UMR IRISA

Activity report 2013

Department 7
DATA AND KNOWLEDGE MANAGEMENT
Singletons

David GROSS-AMBLARD
Israel César LERMAN
Zoltán MIKLÓS

Rennes

David Gross-Amblard
Professor, Université de Rennes 1

1 Overall Objectives

My recent work is focused on database security, social network analysis and crowdsourcing. My main result in 2013 is the proposal of a research team at IRISA, DRUID, that reaches for now the second validation step. My present research project aims at integrating open (as in OpenData), participative (as in Wikipedia), externalized (as in Cloud) and socialized aspects (as in recommandation systems) into classical data management systems. These new viewpoints lead to a deep modification of query optimization, data distribution and data security.

2 Scientific Foundations

Database Watermarking Watermarking techniques allow for invisible and robust information hiding in a digital document, for example the document owner's identity. Many watermarking methods exist for multimedia documents like images, sound files and video. Recently, database watermarking techniques have emerged [10, 3, 21].

I started a database watermarking working group at the Vertigo team, CEDRIC Lab, CNAM Paris. We have proposed a database watermarking model where data hiding must preserve the quality (the result) of a user-defined set of important queries. In this setting, two questions arise : (i) knowing the hiding *capacity* of a given database, that is the largest size of a hidden message, (ii) computing watermarked databases efficiently, that respects the intended result of queries.

From the theoretical point of view, I focused on the relationship between the syntactical form of the query to preserve, and the watermarking capacity. We have shown that, without hypothesis, this watermarking capacity can be null. On the contrary, if the data set fulfills reasonable assumptions, *the watermarking capacity is guaranteed, for any `SQL` (for relational databases) or `XPath` (for XML) queries.* Moreover, corresponding watermarks can be obtained efficiently. These results are published in ACM Principles of Database Systems (2003) [8]. A practical counterpart of this work has been proposed to obtain a full database watermarking prototype, Watermill [6, 9, 16, 7].

This activity has been followed in three directions :

- *Geographical databases watermarking.* This work has been done with GREYC, LAMSADE, and COGIT Labs (French National Cartography Institute) [17, 11, 15, 12, 13, 14, 16] ;
- *Medical images* watermarking under constraints [5] ;
- *Symbolic musical databases* watermarking [2].
- *Formal proof* (à la Coq) of a stylized database watermarking system [4].

Data Provenance and Trust : Classical Web and Web of Objects Faced to the Web, Database techniques have included semi-structured data, navigational query languages, massively distributed query evaluation strategies, to cite a few aspects. Moreover, the Web allows any user to become a data provider, using forums, blogs, tweets, social networks, OpenData architecture or collaborative platforms. Sophisticated on-line content can then be realized by combining data from various distant sources and services calls. In these scenarios, users may require protection methods for the intellectual property of their personal productions, and trust / provenance indicators for the data they query. I would like to consider the following questions :

- How to integrate tools for intellectual property protection in a flow of Web documents, naturally dedicated to exchange, transformation and combination with other documents during their lifecycle.
- How to integrate provenance and trust of data first in a controlled distributed context, then on the generic Web, social networks or sensor networks. The study of trust in a distributed context already produced a prominent literature. Nevertheless, recent works view distributed data as a problem of knowl-

edge management on a large scale. The corresponding tools are then distributed deductive databases (Bloom, WebDamLog), using data production rules. To determine the trust of data produced by such rules, or the trust of the rules themselves, is a new issue.

This approach is now followed in a larger view for the design of security policy, as part of the CominLabs POSEIDON [1] project.

Strategic aspects in participative environment My work on database watermarking naturally leads to the question of the *value of data* : does my information has an (economical, scientific, ...) value for potential users ? What is the best way / time to publish information ? Several recent works focus on these questions, trying to model common behaviors associated with data advertisement systems like Google Smart Pricing and Yahoo Quality Based Pricing [18].

Those questions are hard to apprehend, because the value of a data is no longer a locally defined property, but a property that emerges from user interactions. These users seek to maximize the value of their data according to their own objectives and knowledge of the overall system. From a methodological point of view, these questions are well modeled by game theory. This theory, initially proposed by Von Neumann [22] and popularized by Nash's result [19], allows for the modeling of the behavior of autonomous actors. Its computational counterpart is now very popular, where actors are seen as machines with limited resources [20]. Applications range from crowdsourcing applications to open data publication.

3 New Results

Formal Proofs for Database Watermarking One of the long term goals of the watermarking community is to obtain complete security proofs of watermarking protocols, in a similar spirit as cryptographical protocol proofs. It is sometimes noted that existing proofs for watermarking are limited to specific classes of attacks and simply lead to an "arm race". A better situation is to obtain a proof with the following property : any victorious attacker must have solved an NP-complete problem efficiently, or must have violated a commonly accepted cryptographical hardness hypothesis.

We obtained with David Baelde, Pierre Coutieu, Julien Lafaye, Philippe Audebaud et Xavier Urbain a restricted proof of the Agrawal and Kiernan database watermarking protocol. The result is an ITP publication [4].

Ontology Watermarking Another result is the proposition of a new watermarking algorithm for populated ontologies, that is ontologies with instances of concepts. Those ontologies are currently very successful for the semantic Web, as shown by the huge YAGO and DbPedia ontologies. This work with Fabian Suchanek and Serge Abiteboul, obtained during my visiting period at the WebDam ERC project, is the first to use deletion as a method of watermarking for databases.

[1] http://www.cominlabs.ueb.eu/themes/project/

4 Dissemination of Results

Students

Ph.D students

- (running) Joint direction (33%) with Frédéric Cuppens (33%) and Nora Cuppens-Boulahia (33%) of Anis Bkakria's thesis (Labex CominLabs funding, POSEIDON project), on "security politics for outsourced data", started September 2012.
- Joint direction (33%) with Lylia Abrouk (33%) and Nadine Cullot (33%) of Damien Leprovost's thesis (Bourgogne Young Entrepreneur Funding) entitled "Community discovery by semantic analysis", started September 2009, defended November 30, 2012. Now postdoc in the Axis team at Inria Rocquencourt.
- Joint direction (95%) with Michel Scholl (5%) of Julien Lafaye's thesis (Polytechnique funding), entitled "Database watermarking with constraint preservation", started September 2004, defended November 7, 2007. Now working for the IT company Scimetis.
- Joint direction (30%) with Bernd Amann (70%) of Camélia Constantin's thesis (French research ministry funding), entitled "Web services ranking by utility", started September 2004, defended November 27, 2007. Camélia is now a research assistant at the LIP6 Lab, Paris VI University.

Research Master students

- Adam Kammoun (2014)
- Julien Lafaye (2004)
- Camelia Constantin (2004)
- Ammar Mechouche (2005)
- Jean Béguec (2006)
- Damien Leprovost (2009)

Engineer students

- Camélia Constantin (2003), Meryem Guerrouani (2005), Guillaume Chalade (2006), Karine Volpi (2006), Robert Abo (2006), Mai Hoa Guennou (2007), Juan Pablo Stocca (2013).

Funded projects

Labex CominLabs POSEIDON (member) This project, started in 2012, concerns the security of outsourced data (2 PhD thesis, 1 18-month postdoc, funded for 49 KE, non-staff costs).

PEPS CNRS STRATES (head) This 2010 project, funded for 10 KE studied keyword pricing in search engines, with two economists from École d'économie de Paris.

ANR CONTINT Neuma (member) [2] This 3-years project, started end 2008, funded for 620 kE, focuses on wide musical symbolic databases. This project gathers musicologists from CNRS (IRPFM), along with computer sciences labs (LAMSADE, LE2I) and an IT company (ARMADILLO).

ACI Sécurité Tadorne (head) [1] This 4-years project started in 2005, funded for 61 kE, concerns database watermarking under constraints. Participant labs are CEDRIC, GREYC, LAMSADE and COGIT (French National Cartography Agency) ;

National collaborations

- Visitor of the Wisdom group (`http://wisdom.lip6.fr`), a database group gathering the database groups from LIP6, LAMSADE and CEDRIC labs (PPF - plan pluri-formation) ;
- External participant of SemWeb and SCALP projects.
- Co-authors and collaborators : Serge Abiteboul, Fabian Suchanek, Cristina Bazgan, Bernd Amann, Philippe Rigaux, Richard Chbeir, Anne Ruas, Julien Lafaye, Camelia Constantin, Michel de Rougemont.

Invited talk

- PresDB 2007 (International Workshop on Databases Preservation, Edinburgh, March 23, 2007), "Database watermarking : protection by alteration".

Program committee

- Program chair of the national database conference BDA 2014.
- PC member of the international conference EDBT 2014.
- PC member of the workshop on Open Data WOD 2012 and 2013.
- PC member of international conferences CSTST 2008 and ICDIM 2008 ;
- Demo chair of the national conference Bases de données avancées (BDA) 2008 ;
- PC Chair of SWAN 2006 (1st Workshop on Security and Trust of Web-oriented Application Networks) ;
- PC member of the national conferences Bases de données avancées (BDA) 2005, 2008 and 2009 ;
- Reviewer for journals JOT (2012), JCSS (2005), TKDE (2005, 2006), Information systems (2007), TDSC (2005), TISSEC (2005), WWWJournal (2005), Acta Informatica (2005), Infosec (2004) and TODS (2003), external reviewer for conferences ACNS 2007, ASIACCS 2007, ICDE 2007, ICDIM 2006 et 2007, ASIAN 2005, PODS 2005, SOFSEM 2005, VLDB 2005, EDBT 2004, VLDB 2003.

Références

[1] Projet Tadorne (tatouage de données contraintes). `http://cedric.cnam.fr/vertigo/tadorne`.

[2] The NEUMA Project. `http://neuma.irpmf-cnrs.fr`.

[3] R. Agrawal and J. Kiernan. Watermarking Relational Databases. In *International Conference on Very Large Databases (VLDB)*, 2002.

[4] D. Baelde, P. Courtieu, D. Gross-Amblard, and C. Paulin-Mohring. Towards provably robust watermarking. In L. Beringer and A. P. Felty, editors, *ITP*, volume 7406 of *Lecture Notes in Computer Science*, pages 201–216. Springer, 2012.

[5] R. Chbeir and D. Gross-Amblard. Multimedia and Metadata Watermarking Driven by Application Constraints. In *IEEE Multi Media Modelling conference (MMM)*, 2006.

[6] C. Constantin, D. Gross-Amblard, and M. Guerrouani. Watermill : an Optimized Fingerprinting System for Highly Constrained Data. In *ACM MultiMedia and Security Workshop*, New York City, New York, USA, January 1–2 2005.

[7] C. Constantin, D. Gross-Amblard, M. Guerrouani, and J. Lafaye. Logiciel Watermill. `http://watermill.sourceforge.net`.

[8] D. Gross-Amblard. Query-Preserving Watermarking of Relational Databases and XML Documents. In *Symposium on Principles of Databases Systems (PODS)*, pages 191–201, 2003.

[9] M. Guerrouani. Tatouage de documents xml contraints. Technical report, Rapport scientifique CEDRIC - Mémoire d'ingénieur CNAM, 2005.

[10] S. Khanna and F. Zane. Watermarking maps : hiding information in structured data. In *Symposium on Discrete Algorithms (SODA)*, pages 596–605, 2000.

[11] J. Lafaye. Enhancing security of Web Services Workflows using Watermarking. Technical report, Rapport scientifique CEDRIC - Master Thesis Report, 2004.

[12] J. Lafaye. An analysis of database watermarking security. In *IAS*, pages 462–467. IEEE Computer Society, 2007.

[13] J. Lafaye. On the complexity of obtaining optimal watermarking schemes. In *6th International Workshop on Digital Watermarking (IWDW'07)*, pages 462–467, Guangzhou, China, December 2007.

[14] J. Lafaye, J. Béguec, D. Gross-Amblard, and A. Ruas. Invisible graffiti on your buildings : Blind and squaring-proof watermarking of geographical databases. In D. Papadias, D. Zhang, and G. Kollios, editors, *SSTD*, volume 4605 of *Lecture Notes in Computer Science*, pages 312–329. Springer, 2007.

[15] J. Lafaye and D. Gross-Amblard. XML streams watermarking. In *IFIP WG 11.3 Working Conference on Data and Applications Security (DBSEC)*, 2006.

[16] J. Lafaye, D. Gross-Amblard, C. Constantin, and M. Guerrouani. Watermill : An optimized fingerprinting system for databases under constraints. *IEEE Trans. Knowl. Data Eng. (TKDE)*, 20(4) :532–546, 2008.

[17] A. Mechouche. Tatouage de données géographiques. Technical report, Rapport scientifique CEDRIC - Rapport de master, 2005.

[18] B. Mungamuru and H. Garcia-Molina. Predictive pricing and revenue sharing. In C. H. Papadimitriou and S. Zhang, editors, *WINE*, volume 5385 of *Lecture Notes in Computer Science*, pages 53–60. Springer, 2008.

[19] J. F. Nash. Equilibrium points in n-person games. *Proc. of the National Academy of Sciences*, 1950.

[20] N. Nisan, T. Roughgarden, E. Tardos, and V. V. Vazirani, editors. *Algorithmic Game Theory*. Cambridge university Press, 2007.

[21] R. Sion, M. Atallah, and S. Prabhakar. Rights protection for relational data. In *International Conference on Management of Data (SIGMOD)*, 2003.

[22] J. von Neumann and O. Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, 1944.

David Gross-Amblard
Major Publications in Recent Years

International journals

1. Anis Bkakria, Frédéric Cuppens, Nora Cuppens-Boulahia, José M. Fernandez, and David Gross-Amblard. Preserving Multi-relational Outsourced Databases Confidentiality using Fragmentation and Encryption. In *Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications (JoWUA)*, 4(2) : 39-62, June 2013.
2. **David Gross-Amblard. Query-Preserving Watermarking of Relational Databases and XML Documents. *ACM Transactions on Database Systems (ACM TODS)*, 36(1) :3 (2011).**
3. **Julien Lafaye, David Gross-Amblard, Camélia Constantin and Meryem Guerrouani. Watermill : an optimized fingerprinting system for highly constrained data. IEEE Transactions on Knowledge and Data Engineering (TKDE) (accepted 9/2007), April 2008 (Vol. 20, No. 4) pp. 532-546.**
4. **David Gross-Amblard and M. de Rougemont. Uniform generation in spatial constraint databases and applications. In *Journal of Computer and System Sciences (JCSS)*, 72(4) : 576-591, June 2006.**

National journals

1. **Sonia Guéhis, David Gross-Amblard, Philippe Rigaux. Un modèle de production interactive de programmes de publication. Ingénierie des Systèmes d'Information (Networking and Information Systems),revue des sciences et technologies de l'information (RTSI) série ISI, 13 (5), pp. 107-130, octobre 2008.**
2. **Camelia Constantin, Bernd Amann and David Gross-Amblard. Un modèle de classement de services par contribution et utilité. In *Revue des sciences et technologies de l'information* (numéro special "Recherche d'information dans les systemes d'information avances") (1633-1311) - 12(1), pp.33-60, 2007.**

International conferences with peer review

1. David Baelde, Pierre Courtieu, David Gross-Amblard and Christine Paulin-Mohring. Towards Provably Robust Watermarking. In Interactive Theorem Proving, Princeton, USA, August 2012.
2. Fabian M. Suchanek, David Gross-Amblard, Serge Abiteboul : Watermarking for Ontologies. In Proceedings of International Semantic Web Conference (1) 2011 : 697-713.
3. Sonia Guehis, David Gross-Amblard and Philippe Rigaux. Publish By Example. In Proceedings of IEEE International Conference on Web Engineering (ICWE'08), 14-18 Juillet 2008, Yorktown Heights, New York.
4. Julien Lafaye, Jean Béguec, David Gross-Amblard and Anne Ruas. Invisible Graffiti on your Buildings : Blind & Squaring-proof Watermarking of Geographical Databases. In *10th International Symposium on Spatial and Temporal Databases (SSTD)*, July 16-18, 2007, Boston. LNCS 4605, pages 312-329.
5. Julien Lafaye and David Gross-Amblard. XML Streams Watermarking. In *20th Annual IFIP WG 11.3 Working Conference on Data and Applications Security (DBSec2006)*, Sophia Antipolis, France, 7/31 - 8/02 2006, pages 74–88.
6. Camélia Constantin, Bernd Amann, David Gross-Amblard. A Link-Based Ranking Model for Services. In *Cooperative Information Systems (CoopIS) International Conference, 2006*, pages 327-344.
7. Multimedia and Metadata Watermarking Driven by Application Constraints, avec Richard Chbeir, In IEEE Multi Media Modelling conference (MMM), 8 pp., 2006.

National conferences with peer review, informal proceedings

1. Publication de données par l'exemple. Sonia Guéhis, David Gross-Amblard et Philippe Rigaux. In *Journées nationales Bases de données avancées (BDA 2007)*, Marseille, France, 23/26-10 2007.
2. Invisible Graffiti on your Buildings : Blind & Squaring-proof Watermarking of Geographical Databases. Julien Lafaye, Jean Béguec, David Gross-Amblard and Anne Ruas. In *Journées nationales Bases de données avancées (BDA 2007)*, Marseille, France, 23/26-10 2007.
3. Camélia Constantin, Bernd Amann, David Gross-Amblard. A Link-Based Ranking Model for Services. In *Journées nationales Bases de données avancées*, Lille, France, 10/17-20 2006.

Softwares

1. Camélia Constantin, David Gross-Amblard, Meryem Guerrouani et Julien Lafaye. *`Watermill` : database watermarking with optimized constraint preservation.*
   `http://watermill.sourceforge.net`
2. Julien Lafaye et Jean Béguec. Geographic data watermarking library Watergoat (OpenJump plugin).
   `http://cedric.cnam.fr/~lafaye_j/index.php?n=Main.WaterGoatOpenJumpPlugin`
3. Sonia Guehis. Web publishing-by-example DocQL suite.
   `http://www.lamsade.dauphine.fr/~guehis/docql/`

Contribution to the Activity Report of the Department *Data and Knowledge Management*

January 17, 2014

Israel César Lerman
Professeur émérite, Université de Rennes 1, Irisa
Département Data and Knowledge Management, Irisa

1 Association Rules, Clustering and Data Mining

1.1 Association Rules and Data Mining

1.1.1 Overview; Position of the Problem

Building a relevant interestingness measure for association rules is a fundamental problem in *Data Mining* [GHe07]. We assume a context defined by a data table crossing a set $\mathcal{A}$ of descriptive attributes with a set $\mathcal{O}$ of objects described. The latter is generally given by a training set provided from a universe $\mathcal{U}$ of objects. The most important and basic case is that where $\mathcal{A}$ is constituted by Boolean attributes. Extension to other types of descriptive attributes is also studied in many research works.

Let a and b be two Boolean attributes from $\mathcal{A}$, a statistical *association rule* (also called *implication rule*) is denoted symbolically by $a \rightarrow b$. Intuitively, it means: "If the attribute a is *true* on a given object o belonging to $\mathcal{O}$, then, generally but not absolutely, b is *true* on o. In these conditions, the matter is to assess this statistical tendency. As in logics, a and b are called *premise* and *conclusion*, respectively. This evaluation is obtained by means of a numerical index. Many indices have been proposed in the literature. All of them consider only the two attributes a and b to be compared. One important facet of the originality of our approach consists in taking into account the strength of the

[GHe07] F. GUILLET and H.J. HAMILTON eds. *Quality measures in data mining, Studies in Computational Intelligence, vol. 43.* Springer, 2007.

association $a \to b$ in a relative manner, with respect to the set $\mathcal{A} \times \mathcal{A}$ of all ordered attribute pairs.

Likelihood Linkage Analysis Classification approach leads to a powerful and fine tool for clustering and data analysis of complex data [5, 4, 11, 10, 7] [IL06]. All mathematical types of data can be processed by this method. It is based on two principles:

1. Set theoretic and relational mathematical representation of the descriptive attributes with respect to the object set $\mathcal{O}$;

2. Probabilistic evaluation of the associations between descriptive attributes and of the similarities between objects or categories.

In [5] a very large range of data types are clearly specified, according to item 1. The probabilistic evaluation - mentioned in item 2 - is obtained with respect to an adequate independence probabilistic hypothesis between the descriptive attributes. This method provides a probabilistic association coefficient between Boolean attributes. The latter is symmetrical and for an ordered pair of Boolean attributes (a, b), it expresses a measure of statistical equivalence degree between a and b. We can denote this symbolically by $a \leftrightarrow b$.

The idea to adapt this symmetrical index to the asymmetrical implicative case mentioned above, was proposed, studied and applied [GRA79,LGR81]. It is mainly a local version of this index, restricted to the comparison of a single ordered pair (a, b) of Boolean attributes which is considered in the cited references. However, this local form of the probabilistic index tends - when the object set size increases - towards one of two values 0 and 1, 0 in the repulsive case and 1 in the attractive one. These two cases are defined with respect to a statistical independence hypothesis.

Now, generally, the data size is extremely large in *Data Mining* and then, it is imperious to have a discriminant probabilistic index for interestingness measure of association rules.

1.1.2 Association Rules and Data Mining; New Results

Such an index is obtained from a very simple normalization technique, called "Similarity Global Reduction". Mathematical and statistical justifications were provided for this in the case of symmetric comparison of boolean attributes

[IL06] I.-C. LERMAN. Coefficient numérique général de discrimination de classes d'objets par des variables de types quelconques. *Revue de Statistique Appliquée*, (LIV(2)):33–63, 2006.

[GRA79] R. GRAS. *Contribution à l'étude expérimentale et à l'analyse de certaines acquisitions cognitives et de certains objectifs didactiques en mathématiques, Doctorat d'État.* PhD thesis, Université de Rennes 1, 1979.

[LGR81] I.C. LERMAN, R. GRAS, and H. ROSTAM. Élaboration et évaluation d'un indice d'implication pour des données binaires i et ii. *Mathématique et Sciences Humaines*, (74-75):5–35, 5–47, 1981.

[LER84]. On the other hand, experimental analysis has validated this approach. This method was transposed to the asymmetrical implicative case. Its limit behaviour was studied with respect to an increasing model of the object set $\mathcal{O}$, this model being consistent with the *Data Mining* issue [6].

Obtaining a probabilistic discriminant measure of the *Likelihood of the Link* for association rules is also an objective in [RM08]. For this approach the data are summarized by means of a hypothetical sample sized arbitrarily 100. Then, the notion of *TestValue* is applied to the latter sample.

An extensive theoretical, methodological and experimental analysis [8] has been carried out in order to compare different approaches where a probabilistic index of the *Likelihood of the Link* takes part. This analysis is based on increasing models of the number of objects. On the other hand, variations of the level and the nature of the link between *premise* and *conclusion* for a given association rule, are considered in this analysis. The mathematical and experimental results confirm the validity of our normalization method.

Two major aspects of the previous work gave rise to two significant contributions to the EGC2011 conference [9] and [2]. These led to the important article *Comparing two discriminant probabilistic interestingness measures for association rules* [12].

1.1.3 Clustering and Data Mining; Recent and New Results

Let us return to the case where the set $\mathcal{A}$ of Boolean attributes is endowed with a symmetrical association coefficient. The agglomerative construction of a classification tree based on a symmetrical notion of association measure between the built up clusters leads to the discovery of significant behaviour profiles and subprofiles in the universe described [11, 7].

Consider now the case where the attribute set $\mathcal{A}$ is endowed with an index of implication, defining an association rule coefficient on $\mathcal{A}$. The latter is asymmetrical. A requested condition for building a classification tree on $\mathcal{A}$, is to reflect this asymmetry. The formation of an implicative tree is proposed in [GL93]. In this, the link between two clusters is directed (for example, from left to right). In [3] a global analysis of this directed tree structure is provided.

The sought structure called *directed hierarchy* is examined in a complete framework in [13]. In this work, we establish in a constructive way a bijective correspondence between a directed hierarchy and a specific notion of ultrametric distance called *directed ultrametric*. This result establishes the transposition to

[LER84] I.-C. LERMAN. Justification et validité statistique d ' une échelle [0,1] de fréquence mathématique pour une structure de proximité sur un ensemble de variables observées. *Publications de l'Institut de Statistique des Universités de Paris*, (3-4):XXIX, 27–57, 1984.

[RM08] R. RAKOTOMALALA and A. MORINEAU. The tvpercent principle for the counterexamples statistic. In F. Guillet R. Gras, E. Suzuki and F. Spagnolo, editors, *Statistical Implicative Analysis*, pages 449–462. Springer, 2008.

[GL93] R. GRAS and A. LARHER. L'implication statistique, une nouvelle méthode d'analyse des données. *Mathématiques (, Informatique) et Sciences Humaines*, (120):5–31, 1993.

the asymmetrical case of a very known result (the Johnson correspondence) obtained in the classical and much simpler symmetrical case. Thus, the passage from the classical symmetrical tree construction to the asymmetrical one is made explicit.

1.1.4 Identification of proteic families; New Results

In the Abbassi work (supervised by R. Andonov with my contribution) [1], the Clustering and Classification problems are very important. Facing the very considerable size and increasing of the databases storing macro molecular structures, unsupervised Clustering and supervised Classification take fundamental parts. For a given protein data representation, the matter consists of identifying by a clustering process, families and even super families of proteins. Moreover, for a given unknown query protein, it is essential to recognize if there exists, in the database concerned, an identified family to which the query belongs. For these two objectives the software $CHALH$ (see below) has proved to be very adapted and efficient. To show that on the basis of real data, two original methodological developments were carried out:

1. Protein clustering in the case where for each protein pair, only a *lower* and an *upper* bounds are available for estimating their similarity;

2. Use of $CHALH$ - after a specific process - as a tool of supervised Classification in order to recognize the proteic family to which an unknown query protein belongs.

1.1.5 Work of reflection and synthesis

The largest part of my work this year has been to study in depth several important chapters of my previous book *Classification et Analyse Ordinale des Données* published by Dunod - with the support of the Centre National de la Recherche Scientifique - in 1981. Indeed, considerable research and development has taken place in recent years following the topics and ideas covered in these chapters.

1.2 Software

$CHAVLH$ (Classification Hiérarchique par Analyse de la Vraisemblance des Liens en cas de données Hétérogènes) [PLL05] is the software which implements the Likelihood Linkage hierarchical agglomerative clustering. For a description of an object set $\mathcal{O}$ the following types of descriptive attributes are provided:

[PLL05] P. PETER, H. LEREDDE, and I.C. LERMAN. Notice du programme CHAVLH (Classification Hiérarchique par Analyse de la Vraisemblance des Liens en cas de variables Hétérogènes). Dépôt APP (Agence pour la Protection des Programmes) IDDN.FR.001.240016.000.S.P.2006.000.20700, Université de Rennes 1, Décembre 2005.

1. Numerical;
2. Boolean;
3. Nominal categorical;
4. Ordinal categorical;
5. Categorical, endowed with an ordinal or numerical similarity between its values.

For the latest type, the attribute is called *preordonance* attribute.

Such a description is represented by a classical data table crossing the object set $\mathcal{O}$ with an attribute set $\mathcal{A}$. Clustering $\mathcal{O}$ can be carried out when the attribute set $\mathcal{A}$ is constituted by attributes of

- one single type;
- different types.

Clustering the attribute set $\mathcal{A}$ requires a single type for all of the attributes. However, *preordonance* coding can be considered for all of the descriptive attributes [OA91].

The software *AVARE* (*A*ssociation entre *VA*riables *RE*lationnelles) calculates the symmetrical association coeficients table between such attributes [OA00]. This software has been integrated in *CHAVLH* in 2011 by Philippe Peter.

Two other types of a data table can be handled by *CHAVLH*:

- Pairwise dissimilarity table between objects, directly provided by expert knowledge or other sources;
- Horizontal juxtaposition of contingency tables.

CHAVLH is very used. More particulary, it has been applied in many research works at the IRISA institute. It has played an important role in the validation of the results of the thesis of Noel Malod-Dognin: “Protein Structure Comparison: From Contact Map Overlap Maximization to Distance-based Alignement Search Tool”, defended in 2010.

CHAVLH is implemented in “GenOuest Bioinformatics Platform” of *Symbiose* project, as a clustering tool. Interfacing project is envisaged in order to optimize its use.

Since July 2007, an ergonomic and simplified version of *CHAVLH*, called *LLAhclust* (*L*ikelihood *L*inkage *A*nalysis *h*ierarchical *clust*ering), is implemented

[OA91] M. OUALI-ALLAH. *Analyse en préordonnance des données qualitatives. Application aux données numériques et symboliques.* PhD thesis, Université de Rennes 1, décembre 1991.

[OA00] M. OUALI ALLAH. Programme de calcul de coefficients d'association entre variables relationnelles. *La Revue de Modulad*, (25):63–74, 2000.

in the **R** software environment (I. Kojadinovic (École Polytechnique de l'Université de Nantes), I.-C. Lerman, P. Peter and N. Le Meur de l'Irisa).

CHAVLH is written in Fortran77. A *C* language version is planned by Philippe Peter.

1.3 Scientific Committees and Editorial Boards

I.-C. Lerman was a PC member of the *EGC2012* conference, *Extraction et Gestion de Connaissances*, January 2011, Bordeaux, France.

I.-C. Lerman is a member of the editorial board of the journal "Mathématiques et Sciences Humaines, *Mathematics and Social Sciences*", Paris.

I.-C. Lerman was in 2011 "Special Reviewer" of the Journal of Classification", New York.

1.4 National Collaborations

- Sylvie Guillaume, Université de Clermont, Auvergne, LIMOS, Clermont Ferrand ;
- Philippe Peter, Université de Nantes, Laboratoire d′Informatique de Nantes Atlantique, Equipe COD, Site Polytech ′ Nantes.

Major publications in recent years (2007-2013)

References

[1] N. ABBASSI. Identification de familles protéiques. Rapport de stage master 2, Université de Rennes 1, Juin 2013.

[2] S. GUILLAUME and I.-C. LERMAN. Analyse du comportement limite d'indices probabilistes pour une sélection discriminante. In A. Khenchaf et P. Poncelet, editor, *Revue de l'Information et des Nouvelles Technologies, RNTI E.20, EGC'2011*, pages 657–664. Hermann, 2011.

[3] I.-C. LERMAN. Analyse logique, combinatoire et statistique de la construction d'une hiérarchie binaire implicative; niveaux et noeuds significatifs. *Mathématiques et Sciences Humaines, Mathematics and Social Sciences*, (184):47–103, 2008.

[4] I.-C LERMAN. Analyse de la vraisemblance des liens ; une méthodologie d'analyse classificatoire de données relationnelles : le cas symétrique d'abord, le cas orienté ensuite. Séminaire, IRISA-INRIA, October 2012.

[5] I.-C LERMAN. Facets of the set theoretic representation of categorical data. Publication Interne 1988, IRISA-INRIA, January 2012.

[6] I-C. LERMAN and J. AZÉ. A new probabilistic measure of interestingness for association rules, based on the likelihood of the link. In F. Guillet and H.J. Hamilton, editors, *Quality measures in data mining, Studies in Computational Intelligence, vol. 43*, pages 207–236. Springer, 2007.

[7] I.-C. LERMAN and K. BACHAR. Comparaison de deux critères en classification ascendante hiérarchique sous contrainte de contiguïté. *Journal de la Société de Statistique de Paris et Revue de Statistique Appliquée*, (149, 2):45–74, 2008.

[8] I.-C. LERMAN and S. GUILLAUME. Analyse comparative d'indices discriminants fondés sur une échelle de probabilité. Rapport de Recherche PI Irisa 1942, RR Inria 7187, IRISA-INRIA, Février 2010.

[9] I.-C. LERMAN and S. GUILLAUME. Comparaison entre deux indices pour l' évaluation probabiliste discriminante des règles d'association. In A. Khenchaf et P. Poncelet, editor, *Revue de l'Information et des Nouvelles Technologies, RNTI E.20, EGC'2011*, pages 647–656. Hermann, 2011.

[10] I.-C. LERMAN and P. PETER. Representation of concept description by multivalued taxonomic preordonance variables. In G. Cucumel P. Brito, P. Bertrand and F. Carvalho (eds), editors, *Selected Contributions in Data Analysis and Classification*, pages 271–284. Springer, 2007.

[11] I.C. LERMAN. Analyse de la vraisemblance des liens relationnels une méthodologie d ' analyse classificatoire des données. In Younès Benani and Emmanuel Viennet, editors, *RNTI A3, Revue des Nouvelles Technologies de l'Information*, pages 93–126. Cèpaduès, 2009.

[12] Israël-César Lerman and Sylvie Guillaume. Comparing Two Discriminant Probabilistic Interestingness Measures for Association Rules. In Fabrice Guillet, Bruno Pinaud, Gilles Venturini, and Djamel Abdelkader Zighed, editors, *Advances in Knowledge Discovery and Management*, volume 471 of *Studies in Computatinal Intelligence*, pages 59–83. Springer, 2013.

[13] Israël-César Lerman and Pascale Kuntz. Directed Binary Hierarchies and Directed Ultrametrics. *Journal of Classification*, 28(3):page 272–296, 2011.

Zoltán Miklós

Maitre de conferences, Université de Rennes 1

1 Overall objectives

My recent work has focused on semantic interoperability establishment techniques in a network setting, notably on the Web as well as in business-oriented context. This work is a collaboration with my pervious group, where I co-supervised 2 PhD students: Surrender Reddy Yerva and Nguyen Quoc Viet Hung. Both of them defended their thesis in 2013 (in July and in December, respectively). I am continuing my collaboration with Nguyen Quoc Viet Hung.

At the same time, I am gradually building up research collaboration with David Gross-Amblard and other members of IRISA. In particular, I was helping David Gross-Amblard in the preparations of various proposals submissions (such as 2 ANR proposals, Mathise proposal, Master research) as well as in the construction of the research team DRUID. We were also active in the mastodons Aresos project, where we had a research intern Juan-Pablo Stocca, with whom we worked on the reconstruction of science phylomemetic networks from a large corpus of scientific articles. We have developed a MapReduce variant of the existing state-of-the-art algorithms. I also worked on some questions of database theory, namely the containment of conjunctive queries under bag semantic.

My other activities include the co-creation et coordination of an informal working group on the themes Open Data and Crowdsourcing. I also started coordinating the preparation of a proposal for an ANR call 2014 (initial partners INRIA and University of Lille/CNRS) and have made some initial contacts with local industrial partners for potential partnership (CIFRE). I am following the H2020 calls to potentially participate in a collaborative project.

2 Scientific Foundations

2.1 Semantic interoperability

The large body of research on schema matching (or ontology alignment) focuses on identifying attribute correspondences between two schemas, while in the business world the databases do not exist in isolation, but in the context of a network. The presence of such networks is often not considered in schema matching, even this could be an important source of information (in case it is known).

The schema matching process is inherently uncertain, but the business requirements w.r.t. to the quality of matchings is usually high. Thus, often in practice there is phase, where expert users fix the errors produced by the automatic matching tools. We call this phase the reconciliation phase. We study this reconciliation phase of schema matching in the presence of a matching network. In fact, the reconciliation is the real cost of schema matching in an enterprise, as this involves human experts. Thus, there is a high need to reduce this effort.

2.2 Phylomemetic networks

Large collections of scientific articles are a rich source of information if we would like to understand the evolution of ideas in scientific thought. Recent papers describe automated techniques to reconstruct a phylomemetic tree, a structure that shall represent the lineage between scientific fields. The constructed structures have largely extended our knowledge about the developments of our understanding of the corresponding domains, so one would like to reconstruct the phylomemetic trees even larger corpora of scientific articles. This raises a number of computational issues, including the reconstruction of large co-occurrence graphs, efficient discovery of dense structures in a large graph.

Our ongoing work in this are focuses on the development of techniques that enable the social scientists to analyse the temporal evolution of scientific ideas, based on large corpora of scientific articles. We have developed variants of the algorithms from the literature, relying on the MapReduce paradigm.

2.3 Conjunctive queries under bag semantic

Real databases often contain multiple copies of tuples. Equally, the result of an SQL query can contain multiple copies of the same tuple (if the DISTINCT command is not used). Database theory has developed mathematical models for modelling this situations. One of these models is called bag semantics. While this model does not faithfully model real SQL queries, and much more complete models exists, there is a number of open problems even for this setting. For example, one of the most famous open questions is to show whether the conjunctive query containment under bag semantic is decidable. This question reappears in the more complete models (for example, combined semantic), thus it is important to answer this question. There is a number of other database problems, where the same mathematical question arises: most importantly, the provenance of data tuples.

3 New results

3.1 Semantic Interoperability

We have obtained several interesting results concerning the reconciliation process in a schema matching network. In our work, we exploit the presence of this matching network: we represent the natural expectations that one has from a network of databases in the form of a constraint satisfaction problem (in the formalism of Answer Set Programming) and use this formalism for various purposes.

In particular, we could use the constraint representation, to reduce the necessary human intervention [C2]. This paper develops a general model of schema matching networks, and uses the reasoning on the user input (in the presence of consistency constraints). While the reasoning techniques can systematically reduce the efforts, in real settings the expert who is working on the reconciliation has only a limited time budget, thus in practice he can only provide partial inputs. In order to cope with this situation we have developed a pay-as-you-go variant of the reconciliation process where we apply advanced probabilistic sampling techniques to obtain the most probable set of attribute correspondences from an incomplete set of assertions.

In [C2] we assume that there is a single expert working on the reconciliation. We have also analysed, how could a crowd (of non-experts) realize the same task [C1]. In this case we need to handle the imprecision of user input, where we apply the expectation maximization techniques. In the case of crowd, one would like to minimize the financial cost that is needed to realize the task by the crowd. Here we again exploit the presence of the network and the network-level consistency conditions. With the help of a simulation (and a theoretical analysis) we could demonstrate that the constraints can be exploited to lower the expected error, thus for achiving a given (estimated) error rate, we need less human efforts.

We have also studied the situation, where a small number of experts would like to eliminate these problems, in which case these experts might have conflicting views. For this case we have developed a model based on augmentation techniques [C4], that helps the participating experts identifying the implications of their input, together with a tool [O1] that incorporates our techniques.

In our work we also analysed techniques that help to work with large networks [C3]. For this case, we have developed a schema covering technique, that uses tools and models from mathematical programming and operations research to decompose the schema into smaller units.

3.2 Bag semantics

We have analyzed the conjunctive query containment under bag semantics. Our paper is under review. While in the case of set semantic, the existence of a homomorphism from Q_2 to Q_1 is sufficient and necessary for query containment, in a bag semantic setting this is not sufficient. Our work uses a set of homomorphisms satisfying some special conditions, to analyze the problem.

4 Dissemination of results

4.1 Talks

I gave the following talks:

- On Leveraging Crowdsourcing Techniques for Schema Matching Networks, BDA 2013, Nantes, October 2013
- Schema matching networks, European Commission, Brussels, NisB projet review, March 2013

4.2 Journal, Conference and Research Project Proposal Reviewing

I was acting as a referee for the following journals and conferences. I list here also other related refereeing activities.

Journals: Computer, Future Generation Computing Systems (Elsevier)

Conferences: ECIS'2013 (PC member), LinkedScience workshop at ISWC'2013 (PC member), OnToContent workshop 2013 (PC member), BDA'2013 Bases de données avancées, EDBT'2013

Project proposals:

- Expert evaluator of EU project proposals on behalf of the European Commission for the FP7-ICT-2013-11 Call - Objective 4.2 ? Scalable Data Analytics, May 2013.
- Expert, research project evaluation for the SNF Swiss National Science Foundation

4.3 Major publications in 2013

4.3.1 International Conference with PC

The paper [C1] won the Best Student paper award at DASFAA'2013. It also has been published in the (informal) proceedings of the conference Bases de donnees avancees (BDA'2013), the French national database conference.

[C1] Nguyen Quoc Viet Hung, Nguyen Thanh Tam, Zoltán Miklós, and Karl Aberer. On Leveraging Crowdsourcing Techniques for Schema Matching Networks. In *DASFAA 2013*, 2013.

[C2] Quoc Viet Hung Nguyen, Tri Kurniawan Wijaya, Zoltan Miklos, Karl Aberer, Eliezer Levy, Victor Shafran, Avigdor Gal, and Matthias Weidlich. Minimizing Human Effort in Reconciling Match Networks. In *32nd International Conference on Conceptual Modeling (ER 2013)*, 2013.

[C3] Avigdor Gal, Michael Katz, Tomer Sagi, Matthias Weidlich, Karl Aberer, Zoltan Miklos, Nguyen Quoc Viet Hung, Eliezer Levy, and Victor Shafran. Completeness and Ambiguity of Schema Cover. In *21st International Conference on Cooperative Information Systems (CoopIS 2013)*, 2013.

[C4] Nguyen Quoc Viet Hung, Xuan Hoai Luong, Zoltan Miklos, Tho Thanh Quan, and Karl Aberer. Collaborative Schema Matching Reconciliation. In *21st International Conference on Cooperative Information Systems (CoopIS 2013)*, 2013.

[C5] Nguyen Quoc Viet Hung, Nguyen Thanh Tam, Zoltan Miklos, Karl Aberer, Avigdor Gal, and Matthias Weidlich. Pay-as-you-go Reconciliation in Schema Matching Networks. In *30th International Conference on Data Engineering (ICDE 2014)*, 2014.

4.3.2 Other

[O1] Nguyen Quoc Viet Hung, Xuan Hoai Luong, Zoltan Miklos, Tho Quan Thanh, and Karl Aberer. An MAS Negotiation Support Tool for Schema Matching (Demonstration). In *Twelfth International Conference on Autonomous Agents and Multiagent Systems (AAMAS'2013)*, 2013.