Project-Team DRUID

Declarative & Reliable management of Uncertain, user-generated & Interlinked Data

Lannion - Rennes
Activity Report
2016
1 Team

DRUID is a newly created team at IRISA, starting October 2014. DRUID is supported by 5 active members from distinct IRISA sites, Rennes and Lannion. The team is completed by 4 associated researchers and several Ph.D students. 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/Rennes
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
Salma Ben Dhaou, IHEC, Tunisia, PhD student, March-June and September-December 2016
Amr El Abadi, Full professor, University of California Santa Barbara, October 2016

PhD students
Dorra Attiaoui, Tunisian grant, ATER IUT Lannion (2016-2017), since June 2013
Siwar Jendoubi, Tunisian MOBIDOCS grant, since June 2013 - defended December 2016
Kuang Zhou, Chinese grant, since October 2013 - defended July 2016
Panagiotis Mavridis, MENRT/Rennes 1 grant, since October 2014
Yann Dauxais, MENRT/Rennes 1 grant, since October 2015, co-advised with Thomas Guyet (LACODAM) and André Happe (CIC-Inserm Rennes)
Yiru Zhang, LTC/ARED, since November 2016

Master students
Hosna Ouni, M2, Polytechnic, Tunisia, March-June 2016
Shashank Sharma, M2 Big Data, ENSAI, France, February-June 2016
Louis Béziaud, L3 RI, Université Rennes 1 and ENS Rennes, France, May-August 2016
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. 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 company’s 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.

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.

---

1. [https://www.mturk.com](https://www.mturk.com)
2. See for example participative journalism platforms, [https://witness.theguardian.com/moreabout](https://witness.theguardian.com/moreabout)
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. Indeed, Facebook is using a NoSQL, graph-oriented database for its core data, Neo4j, [http://www.zdnet.com/facebook-neo4j-7000009866/](http://www.zdnet.com/facebook-neo4j-7000009866/).
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).

**[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

---

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](https://requester.mturk.com/tour).

---

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\[^{FFK+11,PPP+13}\]. 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. 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 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 literature 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 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\textsuperscript{RLT+13}, and game theory for the study of strategic aspects in crowdsourcing, data pricing\textsuperscript{LLMS13} and data publication\textsuperscript{JP13}. For the modeling of complex tasks in crowdsourcing, we envision to extend declarative approaches for business processes\textsuperscript{DM12}, such as the collaborative business artifact model\textsuperscript{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\textsuperscript{Sme97}.


In particular, the theory of belief functions\cite{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\cite{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)\cite{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, 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


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 Twitter 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 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

http://craft.irisa.fr
to affect the tasks. We have presented the software at the BDA’2014 conference [CGAG+14].

6 New Results

6.1 Social Network Analysis

Participants: Siwar Jendoubi, Arnaud Martin, Kuang Zhou.

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.

Influencers characterization in a social network for viral marketing

The Siwar Jendoubi’s thesis [1] is focus on influencers characterization. The Viral Marketing is a relatively new form of marketing that exploits social networks in order to promote a product, a brand, etc. It is based on the influence that exerts one user on another. The influence maximization is the scientific problem for the Viral Marketing. In fact, its main purpose is to select a set of influential users that could adopt the product and trigger a large cascade of influence and adoptions through the network. In this thesis, we propose two evidential influence maximization models for social networks. The proposed approach uses the theory of belief functions to estimate users influence. Furthermore, we introduce an influence measure that fuses many influence aspects, like the importance of the user in the network and the popularity of his messages. Next, we propose three Viral Marketing scenarios. For each scenario we introduce two influence measures. The first scenario is about influencers having a positive opinion about the product. The second scenario searches for influencers having a positive opinion and influence positive opinion users and the last scenario looks for influencers having a positive opinion and influence negative opinion users. In 2016, this part of the work has been published in [9]. On the other hand, we turned to another important problem which is about the prediction of the social message topic. Indeed, the topic is also an important parameter in the influence maximization problem. For this purpose, we introduce four classification algorithms that do not need the content of the message to classify it, they just need its propagation traces. In our experiments, we compare the proposed solutions to existing ones and we show the performance of the proposed influence maximization solutions and the proposed classifiers. The main result of the thesis as been accepted for a publication in 2017, in [3].

Belief relational clustering and its application to community detection

The Kuang Zhou’s thesis [2] is focus on the matter of community detection by the way of clustering algorithms taking into account the uncertainty.

Clustering, also called unsupervised learning, is an important technique in the field of data mining. According to the type of data sets, clustering algorithms can be divided into two kinds. One is for object data in the distance space, where the objects to be clustered are described by feature vectors. The other is for proximity data, where only the relationship values such as similarities or dissimilarities between objects are known. The latter is a more general case, as the relationship could also be got for the data represented by feature vectors. On the contrary, many real-world data sets can only be represented by relational data for which object-based clustering algorithms could not be applied directly.

Communities are groups of nodes (vertices) which probably share common properties and/or play similar roles within the graph. They can extract specific structures from complex networks, and consequently community detection has attracted considerable attention crossing many areas where systems are often represented as graphs. Community detection is in fact a clustering problem on graphs, and the available information in this problem is often in the form of similarities or dissimilarities (between nodes).

We consider in this work to represent graphs as relational data, and propose models for the corresponding relational data clustering. Four approaches are brought forward to handle the community detection problem under different scenarios.

We start with a basic situation where nodes in the graph are clustered based on the dissimilarities and propose a new c-partition clustering approach named Median Evidential C-Means (MECM) algorithm. This approach extends the median clustering methods in the framework of belief function theory. Moreover, a community detection scheme based on MECM is presented. The proposed approach could provide credal partitions for data sets with only known dissimilarities. The dissimilarity measure could be neither symmetric nor fulfilling any metric requirements. It is only required to be of intuitive meaning. Thus it expands application scope of credal partitions. In addition, some practical issues about how to apply the method into community detection problems such as how to determine the initial prototypes and the optimum community number in the sense of credal partitions are discussed. This makes the approach appropriate for graph partitions and enables us to gain a better understanding of the analyzed networks, especially for the uncertain and imprecise structures.

In MECM, one single representative object in the original data set is used to describe each of the individual classes. However, in some cases the way of using only one node to describe a community may not be sufficient enough. In order to capture various aspects of the community structures, more members rather than one should be referred as the prototypes of an individual group. Motivated by this idea, a Similarity-based Multiple Prototype (SMP) community detection approach is proposed. The centrality values are used as the criterion to select multiple prototypes to characterize each community. The prototype weights are derived to describe the degree of representativeness of objects for their own communities. Then the similarity between each node and community is defined, and the nodes are partitioned into divided communities according to these similarities. Crisp and fuzzy partitions could be obtained by the application of SMP.
Following, we extend SMP in the framework of belief functions to get credal partitions so that we can gain a better understanding of the data structure. The prototype weights are incorporated into the objective function of evidential clustering. The mass membership and the prototype weights could be updated alternatively during the optimization process. In this case, each cluster could be described using multiple weighted prototypes. As we will show, the prototype weights could also provide us some useful information for structure analysis of the data sets. This work has been published in [1].

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. This work has been published in 2016, in [12] and [13]. The last publication obtain the best student paper of the conference.

The performance of the proposed approaches is evaluated using benchmark graph data sets and generated graphs. Our experimental results illustrate the effectiveness of the proposed clustering algorithms and community detection approaches.

6.2 Characterization of experts in crowdsourcing platforms

Participants: Amal Ben Rjab, Mouloud Kharoune, Arnaud Martin, Zoltan Miklos, Hosna Ouni.

As an extension of the paper [BRKM+15], we have proposed in [5] to add a precision degree to the exactitude degree in order to better characterize the experts on a crowdsourcing platform without gold data. The proposed method is based on the theory of belief functions that allows to combine the basic belief assignment coming form both degrees.

Based on the model of responses given by the mass functions $m_{U_j}$ (see previous work [BRKM+15]), we can define a degree of precision. Here we allow the participants to give partial answers, that is crucial for calculating the precision degree. The usual model of responses (that is, the worker must give a complete answer), we could not define a such degree.

We note $\delta_{U_j}^{\Omega_k}$ the specificity degree of the mass function $m_{U_j}^{\Omega_k}$. It is defined by [SMO11] as follows:

$$\delta_{U_j}^{\Omega_k} = 1 - \sum_{X \in 2^{\Omega_k}} m_{U_j}^{\Omega_k}(X) \frac{\log_2(|X|)}{\log_2(|\Omega_k|)}$$  \hspace{1cm} (1)$$

This specificity degree allows to translate the precision level of each response independently of the other participant’s responses. To measure the degree of precision of each participant $IP_{U_j}$, we propose to calculate the average of the specificity degrees for all the $k^{th}$ questions. Such as:

$$IP_{U_j} = \frac{1}{r(U_j)} \sum_{k=1}^{K} \delta_{U_j}^{\Omega_k}$$  \hspace{1cm} (2)$$

We determine the experts by using $k$-means (with $k = 2$). We do not need the assumption on the majority of participant’s answers.

The results show the interest of the model. In another work [11], 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 automatable 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.

### 6.3 Using Hierarchical Skills for Optimized Task Assignment in Knowledge-intensive Crowdsourcing

**Participants:** Panagiotis Mavridis, David Gross-Amblard, Zoltan Miklos.

Besides the simple human intelligence tasks such as image labeling, crowdsourcing platforms propose more and more tasks that require very specific skills, especially in participative science projects. In this context, there is a need to reason about the required skills for a task and the set of available skills in the crowd, in order to increase the resulting quality. Most of the existing solutions rely on unstructured tags to model skills (vector of skills). In this paper

we propose to finely model tasks and participants using a skill tree, that is a taxonomy of skills equipped with a similarity distance within skills. This model of skills enables to map participants to tasks in a way that exploits the natural hierarchy among the skills. We illustrate the effectiveness of our model and algorithms through extensive experimentation with synthetic and real data sets. In a subsequent work we are about to extend this model.

6.4 Task Assignment in Crowdsourcing under Differential Privacy

Participants: Louis Béziaud (ENS Rennes), Tristan Allard [contact point], David Gross-Amblard.

Crowdsourcing is a technique that engages individuals in the act of completing outsourced tasks. Crowdsourcing platforms propose more and more tasks requiring specific skills. However, current task assignment solutions requires that the participant discloses its skills to untrustworthy entities. In this paper, we introduce a framework to compute task assignments in a privacy-preserving way. We investigate multiple strategies based on the randomized response perturbation scheme to provide differential privacy.

6.5 Lightweight Privacy-Preserving Averaging for the Internet of Things

Participants: Julien Lepiller (Supélec Rennes), Davide Frey (Inria) [contact point], Tristan Allard, Georges Giakkoupis (Inria).

The number of connected devices is growing continuously, and so is their presence into our everyday lives. From GPS-enabled fitness trackers, to smart fridges that tell us what we need to buy at the grocery store, connected devices - things - have the potential to collect and make available significant amounts of information. On the one hand, this information may provide useful services to users, and constitute a statistical gold mine. On the other, its availability poses serious privacy threats for users. In this paper we propose a novel protocol that makes it possible to aggregate personal information collected by smart devices in the form of an average, while preventing attackers from learning the details of the non-aggregated data. This work has been presented at M4IOT 2016 [5].

6.6 A new privacy-preserving solution for clustering massively distributed personal times-series

Participants: Tristan Allard [contact point], Georges Hébrail (EDF lab), Florent Masseglia (Inria), Esther Pacitti (Lirmm), Matthieu Simonin (Inria SED).

New personal data fields are currently emerging due to the proliferation of on-body/at-home sensors connected to personal devices. However, strong privacy concerns prevent individuals to benefit from large-scale analytics that could be performed on this fine-grain highly sensitive wealth of data. We propose a demonstration of Chiaroscuro, a complete solution for clustering massively-distributed sensitive personal data while guaranteeing their privacy. The demonstration scenario highlights the affordability of the privacy vs. quality and privacy vs. performance tradeoffs by dissecting the inner working of Chiaroscuro - launched over energy.
consumption times-series -, by exposing the results obtained by the individuals participating in the clustering process, and by illustrating possible uses.

This demonstration has been presented at IEEE ICDE 2016 [6].

6.7 A Differentially Private Index for Range Query Processing in Clouds

Participants: Tristan Allard [contact point], Cetin Sahin (UCSB), Reza Akbarinia (Inria), Amr El Abbadi (UCSB).

Despite the benefits of cloud services, legitimate privacy concerns continue to hinder their adoption for outsourcing sensitive data. Although performing range queries is a fundamental operation, achieving both privacy and efficiency when the cloud is not fully trusted remains a challenging problem today. This work is the first to construct a differentially private index to an outsourced encrypted dataset. Efficiency is enabled by the fact that the cloud uses a cleartext index structure to perform range query processing. 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 while minimizing privacy budget consumption. Query processing mechanisms are developed that reduce the inherent false positives/false negatives due to differential privacy. All this while maintaining the index structure without exhausting the privacy budget and allocating it smartly. The security of PINED-RQ is proved and its efficiency is assessed by an experimental validation using both real and synthetic datasets.

7 Contracts and Grants with Industry

7.1 CROWDGUARD

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

<table>
<thead>
<tr>
<th>Acronym</th>
<th>CROWDGUARD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Call</td>
<td>ANR JCJC</td>
</tr>
<tr>
<td>Year</td>
<td>2016</td>
</tr>
<tr>
<td>Title</td>
<td>GUARanteed confidentiality and efficiency in CROWDsource platforms</td>
</tr>
<tr>
<td>Coordinator</td>
<td>Tristan Allard</td>
</tr>
<tr>
<td>Funding</td>
<td>144 720 euros</td>
</tr>
<tr>
<td>Length</td>
<td>42 months</td>
</tr>
</tbody>
</table>

Crowdsourcing platforms offer the unprecedented opportunity to connect easily on-demand task providers, or taskers, and on-demand task solvers, or workers, locally or world-wide, for paid or 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 21st 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 \[6\]) 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 \[7\]). 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], David Gross-Amblard, Tristan Allard, Virginie Sans.

\[6\]https://tinyurl.com/wp-priv  
\[7\]https://tinyurl.com/waj-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.

7.3 HEADWORK

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

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 themselves. 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.

7.4 PROFILE

Participants: Tristan Allard, Zoltan Miklos.

<table>
<thead>
<tr>
<th>Acronym</th>
<th>PROFILE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Call</td>
<td>Labex CominLabs</td>
</tr>
<tr>
<td>Year</td>
<td>2016</td>
</tr>
<tr>
<td>Title</td>
<td>Analyzing and mitigating the risks of online profiling: building a global perspective at the intersection of law, computer science and sociology</td>
</tr>
<tr>
<td>Coordinator</td>
<td>Benoît Baudry (DiverSE)</td>
</tr>
<tr>
<td>Funding</td>
<td>480 000 euros</td>
</tr>
<tr>
<td>Length</td>
<td>36 months</td>
</tr>
</tbody>
</table>

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 P4CROWD

Participants: Tristan Allard, Tassadit Bouadi.

<table>
<thead>
<tr>
<th>Acronym</th>
<th>P4-CROWD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Call</td>
<td>PEPS JCJC INS2I</td>
</tr>
<tr>
<td>Year</td>
<td>2016</td>
</tr>
<tr>
<td>Title</td>
<td>Privacy-Preserving Personal Preference Acquisition and Learning in the Crowd</td>
</tr>
<tr>
<td>Coordinator</td>
<td>Tristan Allard</td>
</tr>
<tr>
<td>Funding</td>
<td>12 000 euros</td>
</tr>
<tr>
<td>Length</td>
<td>12 months</td>
</tr>
</tbody>
</table>

The goal of the P4-Crowd project is to initiate the definition, design, and implementation of a crowdsourcing platform that allows the representation of personal preferences and their acquisition by querying a cohort of individuals (also called the crowd below) and by learning multi-user preferences, while offering robust privacy guarantees. Made of three sub-objectives, it promotes an original approach of co-design of preference representation models, crowd querying algorithms, and multi-user learning methods on the one hand, together with privacy-preserving mechanisms that mix encryption and sanitization. The expected benefits of P4-Crowd are both scientific and technical (co-design preferences / crowd / privacy), industrial (trust economy), and societal (fundamental right to privacy). P4-Crowd’s partners are two young research faculty members of a newly formed Irisa team.

7.6 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.
7.7 ExPRESS

**Participants**: Tassadit Bouadi, Arnaud Martin, Yiru Zhang.

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

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.

7.8 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.
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.

8 Other Grants and Activities

8.1 International Collaborations

- Regular collaboration with LSIR/EPFL (Switzerland), LARODEC (Tunisia) 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.

8.2 National Collaborations

- We are part of the POSEIDON project on outsourced data security funded by the LABEX COMINLABS[^5].
- We are currently involved in the CNRS Big Data (MASTODONS) ARESOS[^9] project on semantic social networks analysis, particularly with the database team at LIP6.
- 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.

[^5]: http://www.poseidon.cominlabs.ueb.eu/
9 Dissemination

9.1 Scientific Responsibilities

Awards in 2016

- Panagiotis Mavridis, PhD student in DRUID, co-supervised by David Gross-Amblard and Zoltan Miklos, received the “Best student paper award” at WWW’2016 conference for his paper “Using Hierarchical Skills for Optimized Task Assignment in Knowledge-intensive Crowdsourcing”.

- Kuang Zhou, PhD student in DRUID team, co-supervised by Arnaud Martin and Quan Pan, received the “Best student paper award” at Belief 2016 conference for his paper “Semi-supervised Evidential Label Propagation Algorithm for Graph Data”.

Phd defense in DRUID in 2016


Jury of Phd defense in 2016

- D. Gross-Amblard:
  - Dai Hai Ton That, UVSQ, 2016 (president).

- A. Martin:
  - Z. Meng (Universite Nice Sophia Antipolis, 2016) (member)

Lab scientific committees and evaluations in 2016

- A. Martin: LGI2A, I3S, Heudiasyc

Steering committees in 2016

- A. Martin: international conference on belief functions (Belief), Extraction et Gestion de Connaissances (EGC) national conference.
• T. Bouadi and A. Martin: co-direction with Fabien Gandon (INRIA Sophia Antipolis) of the special issue "Analyse Intelligente des Réseaux Sociaux" of the journal Revue d'Intelligence Artificielle

Organizing committees in 2016
• D. Gross-Amblard:
  – Organization of the Summer School MDD 2016\(^{10}\) in Urrugne, France.
• A. Martin:
  – Organization of the registrations to Belief 2016\(^{11}\) Prague, CZ.
  – Co-founder and organizer with Engelbert Mephu Nguifo (LIMOS, ISIMA) of the workshop “Données participatives et sociales” at EGC 2016, Reims, France.

Program committees in 2016
• Z. Miklos: PC member: WWW’2017
• T. Bouadi: PC member: IFSA-SCIS’2017
• T. Allard : PC member : WISSE’2016

Conferences Reviews in 2016
• Tassadit Bouadi: IJCAI
• T. Allard : BDA, PKC

Journals Reviews in 2016
• Zoltan Miklos: Future Generation Computing Systems, Transactions on Data and Knowledge Engineering (TKDE), Information Systems Frontiers

\(^{10}\)http://cedric.cnam.fr/summer-school/MDD2016/
\(^{11}\)http://belief.utia.cz/
9.2 Involvement in the Scientific Community

- Arnaud Martin: treasurer of BFAS society\textsuperscript{12}
- Arnaud Martin: in charge of economical and business relations for EGC society\textsuperscript{13}

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 mining and data fusion at ENSSAT.
- Privacy-preserving data publishing course at ENSAI (Ecole Nationale de la Statistique et de l’Analyse de l’Information)

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

- 3 new ANR projects

10.4 Threats

- Huge teaching duties, causing difficult meeting schedule

\textsuperscript{12}\url{http://www.bfasociety.org}
\textsuperscript{13}\url{http://www.egc.asso.fr}
11 Bibliography

Major publications by the team in recent years


Doctoral dissertations and “Habilitation” theses


Articles in referred journals and book chapters


Publications in Conferences and Workshops


