Activity Report 2017

Team DRUID

Declarative & Reliable Management of Uncertain, User-generated & Interlinked Data

D7 – Data and Knowledge Management

1 Team

DRUID is an IRISA team, starting October 2014, supported by 9 active members from distinct IRISA sites, Rennes and Lannion. The team is completed by several Ph.D students. In the sequel, PR means full professor, and MCF means "Maître de conférences", a tenured assistant professor.

Head of the team

David Gross-Amblard, Professor, ISTIC Rennes 1 - Rennes
Arnaud Martin, Professor, IUT Lannion - Lannion

Administrative assistant

Tifenn Donguy, AI CNRS - Rennes

Université Rennes 1 personnel

Tristan Allard, Assistant Professor, ISTIC Rennes 1 - Rennes
Tassadit Bouadi, Assistant Professor, IUT Lannion - Lannion
Jean-Christophe Dubois, Assistant Professor, IUT Lannion - Lannion
Mouloud Kharoune, Assistant Professor, IUT Lannion - Lannion
Yolande Le Gall, Assistant Professor, IUT Lannion - Lannion
Zoltan Miklos, Assistant Professor, ESIR Rennes 1 - Rennes
Virginie Sans, Assistant Professor, ISTIC Rennes 1 - Rennes

Visitors

Amr El Abbadi, Full professor, University of California Santa Barbara, June (last week)-July (first week) 2017
Salma Ben Dhaou, IHEC, Tunisia, PhD student, March-June and September-December 2017

PhD students

Tompoariniaina Andriamilanto, IRT b<>com, Université Rennes 1, France, since September 2017
Dorra Attiaoui, Tunisian grant, ATER IUT Lannion (2016-2017), since June 2013 - defended December 2017
Yann Dauxais, MENRT/Rennes 1 grant, since October 2015, co-advised with Thomas Guyet (LACODAM) and André Happe (CIC-Inserm Rennes)
Joris Degueperoux, ANR/CROWDGUARD, Université Rennes 1, France, since September 2017
Ian Jeantet, ANR/EPIQUE, since December 2017
Na Li, CIFRE Total, since May 2017
Panagiotis Mavridis, MENRT/Rennes 1 grant, since Oct. 2014 - defended Nov. 2017
Yiru Zhang, LTC/ARED, since November 2016

Master students

Raymond Anani Kouame, M1 miage, Université Rennes 1, France, June-August 2017
Tarek Benzima, M2, Polytechnic, Tunisia, March-June 2017
Louis Béziaud, M1 RI, Université Rennes 1 and ENS Rennes, France, May-August 2017
Manel Chehibi, M2, Manouba University, Tunisia, March-June 2017
Joris Degueperoux, M2 MRI, Université Rennes 1 and ENS Rennes, France, February-June 2017
Alexandre Siguret, M1 Miage, Université Rennes 1, France, February-June 2017
Kevin Paratre-Badois, M1 Info, Université Rennes 1, France, February-June 2017
Olivier Pelgrin, M1 Info, Université Rennes 1, France, April-June 2017

2 Overall Objectives

2.1 Overview

Our perception of digital information has completely shifted in recent years, in several ways. First, data are no longer isolated, but are now part of distributed, **interlinked networks**. Such networks include web documents (URLs and URIs), communities in a social graph (*e.g.* FOAF), conceptual networks on the Semantic Web or the continuously growing network of Linked Open Data (RDF). Second, data are now dynamic. Obtaining an up-to-date piece of information is as simple as a Web service call or a syndication (as is RSS or Atom). A large diversity of such dynamic data sources is available, including corporate Web services, wireless sensors in the environment, humans in the participative Web, or workers in crowdsourcing platforms[1]. Hence, what becomes important is the **data source** itself.

The openness and liveliness of such interlinked data networks is a great opportunity. Business Intelligence applications no longer restrict their attention to the companies own data sets or sales records, but try to incorporate data collected from the Web (such as opinions from social networks or Web forums). In this way they can extract useful information about their customers and the reception of their products and services. Another domain is the integration of personal information from multiple devices. The same opportunities arise also in the context of non-profit organizations or societal challenges: there is a lot of information available on health problems (Web forums on health, body area networks), environmental issues (environmental sensors) or in administrative domains (smart cities, Open Data initiatives). A new key issue is also to benefit from the growing "digital presence", that allows **interaction** with users at virtually any moment through mobile phone applications such as Twitter. Feedback loops between users and data managers can now be devised[2].

But the diversity and the dynamic of data sources raise several challenges. One can legitimately question if a data source is reliable or malevolent or if two data sources are independent These problems are strengthened by the mutual links between data items or data sources. Hence fact provenance and sources independence are prominent data annotations that shall be taken into account. For user-generated or crowdsourced content, knowing the skills or the

[1] https://www.mturk.com

[2] See for example participative journalism platforms, https://witness.theguardian.com/moreabout

social relationships between participants allows for a better understanding of the produced raw data. This calls for a powerful **qualification mechanisms** that would integrate these annotations and help data managers in understanding their data and selecting their sources. Furthermore, even if interaction with participants is technically possible, the **orchestration** of complex data acquisition tasks from a mass still remains a black art.

The objective of the DRUID team is to provide models and algorithms for the annotation and management of interlinked data and sources at a large scale. We consider three main goals:

1. To propose well-founded models for interlinked data and, more importantly, interlinked data sources (for example, profiling users in a social network, orchestrating users and tasks in a crowdsourcing platform),
2. To develop theories for the qualification of such data and sources in terms of reliability, certainty, provenance, influence, economical value, trust, etc.
3. To implement systems that are proof-of-concepts of these models and theories. In particular we would like to demonstrate that these systems can overcome specific key problems in real-world applications, such as scalable data qualification and data adaptation to the final users.

More concretely, we would like to address the following challenges:

- to develop integrated and scalable analysis tools for participants in social networks, that encompass the semantics of communications between users, computes user influence or user independence for example.
- to extend existing crowdsourcing platforms with fine user profiles, team building or complex task management abilities, with application for e-science or e-government (smart cities).
- to develop reliability assessment techniques for large sensor networks (uncertainty), heterogeneous data sources or Linked Open Data (quality), or microblog conversations (misinformation).
- to adapt data to its use (data visualization, accessibility of information).

2.2 Key issue 1: Well-Founded Models for Interlinked Data and Sources

The Data Management field aims to build pertinent models for information, expressive query languages at a high level of abstraction for computer engineers or basic users, and efficient evaluation methods. The field was successful with a wide acceptance of solutions at the industrial level (banking, electronic commerce, document management, ticketing, etc.).

But classical approaches are not directly suited for nowadays applications. On the one hand, with the spreading of graph data models such as RDF, data are no longer relational (structured into tables) not even tree-based (XML), but graph-based (reminiscent of the semi-structured data model). Furthermore, these graphs are no longer centralized but interlinked through the network. The success of NoSQL graph database for social networks is an illustration[3]. New models are then required to express queries on such graphs in a well-founded manner[BLLW12].

On the other hand, the very structure of data sources should now be investigated. A first example is sensors (in a broad sense, from specialized sensors to smart phones sensors or personal health monitors). Such devices support severe constraints on their connectivity and their ability to provide data. They are mobile and energy-restricted. A more recent and striking example is to consider also humans as data sources. These sources are related to each others (social network) and they produce data with a rich semantics (see for example post contents in a forum). Hence being able to query such data sources with a clean language, while taking into account their relationships and reasoning about their semantics would greatly impact practical applications. A typical relevant query is to find the central user of a social network (structural query), restricted to the subset of users talking about action movies (semantic query).

Finally, it is now possible to interact with data sources, as in large participative, crowd-based systems that gather information (e.g. participative science) or resolve tasks (Human based computing)[4]. Having a clean framework to organize such interactions would also benefit to these applications.

Our first objective is to provide well-funded models for interlinked data sources (social graphs, microblogs, sensors) and complex workflows of human tasks.

The sub-goals of this scientific axis are listed below, ordered by their priority (short terms are already started, mid terms cover a classical Ph.D duration, and long terms target prospective issues).

[short term] Adding semantics Integration of the semantics of the information flow within networks, as many properties in social graphs are not only syntactical or linked-based. Relevant tools are taxonomies, ontologies but also sentiment analysis and controversy analysis.

[mid term] Querying social graph data Few query languages are available to reason about the structure of a (huge) social graph. For example, selecting nodes that are part of distinct communities is hardly expressible without relying on a ad hoc program.

[mid term] Modeling users Modeling a user as a data source with it's profile, opinion, social network, motivations, personal goals, personal strategies, location and available time (e.g. for real-time crowdsourcing applications).

[3] Indeed, Facebook is using a NoSQL, graph-oriented database for its core data, Neo4j, `http://www.zdnet.com/facebook-neo4j-7000009866/`.

[4] As an example, the micro-tasking platform AMT (Amazon Mechanical Turk) had in 2013 more than 300 000 users available at any time to resolve a task, `https://requester.mturk.com/tour`.

[BLLW12] P. Barceló, L. Libkin, A. W. Lin, P. T. Wood, "Expressive Languages for Path Queries over Graph-Structured Data", *ACM Trans. Database Syst. 37*, 4, 2012, p. 31.

[mid term] Mixing queries about data and sources As a new query type, we have for example "give me the average ranking of this movie from users who are absolutely not connected with me, but have shown a long habit in watching and ranking movies". The first part of the query is related to the graph underlying the social network, the second part to the properties of the data source.

[mid term] Crowdsourcing complex tasks Most existing crowdsourcing platforms are not generic and the deployment of complex tasks supposes a huge development cost[ABMK11]. Recently, declarative crowdsourcing systems, mixing database approaches with user interaction have emerged[FKK+11,PPP+13]. These first efforts reduce the development cost of simple data curation tasks. We propose to model complex tasks using declarative workflow models[HMT12, AV13] in order to reason about the correctness of complex, human-based computation processes.

[long term] Integrated interaction model Our goal is to break the discrepancy between the content on the one side, and the users generating this content on the other side. We would like to achieve a generic model where one can reason both on the graph structure of data (paths, clusters), the social structure of the users (skills, friendship, team structure, centrality), and on the social workflows that connect them.

2.3 Key issue 2: Interlinked Data and Sources Qualification

By qualification we mean any type of qualitative and quantitative indicators on data and sources of data. We can evoke as examples data uncertainty, imprecision, economical and strategic value, privacy, accessibility, data provenance and also reliability, expertise, independence, conflict of sources of data.

> Our second objective is to provide qualification mechanisms for interlinked data and sources, taking into account their mutual interactions and the available information, even if this information is unsure and imprecise.

Data and sources qualification is a great social need for the contemporary Web. Users shall be enlighten by the provenance, quality and accessibility of their information sources.

[ABMK11] S. Ahmad, A. Battle, Z. Malkani, S. Kamvar, “The jabberwocky programming environment for structured social computing”, *in: Proceedings of the 24th annual ACM symposium on User interface software and technology, UIST '11*, p. 53–64, 2011.

[FKK+11] M. J. Franklin, D. Kossmann, T. Kraska, S. Ramesh, R. Xin, “CrowdDB: answering queries with crowdsourcing”, *in: Proceedings of the 2011 ACM SIGMOD International Conference on Management of data, SIGMOD '11*, p. 61–72, 2011.

[PPP+13] H. Park, R. Pang, A. Parameswaran, H. Garcia-Molina, N. Polyzotis, J. Widom, “An overview of the deco system: data model and query language; query processing and optimization”, *SIGMOD Rec. 41*, 4, January 2013, p. 22–27.

[HMT12] R. Hull, J. Mendling, S. Tai, “Business process management”, *Inf. Syst. 37*, 6, 2012, p. 517.

[AV13] S. Abiteboul, V. Vianu, “Models for Data-Centric Workflows”, *in: In Search of Elegance in the Theory and Practice of Computation*, V. Tannen, L. Wong, L. Libkin, W. Fan, W.-C. Tan, M. P. Fourman (editors), *Lecture Notes in Computer Science*, *8000*, Springer, p. 1–12, 2013.

We mention three important directions:

[short term] Assessing social network reliability Using social networks can exhibit some risks users are not aware of. Indeed, erroneous information can be send deliberately or involuntarily (by lack of scrutiny, of from hacked accounts). Information in social networks can easily be distorted and amplified according to relationships between relaying users. Even if information is corroborated by several contacts, its source can be unique and erroneous. There is a crucial need for tools to evaluate social networks reliability and weaknesses, in order to take valuable decisions. The relationships between users has to be taken into account, along with the quality and amount of independent data sources. A great challenge is to identify relevant information in a mass of data exchange and to predict real events from purely electronic activities.

[short term] Interlinked data integration Schema integration is a long standing problem in information systems. This problematic is amplified by the relationship between data sets. We propose the notion of schema networks to model this situation, and techniques to provide schema mappings as an equilibrium within this network.

[mid term] Interlinked data fusion More and more information systems gather information from network-organized sensors. The vanishing price of such sensors allows their use in everyday life and tools. They are also used in dedicated applications such as military watching, aerial, terrestrial or oceanic missions (sensor swarms). In such complex networks, sensors do not play an equal role: some may be dedicated to observation, others to positioning, or communication. The flow of information can be altered because of sensor deficiency or the structure of the network itself. In such scenarios, data fusion must be preceded by a correct qualification of data and sources. Such qualification also leverages reliability [KM13], independence [CMBY12] and conflict measurement [Mar12]. While mature approaches for data fusion already exist, the network structure is rarely studied. One of our goal is then to propose efficient methods that incorporate this structure.

[mid term] Crowdsourcing quality optimization In crowdsourcing applications, data quality is a central concern. Many techniques have been envision to enhance this quality, by, for example, performing majority voting between redundant tasks. Since our first goal is to go beyond simple query-answer tasks, that is to encompass their composition, adapted quality enhancement mechanisms have to be designed accordingly. As an example, we will consider models of user motivation to select which part of a complex task is more suited to a given user. Other directions concern designing incentive to motivate users, and taking the reputation of

[KM13] M. Kharoune, A. Martin, "Mesure de dépendance positive et négative de sources crédibilistes", *in: EGC - Atelier Fouille de données complexes*, Toulouse, France, January 2013.

[CMBY12] M. Chebbah, A. Martin, B. Ben Yaghlane, "Positive and negative dependence for evidential database enrichment", *in: IPMU*, p. 575–584, Italy, July 2012.

[Mar12] A. Martin, "About conflict in the theory of belief functions", *in: International Conference on Belief Functions*, France, 8-10 May 2012.

users into account. The mixing of users skills and the knowledge of their social network is also a natural direction. We also would like to allow the user to provide a self estimation of his input accuracy. This feedback would aim to estimate the imprecision and the level of certainty of his answers, in order to optimize the decision process.

[long term] Integrated annotation model Our vision is a transparent data model that accepts and triggers any kind of source contribution in a non-blocking way, while offering a coherent, qualified view of the data set at any time and from any user perspective. Our goal is to keep the model simple in order to promote its adoption by industry.

2.4 Key issue 3: Data & Sources Management: Large Scale, High Rate, Ease of Use

The two previous goals we just introduced will provide models for interlinked data and sources, along with rich qualification mechanisms.

Our third objective is to provide fully integrated systems that allow for the manipulation of interlinked data and interlinked sources, along with rich qualification indicators, while being efficient and adapted to users.

Two ingredients are needed for the success of such systems: scalability and ease of use. We discuss these two issues in the sequel.

Optimization and scalability From the efficiency point of view, many problems arise. Qualification indicators may appear as meta-data in the core of a data management system, with the difficulty of their storage. But the main challenge is the algorithmic complexity of their computation. In order to deal with high volume or rate of data, the proposed algorithms should be designed for scalability. Several directions are envisioned:

- **[short term]** Using distributed computation paradigms such as MapReduce, and iterative computations such as PageRank and variants.
- **[mid term]** Relying on controlled approximation algorithms: only an estimate of the correct data qualification will be obtained, but with a small error (say 5%), with a small failure probability (say one chance over 1 billion), but with a rapid computation time.
- **[mid term]** Filtering relevant information with coarse-grain qualification estimate, in order to reduce the amount of data (for example, to reduce the number of focal elements for belief functions approaches).
- **[long term]** Using streaming algorithms, where computations are done on the fly, also with a controlled error.

Security The crowdsourcing litterature is growing at a fast pace. However, security and privacy have been ignored until now in crowdsourcing contexts despite their importance. Indeed, crowdsourcing processes involve (1) exporting data and workflows to the crowd, and/or (2) collecting data and results from the crowd. We plan to study privacy and security issues that arise in these contexts.

- **[mid term]** Exporting Data and Workflows to the Crowd. Most works have focused on exploiting the crowd by delegating specific tasks to workers. Usually, the task specification involves sending data to workers together with the task specification. For example, matching pairs of similar items [WLK+13] requires sending the items to workers, or planning a schedule [KLMN13] requires sending the objects or actions to be planned and their constraints. However, sensitive data cannot be sent in the clear to participants. In a traditional context, where only machines participate to the computation, strong cryptographic protocols can let machines participate without accessing non-encrypted sensitive data. In a crowdsourcing context, such cryptographic protocols cannot be used anymore because they would simply preclude humans to participate. Data has to be disclosed to humans. How can we disclose sensitive data to humans in crowdsourcing processes while still guaranteeing its privacy? Similarly, involving the crowd in a complex workflows implies disclosing each task to the human workers to which it is assigned. How can we guarantee the confidentiality of workflows, *e.g.*, for intellectual property reasons, while still allowing the participation of workers?

- **[mid term]** Collecting Data and Results from the Crowd. The crowd can be viewed as a specific database that can be, *e.g.*, queried [PW14], indexed [ADM+14], or mined [AAM14,ADM+14]. However, data that is collected by such algorithms is individual data and may be consequently identifying or sensitive. How can we guarantee the privacy of individual data in crowdsourcing data-oriented processes? Protecting such obviously-sensitive data is however not sufficient. Indeed, covert channels may exist and lead to the disclosure of sensitive data. For example, a worker may answer intriguingly fastly

[WLK+13] J. Wang, G. Li, T. Kraska, M. J. Franklin, J. Feng, "Leveraging Transitive Relations for Crowdsourced Joins", *in: Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data, SIGMOD '13*, ACM, p. 229–240, New York, NY, USA, 2013, http://doi.acm.org/10.1145/2463676.2465280.

[KLMN13] H. Kaplan, I. Lotosh, T. Milo, S. Novgorodov, "Answering Planning Queries with the Crowd", *Proc. VLDB Endow. 6*, 9, July 2013, p. 697–708, http://dx.doi.org/10.14778/2536360.2536369.

[PW14] H. Park, J. Widom, "CrowdFill: Collecting Structured Data from the Crowd", *in: Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data, SIGMOD '14*, ACM, p. 577–588, New York, NY, USA, 2014, http://doi.acm.org/10.1145/2588555.2610503.

[ADM+14] Y. Amsterdamer, S. B. Davidson, T. Milo, S. Novgorodov, A. Somech, "OASSIS: Query Driven Crowd Mining", *in: Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data, SIGMOD '14*, ACM, p. 589–600, New York, NY, USA, 2014, http://doi.acm.org/10.1145/2588555.2610514.

[AAM14] A. Amarilli, Y. Amsterdamer, T. Milo, "On the Complexity of Mining Itemsets from the Crowd Using Taxonomies", *in: Proc. 17th International Conference on Database Theory (ICDT), Athens, Greece, March 24-28, 2014.*, p. 15–25, 2014, http://dx.doi.org/10.5441/002/icdt.2014.06.

to questions related to a given disease or to a given place. This may reveal a surprising strong connection between the worker and this disease or place. How can we protect workers from covert channel attacks in crowdsourcing processes?

Data presentation Beside efficiency, there is a tremendous need from end-user for an adapted presentation of information. The amount of available data along with the rich annotations we will add are certainly overwhelming for any user. We will consider in this axis also

- **[short term]** Data adaptation methods, that filter information according to the user's needs and capabilities: on a static or mobile environment, on-line or off-line, with or without real-time needs, disabled persons, seniors, and so on.
- **[long term]** Data visualization methods, that present a visual and navigational summary in qualified data.

3 Scientific Foundations

3.1 Data management

To achieve our goals we will rely on techniques of two scientific domains: data management and data qualification. For data management we will naturally elaborate on classical techniques: finite model theory, complexity theory, approximation algorithms, declarative or algebraic languages, execution plans, costs models, indexing. We intend to explore new models such as schema networks for data integration, user modeling for crowdsourcing application[RLT+13], and game theory for the study of strategic aspects in crowdsourcing, data pricing[LLMS13] and data publication[JP13]. For the modeling of complex tasks in crowdsourcing, we envision to extend declarative approaches for business processes[DM12], such as the collaborative business artifact model[AV13].

3.2 Data qualification

For data qualification, our first focus will be on uncertainty. Many frameworks are available, but all are based on the theories of uncertainty that are able to model imperfect data. Two

[RLT+13] S. B. Roy, I. Lykourentzou, S. Thirumuruganathan, S. Amer-Yahia, G. Das, "Crowds, not Drones: Modeling Human Factors in Interactive Crowdsourcing", *in: DBCrowd*, p. 39–42, 2013.

[LLMS13] C. Li, D. Y. Li, G. Miklau, D. Suciu, "A theory of pricing private data", *in: ICDT*, W.-C. Tan, G. Guerrini, B. Catania, A. Gounaris (editors), ACM, p. 33–44, 2013.

[JP13] S. Jain, D. C. Parkes, "A game-theoretic analysis of the ESP game", *ACM Trans. Econ. Comput. 1*, 1, January 2013, p. 3:1–3:35.

[DM12] D. Deutch, T. Milo, *Business Processes: A Database Perspective*, *Synthesis Lectures on Data Management*, Morgan & Claypool Publishers, 2012.

[AV13] S. Abiteboul, V. Vianu, "Models for Data-Centric Workflows", *in: In Search of Elegance in the Theory and Practice of Computation*, V. Tannen, L. Wong, L. Libkin, W. Fan, W.-C. Tan, M. P. Fourman (editors), *Lecture Notes in Computer Science*, *8000*, Springer, p. 1–12, 2013.

main aspects of imperfection are classically distinguished: uncertainty and imprecision[Sme97].

In particular, the theory of belief functions[Dem67,Sha76] (also commonly referred to as evidence theory or Dempster-Shafer theory) allows to take simultaneously into account both uncertainty and imprecision. 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[DPS96]. Its strength lies in (1) its richer representation of uncertainty and imprecision compared to probability theory and (2) its higher ability to combine pieces of information. In particular, a crucial task in information fusion is the management of conflict between different (partially or totally) disagreeing sources. The origins of conflict can come from the source reliability, disinformation, truthfulness, etc. For interlinked data such as posts flowing through a social network, we also have to consider the quality of the data, especially its uncertainty and imprecision. We can also see each node of the network as a node of information fusion. The framework of belief functions is therefore well adapted.

Two main difficulties are to be underlined: first, to find a correct definition of data quality (in order to encompass reliability, truthfulness, disinformation) combined with source quality (reliable experts, liars, collusion,

trollers) [PDD12,Sme93] ;

second to possibly resolve the problem of scalability associated with belief function approaches (still, the corresponding complexity is lower than for other approaches such as imprecise probability theories or random set theories).

4 Application Domains

4.1 Generic Crowdsourcing Platform, Data Annotation and Sensing

Participants: Tristan Allard, Tassadit Bouadi, David Gross-Amblard, Jean-Christophe Dubois, Mouloud Kharoune, Yolande Le Gall, Panagiotis Mavridis, Arnaud Martin, Zoltan Miklos, Virginie Sans.

The models we develop for crowdsourcing provide a strong basis for the development of a generic crowdsourcing platform that can be adapted to various uses. We envision for now to target two kinds of applications: data annotations and data sensing. In data annotation,

[Sme97] P. Smets, "Imperfect information: Imprecision - Uncertainty", *in: Uncertainty Management in Information Systems*, A. Motro and P. Smets (editors), Kluwer Academic Publishers, 1997, p. 225–254.

[Dem67] A. P. Dempster, "Upper and Lower probabilities induced by a multivalued mapping", *Annals of Mathematical Statistics 38*, 1967, p. 325–339.

[Sha76] G. Shafer, *A mathematical theory of evidence*, Princeton University Press, 1976.

[DPS96] D. Dubois, H. Prade, P. Smets, "Representing Partial Ignorance", *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans 26*, 3, 1996, p. 361–377.

[PDD12] F. Pichon, D. Dubois, T. Denoeux, "Relevance and truthfulness in information correction and fusion", *International Journal of Approximate Reasoning 53*, 2, 2012, p. 159 – 175, Theory of Belief Functions (BELIEF 2010).

[Sme93] P. Smets, "Belief Functions: the Disjunctive Rule of Combination and the Generalized Bayesian Theorem", *International Journal of Approximate Reasoning 9*, 1993, p. 1–35.

the crowd is asked to tag a set of resources (images, videos, locations, etc.) using a free or controlled vocabulary. In data sensing, the crowd is consulted to obtain any kind of data, say for example environmental measurements (temperature, weather, water quality, etc.) or personal information (location, speed, feelings about a place, etc.). The role of the crowdsourcing platform is to orchestrate crowd interactions and to protect (sanitize) the collection of private information.

4.2 Social Network Analysis for Humanities and Marketing

Participants: Dorra Attiaoui, Salma Ben Dahou, Tassadit Bouadi, David Gross-Amblard, Mouloud Kharoune, Arnaud Martin, Zoltan Miklos, Virginie Sans, Yiru Zhang.

We consider social network analysis by the way of heterogeneous social networks where we integrate the models of imperfect linked data. Therefore, we consider several problems for social network analysis such as the community detection, experts and trolls identification and message's propagation for example for viral marketing applications. Hence, we consider different kinds of social network such as Twiter and dblp. We also test our models on generated networks.

5 Software

5.1 ibelief

Participants: Kuang Zhou, Arnaud Martin [contact point].

The R package ibelief aims to provide some basic functions to implement the theory of belief functions, and it has included many features such as:

1. Fast Mobius Transformation to convert any of the belief measures (such as basic belief assignment, credibility, plausibility and so on) to another type;
2. Some commonly used combination rules including DS rule, Smets' rule, Yager's rule, DP rule, PCR6 and so on;
3. Some rules for making decisions;
4. The discounting rules in the theory of belief functions;
5. Different ways to generate random masses.

The stable version of package ibelief could be found on CRAN (common R code repository). In 2016 a new rule has been added in order to combine a large number of basic belief assignments.

5.2 Crowd

Participants: Tristan Allard, Tassadit Bouadi, David Gross-Amblard [contact point], Panagiotis Mavridis, Zoltan Miklos, Virginie Sans.

We have realized in 2014 [CGAG+14] the prototype of a crowdsourcing platform that can execute complex tasks that one can obtain as a composition of simple human intelligence tasks. The platform uses a skill model to assign tasks. A better founded platform is under developpement in the ANR HEADWORK project (see below).

6 New Results

6.1 Social Network Analysis

Participants: Dorra Attiaoui, Salma Ben Dahou, Tassadit Bouadi, Mouloud Kharoune Arnaud Martin, Yiru Zhang.

The web plays an important role in people's social lives since the emergence of Web 2.0. It facilitates the interaction between users, gives them the possibility to freely interact, share and collaborate through social networks, online communities forums, blogs, wikis and other online collaborative media.

Expert detection in Question Answering communities

The Dorra Attiaoui's thesis [**?**] is focus on persons' characterization in question answering communities on Stack Owerflow.

During the last decade, people have changed the way they seek information online. Between question answering communities, specialized websites, social networks, the Web has become one of the most widespread platforms for information exchange and retrieval. Question answering communities provide an easy and quick way to search for information needed in any topic. The user has to only ask a question and wait for the other members of the community to respond. Any person posting a question intends to have accurate and helpful answers. Within these platforms, we find experts. They are key users that share their knowledge with the other members of the community. Expert detection in question answering communities has become important for several reasons such as providing high quality content, getting valuable answers, etc. In this thesis, we are interested in proposing a general measure of expertise based on the theory of belief functions. Also called the mathematical theory of evidence, it is one of the most well known approaches for reasoning under uncertainty. In order to identify experts among other users in the community, we have focused on finding the most important features that describe every individual. Next, we have developed a model founded on the theory of belief functions to estimate the general expertise of the contributors . This measure will allow us to classify users and detect the most knowledgeable persons. Therefore, once this metric defined, we look at the temporal evolution of users' behavior over time. We propose an analysis of users activity for several months in community. For this temporal investigation, we will describe how do users evolve during their time spent within the platform. Besides, we are also interested on

[CGAG+14] A. Chettih, D. Gross-Amblard, D. Guyon, E. Legeay, Z. Miklós, "Crowd, a platform for the crowdsourcing of complex tasks", *in : BDA 2014 : Gestion de données - principes, technologies et applications*, p. 51–55, Autrans, France, October 2014, https://hal.archives-ouvertes.fr/hal-01163824.

detecting potential experts during the beginning of their activity. The effectiveness of these approaches is evaluated on real data provided from Stack Overflow.

The results of the thesis have been published in 2017, in [?] and in [?].

Influencers characterization in a social network for viral marketing

The Siwar Jendoubi's thesis was focus on influencers characterization. In 2017, we continue this work. Influence maximization is the problem of selecting a set of influential users in the social network. Those users could adopt the product and trigger a large cascade of adoptions through the "word of mouth" effect. We propose two evidential influence maximization models for Twitter social network. The proposed approaches use the theory of belief functions to estimate users influence. Furthermore, the proposed influence estimation measure fuses many influence aspects in Twitter, like the importance of the user in the network structure and the popularity of user's tweets (messages). In our experiments, we compare the proposed solutions to existing ones and we show the performance of our models. This result has been accepted in [?] and in [?].

Evidential Attributes in Social Network

Currently, there are many approaches designed for the task of detecting communities in social networks. Among them, some methods only consider the topological graph structure, while others can take use of both the graph structure and the node attributes. In real-world networks, there are many uncertain and noisy attributes in the graph. In this paper, we will present how we can detect communities for graphs with uncertain attributes in the first step. The numerical, probabilistic as well as evidential attributes are generated according to the graph structure. In the second step, some noise will be added to the attributes. We perform experiments on graphs with different types of attributes and compare the detection results in terms of the Normalized Mutual Information (NMI) values. The experimental results show that the clustering with evidential attributes give better results comparing to those with probabilistic and numerical attributes. This illustrates the advantages of evidential attributes. This result has been published in [?].

Preference fusion in social network

Facing an unknown situation, a person may not be able to firmly elicit his/her preferences over different alternatives, so he/she tends to express uncertain preferences. Given a community of different persons expressing their preferences over certain alternatives under uncertainty, to get a collective representative opinion of the whole community, a preference fusion process is required. The aim of this work is to propose a preference fusion method that copes with uncertainty and escape from the Condorcet paradox. To model preferences under uncertainty, we propose to develop a model of preferences based on belief function theory that accurately describes and captures the uncertainty associated with individual or collective preferences. The benefits of our contribution are twofold. On the one hand, we propose a qualitative and expressive preference modeling strategy based on belief-function theory which scales better with the number of sources. On the other hand, we propose an incremental distance-based algorithm (using Jousselme distance) for the construction of the collective preference order to avoid the Condorcet Paradox. This result has been published in [?].

6.2 Characterization of experts in crowdsourcing platforms

Participants: Tarek Benzina, Jean-Christophe Dubois, Mouloud Kharoune, Yolande Le Gall, Arnaud Martin, Zoltan Miklos.

In [?], we have propose an expertise measure on real data given by orange labs. The goal is here to evaluate the work quality of the participants, a major issue in crowdsourcing. Indeed, contributions must be controlled to ensure the effectiveness and relevance of the campaign. We are particularly interested in small, fast and not automatic tasks. Several methods have been proposed to solve this problem, but they are applicable when the "golden truth" is not always known. This work has the particularity to propose a method for calculating the degree of expertise in the presence of gold data in crowdsourcing. This method is based on the belief function theory and proposes a structuring of data using graphs. The proposed approach will be assessed and applied to the real data established by several music tracks for which the participants have to note.

The main idea of the proposed approach is to build a kind of preference graph from the notes given by the participants. This graph is compared by a new similarity measure to a known graph of the real notes. This similarity is based on four degrees (exactitude degree, confusion degree, false order with the previous node and following node degrees). These degrees are modeling by a basic belief assignment and combine in order to give an expert measure. This measure can be used to consider some notes on tracks without knowledge. An extended version has been submitted.

6.3 Combination in the theory of belief functions

Participants: Arnaud Martin.

The combination of information in the theory of belief functions can still be a problem according to the data and the waiting properties of the combination rule. That is the reason why a new rule has been proposed in [?] adapted for large number of sources. Two different works have been published in [?] and in [?] in order to better classify the data according to the noise of the data and to the sources. In [?], a state of art of conflict management in information fusion in proposed.

6.4 A datawarehouse for simulation data

Participants: Tassadit Bouadi.

Spatially distributed agro-hydrological models allow researchers and stakeholders to represent, understand and formulate hypotheses about the functioning of agro-environmental systems and to predict their evolution. These models have guided agricultural management by simulating effects of landscape structure, farming system changes and their spatial arrangement on stream water quality. Such models generate many intermediate results that should be managed, analyzed and transformed into usable information. In [?], we describe a data warehouse (N-Catch) built to store and analyze simulation data from the spatially distributed agro-hydrological model TNT2. We present scientific challenges to and tools for building data warehouses and describe the three dimensions of N-Catch: space, time and an original hierarchical description of cropping systems. We show how to use OLAP to explore and extract all kinds of useful high-level information by aggregating the data along these three dimensions and how to facilitate exploration of the spatial dimension by coupling N-Catch with GIS. Such tool constitutes an efficient interface between science and society, simulation remaining a research activity, exploration of the results becoming an easy task accessible for a large audience.

6.5 Elicitation of personal preferences

Participants: Tristan Allard, Tassadit Bouadi, Joris Duguépéroux, Virginie Sans.

Ever-increasing quantities of personal data are generated by individuals, knowingly or unconsciously, actively or passively (*e.g.*, bank transactions, geolocations, posts on web forums, physiological measures captured by wearable sensors). Most of the time, this wealth of information is stored, managed, and valorized in isolated systems owned by private companies or organizations. Personal information management systems (PIMS) propose a groundbreaking counterpoint to this trend. They essentially aim at providing to any interested individual the technical means to recollect, manage, integrate, and valorize his/her own data through a dedicated system that he/she owns and controls. In [?], we consider personal preferences as first-class citizens data structures. We define and motivate the threefold preference elicitation problem in PIMS-elicitation from local personal data, elicitation from group preferences, and elicitation from user interactions. We also identify hard and diverse challenges to tackle (*e.g.*, small data, context acquisition, small-scale recommendation, low computing resources, data privacy) and propose promising research directions.

6.6 Privacy-Preserving Crowdsourcing

Participants: Tristan Allard, Louis Béziaud, David Gross-Amblard.

Crowdsourcing platforms dedicated to work are used by a growing number of individuals and organizations, for tasks that are more and more diverse, complex, and that require very specific skills. These highly detailed worker profiles enable high-quality task assignments but may disclose a large amount of personal information to the central platform (*e.g.,* personal preferences, availabilities, wealth, occupations), jeopardizing the privacy of workers. In this work, we propose a lightweight approach to protect workers privacy against the platform along the current crowdsourcing task assignment process. Our approach (1) satisfies differential privacy by letting each worker perturb locally her profile before sending it to the platform, and (2) copes with the resulting perturbation by leveraging a taxonomy defined on workers profiles. We overview this approach below, explaining the lightweight upgrades to be brought to the participants. We have also shown formally that our approach satisfies differential privacy, and empirically, through experiments performed on various synthetic datasets, that it is a promising research track for coping with realistic cost and quality requirements.

7 Contracts and Grants with Industry

7.1 CROWDGUARD

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

Acronym	**Crowdguard**
Call	**ANR JCJC**
Year	2016
Title	**GUARanteeD confidentiality and efficiency in CROWDsourcing platforms**
Coordinator	Tristan Allard
Funding	144 720 euros
Length	42 months

Crowdsourcing platforms offer the unprecedented opportunity to connect easily on-demand task providers, or taskers, and on-demand voluntary work, and for various kinds of tasks. By facilitating the accurate search of specific workers, otherwise unavailable, they have the potential to reduce costs as well as to accelerate and even democratize innovation. Their growing importance has made them unavoidable actors of the 21^{st} century economy. However, abusive behaviors from crowdsourcing platforms against taskers or workers are frequently reported in the news or on dedicated websites, whether performed willingly or not, putting them at the epicenter of a burning societal debate. Real-life examples of such abusive behaviors range from strong concerns about private information accesses and uses (see, *e.g.,* the privacy scandals due to illegitimate accesses to the location data of a well-known drivers-riders company [5]) to blatant denials of workers' independence (see, *e.g.,* the complaints of micro-task workers or of drivers about the strong work control and monitoring imposed by their respective platforms [6]). This fuels the growing concern of individuals, overshadowing the possible benefits that crowdsourcing processes can bring to societies. In addition to obvious legal and ethical reasons, protecting both taskers and workers - *i.e.,* the two sides of a crowdsourcing platform - from the platform itself, is thus crucial for establishing sound trust foundations.

The goal of the CROWDGUARD project is to design sound protection measures of the taskers and workers from threats coming from the platform, while still enabling the latter to perform efficient and accurate tasks assignments. In CROWDGUARD, we advocate for an approach that uses confidentiality and privacy guarantees as building blocks for preventing a large variety of abusive behaviors. First, the enforcement of privacy and confidentiality guarantees directly prevents the first kind of abuse that we consider, i.e., the abusive usage of the personal or confidential information that taskers and workers disclose to the platform for the assignment of tasks. Second, through their obfuscation abilities, privacy and confidentiality guarantees carry the promise, in an extended form, to be also efficient for preventing a large variety of abusive behaviors (e.g, non-discrimination, or workers' independence).

[5] `https://tinyurl.com/wp-priv`

[6] `https://tinyurl.com/wsj-ind` and `https://tinyurl.com/trans-ind`

The CROWDGUARD project will specify relevant use-cases, extracted from real-life situations and illustrating the need to protect the crowd from various abusive behaviors from the platform. The project will propose secure distributed algorithms for allowing workers (resp. taskers) to collaboratively compute a privacy-preserving version of their profiles (resp. a confidentiality-preserving version of their tasks) which will then be sent to the platform. The resulting tasks and profiles will enable highly efficient and accurate crowdsourcing processes while being protected by sound confidentiality and privacy guarantees. CROWDGUARD will also identify and formalize the possible abusive behaviors that the platform may perform, and propose sound models/algorithms to prevent them. Finally, the project will develop a prototype that will be used for evaluating the efficiency of the techniques proposed.

The main scientific outcomes of CROWDGUARD will advance the state-of-the-art on sound models and algorithms for the definition and prevention of abusive behaviors from crowdsourcing platforms. They will enable the development of respectful crowdsourcing processes by private companies or associations.

7.2 EPIQUE

Participants: Zoltan Miklos [Contact point], David Gross-Amblard, Tristan Allard, Arnaud Martin, Virginie Sans, Ian Jeantet, Mickael Foursov.

Acronym	**EPIQUE**
Call	**ANR Generic**
Year	2016
Title	**Large-scale phylomemetic networks**
Coordinator	Zoltan Miklos
Funding	142560 euros (IRISA) / 599800 euros (Total project)
Length	42 months

The evolution of scientific knowledge is directly related to the history of humanity. Document archives and bibliographic sources like the “Web Of Science” or PubMed contain a valuable source for the analysis and reconstruction of this evolution. The proposed project takes as starting point the contributions of D. Chavalarias and J.P. Cointet about the analysis of the dynamicity of evolutive corpora and the automatic construction of “phylomemetic” topic lattices (as an analogy with genealogic trees of natural species). Currently existing tools are limited to the processing of medium sized collections and a non interactive usage. The expected project outcome is situated at the crossroad between Computer science and Social sciences. Our goal is to develop new highly performant tools for building phylomemetic maps of science by exploiting recent technologies for parallelizing tasks and algorithms on complex and voluminous data. These tools are conceived and validated in collaboration with experts in philosophy and history of science over large scientific archives.

In 2017 we started our reflections on the design of possible quality metrics for phylomemetic structures. Our analytical work was complemented by the data exploration. As we still do not have access to the entire WebOfScience dataset, our exploration was focused on a smaller but available dataset from medline. Two interns, who worked with us for 3 months during the summer months, helped us to explore the data and to better understand the challenges in the phylomemetic structure reconstruction and in user interaction design, that could be involved in this process. Mickael Fourson joined the project in September. He continues the data exploration work that we started with the interns. On possible path that seems promising is to use a vector space representation of the data and analyze the clustering methods and the analysis of the evolution in this representation. We started this work in December 2017, with Ian Jeantet.

7.3 HEADWORK

Participants: Tristan Allard, Tassadit Bouadi, David Gross-Amblard [contact point], Panagiotis Mavridis, Zoltan Miklos, Virginie Sans.

Acronym	**HEADWORK**
Call	**ANR PRCE**
Year	2016
Title	**Crowdsourcing Management Systems**
Coordinator	David Gross-Amblard
Funding	146 kE (IRISA) / 800 kE (Total project)
Length	48 months

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 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.

In 2017 we have launched the project (Consortium building and signing, kick-off meeting, recruitment). Advancements are listed on a public website[7]. Two internships were recruited to

[7] http://headwork.gforge.inria.fr

develop a skill management system for our future crowdsourcing platform, following the thesis work of Panagiotis Mavridis[?], defended in November.

7.4 PROFILE

Participants: Tristan Allard, Zoltan Miklos.

Acronym	**Profile**
Call	**Labex CominLabs**
Year	2016
Title	**Analyzing and mitigating the risks of online profiling: building a global perspective at the intersection of law, computer science and sociology**
Coordinator	Benoît Baudry (DiverSE)
Funding	480 000 euros
Length	36 months

The practice of online profiling, which can be defined as the tracking and collection of user information on computer networks, has grown massively during the last decade, and is now affecting the vast majority of citizens. Despite its importance and impact, profiling remains largely unregulated, with no legal provisions determining its lawful use and limits under either the French or European law. This has encouraged market players to exploit a wide range of tracking technologies to collect user information, including personal data. Consequently, most online companies are now routinely violating the fundamental rights of their users, especially with respect to their privacy, with little or no oversight. The PROFILE project brings together experts from law, computer science and sociology to address the challenges raised by online profiling, following a multidisciplinary approach. More precisely, the project will pursue two complementary and mutually informed lines of research:

- Investigate, design, and introduce a new right of opposition into the legal framework of data protection to better regulate profiling and to modify the behavior of commercial companies towards being more respectful of the privacy of their users.

- Provide users with the technical means they need to detect stealthy profiling techniques as well as to control the extent of the digital traces they routinely produce. As a case study, we focus on browser fingerprinting, a new profiling technique for targeted advertisement. The project will develop a generic framework to reason on the data collected by profiling algorithms, to uncover their inner working, and make them more accountable to users.

PROFILE will also propose an innovative protection to mitigate browser fingerprinting, based on the collaborative reconfiguration of browsers. The legal model developed in PROFILE will be informed by our technological efforts (e.g., what is technologically possible or not), while our technological research will incorporate the legal and sociological insights produced by the project (e.g., what is socially and legally desirable / acceptable). The resulting research lies at the crossing of three fields of expertise (namely Law, Computer Science and Sociology), and we believe forms a proposal that is timely, ambitious, and immediately relevant to our modern societies.

7.5 ORACULAR

Participants: Tristan Allard, Tassadit Bouadi, David Gross-Amblard, Arnaud Martin, Zoltan Miklos.

Acronym	**ORACULAR**
Call	**Defis scientifiques emergents (2016) University Rennes 1**
Year	2016
Title	**ORganisation de l'interaction Avec des Cohortes d'UtiLisatEuRs**
Coordinator	Tassadit Bouadi
Funding	4 000 euros
Length	24 months

The idea of ORACULAR is to propose declarative approaches for: (1) the description and modeling of input data of a crowdsourcing platform (task building, user modeling: preferences, availability, cost, skills), (2) the definition of optimization methods to organize the acquisition of user cohort contributions, while providing at the same time a reasonable level of interaction, (3) the definition of quality measures to evaluate the relevance and effectiveness of the crowdsourcing data collection process.

7.6 ExPRESS

Participants: Tassadit Bouadi, Arnaud Martin, Yiru Zhang.

Acronym	**ExPRESS**
Call	**ARED/LTC (2016)**
Year	2016
Title	**Evaluation de la qualité des informations restituées par une analyse à base de PRéférences. Application aux rESeaux Sociaux**
Coordinator	Tassadit Bouadi and Arnaud Martin
Funding	90 000 euros
Length	36 months

The application context of this project concern social network analysis. The theoretical context is the preference queries applied to very large databases.

The concept of preference queries has been established in the database community and was intensively studied in the last decade. These queries have dual benefits. On the one hand, they allow to interpret accurately the information needs of a given user. On the other hand, they constitute an effective method to reduce very large datasets to a small set of highly interesting results and to overcome the empty result set. A query is personalized by applying related user preferences stored in the user's profile.

However, with the advent of social networks such as Facebook, Twitter, Instagram, or more locally Breizbook, the user is no longer considered as an individual entity, at least more only. In this context, the user designates an interconnected social entity and is the author of significant information flow. The objective of this project is the development of a collaborative system for personalizing analyzes (*i.e. preference queries*) based on profiles of social network users.

7.7 MetaTNT2

Participants: Tassadit Bouadi, Véronique Masson (Lacodam Team).

Acronym	**MetaTNT2**
Call	**AMI EAU (2016)**
Year	2016
Title	**Un modèle agro-hydrologique simplifié et interactif pour l'analyse de scénarios de réduction des flux d'azote dans les bassins versants**
Coordinator	Tassadit Bouadi (responsible of the scientific program of the IRISA partner)
Funding	7 000 euros
Length	36 months

Spatially distributed agro-hydrological models allow researchers and stakeholders to represent, understand and formulate hypotheses about the functioning of agro-environmental systems and to predict their evolution. These models have guided agricultural management by simulating effects of landscape structure, farming system changes and their spatial arrangement on stream water quality.

The objective of this project is to develop a meta-model based on simulations of the spatially distributed agro-hydrological model TNT2 (Topography-based Nitrogen Transfer and Transformations) in agricultural catchments, and to propose a conceptual guidance tool as a means of building and testing environmental management scenarios.

7.8 TOTAL

Participants: Na Li, Arnaud Martin.

Acronym	**TOTAL**
Call	**Total**
Year	2017
Title	**Fusion d'images sattelitaires**
Coordinator	Arnaud Martin
Funding	19 485 euros
Length	4 months

The main goal of this short project was to begin the state of art on the methods on classifiers fusion in order to characterize the forest from satellite images.

7.9 CIFRE TOTAL

Participants: Na Li, Arnaud Martin.

Acronym	**CIFRE TOTAL**
Call	**Total**
Year	2017
Title	**Tests et analyses de données de plateformes de crowdsourcing**
Coordinator	Arnaud Martin
Funding	67 252 euros
Length	36 months

The objective of this CIFRE project is to develop a method that allows to obtain the best model of cost-trip for a campaign in order to acquire geophysical data. A first study has been published in [?] in order to propose an automatic water detection approach based on Dempster-Shafer theory for multi spectral images.

7.10 CRE Orange

Participants: Jean-Christophe Dubois, Yolande Le Gall, Mouloud Kharoune, Arnaud Martin.

Acronym	**CRE**
Call	**Orange lab**
Year	2017
Title	**Tests et analyses de données de plateformes de crowdsourcing**
Coordinator	Arnaud Martin
Funding	6 000 euros
Length	6 months

The objective of this project is to realize a crowdsourcing campaign with uncertain and imprecise questions on FouleFactory platform. The obtained data could serve to continue the work on the characterization of persons in crowdsourcing platforms.

8 Other Grants and Activities

8.1 International Collaborations

- Regular collaboration with LSIR/EPFL (Switzerland) and Northwestern Polytechnical University (Xi'an, China).

- Collaboration with University of Sheffield, the group of Prof Gianluca Demartini. Panagiotis Mavridis has spent 3 months at this group in 2016. We continue the collaborations.
- Collaboration with the DSL lab of the University of California Santa Barbara (informal). The collaboration is ongoing.
- Collaboration with University of Québec in Montreal (PROFILE project). The collaboration is ongoing.
- Collaboration with the University of Shenzhen (informal). The collaboration is ongoing.

8.2 National Collaborations

- We have regular collaborations with the SAS INRA research group (Rennes) in the field of environmental decision making
- We have regular informal collaborations with the following teams: Vertigo/CEDRIC/Cnam-Paris, Hadas/LIG-Grenoble, DBWeb/Telecom Paristech-Paris, DAHU/ENS-Cahan, OAK-LRI/Orsay, ONERA, LABSTICC-Telecom Bretagne.
- We have an informal collaboration with the ASCOLA INRIA team (Nantes)
- We have a collaboration with the "Identité et Confiance" team of the IRT b<>com

9 Dissemination

9.1 Scientific Responsabilities

Phd defense in DRUID in 2017

- Dorra Attiaoui, co-directed by Arnaud Martin and Boutheina Ben Yaghlane defenced her Phd entitled "Belief Detection and Temporal Analysis of Experts in Question Answering: case study Stack Overflow" December, 1, 2017 behind the jury members: A. Hadjali, J. Velcin, F. Rousseau, B. Ben Yaghlane, A. Martin, [?].
- Panagiotis Mavridis, co-directed by David Gross-Amblard and Zoltan Miklos, defenced his Phd entitled "Using Hierarchical Skills for Optimized Task Selection in Crowdsourcing", November, 17, 2017 behind the jury members: S. Amer-Yahia, A. Bozzon, P. Cudré-Mauroux, G. Demartini, A.-M. Kermarrec.

Jury of Phd and HDR defense in 2017

- D. Gross-Amblard:
  - Maria ROSSI (Polytechnique, 2017) (reviewer)
  - Olfa SLAMA (ENSSAT, 2017) (president)
- A. Martin:

  - J-P. Attal (Université d'Evry, 2017) (reviewer)
  - J. Petit (Université de Reims, 2017) (reviewer and president)
  - J-C. Risch (Université de Remis, 2017) (reviewer and president)
  - A. Ben Othmane (Université de Nice, 2017) (reviewer and president)
  - S. Elmi (Université de Poitiers, 2017) (reviewer)
  - N. Helal (Univeristé d'Artois, 2017) (reviewer)
  - I. Hamamai (Tunis University, Tunisia, 2017) (reviewer)
  - Germain Forestier (Université de Lorraine, 2017) (HDR reviewer)
  - John Klein (Université de Lille, 2017) (HDR reviewer)

Lab scientific committees and evaluations in 2017

- A. Martin: Lamsade

Steering committees in 2017

- A. Martin: Extraction et Gestion de Connaissances (EGC) national conference.

Organizing committees in 2017

- A. Martin:
  - Organization of the registrations to BFAS summer school 2017[8] in Xi'an, China.
  - Co-organization of a special session on Big Data Information Fusion with the Theory of Belief Functions at the conference Fusion 2017[9] in Xi'an, China.

Program committees in 2017

- T. Bouadi: PC member: IFSA-SCIS'2017, CSA'2018, EGC'2018
- A. Martin: PC member of IJCNN 2017, IAE/AIE 2017, Fusion 2017, BESC 2017, IGARSS 2017, LFA 2017, EGC'2018
- Z. Miklos: PC member: WWW'2018, EGC'2018

Conferences Reviews in 2017

- A. Martin: IINTEC

[8] http://belief2017.nwpu.edu.cn/

[9] http://www.fusion2017.org/

Journals Reviews in 2017

- Arnaud Martin: Fuzzy Sets and Systems, International Journal of Approximate Reasoning, Pattern Recognition, Information Fusion, Computers & Industrial Engineering, IET Signal Processing, Transactions on Information Systems, Revue d'Intelligence Artificielle
- Zoltan Miklos: Transactions on Data and Knowledge Engineering (TKDE), Information Systems, Computer, Future Generation Computing Systems, International Journal of Web Information Systems
- Tassadit Bouadi : Computers and Electronics in Agriculture (COMPAG)
- Tristan Allard : The VLDB Journal (VLDBJ), Transactions on Big Data (TBDSI), Journal of Information Security and Applications (JISA), Transactions on Knowledge and Data Engineering (TKDE)

9.2 Involvement in the Scientific Community

- Arnaud Martin:
  - treasurer of BFAS society[10]
  - in charge of the challenge for EGC society[11]
  - webmaster for AFIA[12]
- David Gross-Amblard, Zoltan Miklos, Tassadit Bouadi, Tristan Allard
  - In charge of the website of the French research in databases community (`http://bdav.org`)

9.3 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.
- Database course (theory and practice) for ENS Rennes (one of the major French “grande ecole‘’).
- Database course at INSA Rennes (also a ”grande ecole“).
- Arnaud Martin is in charge of a M2 research module on data ming and data fusion at ENSSAT.

[10] `http://www.bfasociety.org`
[11] `http://www.egc.asso.fr`
[12] `http://afia.asso.fr/`

- Arnaud Martin is member of the Java Challenge jury[13], Paris, 2017
- Privacy-preserving data publishing course at ENSAI (Ecole Nationale de la Statistique et de l'Analyse de l'Information)
- David Gross-Amblard is co-head of the Research Master in Computer Science (SIF), Rennes 1 University[14]

10 SWOT

10.1 Strengths

- Good dynamism: young team
- Strong link with applications

10.2 Weakness

- Few publication between Lannion and Rennes parts of team

10.3 Opportunities

- Ongoing ANR projects
- Cybersecurity opportunities

10.4 Threats

- Huge teaching duties, causing difficult meeting schedule
- Seminars and doctoral formation at Rennes, not always in visio.

11 Bibliography

Major publications by the team in recent years

[1] T. Allard, T. Bouadi, J. Duguépéroux, V. Sans, "From Self-Data to Self-Preferences: Towards Preference Elicitation in Personal Information Management Systems", *in: International Workshop on Personal Analytics and Privacy (In conjunction with ECML PKDD 2017)* , Skopje, Macedonia, September 2017, https://hal.inria.fr/hal-01578990.

[2] D. Attiaoui, A. Martin, B. Ben Yaghlane, "Belief Measure of Expertise for Experts Detection in Question Answering Communities: case study Stack Overflow", *in: 21st International Conference on Knowledge-Based and Intelligent Information & Engineering Systems, Published by Elsevier B.V.*, Marseille, France, September 2017, https://hal.archives-ouvertes.fr/hal-01568061.

[13] https://editx.eu/it-challenge/java-code-challenge-france

[14] http://master.irisa.fr

[3] D. Attiaoui, A. Martin, B. Ben Yaghlane, "Belief Temporal Analysis of Expert Users: case study Stack Overflow", *in: 19th International Conference on Big Data Analytics and Knowledge Discovery - DaWaK 2017*, Lyon, France, August 2017, https://hal.archives-ouvertes.fr/hal-01576875.

[4] D. Attiaoui, *Belief Detection and temporal analysis of experts in question answering communities: case study Stack Overflow*, PhD Thesis, December 2017.

[5] S. Ben Dhaou, K. Zhou, M. Kharoune, A. Martin, B. Ben Yaghlane, "The Advantage of Evidential Attributes in Social Networks", *in: 20th International Conference on Information Fusion*, Xi'an, China, July 2017. 20th International Conference on Information Fusion, Jul 2017, Xi'an, China, https://hal.archives-ouvertes.fr/hal-01562965.

[6] L. Béziaud, T. Allard, D. Gross-Amblard, "Lightweight Privacy-Preserving Task Assignment in Skill-Aware Crowdsourcing", *in: 19th International Conference on Big Data Analytics and Knowledge Discovery - DaWaK 2017*, *Lecture Notes in Computer Science*, *10439*, p. 18 – 26, Lyon, France, August 2017, https://hal.inria.fr/hal-01580249.

[7] T. Bouadi, M.-O. Cordier, P. Moreau, R. Quiniou, J. Salmon-Monviola, C. Gascuel-Odoux, "A data warehouse to explore multidimensional simulated data from a spatially distributed agro-hydrological model to improve catchment nitrogen management", *Environmental Modelling and Software 97*, November 2017, p. 229 – 242, https://hal.inria.fr/hal-01597840.

[8] Y. Dauxais, D. Gross-Amblard, T. Guyet, A. Happe, "Extraction de chroniques discriminantes", *in: Extraction et Gestion des Connaissances (EGC)*, Grenoble, France, January 2017, https://hal.inria.fr/hal-01413473.

[9] Y. Dauxais, T. Guyet, D. Gross-Amblard, A. Happe, "Discriminant chronicles mining: Application to care pathways analytics", *in: Artificial Intelligence in Medicine, 16th Conference on Artificial Intelligence in Medicine*, Vienna, Austria, June 2017, https://hal.archives-ouvertes.fr/hal-01568929.

[10] S. Jendoubi, A. Martin, L. Liétard, H. Ben Hadji, B. Ben Yaghlane, "Two Evidential Data Based Models for Influence Maximization in Twitter", *Knowledge-Based Systems*, 2017, https://hal.archives-ouvertes.fr/hal-01435733.

[11] S. Jendoubi, A. Martin, "A reliability-based approach for influence maximization using the evidence theory", *in: 19th International Conference on Big Data Analytics and Knowledge Discovery - DaWaK 2017*, Lyon, France, August 2017, https://hal.archives-ouvertes.fr/hal-01551780.

[12] F. Karem, M. Dhibi, A. Martin, M. S. Bouhlel, "Credal Fusion of Classifications for Noisy and Uncertain Data", *International Journal of Electrical and Computer Engineering (IJECE) 7*, 2, 2017, p. 1071–1087, https://hal.archives-ouvertes.fr/hal-01546634.

[13] N. LI, A. Martin, R. Estival, "An automatic water detection approach based on Dempster-Shafer theory for multi spectral images", *in: 20th International Conference on Information Fusion*, XI'AN, China, July 2017, https://hal.archives-ouvertes.fr/hal-01573200.

[14] Z.-G. LIU, Q. Pan, J. Dezert, A. Martin, "Combination of classifiers with optimal weight based on evidential reasoning", *IEEE Transactions on Fuzzy Systems*, 2017, https://hal.archives-ouvertes.fr/hal-01588701.

[15] A. Martin, “Conflict management in information fusion with belief functions”, working paper or preprint, April 2017.

[16] H. Ouni, A. Martin, L. Gros, M. Kharoune, Z. Miklos, “Une mesure d’expertise pour le crowdsourcing”, *in: Extraction et Gestion des Connaissances (EGC)*, Grenoble, France, January 2017, `https://hal.archives-ouvertes.fr/hal-01432561`.

[17] Y. Zhang, T. Bouadi, A. Martin, “Preference fusion and Condorcet’s Paradox under uncertainty”, *in: International Conference on Information Fusion*, Xi’an, China, July 2017, `https://hal.archives-ouvertes.fr/hal-01573217`.

[18] K. Zhou, A. Martin, Q. Pan, “Evidence combination for a large number of sources”, *in: 2017 20th International Conference on Information Fusion (FUSION)*, Xi’an, China, July 2017, `https://hal.archives-ouvertes.fr/hal-01567484`.