UMR IRISA

# Activity Report 2018

## Team DRUID

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

D7 – Data and Knowledge Management

CentraleSupélec · cnrs · rennes · IMT Atlantique Bretagne-Pays de la Loire École Mines-Télécom · Inria · INSA RENNES · Université Bretagne Sud ubs · UNIVERSITÉ DE RENNES 1

# 1 Team

DRUID is an IRISA team, starting October 2014, supported by 9 active members from distinct IRISA sites, Rennes and Lannion. This year a new member joint the team: Mickaël Foursov. 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 - until September 2018
Fanny Banor - Rennes - since September 2018

**Université Rennes 1 personnel**

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

**Post-doc**

Julien Lolive, January 2018-September 2018

**Visitors**

Amr El Abbadi, Full professor, University of California Santa Barbara, July 2018 (two weeks)
Salma Ben Dhaou, IHEC, Tunisia, PhD student, March-June and September-November 2018

**PhD students**

Tompoariniaina Andriamilanto, IRT b<>com/Université Rennes 1, co-advised with Benoit Baudry (KTH Stockholm) since Sept. 2017
Yann Dauxais, MENRT/Rennes 1, co-advised with Thomas Guyet (LACODAM) and André Happe (CIC-Inserm Rennes), since October 2015, defended April 2018

Joris Degueperoux, ANR/CROWDGUARD, Université Rennes 1, since September 2017
Na Li, CIFRE Total, since May 2017
Ian Jeantet, ANR EPIQUE, since December 2017
Gauthier Lyan, CIFRE Keolis, co-advised with Jean-Marc Jezequel (Diverse), since November 2018
Rituraj Singh, ANR HEADWORK, co-advised with Loic Helouet (Sumo), since January 2018
Constance Thierry, CD22/ANR HEADWORK, since October 2018
Yiru Zhang, LTC/ARED, since November 2016

**Other students**

Francesco Bariatti M2, INSA - 2018
François Mentec (M2), with ALTEN - 2018
Constance Thierry, M2 MRI, Université Rennes 1 - February-August 2018
Romain Boitard M1 Miage, Université Rennes 1 - 2018
Diego Cárdenas Cabeza M1 Miage, Université Rennes 1 - 2018
Manon Derocles (M1), with Lacodam - 2018
Hamed Diakité M1 Miage, Université Rennes 1 - 2018
Nellya Zohoun M1 Miage, Université Rennes 1 - 2018
Emerick Morel L3, ISITC, Université Rennes 1 - 2018
Nicholas Peace BS, Northern Kentucky University, USA - May-June 2018

# 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

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

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 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).

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

- 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

[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.

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).

**[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,

[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.

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. 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 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

[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.

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?

[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.

- **[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 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],

[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.

[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.

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 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,


[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.

[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.

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, Joris Duguépéroux, David Gross-Amblard, Jean-Christophe Dubois, Mouloud Kharoune, Yolande Le Gall, Arnaud Martin, Zoltan Miklos, Virginie Sans, Rituraj Singh, Constance Thierry.

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, 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**: Salma Ben Dhaou, Tassadit Bouadi, David Gross-Amblard, Mickaël Foursov, Ian Jeantet, 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].

[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 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 HEADWORK

**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 development in the ANR HEADWORK project (see below).

## 6 New Results

### 6.1 Social Network Analysis

**Participants**: Salma Ben Dhaou, 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. This year we work on several aspect to characterize the social networks: community detection, noise suppression on attributes, independence measurement and preference fusion.

[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.

### Semi-supervised evidential label propagation algorithm

In [?] we propose an extension of [?] that obtain the best student paper of the conference Belief in 2016. With the increasing size of social networks in real world, community detection approaches should be fast. The Label Propagation Algorithm (LPA) is known to be one of the near-linear solutions and benefits of easy implementation, thus it forms a good basis for efficient community detection methods. We extend the original update rule and propagation criterion of LPA in the framework of belief functions. A new community detection approach, called Semi-supervised Evidential Label Propagation (SELP), is proposed as an enhanced version of the conventional LPA. One of the advantages of SELP is that it can take use of the available prior knowledge about the community labels of some individuals. This is very common in real practice. For instance, in the co-authorship network, some domain experts are very easy to be labeled as their research interests are well-known to everyone. In SELP, the nodes are divided into two parts. One contains the labeled nodes, and the other includes the unlabeled ones. The community labels are propagated from the labeled nodes to the unlabeled ones step by step according to the proposed evidential label propagation rule.

### Correcting noisy information in evidential social network

In previous works [?, ?], we proposed a model for evidential social network and we show the interest of evidential attributes in a social network. In [?], attributes on nodes and links are considered and a method is proposed in order to correct the noise on the attributes. We show that even with noise in the network, the proposed algorithm is able to classify the nodes in their initial clusters. In the case of a large noise, the algorithm guarantees the coherence of the information of any network even when it is a network whose nodes and links attributes have been strongly modified. Experimentations are made on real networks and generated networks with several levels of noise.

### Independence on Twitter

Based on the proposed method in [?] for influencers characterization, a new method is proposed in [?] in order to detect independent users in online social networks. In fact, independent users cannot generally be influenced, they are independent in their choices and decisions. Independent users may attract other users and make them adopt their point of view. A user is qualified as independent when his/her point of view does not depend on others ideas. Thus, the behavior of such a user is independent from other behaviors. Detecting independent users is interesting because a part of them can be influencers. Independent users that are not influencers can be directly targeted as they cannot be influenced. The proposed approach is based on three metrics reflecting users behaviors. We propose an useful approach for detecting influencers. Indeed, we consider the independence as a characteristic of influencers even if not all independent users are influencers. The proposed approach is experimented on real data crawled from Twitter.

### Preference fusion in social network

In social network analysis, preference is often applied as an attribute for individuals' representation. In some cases, uncertain and imprecise preferences may appear. Moreover, conflicting preferences can arise from multiple sources. From a model for imperfect preferences we proposed in [?], we study in [?] the clustering quality in case of perfect preferences as well as imperfect ones based on weak orders (orders that are complete, reflexive and transitive). The model for uncertain preferences is based on the theory of belief functions with an appropriate dissimilarity measure when performing the clustering steps. To evaluate the quality of clustering results, we used Adjusted Rand Index (ARI) and silhouette score on synthetic data as well as on Sushi preference data set collected from real world. The results show that our model has an equivalent quality with traditional preference representations for certain cases while it has better quality confronting imperfect cases.

## 6.2 Characterization of experts in crowdsourcing platforms

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

On crowdsourcing platform, the crowd, usually diversified, can include users without qualification and/or motivation for the tasks. In [?] a new method of user expertise characterization in the crowdsourcing platforms is introduced. The proposed method is based on the theory of belief functions in order to identify four profile types: spammer, unqualified, average and expert. To determine the profile of a worker we consider both his knowledge for the task and his behavior. In one hand, to model the knowledge of a worker, we first consider his confidence in his answer. This confidence allow us to measure degrees of accuracy and precision, which are used to estimate the qualification of the worker. This one can be qualified or not for the task. In another hand, the behavior of the worker is considered via the time he takes to answer questions. His answer can be instinctive or reflect. The combination of the knowledge and behavior information is made by the theory of belief functions.

## 6.3 Combination in the theory of belief functions

**Participants**: Na Li, 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 [?] (an extension of [?]) adapted for large number of sources. In [?] an approach in order to weight classifiers to combine is proposed. In [?], a state of art of conflict management in information fusion in proposed. In [?] we propose a combination of supervised and unsupervised classification based on a belief associaiton for land cover classification.

## 6.4 Performing range queries over encrypted personal data

**Participants**: Cetin Sahin (Univ. California Santa Barbara), Tristan Allard [contact

point], Reza Akbarinia (INRIA), Amr El Abbadi (Univ. California Santa Barbara), Esther Pacitti (LIRMM).

Performing non-aggregate range queries on cloud stored data, while achieving both privacy and efficiency is a challenging problem. This paper proposes constructing a differentially private index to an outsourced encrypted dataset. Efficiency is enabled by using a cleartext index structure to perform range queries. Security relies on both differential privacy (of the index) and semantic security (of the encrypted dataset). Our solution, PINED-RQ develops algorithms for building and updating the differentially private index. Compared to state-of-the-art secure index based range query processing approaches, PINED-RQ executes queries in the order of at least one magnitude faster. The security of PINED-RQ is proved and its efficiency is assessed by an extensive experimental validation.

We have presented our results during the IEEE ICDE'18 conference [**?**].

# 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 plat-

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

forms [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).

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], Tristan Allard, David Gross-Amblard, Ian Jeantet, Mickaël Foursov, Arnaud Martin, Virginie Sans.

| | |
|---|---|
| **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 |

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

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 2018 we continued our work of the reconstruction of the phylomemetic structure, using a vector space representation. We can now extract the phylomemetic structures un the basis of word embeddings. We have obtained such structures for various datasets, including a DBLP publication database, PubMed data as well as other datasets in the context of economy and ecology, a shared dataset of the project. However, for this reconstruction we use word embedding for each year that does not enable to analyse the evolution entirely in vector representation, that requires further research efforts. One of the major issues raised by our partners in social sciences is that they would like to understand the evolution of scientific disciplines and sub-disciplines in hierarchical way (for example to understand whether mathematics and biology are getting closer, but also whether for example population genetics, a specific domain in biology is getting closer to some specific sub-fields of mathematics). To address this issues, we first need to reconstruct a such a specific hierarchical representation of the scientific filed. While various hierarchical taxonomical representations exists, they are often created by experts and not extracted from the data (they might even be incompatible with a particular dataset) and the usual hierarchical clustering methods also fall short as they produce disjoint clusters and the scientific disciplines naturally share vocabularies. We have developed a more suitable way of constructing hierarchies. Parallel to this work we also explored the use of graph embeddings in this context. In this case we first construct a word co-occurrence graph and we use graph embedding methods (in particular, the method of Laplacian eigenmaps) to cluster the words. This is an ongoing work, where we have promising preliminary experimental results.

## 7.3 HEADWORK

**Participants**: Tristan Allard, Tassadit Bouadi, Jean-Christophe Dubois, David Gross-Amblard [contact point], Yolande Le Gall, Arnaud Martin, Zoltan Miklos, Virginie Sans, Constance Thierry.

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

| | |
|---|---|
| **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 |

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 2018 we have recruited Rituraj Singh on a the ANR PhD funding, co-advisored by the SUMO team. We started the development of the HEADWORK plateform. Two internships were positioned on software industrialization (git management, source documentation, continuous integration and deployment).

We have also extended our model for affecting crowd workers to tasks, with the help of a hierarchical skill model. Our extended model can handle more richer hierarchies, and not only skill taxonomies. This is a joint work with Panagiotis Mavridis who is now at University of Delft, the Netherlands.

## 7.4 PROFILE

**Participants**: Tristan Allard, Zoltan Miklos.

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

| | |
|---|---|
| **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 |

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.

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.

| | |
|---|---|
| **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 |

### 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).

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.

| | |
|---|---|
| **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 |

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 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. In [**?**] we propose a combination of supervised and unsupervised classification based on a belief association for land cover classification. The goal is to obtain a land map of defined classes as much precise and sure as possible. Then a model of cost-trip will be proposed.

## 7.9 Defi

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

The objective of this project is to propose a method in order to model contributors profile. The experiments was made on the data obtain by the CRE Orange project in 2017. This work

| | |
|---|---|
| **Acronym** | **AAP DS** |
| **Call** | **Défi émergent** |
| **Year** | 2018 |
| **Title** | **Tests et analyses de données de plateformes de crowdsourcing** |
| **Coordinator** | Yolande Le Gall |
| **Funding** | 7 000 euros |
| **Length** | 1 yeat |

was published in [**?**] where a new method of user expertise characterization in the crowdsourcing platforms is introduced.

## 7.10 PEPS IRDICS

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

| | |
|---|---|
| **Acronym** | **IRDICS** |
| **Call** | **PEPS S2IH INS2I 2018** |
| **Year** | 2018 |
| **Title** | **Interface de Recueil de Données Imparfaites pour le CrowdSourcing** |
| **Coordinator** | Jean-Christophe Dubois |
| **Funding** | 10 000 euros |
| **Length** | 1 year |

The objective of this project is to well design a campaign with uncertain and imprecise answer to some questions. The collaboration with LOKI team of CRISTAL lab in Lille, allowed to propose a first experiment with different size of stick. The question is "which one is the biggest" and the worker can give a degree of certainty on a 7-level scale. The answers allow to produce a curve for each worker given the point of discernibility. The experiment will be set up on the Figure eight platform.

## 7.11 PEPS IRDICS

**Participants**: Anne-Isabelle Graux (INRA/ Lacodam), Alexandre Termier (Lacodam), Tassadit Bouadi.

With the simultaneous increase in the questions addressed to cattle breeding, in system knowledge and computing power, simulation models tend to be more and more complex producing increasing volumes of heterogeneous data. These data are generally not fully analysed, due to their huge volume that makes their mining difficult, and to the fact that model users are often just interested in a small part of the data. Futhermore, simulation data are often lost after their partial valorisation although they could help answering other scientific questions.

| | |
|---|---|
| **Acronym** | **Bourse DigiTag** |
| **Call** | **Bourses #DigiTag 2018** |
| **Year** | 2018 |
| **Title** | **Development of a method allowing to mine and analyse huge volumes of simulation results from a crop model** |
| **Coordinator** | Anne-Isabelle Graux |
| **Funding** | 8 000 euros |
| **Length** | 2 years |

This suggests a need for a method allowing a storage of simulation data on the long term, as well as an easier mining and analysis of simulation data with the possibility for model users to identify multi-criteria trade-off solutions. The IRISA-INRIA LACODAM team is developping data mining methods allowing to identify interesting patterns supporting the recommandation of actions. A data warehouse to explore multidimensional simulated data from a spatially distributed agro-hydrological model was recently developed by this team to improve catchment nitrogen management [**?**]. The objective of the work is to adapt this method for the exploration and analysis of the simulated data from the STICS crop model that were produced in the framework of a French study called "Production, exportation d'azote et risques de lessivage".

# 8 Other Grants and Activities

## 8.1 International Collaborations

- Regular collaboration with Northwestern Polytechnical University (Xi'an, China).
- Collaboration with Panagiotis Mavridis (University of Delft, The Netherlands)
- Collaboration with the DSL lab of the University of California Santa Barbara (informal). The collaboration is ongoing.
- Collaboration with the KTH Royal Institute of Technologies in Stockholm. 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.
- Preparation of collaborations with Griffith's University (Australia)

## 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
- We develop some collaboration with Sébastien Destercke from Heudiasyc, (Université Technologique de Compiègne), on the model of fusion of preferences. Yiru Zhang has spent 2 times one week in Heudiasyc.
- From the PEPS project, we have some collaboration with LOKI team of Cristal, Lille.
- Collaboration with the STACK INRIA team (Nantes, informal). The collaboration is ongoing.
- Collaboration with the SHAMAN team (Lannion). The collaboration is ongoing.

# 9 Dissemination

## 9.1 Scientific Responsabilities

### Phd defense in DRUID in 2018

- Yann Dauxais, co-directed by David Gross-Amblard, Thomas Guyet and André Happe defenced her Phd entitled "Discriminant chronicle mining" April, 13, 2018 behind the jury members: S. Bringay, F. Masseglia, P. Papetrou, D. Gross-Amblard, T. Guyet and A. Happe [**?**].

### Jury of Phd and HDR defense in 2018

- D. Gross-Amblard:
  - –
- A. Martin:
  - Rihab Ben Ameur (Communauté université Grenoble, Annecy, 2018) (reviewer)
  - Mahdi Washha (Université de Toulouse, 2018) (reviewer)
  - Aurélien Moreau (Université de Rennes 1, 2018) (president)

### Lab scientific committees and evaluations in 2018

- D. Gross-Amblard: LIX
- A. Martin: LRI

### Evaluation of scientific project proposals

- Zoltan Miklos:
  - Independent expert for the European commission, call H2020-MSCA-RISE-2018 (Marie Sklodowska Curie Action)
  - Independent expert for the European commission, call H2020-WIDESPREAD-2018-3
- Tassadit Bouadi:
  - Evaluator for the ANR (Agence Nationale de la Recherche), call ANR'2018-PRC funding instrument

### Steering committees in 2018

- A. Martin: Belief 2018, Extraction et Gestion de Connaissances (EGC) national conference.

### Program committees in 2018

- T. Bouadi: PC member: ECML/PKDD'2018, EGC'2019
- A. Martin: PC member of Belief'2018, Fusion'2018, IGARSS'2018, IJCNN'2018, LFA 2018, EGC'2018
- Z. Miklos: PC member: WWW'2018, EGC'2018, 2019

### Conferences Reviews in 2018

- T. Bouadi: ICDM'2018

### Journals Reviews in 2018

- Arnaud Martin: Computer, Decision Support Systems, Fuzzy Sets and Systems, IEEE Access, IEEE Transaction on Fuzzy Systems, IEEE Transactions on Systems, Man, and Cybernetics: Systems,IEEE Transactions on Network Science and Engineering, International Journal of Approximate Reasoning, Information Sciences, International Journal of Machine Learning and Cybernetics, Knowledge-Based Systems, Pattern Recognition
- Zoltan Miklos: Transactions on Data and Knowledge Engineering (TKDE), Computer, Applied Computing and Informatics, Future Generation Computing Systems,
- Tassadit Bouadi : Computers and Electronics in Agriculture (COMPAG)
- Tristan Allard : The VLDB Journal (VLDBJ)

## 9.2 Involvement in the Scientific Community

- Arnaud Martin:
  - member of BFAS society[7]
  - in charge of the challenge for EGC society[8]
  - webmaster for AFIA[9]
- David Gross-Amblard, Zoltan Miklos, Tassadit Bouadi, Tristan Allard
  - In charge of the website of the French research in databases community (`http://bdav.org`)
- Zoltan Miklos:
  - member of the ACM, AFIA[10]

## 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'').
- Zoltan Miklos is in charge of a M2 research module Data and knowledge management (advanced course) ISTIC
- Arnaud Martin is in charge of a M2 research module on data mining and data fusion at ENSSAT.
- Arnaud Martin was in charge of a course on algorithms at Al Hussein Technical University, Jordan
- Privacy-preserving data publishing course at ENSAI (Ecole Nationale de la Statistique et de l'Analyse de l'Information)
- Zoltan Miklos is in charge of a the modules Artificial Intelligence, and Data management for BigData at ESIR
- Zoltan Miklos is the responsable of the Information Systems study path, at ESIR
- David Gross-Amblard is co-head of the Research Master in Computer Science (SIF), Rennes 1 University[11]

[7] `http://www.bfasociety.org`
[8] `http://www.egc.asso.fr`
[9] `http://afia.asso.fr/`
[10] `http://afia.asso.fr/`
[11] `http://master.irisa.fr`

# 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
- Few publication for young PhD students

## 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] S. Ben Dhaou, M. Kharoune, A. Martin, B. Ben Yaghlane, "An Evidential Method for Correcting Noisy Information in Social Network", *Online Social Networks and Media Volume 7*, September 2018, p. 30–44.

[2] M. Chehibi, M. Chebbah, A. Martin, "Independence of Sources in Social Networks", *in : Information Processing and Management of Uncertainty in Knowledge-Based Systems. Theory and Foundations - 17th International Conference, IPMU*, Cadiz, Spain, June 2018.

[3] Y. Dauxais, *Extraction de chroniques discriminantes*, PhD Thesis, April 2018.

[4] S. Destercke, T. Denoeux, F. Cuzzolin, A. Martin (editors), *Belief Functions: Theory and Applications - 5th International Conference, BELIEF 2018, Compiègne, France, September 17-21, 2018, Proceedings*, *Lecture Notes in Computer Science*, *11069*, France, Springer, 2018.

[5] S. Jendoubi, M. Chebbah, A. Martin, "Evidential Independence Maximization on Twitter Network", *in : 5th International Conference, Belief 2018*, *Belief Functions: Theory and Applications*, Compiègne, France, September 2018.

[6] N. LI, A. Martin, R. Estival, "Combination of supervised learning and unsupervised learning based on object association for land cover classification", *in : DICTA2018*, Canberra, Australia, December 2018, `https://hal.archives-ouvertes.fr/hal-01922096`.

[7] Z.-G. LIU, Q. Pan, J. Dezert, A. Martin, "Combination of classifiers with optimal weight based on evidential reasoning", *IEEE Transactions on Fuzzy Systems 26*, 3, June 2018, p. 1217_1230.

[8] C. Sahin, T. Allard, R. Akbarinia, A. Abbadi, E. Pacitti, "A Differentially Private Index for Range Query Processing in Clouds", *in : ICDE: International Conference on Data Engineering*, Paris, France, April 2018.

[9] C. Thierry, J.-C. Dubois, Y. Le Gall, A. Martin, "Contributors profile modelization in crowdsourcing platforms", *in : 27èmes rencontres francophones sur la logique floue et ses applications*, Arras, France, November 2018, `https://hal.archives-ouvertes.fr/hal-01920669`.

[10] Y. Zhang, T. Bouadi, A. Martin, "A clustering model for uncertain preferences based on belief functions", *in : DaWaK: Data Warehousing and Knowledge Discovery*, Regensburg, Germany, September 2018.

[11] Y. Zhang, T. Bouadi, A. Martin, "An empirical study to determine the optimal k in EkNNclus method", *in : 5th International Conference on Beleif Functions (BELIEF2018)*, BFAS, Compiègne, France, September 2018.

[12] K. Zhou, A. Martin, Q. Pan, Z. Liu, "SELP: Semi-supervised evidential label propagation algorithm for graph data clustering", *International Journal of Approximate Reasoning 92*, 2018, p. 139–154.

[13] K. Zhou, A. Martin, Q. Pan, "A belief combination rule for a large number of sources", *Journal of Advances in Information Fusion 13*, 2, December 2018.

[14] K. Zhou, Q. Pan, A. Martin, "Evidential community detection based on density peaks", *in : BELIEF 2018 - The 5th International Conference on Belief Functions*, Compiègne, France, September 2018, `https://hal.archives-ouvertes.fr/hal-01882803`.