021)
Gunter Tessier (from December 2021)

2 Overall objectives

2.1 Overview

Recently, there is an increased interest in data management methods. Statistical machine learning techniques, empowered by the available pay-as-you-go distributed computing power, are able to extract useful information from certain data. The international press, being specialized or not, has echoed these remarkable results as a new Spring for Artificial Intelligence in a broad sense. The data is sometimes even referred to as the "gold of the 21st century'. In any area of business and science, one tries to construct huge datasets to be able to profit from the benefits of the Artificial intelligence revolution.

However, when datasets contain personal data, their collection and usage may lead to undesirable practices. In particular, there is a growing interest in privacy, mirroring the still-growing interest in analytics over personal data. Machine Learning and Privacy can indeed be seen as two sides of the same coin: machine learning tries to extract relevant information from data, while privacy tends to blur information in order to hide identifying or sensitive individual information. In addition to the protection of the personal data input by machine learning algorithms, guaranteed by privacy models and privacy-preserving algorithms, the fairness of the output is critical for mitigating discrimination issues within automatic or "semi-automatic" high-stake decisions about individuals (e.g. laws, social rights, police).

Unfortunately, both these desirable needs –seamless machine learning and privacy– are not supported elegantly for now, in the data management dogma. For example, Machine Learning operators are seen for now as external procedures outside the query language, barely accounted for by the optimizer. Moreover, the knowledge extraction tasks are hard to design without understanding the available data, thus one should consider knowledge extraction as an interactive process, where users influence the process. Privacy-preserving algorithms often make extensive use of cryptography, incurring prohibitive costs when considering typical volumes in data management use-cases. Additionally, the choice of a privacy model and of its parameters, among a large number of possible models, is barely understandable for non-expert database administrators. Finally, privacy and fairness are usually considered apart without analysing their mutual impacts.

The above listed observations lead us to define the following goals for the DRUID team:

- Propose new query mechanisms, in particular for network oriented data and to better integrate Machine Learning methods with the database logic and engines
- Propose interactive, human-in-the-loop data analysis and knowledge extraction methods even with uncertain data
- Make privacy-preserving techniques meet real-life constraints within data-centred systems, with a special focus on performance and intelligibility
- Design data-centred systems that are both private and fair.

2.2 Scientific foundations

Our team gathers specialists from data management, privacy, information extraction and belief functions, various bricks that contribute to our goal. As a common ground, for data management we will naturally elaborate on classical techniques: finite model theory, complexity theory, declarative or algebraic languages, execution plans, costs models, storage and indexing strategies. The theory of belief functions (also commonly referred to as Dempster-Shafer theory) allows to take simultaneously into account both uncertainty and imprecision on the data but also on the models. This theory is one of the most popular one among the quantitative approaches because it can be seen as a generalization of both classical probabilities and possibilities theories. Belief functions are especially developed for information fusion, pattern recognition and clustering.

Privacy & fairness in data-centered systems Despite continuous improvements and major breakthroughs along the last two decades, elaborate privacy-preserving techniques (*e.g.* privacy-preserving computing, privacy-preserving data publishing, privacy-preserving data querying) often suffer from low adoption rates within real-world systems. We believe that this is due, at least partially, to the difficulty of privacy-preserving techniques for satisfying diverse and possibly contradictory real-world needs.

Performance is a major issue. Privacy-preserving techniques often make an extensive use of cryptography. Although the resulting security guarantees can be strong (*e.g.* semantic security against computationally bounded adversaries, or even information theoretic guarantees), prohibitive performance costs (*e.g.* rarely sublinear in the size of data, strong computation and/or communication overheads due to the extensive use of encryption schemes) most often deter their deployments in real-life data-centered systems. We propose to follow two promising tracks for contributing to reaching affordable performance costs for privacy-preserving data-centered systems. First, we will **develop the joint use of perturbation schemes with encryption schemes**. The use of perturbation results in a relaxation of the privacy guarantees but sound *differentially private* guarantees can still be satisfied (*e.g.* computational differential privacy Using differentially private perturbation as a building block both builds on our previous works (*e.g.* [AHMP15,SAA+18]) and is a promising and original research track. Second, we will **explore distribution and parallelization techniques** for avoiding the creation of bottlenecks (and at the same time avoiding a systematic centralization of personal data). This may require to distribute the privacy-preserving computations (*e.g.* by using infinitely divisible distributions [KKP12], homomomorphic encryption schemes with threshold decryption features [DJ01], group signatures [CG04]). This track also builds

[AHMP15] T. Allard, G. Hébrail, F. Masseglia, E. Pacitti, "Chiaroscuro: Transparency and Privacy for Massive Personal Time-Series Clustering", *in : SIGMOD '15*, p. 779–794, 2015.

[SAA+18] C. Sahin, T. Allard, R. Akbarinia, A. El Abbadi, E. Pacitti, "A Differentially Private Index for Range Query Processing in Clouds", *in : ICDE '18*, p. 857–868, 2018.

[KKP12] S. Kotz, T. Kozubowski, K. Podgorski, *The Laplace distribution and generalizations: a revisit with applications to communications, economics, engineering, and finance*, 2012.

[DJ01] I. Damgård, M. Jurik, "A generalisation, a simplification and some applications of paillier's probabilistic public-key system", *in : International Workshop on PKC*, p. 119–136, 2001.

[CG04] J. Camenisch, J. Groth, "Group signatures: Better efficiency and new theoretical as-

on previous works (*e.g.* [DA,AHMP15,TAdE19,ADA+20]). Its originality comes from the desired features achievable by composing the above techniques, the specific properties that arise from the resulting conjunction, and the distribution of techniques that are usually centralized (*e.g.* differentially private perturbation).

Intelligibility is a major issue. Privacy-preserving data publishing techniques are complex: choosing, tuning, and parameterizing a model and an algorithm, while anticipating the real-life impact on the privacy of individuals is hard for non-expert users (*e.g.* database administrators). We propose to follow an empirical approach based on stress-testing the privacy-preserving schemes through batteries of attacks. In particular, we will systematize the known attacks, develop new attacks (possibly based on machine learning algorithms, *e.g.* [PTC18]), and design algorithms for generating relevant attackers given a dataset planned to be published. We believe that stress-tests can give to database administrators *by-example* overviews on the possible impacts of publishing their data.

The mutual impact of privacy and fairness guarantees is still under-studied [PMK+20]. We propose to contribute to this emerging field by studying both the privacy and fairness formal models and their experimental behaviors (*e.g.* data-dependent issues – distribution – may strongly impact the results). Moreover fairness is for now an evasive goal, as this concept lacks of a proper definition. As it is strongly connected with skewed distributions, we are confident that our expertise in statistical tweaking will play a role. For example, belief functions theory can be used to better capture information leaks.

Natural applications of our techniques are enabling fairness and privacy for modern crowdsourcing/future of work platforms (leading the not-yet-accepted FairWork ANR proposal) or for AI agent for Human Resource Management (*e.g.* FranÃ§ois Mentec's thesis). Fairness and privacy are important properties for legal technologies as well (*e.g.* Louis Beziaudś PhD thesis with UQAM). Intelligibility of privacy-preserving data publishing and real-life attacks are crucial in today's open data systems (*e.g.* Antonin Voyezś PhD thesis with Enedis, engineer position within the UIA RUDI project led by Rennes Metropole). And performance issues of privacy-preserving computing techniques are important in distributed systems for a wide range of applications (*e.g.* ongoing informal collaboration with UCSB).

---

pects", *in : International Conference on Security in Communication Networks*, Springer, p. 120–133, 2004.

[DA] J. Duguépéroux, T. Allard, "Differentially Private Space Partitioning Algorithm for Privacy-Preserving Crowdsourcing Platforms", *research report*, Univ Rennes, CNRS, IRISA.

[AHMP15] T. Allard, G. Hébrail, F. Masseglia, E. Pacitti, "Chiaroscuro: Transparency and Privacy for Massive Personal Time-Series Clustering", *in : SIGMOD '15*, p. 779–794, 2015.

[TAdE19] H. V. Tran, T. Allard, L. d'Orazio, A. El Abbadi, "Range Query Processing for Monitoring Applications over Untrustworthy Clouds", *in : EDBT*, p. 666–669, 2019.

[ADA+20] M. J. Amiri, J. Duguépéroux, T. Allard, D. Agrawal, A. El Abbadi, "SEPAR: A Privacy-Preserving Blockchain-based System for Regulating Multi-Platform Crowdworking Environments", *arXiv preprint arXiv:2005.07850*, 2020, Submitted to VLDB '21.

[PTC18] A. Pyrgelis, C. Troncoso, E. D. Cristofaro, "Knock Knock, Who's There? Membership Inference on Aggregate Location Data", *in : NDSS '18*, 2018.

[PMK+20] D. Pujol, R. McKenna, S. Kuppam, M. Hay, A. Machanavajjhala, G. Miklau, "Fair decision making using privacy-protected data", *in : FAT* '20*, p. 189–199, 2020.

Analytics in databases Making sense of large amounts of data and extracting useful information is a problem in various fields, in business context as well in various scientific domains. One needs to rely on a wide range of techniques (regression, clustering, embeddings, ...). A classical data analytics workflow is 1) to extract, to model imperfection and clean a data set, 2) to learn a model and to consider imperfection and 3) to make predictions. Such workflows are now very well handled in procedural languages such as Python or Scala, at various scales (*e.g.* Big Data in Spark).

While this approach works well, it it does not make use the numerous achievements in the database field: when the data set is updated, the workflow has to be re-run (dynamicity problem), data are now much more evolved than classical numerical or categorical ones, such as graphs (data type problem), and machine learning operators are not first class citizen in database query languages (closure problem). Moreover, it is often impossible to formulate the "right" knowledge extraction or machine learning tasks, as it would require the knowledge of the large and heterogeneous datasets, *a priori*.

Our specific goal is to develop data management and -possibly interactive- data analysis methods for generic, uncertain and time-varying data (*e.g.* large evolving graphs). We will rely on graph signal processing [SNF+13], spectral graph theory [Chu97], graph neural networks [WPC+20], graph databases [RWE15], [BFVY18] and graph embedding techniques. We aim at modelling and querying graphs, with (temporal) integrity constraints, where graph analytics is first used to optimize data storage and evaluation. Machine learning techniques also allow to build realistic huge benchmark data sets, that do not exist for all domains.

[BN03]

On a longer perspective, we would like to work on other aspects of database and machine learning integration. In particular, databases have efficient mechanisms for indexing and loading data to main memory and these could be better exploited to realize machine learning tasks. In some cases one could envisage that the machine learning tasks are realized inside the database systems and machine learning methods use database primitives [GR17]. Other potential direction is to consider a vector-space embedding of entire relational databases [Gro20] that could open entire new ways to

[SNF+13] D. I. Shuman, S. K. Narang, P. Frossard, A. Ortega, P. Vandergheynst, "The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains", *IEEE Signal Processing Magazine 30*, 3, May 2013, p. 83–98.

[Chu97] F. R. K. Chung, *Spectral Graph Theory*, American Mathematical Society, 1997.

[WPC+20] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, P. S. Yu, "A Comprehensive Survey on Graph Neural Networks", *IEEE Transactions on Neural Networks and Learning Systems*, 2020, p. 1–21.

[RWE15] I. Robinson, J. Webber, E. Eifrem, *Graph Databases: New Opportunities for Connected Data*, edition 2nd, O'Reilly Media, Inc., 2015.

[BFVY18] A. Bonifati, G. H. L. Fletcher, H. Voigt, N. Yakovets, *Querying Graphs*, *Synthesis Lectures on Data Management*, Morgan & Claypool Publishers, 2018, https://doi.org/10.2200/S00873ED1V01Y201808DTM051.

[BN03] M. Belkin, P. Niyogi, "Laplacian Eigenmaps for Dimensionality Reduction and Data Representation", *Neural Comput. 15*, 6, June 2003, p. 1373–1396, https://doi.org/10.1162/089976603321780317.

[GR17] M. Grohe, M. Ritzert, "Learning first-order definable concepts over structures of small

analyze data stored in such systems.

2.3 Application domains

Our natural applications are storing and querying large-scale semantic graphs for IOT (*e.g.* Maria Massri's thesis, GraphStore project), Digital humanities (epistemology, understanding the evolution of ideas and scientific fields, *e.g.* EPIQUE ANR), human resources management (*e.g.* François Mentec's thesis), and crowd management systems (HEADWORK ANR) for "artificial artificial intelligence". Our work and results can be used to analyze the evolution of other types of networks (e.g. transportation network, e.g. Gauthier Lyan's thesis).

Further applications of our techniques are enabling fairness and privacy for modern crowdsourcing/future of work platforms (we lead the ANR FairWork, under review) or for AI agent for Human Resource Management (*e.g.* François Mentec's thesis). Fairness and privacy are important properties for legal technologies as well (*e.g.* Louis Beziaudś PhD thesis with UQAM). Intelligibility of privacy-preserving data publishing and real-life attacks are crucial in today's open data systems (*e.g.* Antonin Voyezś PhD thesis with Enedis, engineer position within the UIA RUDI project led by Rennes Metropole). And performance issues of privacy-preserving computing techniques are important in distributed systems for a wide range of applications (*e.g.* ongoing informal collaboration with UCSB).

3 Scientific achievements

3.1 Temporal graph databases

Participants: Zoltan Miklos, David Gross-Amblard, Maria Massri.

Our work is focused on questions related to temporal graph databases. We further improved our δ-Copy+Log storage method that we conceived in 2020. We simplified our definitions and completed the evaluation with further aspects. The article that describes thus work is under review.

Besides the storage mechanisms, we also worked on temporal graph query languages. We developed a new query language for temporal graphs: T-Cypher. It is an extension of the well-known query language Cypher for temporal graphs. The article presenting this work is also under review.

We also work on query execution: we analyse how to evaluate T-cypher queries efficiently, especially if we use the Clock-G storage mechanisms. This is work in progress.

---

degree", *in : 32nd Annual ACM/IEEE Symposium on Logic in Computer Science, LICS 2017, Reykjavik, Iceland, June 20-23, 2017*, IEEE Computer Society, p. 1–12, 2017, https://doi.org/10.1109/LICS.2017.8005080.

[Gro20] M. Grohe, "word2vec, node2vec, graph2vec, X2vec: Towards a Theory of Vector Embeddings of Structured Data", *CoRR abs/2003.12590*, 2020, https://arxiv.org/abs/2003.12590.

3.2 AI for human resource applications

Participants: Zoltan Miklos, Francois Mentec.

We have designed a conversational recommender system for job recruiters. The basis of the recommender system is the recommendation that one can obtain through matching to hierarchical skill representations. We use such a skill representation (in particular, the European skill ontology ESCO) and we match job offers and candidate profiles to this structure. Besides a simple recommendation, the users can interact with the system and refine their search. This work is presented at the Conversational Recommender Systems workshop KaRS'21 [21].

Our work consists also to develop another recommender system that also takes into account the user's history and other input that we collect through interaction. The recommendation tool will be deployed in 2021, for recruiters of the company. The publication on this work is in preparation.

3.3 Complex workflows and aggregation in crowdsourcing

Participants: Loic Helouet (SUMO team), Zoltan Miklos, Rituraj Singh.

Crowdsourcing is a way to solve problems that need human contribution. Crowdsourcing platforms distribute replicated tasks to workers, pay them for their contribution, and aggregate answers to produce a reliable conclusion. Our team leads the ANR HEADWORK project, which aim is to build a fully functional crowd management system.

A fundamental problem is to infer a consensual answer from the set of returned results. Another problem is to obtain this answer at a reasonable cost: unlimited budget allows hiring experts or large pools of workers for each task but a limited budget forces to use resources at best. Last, crowdsourcing platforms have to detect and ban malevolent users (also known as "spammers") to achieve good accuracy of their answers.

This paper considers crowdsourcing of simple Boolean tasks. We first define a probabilistic inference technique, that considers difficulty of tasks and expertise of workers when aggregating answers. We then propose CrowdInc, a greedy algorithm that reduces the cost needed to reach a consensual answer. CrowdInc distributes resources dynamically to tasks according to their difficulty. The algorithm solves batches of simple tasks in rounds that estimate workers expertize, tasks difficulty, and synthesizes a plausible aggregated conclusion and a confidence score using Expectation Maximization. The synthesized values are used to decide whether more workers should be hired to increase confidence in synthesized answers. We show on several benchmarks that CrowdInc achieves good accuracy, reduces costs and we compare its performance to existing solutions. We then use the estimation of CrowdInc to detect spammers and study the impact of spammers on costs and accuracy.

Transactions on Large-Scale Data and Knowledge-Centered Systems [8]. (extended version of our ICWS'2020 paper)

3.4 Analysis of the evolution of science

Participants: Zoltan Miklos, Ian Jeantet, David Gross-Amblard, Mickael Foursov.

3.4.1 EvoMaps

Participants: Ian Jeantet, Zoltan Miklos, David Gross-Amblard.

To describe the evolution of quasi dendograms, the structures that we can obtain through overlapping hierarchical clustering, we have developed the notion EvoMaps. These are graphical structures which allow to visualize the evolution of quasi-dendograms in time and which serve as a visualization of the evolution of scientific domains. This work is discussed in the PhD thesis of Ian Jeantet.

3.4.2 Spectral analysis of evolving graphs

Participants: Zoltan Miklos, Mickael Foursov.

We have continued to develop methods to describe evolving graphs through spectral techniques. This work is ongoing.

3.5 AI and belief functions

Participants: Arnaud Martin, Zuowei Zhang, Yiru Zhang, Na Li, Constance Thierry, Daniel Zhu, Yolande Le Gall, Jean-Christophe Dubois.

3.5.1 Belief clustering

Participants: Arnaud Martin, Zuowei Zhang, Yiru Zhang, Na Li.

Clustering is an essential part of data mining, which can be used to organize data into sensible groups. We continue to develop some research to propose new clustering method adapted to imperfect data.

Traditional evidential clustering tends to build clusters where the number of data for each cluster fairly close to each other. However, it may not be suitable for imbalanced data. In [25] we propose a new method, called credal clustering (CClu), to deal with imbalanced data based on the theory of belief functions. Consider a dataset with C wanted classes, the credal c-means (CCM) clustering method is employed at first to divide the dataset into some (i.e., S (S $>$ C)) clusters. Then these clusters are gradually merged following a given principle based on the density of meta-clusters and the associated singleton clusters. The merging is finished when C singleton wanted classes are obtained. During this merging procedure, the objects in each singleton cluster will be assigned to one new singleton class. Moreover, a weighted mean vector rule is developed

to classify the objects in the unmerged meta-cluster to the associated new classes using the K-Nearest neighbor technique.

Moreover, traditional evidential clustering is computationally time-consuming due to the premature inclusion of meta-clusters into the iterative process. In this paper, a simple and fast method is proposed in [26] to extract the credal partition structure in evidential clustering based on modifying the iteration rule, which is specifically described as the following two steps: 1) The iterations are only allowed under the frame of discernment Ω until the termination condition is first triggered; 2) Then only once iteration is needed under the power-set 2^{Ω} to capture the final credal partitions. By doing this, the invalid calculations related to meta-clusters are effectively exempted. It is shown to outperform the known partition detection methods in terms of computation time. Moreover, the quality of the partitions detected is similar to that of existing evidential clustering. The simulation results implemented with synthetic and real patterns illustrate the potential of this method, especially in large data.

Among the various clustering algorithms, the prototype-based methods have been most popularly applied due to the easy implementation, simplicity and efficiency. However, most of them such as the c-means clustering are no longer effective when the data is insufficient and uncertain. While the data for the current clustering task may be sparse, there is usually some useful knowledge available in the related scenes. Transfer learning can be adopted to address such cross domain learning problems by using information from data in a related domain and transferring that data/knowledge to the target task. The inconsistency between different domains can increase the uncertainty in the data. To handle the insufficiency and uncertainty problems in the clustering task simultaneously, a prototype-based evidential transfer clustering algorithm, named transfer evidential c-means (TECM), is introduced in the framework of belief functions. The proposed algorithm in [27] employs the cluster prototypes of the source data as references to guide the clustering process of the target data.

We also focus on a special application of clustering: preference clustering. Evidential preference based on belief function theory has been proposed recently, simultaneously characterizing preference information with uncertainty and imprecision. However, traditional distances on belief functions do not adapt to some intrinsic properties of preference relations, especially when indifference relation is taken into comparison, therefore may cause inconsistent results in preference-based applications. In order to solve this issue, Unequal Singleton Pair (USP) distance has been proposed in [9], with applications limited in preference aggregation. Hence in [10] we explore forward the effectiveness of USP distance in preference clustering, especially confronting multiple conflicting sources

3.5.2 Belief learning

Participants: Arnaud Martin, Zuowei Zhang, Na li, Daniel Zhu, Yolande Le Gall, Jean-Christophe Dubois.

In real-life machine learning applications, a common problem is that raw data (e.g. remote sensing data) is sometimes inaccessible due to confidentiality and privacy constrains of corporations, making classification methods arduous to work in the supervised context. Moreover, even though raw data is accessible, limited labeled samples can also

seriously affect supervised methods. Recently, supervised and unsupervised classification (clustering) results related to specific applications are published by more and more organizations. Therefore, combination of supervised classification and clustering results has gained increasing attention to improve the accuracy of supervised predictions. Incorporating clustering results with supervised classifications at the output level can help to lessen the recline on information at the raw data level, so that is pertinent to improve the accuracy for the applications when raw data is inaccessible or training samples are limited. In [6], we focus on the combination of multiple supervised classification and clustering results at the output level based on belief functions for three purposes: (1) to improve the accuracy of classification when raw data is inaccessible or training samples are highly limited; (2) to reduce uncertain and imprecise information in the supervised results; and (3) to study how supervised classification and clustering results affect the combination at the output level. Our contributions consist of a transformation method to transfer heterogeneous information into the same frame, and an iterative fusion strategy to retain most of the trustful information in multiple supervised classification and clustering results.

The classification analysis of missing data is still a challenging task since the training patterns may be insufficient and incomplete in many fields. To train a high-performance classifier and pursue high accuracy, we learn a credal classifier based on an optimized and adaptive multi-estimation (OAME) method for missing data imputation on training and test sets, introduced in [11]. In OAME, some incomplete training patterns are estimated as multiple versions by a global optimization method thereby expanding the training set. On the other hand, the test pattern is adaptively estimated as one or multiple versions depending on the neighbors. For the test pattern with multiple versions, the corresponding outputs with different discounting factors (weights), represented by the basic belief assignments (BBAs), are fused for final credal classification based on evidence theory. The discounting factor contains two aspects: the importance and reliability factors that are used respectively to quantify the importance of the edited version itself and to represent the reliability of the classification result of the version. The effectiveness of OAME is widely validated on several real datasets and critically compared to other related methods.

This year, we also study active learning with belief function in [28] and [29]. Active learning is a subfield of machine learning which allows to reduce the amount of data necessary to train a classifier. The training set is built in an iterative way such that only the most significant and informative data are used and labeled by an external person called oracle. It is furthermore possible to use active learning with the theory of belief functions in order to take erroneous labels due to the oracle's uncertainty and imprecision into account in order to limit their influence on the classifier's performance. In [29], we compare the classifier of the k nearest neighbours (kNN) to a variant based on belief functions from the theory of belief functions (EkNN), in a situation where some labels have been noised in order to model uncertain labels. We show that although the superiority of EkNN over kNN is not systematic, there are some interesting and modest results supporting the relevance of belief functions in active learning.

3.5.3 Belief functions and crowdsourcing

Participants: Arnaud Martin, Constance Thierry, Yolande Le Gall, Jean-Christophe Dubois.

Crowdsourcing is the outsourcing of tasks to a crowd of contributors on dedicated platforms. The tasks are simple and accessible to all, that is why the crowd is made of very diverse profiles, but this induces con result the contributions of unequal quality. The aggregation method most used in platforms does not take into account the imperfections of the data related to human contributions, which impacts the results obtained. The work of Constance Thierry thesis, defended in December 14, 2021, aims at solving the problem of data quality in crowdsourcing platforms. Thus, we propose a new interface for crowdsourcing offering more expression capacity to the contributor. The experiments carried out allowed us to highlight a correlation between the difficulty of the task, the certainty of the contributor and the imprecision of his answer. In [22], we also validated the hypothesis of Ph. Smets according to which the more imprecise a person is, the more certain he is, and conversely the more precise he is, the less certain he is. Based on this hypothesis, we develop the MONITOR model for the estimation of the contributorś profile and the aggregation of the answers thanks to the theory of belief functions which allows to model imperfections. All our experiments are performed on real data coming from crowdsourcing campaigns.

3.6 Privacy-preserving database management systems

Participants: Tristan Allard, Laurent d'Orazio (SHAMAN), Hoang Van Tran.

We have adapted the PinedRQ index [SAA+18], a differentially-private indexing method for encrypted data that we designed a few years back, to make it able to cope with high ingestion rates. By thoroughly refactoring its software architecture and its internal data structures, we came up with Fresque [23]. We designed Fresque, formally assessed its security guarantees, implemented it, and validated it experimentally. This resulted in ingestion performances not reached yet by the state-of-the-art works. We are pursuing this work by studying more the problem of hiding the leaks resulting from updates to encrypted databases (in particular, the access patterns resulting from updates).

3.7 Privacy-preserving regulation of crowdworking environments

Participants: Tristan Allard, Joris Duguépéroux, Mohamad Amiri (UCSB), Amr El Abbadi (UCSB), Divy Agrawal (UCSB).

We have proposed the Separ system [12] for enabling labor law enforcement in environments made of multiple crowdworking platforms, where workers and requesters register to any subset of platforms. With Separ, we have unveiled the research space

[SAA+18] C. Sahin, T. Allard, R. Akbarinia, A. El Abbadi, E. Pacitti, "A Differentially Private Index for Range Query Processing in Clouds", *in : ICDE '18*, p. 857–868, 2018.

related to the regulation of crowdworking environments in general, and proposed a technical solution tackling a precise point in this space. In particular, we focused on lower-than/greater-than regulations (e.g., a crowdworker works at most 40 hours per week whatever the platform). Separ is technically based on managing budgets made of pseudonymous tokens, on a distributed ledger shared between crowdworking platforms, and on three consensus protocols between platforms. We have formally proven the security of Separ against covert adversaries given a precise set of tolerated leaks, and evaluated experimentally its performances that show its scalability and the low overhead of privacy. We are pursuing this work by studying more generally the problem of constraint enforcement over private data.

3.8 On the Quality of Compositional Prediction for Prospective Analytics on Graphs

Participants: Gauthier Lyan, David Gross-Amblard, Jean-Marc Jézéquel (Diverse team).

Recently, micro-learning has been successfully applied to various scenarios, such as graph optimization (e.g. power grid management). In these approaches, ad-hoc models of local data are built instead of one large model on the overall data set. Micro-learning is typically useful for incremental, what-if/prospective scenarios, where one has to perform step-by-step decisions based on local properties. A common feature of these applications is that the predicted properties (such as speed of a bus line) are compositions of smaller parts (e.g. the speed on each bus inter-stations along the line). But little is known about the quality of such predictions when generalized at a larger scale.

In this work [18] we propose a generic technique that embeds machine-learning for graph-based compositional prediction, that allows 1) the prediction of the behaviour of composite objects, based on the predictions of their sub-parts and appropriate composition rules, and 2) the production of rich prospective analytics scenarios, where new objects never observed before can be predicted based on their simpler parts. We show that the quality of such predictions compete with macro-learning ones, while enabling prospective scenarios. We assess our work on a real size, operational bus network data set.

3.9 On the Quality of Compositional Prediction for Prospective Analytics on Graphs

Participants: Gauthier Lyan, Jean-Marc Jézéquel (Diverse team), David Gross-Amblard, Benoît Combemale (Diverse team).

Models at runtime have been initially investigated for adaptive systems. Models are used as a reflective layer of the current state of the system to support the implementation of a feedback loop. More recently, models at runtime have also been identified as key for supporting the development of full-fledged digital twins. However, this use of models at runtime raises new challenges, such as the ability to seamlessly interact with

the past, present and future states of the system. In this work [19], we propose a framework called DataTime to implement models at runtime which capture the state of the system according to the dimensions of both time and space, here modeled as a directed graph where both nodes and edges bear local states (ie. values of properties of interest). DataTime provides a unifying interface to query the past, present and future (predicted) states of the system. This unifying interface provides i) an optimized structure of the time series that capture the past states of the system, possibly evolving over time, ii) the ability to get the last available value provided by the system's sensors, and iii) a continuous micro-learning over graph edges of a predictive model to make it possible to query future states, either locally or more globally, thanks to a composition law. The framework has been developed and evaluated in the context of the Intelligent Public Transportation Systems of the city of Rennes (France). This experimentation has demonstrated how DataTime can deprecate the use of heterogeneous tools for managing data from the past, the present and the future, and facilitate the development of digital twins.

3.10 Inverse Tone Mapping using FusionNetwork

Participants: Mathieu Chambe, Zoltan Miklos, Ewa Kijak, Kadi Bouatouche.

We develop a deep learning models that can address computer vision tasks, specifically for high dynamic range (HDR) images. Our approach generates HDR images from standard dynamic range (SRD) images, with the help of inverse tone mapping operators. We use then these generated images to train our network. This approach not only enables to train the network with a smaller dataset, but also results a smaller network, while the performance remains comparable to other state-of-the-art approaches. This is a work in progress.

4 Software development

4.1 BrFAST: web authentication and browser fingerprints

Participants: Tristan Allard, Nampoina Andriamilanto, GaÃ«tan Le Guelvouit (BCOM).

We have developped BrFAST [13], an extendible platform for showcasing browser fingerprinting attribute selection methods. We embedded in BrFAST our own proposal, FPSelect [AAG20], baselines based on entropy and conditional entropy, and public browser fingerprint datasets.

[AAG20] N. Andriamilanto, T. Allard, G. L. Guelvouit, "FPSelect: Low-Cost Browser Fingerprints for Mitigating Dictionary Attacks against Web Authentication Mechanisms", *Annual Computer Security Applications Conference*, 2020.

4.2 HEADWORK platform

Participants: David Gross-AmblardRituraj Singh.

Crowdsourcing relies on potentially huge numbers of on-line participants to resolve data acquisition or analysis tasks. It is an exploding area that impacts various domains, ranging from scientific knowledge enrichment to market analysis support. But currently, existing crowd platforms rely mostly on low level programming paradigms, rigid data models and poor participant profiles, which yields severe limitations. The low- level nature of existing solutions prevents the design of complex data acquisition workflows, that could be executed, composed, searched and even be proposed by participants them- selves. Taking into account the quality, uncertainty, inconsistency and representativeness of participant contributions is still an open problem. Methods for assigning a task to the correct participant according to his trust, motivation and expertise, automatically improving crowd execution time, computing optimal participant rewards, are missing. Similarly, usual crowd campaigns produce isolated and rigid data sets: A flexible and common data model for the produced knowledge about data and participants could allow participative knowledge acquisition. To overcome these challenges, Headwork[1] will define:

Rich workflow, participant, data and knowledge models to capture various kind of crowd applications with complex data acquisition tasks and human specificities Methods for deploying, verifying, optimizing, but also monitoring and adapting crowd- based workflow executions at run time.

To reach its goals, Headwork will rely on two experts of large participative knowledge acquisition platforms: Cesco (Museum National d'Histoire Naturelle), Wirk (Foule-Factory), Valda (INRIA Paris), Druid (Rennes 1), Links (Inria-Lille), Sumo (Inria-Bretagne), Spirals (Inria-Lille).

Over the period of this report the platform has gone live in Beta version, holding its first experimental crowd campaigns. The overall project on GitLab has now 681 commits, 29 members, 5500 PHP lines.

5 Contracts and collaborations

5.1 National Initiatives

5.1.1 Project HEADWORK

Participants: Tristan Allard, Jean-Christophe Dubois, David Gross-Amblard [contact point], Yolande Le Gall, Arnaud Martin, Zoltan Miklos, Rituraj Singh, Constance Thierry.

- Project type: ANR
- Dates: 2016-2022

[1] https://headwork.irisa.fr

- Coordinator: David Gross-Amblard
- Funding: 800 000 euros / 146 000 euros (IRISA)
- PI institution: ANR
- Other partners: SUMO/IRISA, Cristal and Inria (Lille), MNHN (Paris), ENS and Inria (Paris), FouleFactory/Wirk (startup).

Crowdsourcing relies on potentially huge numbers of on-line participants to resolve data acquisition or analysis tasks. It is an exploding area that impacts various domains, ranging from scientific knowledge enrichment to market analysis support. But currently, existing crowd platforms rely mostly on low level programming paradigms, rigid data models and poor participant profiles, which yields severe limitations. The low- level nature of existing solutions prevents the design of complex data acquisition workflows, that could be executed, composed, searched and even be proposed by participants them- selves. Taking into account the quality, uncertainty, inconsistency and representativeness of participant contributions is still an open problem. Methods for assigning a task to the correct participant according to his trust, motivation and expertise, automatically improving crowd execution time, computing optimal participant rewards, are missing. Similarly, usual crowd campaigns produce isolated and rigid data sets: A flexible and common data model for the produced knowledge about data and participants could allow participative knowledge acquisition. To overcome these challenges, Headwork project aims: 1) Rich workflow, participant, data and knowledge models to capture various kind of crowd applications with complex data acquisition tasks and human specificities, 2) Methods for deploying, verifying, optimizing, but also monitoring and adapting crowd- based workflow executions at run time.

5.1.2 Project EPIQUE

Participants: Zoltan Miklos [Contact point], David Gross-Amblard, Ian Jeantet, Mickaël Foursov, Arnaud Martin.

- Project type: ANR
- Dates: 2016-2021
- Coordinator: Bernd Amann (LIP 6)
- Funding: 599 800 euros / 142 500 euros (IRISA)
- PI institution: ANR
- Other partners: ISC-PIF, IHPST

The goal of the project is to build global semantic maps of the evolution of science on large scientific domains by applying appropriate scientometric models on large databases like the Web of Science.

5.1.3 Project Clara

Participants: Zoltan Miklos [Contact point], Mickael Foursov, David Gross-Amblard.

- Project type: Cominlabs
- Dates: 2021.12-2024.12
- Coordinator: Patricia Serano (Nantes Université)
- Funding: 113 000 euros (IRISA)
- PI institution: CominLabs
- Other partners: LS2N (Nantes), IRISA (Rennes)

CLARA project aims to empower teachers to facilitate the creation of licensable educational resources based on existing ones. Our approach will suggest a relevant set of educational resources such that these are coherent with a course sketch and have compatible licenses. The main challenges we will face are how to enrich a network of educational resources using AI algorithms, and how to guarantee a minimal set of license-compatible educational resources relevant to a given course goal with query relaxation techniques. We will exploit educational resources provided by the French Ministry of Education and the X5-GON project.

5.1.4 Project CROWDGUARD

Participants: Tristan Allard [contact point], David Gross-Amblard, Zoltan Miklos.

- Project type: ANR JCJC
- Dates: 2016-2021
- Coordinator: Tristan Allard
- Funding: 144 000 euros (IRISA)
- PI institution: ANR

5.2 Bilateral industry grants

5.3 Cifre ALTEN

We collaborate with the company ALTEN, who finance in the form of a CIFRE contract the PhD thesis of Francois Mentec. Our collaboration focuses on the use of artificial intelligence techniques for recruitment and human resources tasks in general. In our collaborations we try to propose new ways how the artificial intelligence methods can support the work of recruiters. We do not intend to replace human recruiters or automatically affect consultants to project, rather to support the work of RH agents.

5.4 Cifre OrangeLabs

We collaborate with the company Orange, who finance in the form of a CIFRE contract the PhD thesis of Maria Massri. The company develops a new platform (called ThingIn) for IoT devices. To realize the proposed services they need to store and query temporal graph data. Our collaboration is on the questions of efficient storage and querying techniques for temporal graph-oriented data.

5.5 Cifre Enedis

We collaborate with the Enedis company through the PhD thesis of Antonin Voyez. We study the privacy guarantees of privacy-preserving data publishing approaches (i.e., anonymization) applied to the fine grain power consumption time series that they collect through the Linky smart meters. Our works include studying empirically the national-scale set of time series collected by Enedis (e.g., general statistical measures, unicity rates) and designing, implementing, validating attacks on real-life and state-of-the-art privacy-preserving data publishing schemes.

6 Dissemination

6.1 Promoting scientific activities

6.1.1 Scientific Events Organisation

General Chair, Scientific Chair

- Z. Miklos : organizing chair, BDA'2021 national conference (37st edition) [2]

Member of the Organizing Committees

6.1.2 Scientific Events Selection

Chair of Conference Program Committees

Member of Conference Program Committees

- T. Allard: PC member: DBSec 2021, BDA 2021, SIGMOD 2022
- D. Gross-Amblard: PC member: BDA'2021, HMData'2021
- A. Martin: PC member: Belief 2021, Fusion 2021, IGARSS 2021, IJCNN 2021, LFA 2021, EGC'2022,
- Z. Miklos: PC member: SIGKDD'2021, CIKM'2021, WSDM'2022, DSAA'2021, EGC'2022,

Reviewer

- T. Allard: KDD 2021

6.1.3 Journal

Member of the Editorial Boards

[2] https://bda2021.inria.fr/

Reviewer - Reviewing Activities

- T. Allard: VLDBJ, Information Science, DAPD, TDSC, Terminal
- A. Martin: PC member: Applied Soft Computing Journal, Computers in Human Behavior Reports, Chinese Journal of Aeronautics, Journal of King Saud University - Computer and Information Sciences, Expert Systems With Applications, IEEE Transactions on Systems, Man, and Cybernetics - Systems, IEEE Transactions on Neural Networks and Learning Systems, Information Fusion, International Journal of Approximate Reasoning, Information Sciences
- Z. Miklos: softwareX, Transactions on Mobile Computing, Applied Sciences

6.1.4 Invited Talks

- T. Allard: Rencontres SHS-Cyber (MSHB, Rennes)

6.1.5 Leadership within the Scientific Community

- Arnaud Martin:
  - president of EGC society[3]
- David Gross-Amblard:
  - member of BDA board[4]
- David Gross-Amblard, Zoltan Miklos, Tristan Allard
  - In charge of the website of the French research in database community (`https://bdav.irisa.fr/`)

6.1.6 Scientific Expertise

- Z. Miklos: Fulbright France

6.1.7 Research Administration

6.2 Teaching, supervision

6.2.1 Teaching

- Our team is in charge of most of the database-oriented courses at University of Rennes 1 (ISTIC department and ESIR Engineering school), with courses ranging from classical databases to business intelligence, database theory, MapReduce paradigm, or database security and privacy.

[3] `http://www.egc.asso.fr`
[4] `http://bdav.org`

- Master Miage: David Gross-Amblard, OLAP and NoSQL databases, ISTIC.
- Master Miage: Tristan Allard, database security and privacy, ISTIC.
- Master: David Gross-Amblard, NoSQL databases, ISTIC.
- Master: Zoltan Miklos, Data and knowledge management (advanced course), M2 research, ISTIC
- Master: Arnaud Martin, research module on data mining and data fusion, M2 research, ENSSAT.
- Master: Tristan Allard, Privacy, ISTIC (Cyberschool).
- Master: Tristan Allard, Database, ISTIC.
- Engineering school (niveau Master) Zoltan Miklos, Artificial Intelligence, ESIR
- Engineering school (Master level) Zoltan Miklos Project in Artificial intelligence, ESIR
- Engineering school (Master level) Zoltan Miklos, Data mining
- Engineering school (Master level): Tristan Allard, Data and systems security, ESIR.
- ENSAI (National School of Statistics) David Gross-Amblard, NoSQL databases
- ENS Rennes: David Gross-Amblard, "préparation à l'agrégation d'Informatique" (databases)
- ENS Rennes: Tristan Allard, "préparation à l'agrégation de MÃ©catronique, SpÃ©cialitÃ© Informatique" (databases)

6.2.2 Administration

- Mickaël Foursov is director of special program in Business Informatics (MIAGE)
- David Gross-Amblard is co-head of the Research Master in Computer Science (SIF) (until spetember 2021), Rennes 1 University[5]
- Yolande Le Gall is the responsable of the alternating training courses of the second year of computer sciences of the DUT
- Arnaud Martin is director of Studies at the Lannion University of Technology
- Arnaud Martin is head of the bachelor development of Web and Mobile applications
- Arnaud Martin is the director od IUT Lannion (September 2021)
- Zoltan Miklos is the responsable of the option Information Systems, ESIR
- Tristan Allard is co-head of the alternating training courses of the first MIAGE master year.

6.2.3 Supervision

- PhD: Ian Jeantet, Hierarchical and temporal analysis of scientific corpora as tool for the history of science, supervisors: Zoltan Miklos, David Gross-Amblard, defended in January 2021
- PhD: Gauthier Lyan, Urban mobility: leveraging machine learning and data masses for building of simulators, Jean-Marc Jézéquel, David Gross-Amblard, defended September 2021
- PhD: Rituraj Singh, Data centric workflows for crowdsourcing applications, Loïc Hélouët, Zoltan Miklos, defended May 2021

[5] `http://master.irisa.fr`

- PhD: Constance Thierry, Evaluation of contribution and workers quality on crowdsourcing platform, Jean-Christophe Dubois, Yolande Le Gall, Arnaud Martin, defended December 2021
- PhD in progress: Arthur Hoarau, Active learning of imprecise and uncertain data, Jean-Christophe Dubois, Yolande Le Gall, Arnaud Martin
- PhD in progress: Louis Béziaud, Vers un développement éthique et respectueux de la vie privée de l'intelligence artificielle en droit et justice, Tristan Allard, Sébastien Gambs, cotutelle UQAM.
- PhD in progress: Antonin Voyez, Privacy-Preserving Power Consumption Time-Series Publishing: Designing Attacks for Stress-Testing Sanitized Datasets, Tristan Allard, Gildas Avoine, Elisa Fromont, CIFRE ENEDIS.
- PhD in progress: Gauthier Lyan, Urban mobility, machine learning for simulator synthesis using Big Data, Jean-Marc Jézéquel, David Gross-Amblard
- PhD in progress: Maria Massri, Gestion distribuée de grands graphes dynamiques, application à l'IOT, Zoltan Miklos, David-Gross-Amblard
- PhD in progress: Francois Mentec, Recommandation sur le recrutement de consultants et Intelligence Artificielle, Zoltan Miklos, David-Gross-Amblard
- PhD in progress: Zuowei Zhang, Network data mining under uncertainty using belief functions, Zhunga Liu, Arnaud Martin, cotutelle NPU, Xi'an, China

6.2.4 Juries

- A. Martin:
  - Q. Laporte-Chabasse (Université de Lorraine, 2021) - PhD thesis - (member) 2021
  - Huiqin Chen (Université Paris Saclay) - PhD thesis - (president) 2021
  - Tao Peng (Aix-Marseille Université Paris) - PhD thesis - (reviewer) 2021
- Z. Miklos:
  - Chi Thang DUONG (EPFL Switzerland) - PhD thesis - (reviewer) 2021

6.3 Popularization

- T. Allard : "Anonymat garanti" (article, Pour la science), "Hands-on anonymization" (coding workshop, Rennes), interviews (LeBlob)
- Z. Miklos: "Qu'est-ce que c'est, l'intelligence artificielle?", elementary school Carle Bahon, Rennes (3/6/2021)

7 Bibliography

Doctoral dissertations and "Habilitation" theses

[1] I. Jeantet, *Hierarchical and temporal analysis of scientific corpora as tools for the history of science*, Theses, Université Rennes 1, January 2021, `https://tel.archives-ouvertes.fr/tel-03108773`.

[2] G. Lyan, *Urban mobility : Leveraging machine learning and data masses for the buildingof simulators*, Theses, Université rennes1, September 2021, https://hal.archives-ouvertes.fr/hal-03355775.

[3] R. Singh, *Data Centric Workflows for Crowdsourcing Application*, Theses, Université de Rennes 1, May 2021, https://hal.inria.fr/tel-03274867.

[4] C. Thierry, *Évaluation de la qualité des contributions et des contributeurs sur plateformes de crowdsourcing*, Theses, Université de Rennes 1, December 2021, https://hal.inria.fr/tel-03537663.

Articles in referred journals and book chapters

[5] N. Andriamilanto, T. Allard, G. Le Guelvouit, a. garel, "A Large-scale Empirical Analysis of Browser Fingerprints Properties for Web Authentication", *ACM Transactions on the Web 16*, 1, 2022, p. 1–62.

[6] N. LI, A. Martin, R. Estival, "Heterogeneous information fusion: combination of multiple supervised and unsupervised classification methods based on belief functions", *Information Sciences 544*, January 2021, p. 238–265, https://hal.archives-ouvertes.fr/hal-02938552.

[7] G. Lyan, D. Gross-Amblard, J.-M. Jézéquel, S. Malinowski, "Impact of Data Cleansing for Urban Bus Commercial Speed Prediction", *Springer Nature Computer Science*, November 2021, p. 1–11, https://hal.inria.fr/hal-03220449.

[8] R. Singh, L. Hélouët, Z. Miklos, "Reducing the Cost of Aggregation in Crowdsourcing", *Transactions on Large-Scale Data- and Knowledge-Centered Systems*, October 2021, p. 1–38, https://hal.archives-ouvertes.fr/hal-03482460.

[9] Y. Zhang, T. Bouadi, Y. Wang, A. Martin, "A distance for evidential preferences with application to group decision making", *Information Sciences*, 2021, https://hal.archives-ouvertes.fr/hal-03171577.

[10] Y. Zhang, A. Martin, "Unequal Singleton Pair Distance for Evidential Preference Clustering", *in : Belief Functions: Theory and Applications*, *Lecture Notes in Computer Science*, *12915*, Springer International Publishing, October 2021, p. 33–43, https://hal.archives-ouvertes.fr/hal-03410134.

[11] Z.-W. Zhang, H.-P. Tian, L.-Z. Yan, A. Martin, K. Zhou, "Learning a credal classifier with optimized and adaptive multi-estimation for missing data imputation", *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2021, https://hal.archives-ouvertes.fr/hal-03271783.

Publications in Conferences and Workshops

[12] M. J. Amiri, J. Duguépéroux, T. Allard, D. Agrawal, A. El Abbadi, "Separ: Towards Regulating Future of Work Multi-Platform Crowdworking Environments with Privacy Guarantees", *in : The Web Conference 2021 (WWW '21)*, Ljubljana, Slovenia, April 2021, https://hal.inria.fr/hal-03199954.

[13] N. Andriamilanto, T. Allard, "BrFAST: a Tool to Select Browser Fingerprinting Attributes for Web Authentication According to a Usability-Security Trade-off", *in : Companion Proceedings of the Web Conference 2021 (WWW '21 Companion)*, Ljubljana, Slovenia, April 2021, https://hal.archives-ouvertes.fr/hal-03202788.

[14] T. Guyet, T. Allard, J. Bakalara, O. Dameron, "An open generator of synthetic administrative healthcare databases", *in : IAS 2021 - Atelier Intelligence Artificielle et Santé*, p. 1–8, Bordeaux (virtuel), France, June 2021, https://hal.archives-ouvertes.fr/hal-03326618.

[15] L. Hélouët, Z. Miklos, R. Singh, "Cost and Quality Assurance in Crowdsourcing Workflows (Extended Abstract)", *in : BDA 2021 - 37 eme Conférence sur la Gestion des Données - Principes, Technologies, Applications*, p. 1–2, Paris, France, October 2021, https://hal.inria.fr/hal-03482426.

[16] L. Hélouët, Z. Miklos, R. Singh, "Cost and Quality in Crowdsourcing Workflows", *in : PETRI NETS 2021 - 42nd International Conference on Applications and Theory of Petri Nets and Concurrency*, *Lecture Notes in Computer Science*, *12734*, Springer International Publishing, p. 33–54, Paris, France, June 2021, https://hal.inria.fr/hal-03482424.

[17] G. Lyan, D. Gross-Amblard, J.-M. Jézéquel, "On the Quality of Compositional Prediction for Prospective Analytics on Graphs", *in : DaWaK 2021 - 23rd International Conference on Big Data Analytics and Knowledge Discovery*, p. 91–105, Linz, Austria, September 2021, https://hal.archives-ouvertes.fr/hal-03356199.

[18] G. Lyan, D. Gross-Amblard, J. Jézéquel, "On the Quality of Compositional Prediction for Prospective Analytics on Graphs", *in : Database and Expert Systems Applications - DEXA 2021 Workshops - BIOKDD, IWCFS, MLKgraphs, AI-CARES, ProTime, AISys 2021, Virtual Event, September 27-30, 2021, Proceedings*, G. Kotsis, A. M. Tjoa, I. Khalil, B. Moser, A. Mashkoor, J. Sametinger, A. Fensel, J. M. Gil, L. Fischer, G. Czech, F. Sobieczky, S. Khan (editors), *Communications in Computer and Information Science*, *1479*, Springer, p. 91–105, 2021, https://doi.org/10.1007/978-3-030-87101-7_10.

[19] G. Lyan, J. Jézéquel, D. Gross-Amblard, B. Combemale, "DataTime: a Framework to smoothly Integrate Past, Present and Future into Models", *in : 24th International Conference on Model Driven Engineering Languages and Systems, MODELS 2021, Fukuoka, Japan, October 10-15, 2021*, IEEE, p. 134–144, 2021, https://doi.org/10.1109/MODELS50736.2021.00022.

[20] G. Lyan, J.-M. Jézéquel, D. Gross-Amblard, B. Combemale, "DataTime: a Framework to smoothly Integrate Past, Present and Future into Models", *in : MODELS 2021 - ACM/IEEE 24th International Conference on Model Driven Engineering Languages and Systems*, p. 1–11, Fukuoka, Japan, October 2021, https://hal.inria.fr/hal-03355162.

[21] F. Mentec, Z. Miklós, S. Hervieu, T. Roger, "Conversational recommendations for job recruiters", *in : Knowledge-aware and Conversational Recommender Systems*, Amsterdam, Netherlands, September 2021, https://hal.inria.fr/hal-03537355.

[22] C. Thierry, A. Martin, J.-C. Dubois, Y. Le Gall, "Validation of Smets' hypothesis in the crowdsourcing environment", *in : 6th International Conference on Belief Functions*, Shanghai, China, October 2021, https://hal.archives-ouvertes.fr/hal-03348663.

[23] H. Van Tran, T. Allard, L. D'orazio, A. El Abbadi, "FRESQUE: A Scalable Ingestion Framework for Secure Range Query Processing on Clouds", *in : EDBT 2021 - 24th International Conference on Extending Database Technology*, Nicosia, Cyprus, March 2021, https://hal.inria.fr/hal-03198346.

[24] A. Voyez, T. Allard, G. Avoine, P. Cauchois, E. Fromont, M. Simonin, "Attaque par inférence d'appartenance sur des séries temporelles agrégées en utilisant la programmation par contraintes", *in : BDA 2021 - 37ème Conférence sur la Gestion de*

Données – Principes, Technologies et Applications, p. 1, Paris, France, October 2021, https://hal.archives-ouvertes.fr/hal-03499977.

[25] Z. Zhang, Z. Liu, K. Zhou, A. Martin, Y. Zhang, "Credal Clustering for Imbalanced Data", *in : 6th International Conference on Belief Functions*, *Lecture Notes in Computer Science*, *12915*, Springer International Publishing, p. 13–21, Shanghai, China, October 2021, https://hal.archives-ouvertes.fr/hal-03394857.

[26] Z. Zhang, A. Martin, Z. Liu, K. Zhou, Y. Zhang, "Fast Unfolding of Credal Partitions in Evidential Clustering", *in : 6th International Conference on Belief Functions*, *Lecture Notes in Computer Science*, *12915*, Springer International Publishing, p. 3–12, Shanghai, China, October 2021, https://hal.archives-ouvertes.fr/hal-03394844.

[27] K. Zhou, M. Guo, A. Martin, "Evidential clustering based on transfer learning", *in : International Conference on Belief Functions*, Shanghai, China, October 2021, https://hal.archives-ouvertes.fr/hal-03405161.

[28] D. Zhu, A. Martin, J.-C. Dubois, Y. Le Gall, V. Lemaire, "Modèle crédibiliste pour l'échantillonnage en apprentissage actif", *in : Rencontres francophones sur la logique floue et ses applications*, Paris, France, October 2021, https://hal.archives-ouvertes.fr/hal-03327140.

[29] D. Zhu, A. Martin, Y. Le Gall, J.-C. Dubois, V. Lemaire, "Evidential Nearest Neighbours in Active Learning", *in : Worksop on Interactive Adaptive Learning (IAL) - ECML-PKDD*, Bilbao, Spain, September 2021, https://hal.archives-ouvertes.fr/hal-03327629.