UNIVERSITÉ DE VERSAILLES SAINT-QUENTIN-EN-YVELINES

Département d'informatique
UFR de Sciences

École doctorale "STV"

Privacy-Centric Data Management

RAPPORT SCIENTIFIQUE

pour l'obtention de l'

Habilitation à Diriger des Recherches en Informatique

Université de Versailles Saint-Quentin

par

Benjamin NGUYEN

Composition du jury

Rapporteurs :	Bernd AMANN	Professeur, Université de Paris VI.
	Amr EL ABBADI	Professeur, University of California Santa Barbara.
	Elena FERRARI	Professeur, University of Insubria.
Examinateurs :	Gilles DOWEK	Directeur de Recherches, INRIA.
	Georges GARDARIN	Professeur Émerite, Université de Versailles Saint-Quentin.
Tuteur :	Philippe PUCHERAL	Professeur, Université de Versailles Saint-Quentin & INRIA.

Laboratoire PRISM

cnrs

informatiques mathématiques Inria

Dedication, to be added later...

Contents

Chapter 1

Introduction

"When you have a lot of jumbled up ideas, they come together slowly over a period of several years."

– *Tim Berners-Lee*

Summary: *In this chapter, I present a brief summary of my research career. I overview my work on data management performed in my previous research group, how it led me to my current interests in data management and privacy. Finally, I describe the projects I participated in.*

Contents

1.1 Historical overview of my research

DATA management, and more specifically *Personal* Data Management, has been the core of my research since 2000, with my DEA and Ph.D. obtained in 2003. One of the reasons, if not the main reason, that pushed me towards this topic, is the fact that data management is an area of computer science in which applications play a crucial role. Indeed, I have been fascinated by computers and electronics since a very young age, from the practical point of view : building useful applications. Although I have switched from writing programs in BASIC or 6502 Assembly language to C++, Java or Python, and have stopped soldering micro-controlers, I have always tried to anchor my research in the real world, by asking the question "will this be of any use to someone ?". Moreover, in all the projects I have participated in or coordinated, not only have I contributed to the global reflection, theoretical analysis and design ; I have also actively participated in prototyping, development and experimentation.

In this first introductory chapter, I will draw a brief overview of my work, since 2000, highlighting the main results obtained. I describe the context and collaborations that made this work possible, and show how my research interests have evolved from the *management of data*, when the Web was still in its infancy, to the *protection of the privacy of this data*, in a world of Facebook and state-scale monitoring of personal information.

Ph.D. and post-doc (2000-2004)

I received my Ph.D. from University of Paris XI in 2003 [125], under the supervision of Serge Abiteboul. During my Ph.D. and following post-doctoral year, I worked on two topics : the first was the Xyleme Monitoring system, [126, 156]. It was a kind of precursor to RSS feeds, where a user could subscribe to XML feeds, and receive notifications when topics of interest were detected. The second was a technique to cluster Web pages into thematic subsets [97, 149], using web links and semantics (ontologies), a collaboration with Michalis Vazirgiannis, Maria Halkidi and Iraklis Varlamis of the Athens University of Economics and Business in Greece.

Associate Professor, Data Integration and Management (2004-2010)

I then joined the University of Versailles and Saint-Quentin-en-Yvelines (UVSQ) in 2004, in the PRiSM Laboratory in Georges Gardarin's "Data Integration and Management" (DIM) team as Associate Professor, until 2010. During my years in DIM, I led my research along three directions, in the field of XML data management : (1) peer-to-peer (P2P) XQuery evaluation [1, 60, 62], (2) applications of XML/XQuery to sociological research [74, 127, 128] and (3) automatic ontology alignment using OWL and XML Schema [45, 46, 48, 47]. I co-supervised (with Georges Gardarin) Bogdan Butnaru's Ph.D. thesis [58] on theme (1) and Ivan Bedini's Ph.D. thesis [41] on theme (3). I participated in three funded research projects on theme (1) (RNTL "WebContent", ACI "SemWeb", ANR "ROSES") and two funded research projects on theme (2) (ACI "NPP", ANR "WebStand", for which I was project leader and coordinator).

During all these years, I had been working on models and systems to manage data that was *produced by humans*, and this was what created a lot of its value : indeed XML peer-to-peer

systems, and the XML Schema based integration relied on schemas built by individuals, and data created by individuals. Similarly, the sociological studies led using our XML based system analyzed email data. Indeed, during the *WebStand* project, we analysed the inner workings of the World Wide Web Consortium (W3C) through the analysis of public email lists, by harvesting and processing this data to discover trends or specific events in the life of a working group. As our institution was a member of the W3C we also had access to private email discussion groups for all the W3C working groups. Obviously, we felt it would have been much more interesting to study these private lists. This led to my first enlightening discussion on privacy, in 2006 (ironically also the year I switched to gmail), with Ralph Swick, the head of the Technology and Society Domain, and Chief Operating Officer of W3C, on how to *anonymize* this *personal data* that we were processing, and on the conflicting needs of data analysts (be they sociologists, medical doctors, advertisers, etc.), who want *high quality data* and of users, who want to remain *private* and *anonymous*. In the end, in our work, such as [128], we went on to use only *publicly* available data – which is not to say that nothing *private* could be deduced from it ! All this got me started thinking on the problem of managing private data, not only from the point of view of anonymization methods, but rather how it would be possible to enforce these methods, and what elements of the global infrastructure could be *trusted*. This led me quite naturally to work with the Secure and Mobile Information Systems (SMIS) team, led by Philippe Pucheral, who was working on data management on secure hardware, with the co-supervision of Tristan Allard's Ph.D. thesis [11], started in 2007. I would formally join the INRIA team in 2010, when Georges Gardarin retired.

Associate Professor, Secure and Mobile Information Systems (since 2010)

And thus we come to the core topics of this document. Since 2010, I have pursued my research at SMIS, focusing on privacy and security aspects of information management systems and applications. More specifically, I have worked on general architectures using *low-power* and *high-security* (trusted) devices [12, 24], on methods to enforce existing privacy models that require global computations (e.g. aggregations) using such devices [18, 19, 21], and on models to represent, quantify and enforce privacy concepts outside the trusted world, such as *limited data collection* [28, 23, 26].

Since it is this last part of my research that this document will focus on, I will next highlight my research during the DIM period (Section 1.2) and give a global overview of my participation in research projects (Section 1.3).

1.2 Research at DIM Highlights (2004-2010)

Personal data is by essence heterogeneous, since it is generated in many different circumstances, by many different systems, and by many different people, all with different needs. The most natural model to represent such data is XML (eXtensible Markup Language) [57], the standard semi-structured data language. XML is in the center of a galaxy of standards on data representation and management, in particular XQuery [54], which is now at its version 3.0 [138].

The other ingredient to manage semantic data is an ontology e.g. expressed in OWL (Web Ontology Language) [72, 39].

During my years in the DIM team, I have worked on three topics : The main topic was XML data management on distributed architectures, the pinnacle of this work being the XQuery on Peer-to-Peer system (XQ2P [62]) developed in collaboration with Bogdan Butnaru and Georges Gardarin. The XQ2P system is a 98.7% conformant XQuery 1.0 engine (with stream processing XQuery 3.0 features), which is *fully distributed*. This experience still proves helpful in my current work, although I am now focused on the *secure* aspects of computation in distributed environments.

The second topic, on which I was leader, was interdisciplinary research with sociologists, using XML to model and XQuery to query sociological data harvested from the Web. The outcome of these interdisciplinary experiences was very positive, since they have led to the design of *WebStand* [127], an email and web pages analysis system currently used by a french sociology and economy laboratory (CNRS LEST – Laboratoire d'Economie et Sociologie du Travail). This motivated me to continue working with other disciplines (jurists, economists, sociologists), once I joined the SMIS team, by analysing their problems of privacy-centric data management.

The last topic, on which I supervised Ivan Bedini's Ph.D. thesis, on ontology alignment [46] by converting OWL to XML Schema [144, 53] and studying semantic and structural metrics to improve alignment quality. This work was a continuation of work I had started during my Ph.D. and I chose to close this field of research when I joined SMIS in 2010, because I wanted to focus on privacy in data management, and it seem difficult for me to keep an orthogonal research direction active.

1.3 Research Projects Overview

Although my research is not limited to funded research projects, I have been involved in many of them since 2000. All my contributions in these projects have been in the field of data management : Indeed, my research started with a strong interest on data management architectures, and semantic aspects. Over the years, I have shifted from semantic aspects to privacy protection, while maintaining a focus on architectures for data management in a dynamic context (XML, relational and NoSQL data). These projects reflect the evolution of my reserach interests and can be divided into the following categories : (1) Data management architectures (Xyleme, IST FET DBGlobe, RNTL WebContent, ANR ROSES, ANR KISS) (2) Interdisciplinary projects involving databases as core technology for data management of the social/legal objects involved (ACI NPP, ANR WebStand, ANR DEMOTIS, Project Lab CAPPRIS, ISN) (3) Ontologies and XML Data Management (RNTL e.dot, ACI SemWeb) I will next give a quick overview of each project, and explain my role in each one. The dates indicated are not necessarily the dates of the project, but rather the dates of my participation in the project. Note that the list of publications related to the project is not exhaustive.

1.3.1 Data Management Architectures

Ever since my Ph.D. I have been working on projects around innovative data management architectures. The focus of the first projects was on XML data, and the focus of my more recent work is on the privacy aspects of data management.

Xyleme (2000-2003)[1] – Monitoring Module *Task Leader*

Xyleme was a project to develop an XML database architecture. I worked on the development of the XML monitoring system [126] during my Ph.D., which was a novel publish/subscribe and continuous query system, managing XML data streams for individuals (*à la* RSS). The Xyleme system (including my module) became a startup and turned into a full fledged company [157].

DBGlobe (2003-2004)[2] – *Contributor*

I worked on this project during my post-doc at INRIA and Athens University of Economics and Business. The DBGlobe project aimed at developing novel data management techniques to deal with the challenge of global computing, in the face of evolving data. During this project, I worked on the design of a service-oriented web warehouse [2] and a better characterization of evolving web data [149].

RNTL WebContent (2006-2009)[3] – *Steering Committee* and Local XML Store *Task Leader*

The objective of the WebContent project was to produce a flexible and generic platform for content management, integrating semantic web technologies in order to show their effectiveness on real applications with strong economic or societal stakes. The platform was used by industrial partners for business intelligence in aeronautics, or more generally strategic intelligence, microbiological and chemical food risk, and seismic events monitoring. During this project, I was in charge of developing the centralized XML database store [129] (project deliverable), on top of the MonetDB system [55]. I also participated in the development of the global architecture [1], and coordinated the development of the *P2PTester* infrastructure [60] deployed and tested on the french Grid5000 cluster.

ANR ROSES (2008-2010)[4]– *Contributor*

The main objective of the ROSES (*Really Open and Simple Web Syndication*) project was to define a set of services for the syndication, localisation, querying, generation and composition of personal data streams (such as RSS). The solutions proposed in this project were built on peer-to-peer data management. I contributed to this project with my Ph.D. student, Bogdan

[1] Now a company, see http://www.xyleme.com/
[2] See : http://softsys.cs.uoi.gr/dbglobe/index1.html for more information.
[3] See : http://www.webcontent-project.org/ for more information.
[4] See : http://www-bd.lip6.fr/roses/doku.php?id=start for more details.

Butnaru [58], subsequently to the development of *P2PTester*. This infrastructure formed the basis for developing the XQuery on Peer-to-Peer (XQ2P) database engine, a fully (98.7%)compliant XQuery 3.0 engine, designed to manage temporal series, such as RSS data [62].

ANR KISS (2012-current)[5] – *Contributor*

The objective pursued by KISS (*Keeping your Information Safe and Secure*) is to provide a credible alternative to a systematic centralization of personal data on third-party servers and to pave the way for new privacy-by-design solutions dedicated to the management of personal data. The idea promoted in KISS is to embed software components in *trusted devices* capable of acquiring, storing and *securely* managing various forms of personal data (e.g., salary forms, invoices, banking statements, medical data, geolocation data) depending on the applications. These software components form a full-fledged personal data server which can interoperate with external servers and services while still remaining under holder's control. My participation in KISS involves access and usage control models and enforcement algorithms [29], large-scale distributed management of databases on trusted devices [24], and technological transfer to the social care field [23].

1.3.2 Data Management Interdisciplinary Projects

I have always had a strong interest in "humanities", which is the reason I started by trying to apply my research to sociology, and have pursued with judicial applications, and am currently starting a collaboration with economists.

ACI NPP (2004-2007)[6] – *Project Coordinator*

I coordinated the NPP (*Web Standards, Regulations and Public Policies*) project, with Ioana Manolescu (INRIA) and François-Xavier Dudouet (Sociologist). The objective of this project was to study the standardization process of the W3C, through the analysis of mailing lists, and their processing using XQuery [74]. Since this project was quite small, we merged its objectives with the ANR WebStand Project, involving more partners, with a more ambitious goal.

ANR WebStand (2005-2009)[7] – *Project Leader*

I was project leader for the WebStand Project, whose goal was to created a customizable application platform to simplify the work of sociologists studying web data, i.e. web pages and mailing lists. The outcome of this project was the WebStand platform, currently used by sociology researchers of the CNRS LEST (Laboratoire d'Economie et Sociologie du Travail) [128] in order to extract and analyse behaviours of individuals through the processing of emails and web pages.

[5] See : https://project.inria.fr/kiss/ for more details.

[6] See : http://cassiopee.prism.uvsq.fr/papers/aci-npp-rapport-final-2008.pdf for more details (in french).

[7] See : http://www.prism.uvsq.fr/~beng/wiki/index.php/WebStand for more details.

ANR DEMOTIS (2010-2012)[8] – EHR Anonymisation *Task Leader*

DEMOTIS (*Define, Evaluate and Model Electronic Health Record Systems*) was a joint project between computer scientists and jurists studying how to design data management systems that are compatible with laws and regulations, and how these laws and regulations should evolve to take into account computer science constraints. The information system studied was the french Electronic Health Record (EHR) system, which is currently centralized, and to study how a distributed and more secure architecture could both improve efficiency of the system and privacy protection the patients. I participated in work on the problem of executing anonymization processes in a distributed setting [21], in the context of Tristan Allard's Ph.D. thesis [11].

Project Lab CAPPRIS (2012-current)[9] – Privacy Reference Architecture *Task Leader*

CAPPRIS' (*Collaborative Action on the Protection of Privacy Rights in the Information Society*) general goal is to foster collaboration between research groups involved in privacy in France, and the interaction between computer science, law and social sciences communities in this field. I am leading the joint research action on the development of a privacy reference architecture. Expected benefits include a better understanding of architectural problems linked to privacy protection, better coverage of solutions and increased interoperability between Privacy Enhancing Technologies (PETs) themselves and between PETs and other services. This interoperability should hopefully favor the development and adoption of PETs and privacy by design in the future.

Digital Society Institute (ISN) (2013-current) – Privacy Working Group *Co-coordinator*

The ISN (*Institut des Sciences du Numérique* or *Digital Society Institute*) has only just been launched. I am co-organiser of its Privacy Working Group (with Fabrice Le Guel, economist), which brings together computer scientists (data management, cryptography, information science), jurists, sociologists and economists (mainly experimental economy). The objective of this institute is to adopt an interdisciplinary analysis of the problems of the digital society, i.e. that stem from the development of computer science techniques, and test their impact on final users, in order to influence laws and individual's behaviours.

1.3.3 Ontologies and Applications

I co-advised Ivan Bedini's Ph.D. [41] and have participated in several research projects around ontologies. I was also member of the W3C Semantic Web Best Practices Working Group until 2006 and eGov Interest Group until 2009. I have however stopped working on this topic since joining the SMIS team.

[8] See : http://www.demotis.org/ for more details.
[9] See : https://site.inria.fr/cappris_institutionnel/ for more details.

RNTL e.dot (2003-2005)[10] – ***Contributor***

e.dot (*Entrepôts de Données Ouverts sur la Toile* or *Data Warehousing of Open Web Data*) was a project whose goal was to curate web data and use it to populate an XML database (at a time where there were no open web data standards). I was in charge of the work package on the design, annotation [149] and organisation of the warehouse, using XML and Web Services [3].

ACI SemWeb (2004-2007) – ***Contributor***

The goal of the SemWeb project was to use XQuery to query the semantic web. Since then (in 2008), SPARQL [135, 33] has become the standard to query RDF data. During this project, I worked on the use of Web Services to publish and integrate semantic web data [22].

1.4 Beyond Research Activities

Before entering the main matter of this document, which will focus on my current research activities, I briefly highlight in this section some of my other noteworthy activities.

Software Development

I take applications very seriously. In most of the research topics I have addressed, I have always tried to produce more than a *proof of concept* demonstration, but rather a full fledged system. Although number of lines of code do not demonstrate the complexity or difficulty of system implementation, they do give an indication on the size of the project. When available online, I have provided references for the source code of each system. In order to access the code hosted on the SVN server, please use the `public/public` login/password combination. For all projects that I have led, I have chosen to deliver the software under the *free* GPL licence.

- Xyleme Monitoring System (approx. 10K lines of C++)[11]: `Designer and developer.` This module was subsequently sold to the Xyleme Startup (2001).
- THESUS (approx. 25K lines of Java)[12] : `Co-designer and developer.` I designed and developed the link semantics module. (2003).
- The WebStand Suite[13] (CV-Crawler, Mailling List extractor, XQuery Module, approx. 30K lines of Java / perl) : `Project Leader.` This system is currently used by some researchers of the CNRS Laboratoire d'Economie et Sociologie du Travail. (2007).
- P2PTester (approx. 20K lines of Java)[14] : `Project Leader.` A development and instrumentation infrastructure to debug and develop P2P applications. (2008).

[10] See : http://leo.saclay.inria.fr/projects/edot/ for more details.
[11] http://www.xyleme.com/
[12] Varlamis, Vazirgiannis, Nguyen, Halkidi, Patent No.: 1004662. Greek Society of Industrial Property.
[13] Available in GPL at https://cassiopee.prism.uvsq.fr:8443/svn/DIM/trunk/WebStand/software/aXess/
[14] Avaivable as a GPL package in the larger XQ2P project, see https://cassiopee.prism.uvsq.fr:8443/svn/XQ2P/

- WebContent Store (approx. 6K lines of Java and scripts)[15] : `Designer and Developer.` I integrated the MonetDB query engine into the WebContent architecture. (2009).
- Janus[16] (approx. 30K lines of Java) : `Co-designer.` An automatic multiple ontology alignment system. (2010).
- XQuery on Peer-to-Peer (XQ2P) system[17] (approx. 50K lines of Java) : `Project Leader.` A fully compliant distributed XQuery 1.0 database engine integrating some 3.0 features. (2011).
- The XQuery 3.0 Test Suite[18] (approx. 5K lines of XSLT, and over 27.000 unit tests written in XML) : `Upgrade.` I worked on the upgrading and refactoring of the 1.0 test suite. I also participated in the development of a new architecture for the unit tests. Due to my leaving the W3C, the project was completed by Michael Kay and O'Neil Delpratt (Saxonica) in 2013. (my participation ended in 2011).
- MinExpCard[19] (approx. 7K lines of Java and 2K lines of C) : `Designer and co-developer.` A prototype demonstrator for Conseil Général des Yvelines. (2013).
- Trusted Cell Global Queries[20] (currently under development) : `Project Leader.` A SQL-92 compliant DBMS running on the Asymmetric Architecture. Developed in the context of Cuong Quoc To's Ph.D. thesis, based on a first prototype by Tristan Allard.
- Coster[21] (approx. 8K lines of PHP and SQL) : `Designer and developer.` This system is not a research application system : it is a system to manage the teaching activities of university departments. Currently used in three departments of the University of Versailles. (2006-2009).

Standardization

I contributed for several years in W3C working groups which linked to my field of research. The objective was to be in contact with industrial companies actually using the research technologies I was working on, and advocate for what seemed to me as better technical choices when standards were concerned. I participated in three kind of activities in the W3C :

- W3C Advisory Committee (2008-2011), representative for University of Versailles.
- Semantic Web and interdisciplinary standards activities: Semantic Web Best Practices WG (2004-2006), e-gov IG (2008-2009), Social Web XG (2009).
- XQuery WG (2008-2011).

[15] Available in LGPL at `https://cassiopee.prism.uvsq.fr:8443/svn/DIM/trunk/WebContentStore/`
[16] See `http://bivan.free.fr/Janus/index.html`
[17] Available in GPL at `https://cassiopee.prism.uvsq.fr:8443/svn/XQ2P/`
[18] See `http://dev.w3.org/2011/QT3-test-suite/`
[19] Demo system at `https://project.inria.fr/minexp/software/`
[20] See `https://scm.gforge.inria.fr/svn/protocg78-inr-g/branches/SGBD_Cuong`
[21] Available in GPL at `http://sourceforge.net/projects/coster/`

These activities were technically very stimulating, and the people involved in the W3C WGs were all highly competent. However, I closed these activities in 2011, when I decided to focus on privacy aspects of data management.

Introducing Computer Science in High Schools

I had the chance of benefiting from the French "Informatique pour tous" program, launched in the 80's for junior school pupils, and the french "Informatique" option in high school in the mid 90's. However surprising it may seem, computer science since then disappeared from french high school curriculums. In 2012, it finally made a come back, but unfortunately very few teachers had the required skills. In 2010, Franck Quessette and I designed a curriculum[22] and started training high school (Maths and Industrial Science) teachers at University of Versailles. To this end, I participated in a textbook for high school teachers of computer science [34]. Since then, we have trained over 100 teachers, all of which now teach computer science in their respective high schools.

Teaching and administration

Finally, and more classically, as a *Maître de Conférences*, I have given over 192 hours of teaching per year since 2004, in Licence (Bachelors) and Masters levels. I also actively participate in the management of the University of Versailles Computer Science teaching department (department associate director, Bachelors and Masters course director). See Appendix B for more details on my administrative positions.

1.5 Document Structure

This document will focus on my *core computer science research* since 2010, covering the topic of data management and privacy. More specifically, I will present the following topics :

- A new paradigm, called *Trusted Cells* for privacy-centric personal data management based on the Asymmetric Architecture composed of *trusted* or *open* (low power) distributed hardware devices acting as personal data servers and a highly powerful, highly available supporting server, such as a cloud. (Chapter 2).
- Adapting aggregate data computation techniques to the *Trusted Cells* environment, with the example of Privacy-Preserving Data Publishing (Chapter 3).
- Minimizing the data that leaves a *Trusted Cell*, i.e. enforcing the general privacy principle of *Limited Data Collection* (Chapter 4).

[22] Based on the May 2010 proposal by Jean-Pierre Archambault, Gérard Berry, Gilles Dowek and Maurice Nivat.

This document contains only results that have already been published. As such, rather than focus on the details and technicalities of each result, I have tried to provide an easy way to have a global understanding of the context behind the work, explain the problematic of the work, and give a summary of the main scientific results and impact.

Each chapter will be structured as follows :

- Context and Motivations
- Related Works
- Approach and Scientific Results
- Future Work

I will conclude the document with Chapter 5 to present my future research perspectives. Three appendices accompany this document : Appendix A where I list my Ph.D. students, my *curriculum vitae* in Appendix B, and my personal bibliography in Appendix C

Note that all my research has always been conducted through *collaborations*, therefore in the rest of the document I will use "we" when describing it.

Chapter 2

The *Trusted Cells* Privacy-Centric Data Management Paradigm

"This is a story about oil and data, two resources basically useless in their raw state, but that can be very valuable when refined."

– *Paul Lambert, Point B. Consulting CEO*

Summary: *How can you keep a secret about your personal life in an age where your daughter's glasses record and share everything she sees, your wallet records and shares your financial transactions, and your set-top box records and shares your family's energy consumption? Your* personal *data has become a prime asset for many companies on the Web, but can you avoid – or even detect – abusive use? Today, there is a wide consensus that individuals should have increased control on how their personal data is collected, managed and shared. Yet there is no appropriate technical solution to implement such personal data services: centralized solutions sacrifice security for innovative applications, while decentralized solutions sacrifice innovative applications for security. In this chapter, we argue that the advent of secure hardware in all personal IT devices, on the edges of the Internet, could trigger a sea change. We propose the* Trusted Cells *paradigm: personal data servers running on secure smart phones, set-top boxes, secure portable tokens or smart cards to form a global, decentralized data platform that both provides security and encourages innovative applications. We motivate our approach, describe the* Trusted Cells *architecture and define a range of challenges linked to the approach, some of which are detailed in later chapters of this document.*

Contents

This chapter overviews the scientific results presented in the following papers :

Scientific Contributions :

- [12] which presents our introductory global vision of what a Secure Personal Data Server should be, and introduces many different research challenges, based on the use of Secure and Portable Tokens to manage one's private data.
- [24] which generalizes this vision to many different kinds of devices with different levels of security (and therefore of trust).
- [27] a tutorial on the architectures and techniques to securely manage one's personal data.

This work was led in collaboration with Indrajit and Indrakshi Ray, from Colorado State University (US), and Philippe Bonnet from Technical University of Copenhagen (Danemark). It led to the funding of the ANR KISS project in 2011 and in Philippe Bonnet's Marie Curie grant in 2013/2014.

2.1 Context and Motivations

The New Oil

With the convergence of mobile communication, sensors and online social networks technologies, we are witnessing an exponential increase in the creation and consumption of personal data : in 2013, the average *homo-internetus* receives 112 emails per day, appears on 800 social networking web pages, and interacts on a daily basis with numerous search engines, online markets, administrations, etc. which are all tracing him : Paper-based interactions (e.g., banking, health), analog processes (e.g., photography, resource metering) or mechanical interactions (e.g., as simple as opening a door) are now sources of digital data that can be linked to one or several individuals (see Figure 2.1), and stored on central servers. This personal data is recognized by the World Economic Forum as a most valuable resource, since they call it "the new oil" [143], creating an unprecedented potential for applications and business. Indeed, the comparison is striking : $2 billion a year are spent by US companies on third-party data about individuals, with an estimated return around $30 for $1 invested [76], where oil yields a maximum return of $0.5 per year and per dollar, for globally equivalent benefits. This personal data provides value to the companies managing it, e.g., Facebook, is valued at approximately $50 per account. The value of this data is fundamental for many different types of companies, e.g. well known companies such as Google and Amazon, but also banks and insurrances, employment market companies, travel and transportation, "love" market, etc.

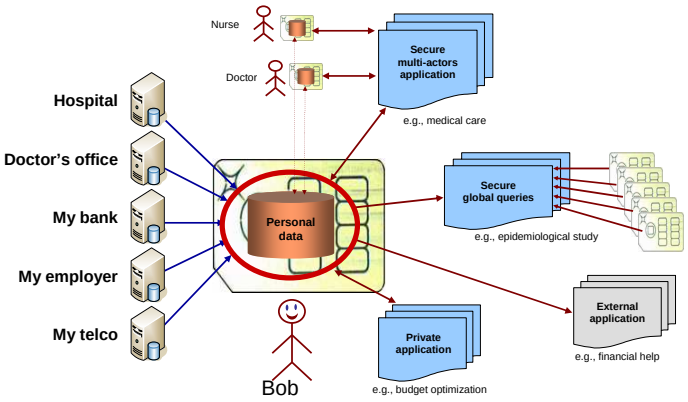

Figure 2.1: The Trusted Cells approach

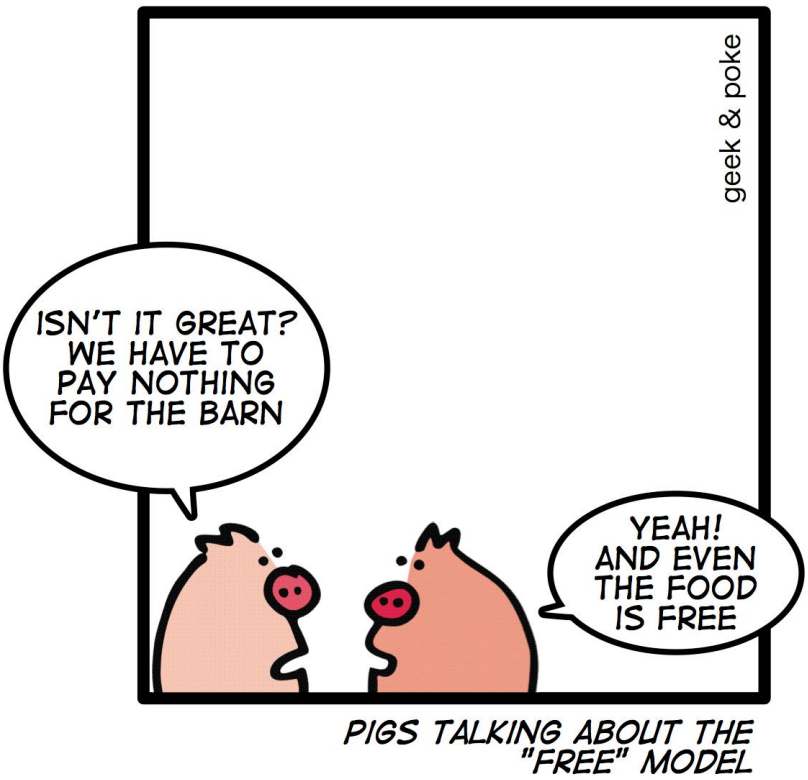

Figure 2.2: The "Free Model" by Geek and Poke (reprinted with permission)

One question is seldom asked : in the case of new oil, who will reap the benefits of its exploitation, the data owner, or the data exploitation company ? Indeed, how are these *new oil* producers behaving, given that each individual is sitting on a new-oil field ? They are offering to exploit the oil field for free. They are offering free services to the individual, that cost them only a few cents (e.g. hardware and software email management). They provide services which are neither targetted at, nor useful to, and not advertized, to the individual, but that yield healthy returns to them (advertising, profiling, location tracking, etc.).

Until now, enthusiasm for these new opportunities has thwarted privacy concerns. Individuals conscientiously build Facebook pages, conduct their communications via Gmail, send and receive megabytes of personal information to and from administrations or commercial services. The PRISM[23] affair is unveiling a situation imagined in the worst dystopias of science fiction literature. Current practices are often not compliant with basic privacy laws and directives. Data leaks are legion[24]. Justified by business interests, governmental pressure or simple inquisitiveness among people, some underlying business models are even based on breaches of users'

[23]Not to be confused with the eponymous French research laboratory...

[24]Many sites and reports exist. An interesting site is http://www.privacyrights.org/data-breach which lists all the data breaches made public since 2005.

privacy, such as the award winning intellius.com company. In general[25], privacy policies of dominant actors are invalid with regards to EU and US standards, in particular they are too vague about purposes for which personal data is collected. Basic security is not guaranteed[26]. Personal data is collected, transferred and used without user's consent, and even personal data of *non-users* is collected[27]. Policies change frequently without requesting users' consent. Openness (i.e. the right to correct false data) although advertized, is not provided in practice[28]. Data retention limits are not applied[29]. In consequence, anyone can exploit weak privacy policies or cross-analyze sensed data with data conscientiously registered on social networks.

Is privacy really required ?

Given the success of privacy intrusive applications, it may seem, to quote Mark Zuckerberg, that "*Privacy is no longer the norm*". Indeed, it look as if today's teenagers are opening up their life to the public by putting it online. However, a finer social analysis [56] indicates that teenagers are in fact building their own private sphere, hidden from the household sphere, which is private for their parents, but not for them. Moreover, there are many cases which advocate in favor of privacy :

- Vulnerable citizens remain under threat : A study [96] conducted on a large sample of 33 european children (approx. 1000 per country) shows that many of them received contact from strangers via online profiles (approx. 25%) and are highly exposed, since 10% of them have met strangers face to face. The current practice "Accept the policy or quit" is not the good option.
- Blatant failures of emblematic applications due to privacy concerns: National EHRs failed in many countries (e.g. The Netherlands [102]) because doctors feel spied and patients fear being discriminated. Prejudice is both economic and social.
- A new digital divide: applications whishing to follow UN charters, such as NGOs, Healthcare companies, etc., must build their applications on infrastructures complying with worldwide privacy laws (See [139]).
- An increasing amount of privacy complaints (+30% in France in 2011): more citizens feel that their privacy online is not sufficiently protected (18/24 years have become the dominant category with 78%) [69]

Even if we look beyond economic and pricing questions, many laudable projects cannot base themselves on weak privacy foundations. Should humanitarian aid workers build their applications if they cannot provide privacy guarantees? Imagine an application whose goal

[25] Many of these examples are drawn from reports from the Article 29 Data Protection Working Party, see http://ec.europa.eu/justice/data-protection/article-29/

[26] To quote the Facebook Terms of Service : "We will do our best to keep Facebook safe, but we cannot guarantee it."

[27] See : http://usatoday30.usatoday.com/tech/news/story/2011-11-15/facebook-privacy-tracking-data/51225112/1

[28] See "EU versus Facebook" affair, http://europe-v-facebook.org/

[29] See "CNIL vs Google" affair.

is to help the homeless, by providing medical monitoring. Managing critical information on potentially discriminated people under weak privacy guarantees could be seen as too strong a danger to create this application. Many other applications, undoubtedly useful, but placing respect for human dignity and privacy upfront, are thus sometimes left by the wayside.

A *sea change* for personal data services

Let us stress that we advocate neither for nor against any type of application, be it the *most trivial* Facebook app, or a *hand-on-heart* low-cost electronic health record for the homeless. We do claim however that the current "centralized data" model (both in terms of applications and servers) is the reason for two major intrinsic problems :

1. personal data is exposed to sophisticated attacks[30] and
2. personal data is hostage of sudden privacy changes, since centralised administration of data means delegation of control.

The nature of the solution is consensual: it is necessary to increase the control that individuals have over their personal data [133, 130, 142]. The World Economic Forum even claims that "*increasing the control that individuals have over the manner in which their personal data is collected, managed and shared will spur a host of new services and applications*". We argue that the advent of secure hardware embedded in all forms of personal devices, on the edges of the Internet, will cause a sea change to the management of personal data by solving the two intrinsic problems of the centralized model : *drastically reducing the benefits/cost ratio* of an attack and effectively returning *complete control* of users on their data, *anywhere and anytime*.

State-of-the-art secure hardware is currently on the market : AMD has recently announced that it will incorporate a secure Trust Zone-based ARM processor[31] on all its chips to be included into smart phones, set-top boxes and laptops. Such secure tamper-resistant micro-controllers provide tangible security guarantees in the context of well-known environments[32] for only a couple of dollars. We can now imagine that whenever one takes a picture, one's smart phone will securely contact the personal services of all individuals in the frame of the picture, and automatically blur the face of those who request it. We can also imagine that one's car's GPS tracker will give detailed turn-by-turn guidance, but hides those details to insurance companies, only delivering overall pay-per-drive pricing results.

Motivation

We now present a *not-so-futuristic* example, to introduce the *trusted cell* paradigm, i.e. personal data servers running on secure devices and interacting with non-trusted supporting servers to form a decentralized privacy centric data management platform.

[30] An attack is always characterized by its benefits/cost ratio. In the case of a central server, the benefits will be very high, so it is feasible to spend a proportional amount of resources to perform the attack.

[31] http://www.arm.com/products/processors/technologies/trustzone.php

[32] The adoption of a standard API for secure micro-controllers [90] and the availability of an open source embedded secure operating system based on it (Open Virtualization) now enable higher level services.

Example : Alice lives in France with Bob and their two children. Their house is now one of the 35 million households equipped with a Linky[33] power meter. The power meter reports once a day to the distribution company, a certified time series of readings for verification, billing and network operation [105]. Alice and Bob have installed an energy butler app on their secure home gateway, a *trusted cell* managing all smart appliances in their home and storing their data. That award-winning app relies on external feeds from their utility and local weather prediction, as well as a feed of readings received every second from the Linky, to control their heat pump and the charge of their electrical vehicle. This app minimizes overall load on the distribution network and saves 30% on their bill. In addition, Alice is engaged in a social game (a follow-up to `simpleEnergy.com`) where she competes with some friends on their energy savings, reducing consumption by an additional 20%.

At the 1Hz granularity provided by the Linky, most electrical appliances have a distinctive energy signature. It is thus possible to infer from the power meter data which activities Alice and Bob are involved in at specific points in time [110]. How do Alice and Bob configure the home gateway trusted cell to preserve privacy while preserving the benefit of their applications? They have a shared account on this trusted cell. Bob, Alice and their children have agreed that they do not want to fully disclose all their activities to each other : they would rather have access to 15 min aggregates via a visualization app – indeed at that granularity one cannot detect specific activities, but it is still possible to infer a daily routine. At the same time, daily statistics feed their social game, monthly statistics are delivered to the distribution company and time series at required granularity are securely exchanged with other *Trusted Cells* in their neighborhood to achieve consumption peak load sharing.

None of this data leaves the *Trusted Cell* application in clear unless it is accessed via a predefined set of aggregate queries. The *Trusted Cell* guarantees that no *malware* can tamper with the data. If the *Trusted Cell* gets stolen, an elaborate attack would need to be mounted to breach the secure hardware and gain access to their personal data.

This scenario can be easily transposed to different types of personal data like GPS traces, Internet traces, mobile phone data, bills, pay slips, photos as well as health, administrative or school records. We classify the data that could be managed with *Trusted Cells*, based on how and who actually produces it:

- Data produced by smart sensors installed by companies in the user's home (e.g., power-meter, heat sensor) or in the user's environment (e.g., user's car GPS tracking box for a Pay As You Drive application) on which the user has full or shared ownership, accepting to externalize aggregated data. Users may opt-in for small-scale sharing (e.g., local traffic optimization) or larger-scale sharing (e.g., social games or traffic optimization).

- Data produced or inferred by external systems (e.g., purchase receipts obtained by near field communication or medical data sent by the hospital or labs). Small-scale sharing allows the user to optimize her buying habits or to compare her medical treatment with people having the same disease. Larger-scale sharing brings public health insights (e.g., epidemiological study cross-analyzing diseases and nutrition).

[33] See : `http://www.erdfdistribution.fr/EN_Linky`

- Data authored by the user herself (e.g., a photo, a mail, a document) on which she has complete ownership. Small-scale sharing benefit is obvious here. Larger-scale sharing of partial data (e.g., photo location only, number of exchanged mails) is undoubtedly a source of precious information (e.g., most interesting places on Google maps).

What personal data services actually run on a *Trusted Cell*? How do these services allow a user to control whom she shares her secrets with? How do applications access these services? What kind of guarantees do *Trusted Cells* offer about the security of the data they manage?

There are many questions, some still open, around these novel applications and this novel architecture. Our goal in this Chapter is modestly to draw the outlines of an architecture based on *Trusted Cells* interconnected via an *untrusted infrastructure*, that we call the Asymmetric Architecture.

Thus, we start by reviewing existing decentralized architecture solutions in Section 2.2, then describe the envisioned system, the *Trusted Cell* in Section 2.3. We conclude in Section 2.4 with an overview of challenges. In the following chapters, we will discuss how to provide distributed computation of aggregations on this architecture (Chapter 3), and we will investigate what can be done when data exits the *Trusted Cell* world, by studying the enforcement of the *Limited Data Collection* principle (Chapter 4).

2.2 Related Works

Centralized Solutions

Centralized solutions, including emerging cloud-based personal data vault management platforms, trade security and protection for innovative services. Many commercial solutions exist[34]. At best, these approaches formulate sound privacy policies, but none of them propose mechanisms to automatically enforce them [8]. Even TrustedDB [38], which proposes tamper-resistant hardware to secure outsourced centralized databases, does not solve the two intrinsic problems of centralized approaches. First, users are hostages of sudden changes in privacy policies, their data can also be unexpectedly exposed by negligence or because it is regulated by too weak policies. Second, users are exposed to sophisticated attacks, whose cost-benefit is high on a centralized database.

Decentralized Architectures

Decentralized solutions are promising because they do not exhibit these intrinsic limitations. We refer to [124] for a survey on decentralized architectures whose goal is to promote privacy. However, [124] are critical of the ambitions of these architectures, in particular they are suspicious of hardware backdoors, and invoke the problem of "downstream abuse" [140] which means the user cannot control what is done with her data once it leaves the sphere of trust. On the

[34]For instance the french Digiposte, https://www.digiposte.fr/ but it is beyond the scope of this document to draw a list of such companies.

contrary, we argue in this chapter in favor of usage control, and we claim that downstream abuse must be tackled via *Limited Data Collection* (See Chapter 4), therefore we do not agree with the negative criticism of [124].

Hardware for Decentralized Architectures

Open Plug Computers

Low-cost plug computers and open software are provided to users to enable anonymous and independent communication networks. FreedomBox [84] is a good representative of this movement. It provides low-cost hardware, running customizable open software, including privacy-centric applications such as running TOR [73] by default. We are seeing a general interest from the public for open plug computers, such as Raspberry Pi [136], or cards such as Arduino [36], which can be used as building blocks for household applications. However, these systems do not natively take security into account.

The Secure Portable Token

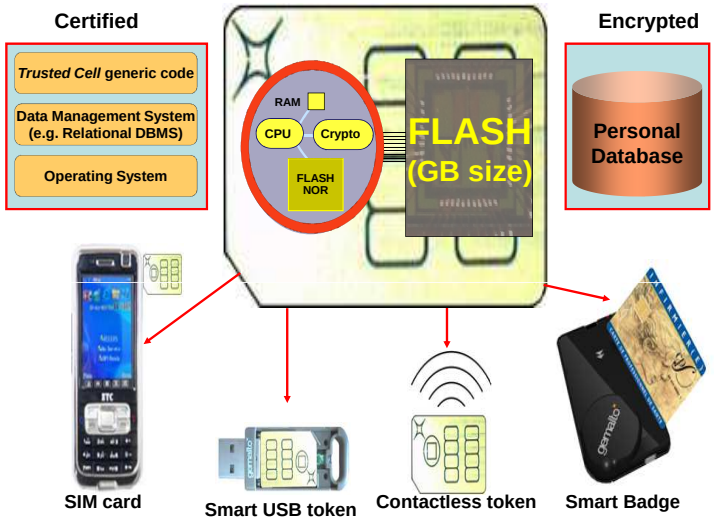

Figure 2.3: A Trusted Cell using a Secure Portable Token

The global architecture for managing personal data revolves around a key element, the *Trusted Cell*. Many different systems can be used to implement a Trusted Cell([12, 80, 89]). Shown in Figure 2.3, our current hardware[35] uses a *Secure Portable Token* (SPT) as secure

[35] We use two kinds of hardware : a proprietary token provided by Gemalto, world leader in Smartcards, and our own custom made token developed using STM components.

element. SPTs appear today in a wide variety of form factors ranging from SIM cards to various forms of pluggable secure tokens. Whatever the form factor, SPTs share several hardware commonalities. Their microcontroller is typically equipped with a 32 bit RISC processor (clocked at about 50-120 MHz today), memory modules composed of ROM, static RAM (about 64KB), a small internal stable storage (about 1MB of NOR Flash) and security modules providing tamper-resistance. The microcontroller is connected by a bus to a large external mass storage (Gigabytes of NAND Flash). However, this mass storage does not benefit from the microcontroller tamper resistance, and therefore any data stored there must be cyphered.

Hardware progresses are fairly slow in the secure chip domain because the size of the market (billions of units), and the requirement for a high tamper-resistance leads to adopt cheap and proven technologies [80]. Nonetheless, SPT manufacturers forecast a regular increase of the CPU power, stable storage capacity and the support of high communication throughputs (up to 480 Mb/s). The RAM, unfortunately, will remain a scarce resource for the foreseeable future owing to its poor density : Indeed, the smaller the silicon die, the more difficult it is to snoop or tamper it during processing, and RAM competes with CPU, ROM and NOR in the same silicon die.

In summary, a SPT can be seen as a low power but very cheap (a few dollars), highly portable, highly secure computer with reasonable storage capacity for a personal usage.

Introducing the problematic

We advocate the use of decentralized solutions, since they do not exhibit the intrinsic limitations of centralized solutions (privacy, security, etc. . .). They raise important, and interesting challenges : (1) economic, in order to build viable business models compatible with privacy, which we will not discuss here and, (2) technical, in order to design a secure personal data server managing secure storage of personal data (*local requirements*) and provide the same level of functionality, responsiveness and availability as a centralized solution (*global requirements*).Most of my work has focused on studying global requirements, and on devising the global architecture, that we will introduce next.

2.3 Approach and Scientific Results

We introduce and describe the Asymmetric Architecture, composed of *Trusted Cells* and a *Supporting Server Infrastructure*. We detail the threat model that must be taken into account when studying protocols running on this architecture. We then present the four major research directions that the *Trusted Cells* approach opens.

2.3.1 The Asymmetric Architecture

The *Trusted Cell*

A *Trusted Cell* implements a client-side reference monitor [132] on top of secure hardware. At the very least, the hardware must guarantee a clear separation between secure and non-

secure software. We abstract a Trusted Cell as (1) a Trusted Execution Environment, (2) a tamper-resistant memory where cryptographic secrets are stored, (3) an optional and potentially untrusted mass storage and (4) communication facilities. Physically, a trusted cell can either be a stand-alone hardware device (e.g., a smart token) or be embedded in an existing device (e.g., a smartphone based on ARM's TrustZone architecture). The very high security provided by trusted cells comes from a combination of factors: (1) the obligation to physically be in contact with the device to attack it, (2) the tamper-resistance of (part of) its processing and storage units making hardware and side-channel attacks highly difficult, (3) the certification of the hardware and software platform, or the openness of the code, making software attacks (e.g., Trojan) also highly difficult, (4) the capacity to be auto-administered, contrary to high-end multi-user servers, avoiding insider (i.e., DBA) attacks, and (5) the impossibility even for the trusted cell owner to directly access the data stored locally or spy the local computing (she must authenticate and only gets data according to her privileges). In terms of functionality, a full-fledged trusted cell should be able to (1) acquire data and synchronize it with the user's digital space, (2) extract metadata, index it and provide query facilities on it, (3) cryptographically protect data against confidentiality and integrity attacks, (4) enforce access and usage control rules, (5) make all access and usage actions accountable, (6) participate in computations distributed among trusted cells. Basic (e.g., sensor-based) trusted cells may implement a subset of this. Also note that a *Trusted Cell* can be in many cases highly disconnected. This will obviously be the case when using secure portable tokens, and with many other kinds of mobile trusted cells providing no connection guarantees.

The Supporting Server Infrastructure

The Supporting Server Infrastructure (SSI) provides the storage, computing and communication services, which expand the resources of a single trusted cell and form the glue between trusted cells. By definition, the SSI does not benefit from the hardware security of the trusted cell and is therefore considered *untrusted* (See Section 3.2 for more details on why we do not believe that central server secure elements provide an adequate solution to trust either). We consider that the SSI is implemented by a Cloud-based service provider[36]. In terms of functionality, the SSI is assumed to: (1) ensure a highly available and resilient (in the database sense) store for all data outsourced by trusted cells, (2) provide communication facilities among cells and (3) participate in distributed computations (e.g., store intermediate results), provided this participation can be guaranteed harmless by security checks implemented at the trusted cells side.

Since this architecture is composed on the one hand of a very large number of low power, weakly connected but highly trusted devices and on the other hand of powerful external computing and communication resources provided by untrusted recipients, we call it an Asymmetric Architecture.

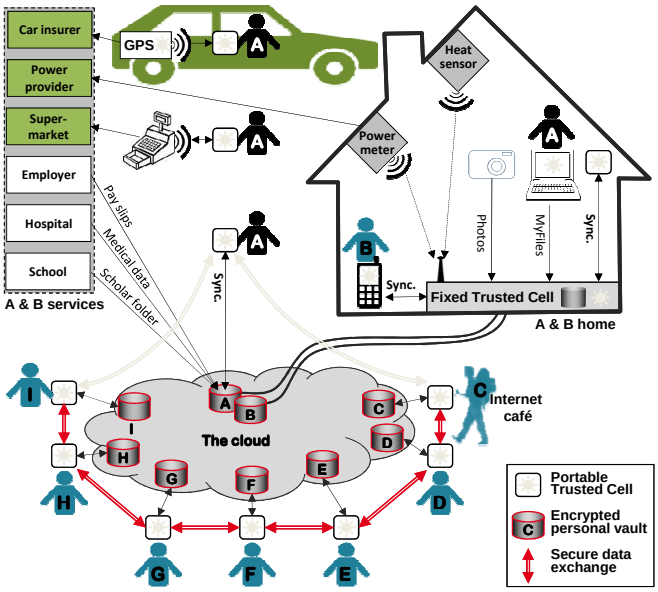

Figure 2.4: The Trusted Cells Architecture

Example

Figure 2.4 illustrates how trusted cells and the SSI can collaborate to implement scenarios meeting the privacy requirements stated above:

Alice (A) and Bob (B) are equipped with fixed and portable trusted cells, acquiring data from several data sources, synchronizing with their encrypted personal digital space on the cloud. Charlie (C) is travelling around the world and can securely access all his data from any (unsecure) terminal thanks to his portable trusted cell. All users equipped with trusted cells can securely share their encrypted data through the cloud.

A threat model for the Asymmetric Architecture

In the Asymmetric architecture, the main source of vulnerability comes from the SSI (they can be dishonest or be themselves attacked). Each trusted cell, composed of secure hardware, is presumably of high trust. However, even secure hardware can be breached, though at very high cost. Therefore we take into account the possibility that a very small number of participating cells might be compromised. We thus consider three attack models of increasing

[36] A P2P approach to implement the SSI, where each peer would itself be a *Trusted Cells* would be interesting but raises other challenges and is left for future work.

strength. *Honest-but-curious* is a classical attacker, and we detail two kinds of *Malicious* attackers. In all cases, the attacker is the SSI. Example supporting servers would be "privacy-friendly" data hosting companies, public administrations, supposedly respecting local privacy legislation, or medical laboratories, whose "business model" or research objective involves using studies based on anonymous data. We therefore consider that an attacking SSI does not want to be discovered[37], since this would draw very negative publicity, and that techniques that detect such attacks with an arbitrarily high probability will suffice to provide protection via deterrence, be it legal action, or simply because nobody will participate in any subsequent applications conducted by a convicted company or administration. We use the term "protocol" to denote a generic interaction between a *Trusted Cell* and the SSI.

- *Honest-but-curious*: the attacker does not deviate from the protocol it is participating in but tries to infer confidential data (in principle hidden by the protocol) by exploiting in any computationally feasible way the results of each step of the protocol. In this model, the attacker (also called *semi-honest* or *passive*) is *only* the SSI.
- *Malicious*: in this model, the attacker is still the SSI, therefore we do not consider denial of service attacks : the attacker will cheat the protocol with the sole objective to disclose confidential data. This model assumes that the recipient can now conduct passive *and* active attacks (i.e., modify the results of some steps of the protocol). Avoiding active attacks is impossible since the SSI plays a role in the protocol but being able to detect active attacks is considered as the best way to deter them[38]. Hence, the recipient can be said a *malicious* adversary having *weakly-malicious* intent [160] : it will perform active attacks only if (1) it will not be convicted as adversary by the trusted parties and (2) the final result is correct. Note that the secure devices form the trusted parties.
- $Malicious_{Hard}$: the attacking SSI is *malicious* and is able to break the tamper-resistance of *at least one* secure device. In this model, the attacker is the SSI colluding with the *compromised cell*.

We now detail four major research directions and corresponding challenges, some of which remain future work, for the user to actually control how the data entering her personal digital space is collected, protected, shared and *in fine* used.

2.3.2 Secure private store

All data must be made highly available, resilient to failure and protected against confidentiality and integrity attacks. Accessing this data from any terminal, including those outside the user's ownership sphere (e.g., internet café), should leave no access trace.

Cryptographic techniques (i.e., encryption, hashing, signatures) are used to protect trusted cell's data, keeping cryptographic keys in a tamper-resistant memory. The data is then stored in the cloud and potentially cached in the trusted cell local mass storage. At a minimum, trusted

[37] If it is itself the victim of an attack, it will be happy to detect it.

[38] The detection of an attack puts the attacker in an awkward position. If the data leak is revealed in a public place, participants are likely to refuse to participate in further studies with an irreversible political/financial damage and they can even engage a class action.

cells keep locally extended metadata: access information, indexes, keywords, and cryptographic keys. Metadata should be sufficient to allow performing queries before accessing the Cloud to retrieve the data of interest. Cryptographic keys never leave the trusted cells tamper-resistant memory. Hence a trusted cell can be used to get securely data from any (untrusted) terminal it is connected with.

A significant amount of data and metadata is likely to be embedded in some trusted cells and may need to be queried efficiently. While it does not seem a major issue in powerful trusted cells (e.g., a smart phone), it appears much more challenging when facing low-end hardware devices like secure tokens (e.g., a microcontroller with tiny RAM, connected to NAND Flash chips or SD cards, possibly with energy consumption constraints). Whatever their complexity, trusted cells should also be designed to support self-tuning, self-diagnosis and self-healing to minimize the management burden put on the trusted cell owner. This research direction is pursued by *other* members of my research group [25].

2.3.3 Global computations

Privacy also has a collective dimension in the sense that preserving one's privacy should not hinder societal benefits (e.g., census, epidemiologic releases, global queries). A trusted cell user is thus expected to participate in global computations assuming her data suffers appropriate transformations (e.g, anonymization, output perturbation) depending on the trustworthiness of the recipient(s) and the expected usage of the data/query. When data needs to be transformed before being delivered, the recipient trusted cell implements the transformation on its own if possible (e.g., filtering, local data perturbation) or in collaboration with other trusted cells if the transformation requires a collective action (e.g., anonymization, global data perturbation). In the latter case, the computation may be implemented in a pure Secure Multi-Party fashion or may require the participation of the untrusted infrastructure (e.g., to store intermediate results).

Such large scale computations may lead to atypical distributed protocols combining security and performance requirements in an asymmetric context made on one side of a very large number of highly secure, low power and weakly available trusted cells and on the other side of a highly powerful, highly available but untrusted infrastructure. We explore this challenge in Chapter 3.

2.3.4 Secure sharing, secure usage and accountability

The user can decide to keep her data private or share it with other users or group of users under certain conditions (e.g., time, location). Works exist in the field of social networks to provide access control mechanisms, such as D-FOAF [109, 65] and in particular [64] which use ontologies to represent complex relationships in between users, and enforce access control on this basis. However, we insist that the user must get a proof of legitimacy for the credentials exposed by the participants of a data exchange and must trust the evaluation of the exchange conditions (if any).

Practically, sharing data means sharing the associated metadata (so that the recipient user can get the referenced data in the Cloud), the cryptographic keys (so that her trusted cell can decrypt them) and the sticky policy (so that her trusted cell can enforce the expected access control rules). Hence, thanks to its security properties, including the protection against illegitimate actions of the recipient user, the recipient trusted cell can enforce all the conditions appearing in the access control rules (user's credential, contextual conditions).

Beyond sharing data, usage control usually refers to UCON$_{ABC}$ [132]: obligations (actions a subject must take before or while it holds a right), conditions (environmental or system-oriented decision factors), and mutability (decisions based on previous usage)[39]. Similarly to access control rules, usage control rules can be implemented as sticky policies so that they are made cryptographically inseparable from the data to be protected. Hence usage control rules will be enforced by any trusted cell downloading data and cannot be bypassed by the recipient user. Regarding accountability, the recipient trusted cell can maintain an audit log, encrypt it and push it on the Cloud to the destination of the originator trusted cell.

On the other hand, a *Trusted Cell* will be led to interact with services that are *outside* the *Trusted Cell* world. Therefore, we must also consider the case when data will exit the *Trusted Cell* environment. In this case, the *Limited Data Collection* principle must be applied, which means that the *Trusted Cell* must be able to compute the minimum amount of information that it will send to the external environment. We explore the enforcement of *Limited Data Collection* in Chapter 4.

2.3.5 Controlled collection of sensed data

The targeted user(s) should be the unique recipient(s) of raw sensed data and would accept externalizing only aggregates by opting in/out for selected applications/services : at home, the power meter continuously pushes raw measurements to Alice's and Bob's trusted cell gateway, while a certified aggregated time series is sent to the power supplier company and aggregates for a social game are pushed to the Cloud every day. Similarly, the tracking box installed on Alice's car is a trusted cell delivering aggregated GPS data to her insurer and raw data to her trusted cell smartphone that she will synchronize with her personal space for further use when back home. Hence, adding a trusted cell to a sensor, allows defining e.g., the frequency and or precision of the data that should be externalized, thus leading to a trusted source both for the user (in terms of privacy preservation) and the provider (in terms of certification of the output data). This research direction remains future work.

2.4 Future Work

The *Trusted Cells* paradigm provides a new architecture to reconcile individual's privacy with innovative acquisition and sharing of personal data. This paradigm is based on the current trend in the development of ubiquitous and open secure hardware. This paradigm undoubtedly opens

[39]For instance, a photo could be accessed ten times (mutability), during the year 2013 (condition), informing the owner of the precise access date (obligation).

a set of exciting challenges that must be explored by the research community, and provides the setting for the development of countless novel *privacy by design* applications. We now sketch three different research directions, based on the challenges introduced in the previous chapter.

Secure sharing

The trusted cells themselves may be a source of simplification for the secure sharing problem : integration of biometric sensors to automatically authenticate users, automatic production of certified credentials safely computed on a trusted cell, definition of default policies by trusted third parties – e.g., citizen associations – which could be automatically selected depending on a computed individual's profile. Trust management (and automatic negociation) is therefore in the heart of the trusted cell paradigm. Taking into account the fact that no hardware is 100% trustworthy is also crucial, and leads to a quantification of trust (or of risk) that must be factored in the analysis.

Secret management is at the heart of any sharing protocol between trusted cells (i.e., at this level a secret is a cryptographic key) and must be carefully designed (e.g., class-breaking attacks must be prevented, master secrets must be restorable in case of crash/loss of a trusted cell). The study of a trustworthy infrastructure to manage such large a number of keys, belonging to many different cells is also a difficult challenge.

Secure Usage

Many challenges are common with secure sharing. However, trusted cells hold the promise of new usages and new usage controls. Designing efficient such protocols, proving that they are secure and do not leak any private information provides the grounds for interesting future work. An example of such a usage would be : the fact that the trusted cell of an individual A would check that the personal data it produced referencing an individual B should be submitted for approbation to B's trusted cell before being integrated in A's digital space.

Controlled Collection of Sensed Data

Co-design is a primary issue to allow the definition of affordable sensor-based trusted cells, i.e. taking into account hardware and application constraints during design phase. Low-cost is indeed a prerequisite to the generalization of trusted sources, capable of securely filtering and aggregating stream-based spatio-temporal data with tiny hardware resources. Some trusted sources being weakly connected to the Internet; asynchrony problems must also be addressed. Finally, the combination of data streams from multiple sources, each being separately harmless, may generate new privacy risks that must be carefully tackled.

Chapter 3

Global Computation on the Asymmetric Architecture : $\text{Met}_{\mathbb{A}}\text{P}$

Helena : "Love all, trust a few, do harm to none. (...)
Be checked for silence, But never taxed for speech."

– *William Shakespeare*, All's well that ends well

Summary: *In this chapter, we show how to manage distributed computations on the* Asymmetric Architecture *by deploying a classical privacy preserving application using secure portable tokens as secure element of a trusted cell, and achieving the same result as in a centralized context. The application chosen is the well known Privacy-Preserving Data Publishing (PPDP) problem. The goal of PPDP is to generate a sanitized (i.e. harmless) view of sensitive personal data (e.g. a health survey), to be released to some agencies or to the public. However, traditional PPDP practices all make the assumption that the process is run on a trusted central server. We propose* $\text{Met}_{\mathbb{A}}\text{P}$*, a generic fully distributed protocol, to execute various forms of PPDP algorithms on an* Asymmetric Architecture*. We show that this protocol is both correct and secure against honest-but-curious or malicious adversaries, and have conducted an experimental validation showing that this protocol can support PPDP processes scaling up to nation-wide surveys, validating the use of global computations with* Trusted Cells*.*

Contents

This chapter overviews the scientific results presented in the following papers :

Contributions :

- [18, 20] our introductory papers which showed how to adapt k-anonymity family of PPDP algorithms to a secure portable hardware token context and proved the protection against honest-but-curious and weakly malicious adversaries.
- [19] provided proofs of probabilistic detection in the case of malicious adversaries having cracked one or more tokens.
- [21] which describes and proves the correctness of the generic MetAP meta-protocol, which is a meta-algorithm used to implement any kind of PPDP algorithm, including differential-privacy based algorithms, running on the Asymmetric architecture.

Many of these results were obtained during the course of Tristan Allard's Ph.D. thesis [11] that I co-supervised with Pr. Philippe Pucheral, funded by a grant from the french Ministry of Research and Higher Education. Part of this work was funded by the ANR DEMOTIS project (2010-2012).

Prototype : The `Trusted Cell Global Queries` system (a RDBMS running on the Asymmetric Architecture) is currently being developed in the course of Cuong Quoc To's Ph.D. thesis, based on initial developments by Tristan Allard. A first prototype has been demonstrated at [147].

3.1 Context and Motivation

In this Chapter, we show how it is possible to achieve the same results on the Asymmetric Architecture as those obtained using a *trusted* centralized computing architecture, even for global computations, i.e. processes that need to access the local data of many different cells to compute aggregate values. A typical large scale application managing private data with this objective is Privacy-Preserving Data Publishing (PPDP), which we will discuss next. Another application would be the computation of `GROUP BY` SQL queries on the Asymmetric Architecture, which is ongoing work [146].

Current PPDP Practices

In a traditional PPDP process (Figure 3.1), *personal data* (e.g., patients' health data) is collected by a *publisher* who sanitizes it and releases it to various *data recipients* (e.g., research teams, public agencies, drug companies). Different sanitized releases are usually built by the publisher to best match (1) the data utility requirement expressed by each recipient and (2) the privacy requirement related both to the purpose of this release and to the trust level attached to this recipient. Although a lot of work has been done on privacy models trying to reach the best utility/privacy trade-off of the sanitized dataset (see [66, 86] for two surveys), much less effort has been spent on the practical implementation and deployment of a PPDP process. This raises two major issues :

First, most PPDP works rely on the assumption of a trustworthy central publisher [66]. As we have seen in Chapter 2, this trust assumption is unfortunately rarely satisfied in practice. Second, the legislation regulating the management of personal data (e.g. [79]), once sanitized, is very permissive (no retention limit, no encryption of data at rest, no user consent), since it considers that this data is no longer sensitive. Some publishers exploit this flaw to implement highly practical - but somehow deviant - PPDP scenarios. For example, French and UK Electronic Health Record (EHR) providers, build sanitized releases as follows: (1) nominative data is extracted from an OLTP EHR server, (2) the data is simply pseudonymized (a legal form of sanitization!) and stored in a warehouse, and (3) different sanitized datasets are produced using this warehouse for different recipients on demand. Letting the publisher host weakly sanitized data with weak security obligations is a major concern, at least as important as the privacy guarantees provided by the final release.

At a time when the scientific community mobilizes a lot of energy in designing privacy models to return to individuals a better control over their personal data, we advocate a fully decentralized PPDP process, based on the use of Secure devices (i.e. a trusted cell) where individuals stay in the center of the loop (Figure 3.2). For the individual, this means (1) having the ability to opt-in/out of her participation to a given release according to its purpose and her confidence in the recipient and (2) keeping control over the process producing this release.

Requirements to solve the distributed PPDP problem

The challenge addressed in this chapter is to devise a PPDP protocol compatible with the fully distributed setting provided by secure devices on the participant's side of the protocol. The

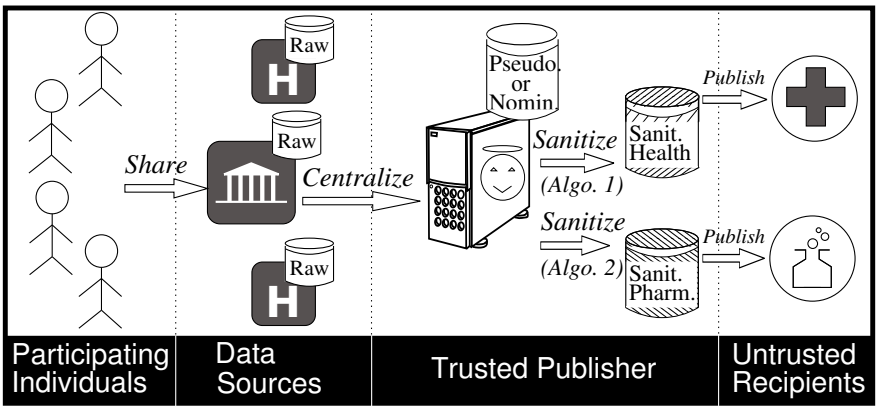

Figure 3.1: Current PPDP Practices

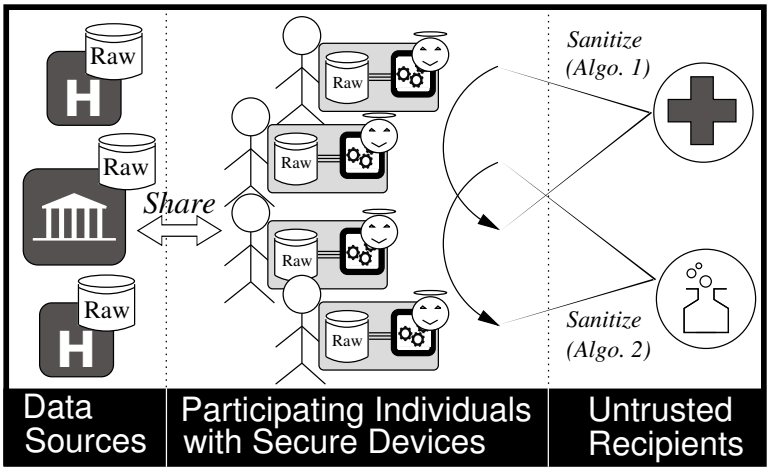

Figure 3.2: Proposed Approach

objective is to get rid of the salient point of vulnerability introduced by a central publisher while still providing equivalent PPDP services. The implication is threefold:

- *Quality of the release:* implementing a distributed PPDP protocol[40] based on *local perturbation* [9] would be rather trivial (i.e., participants perturb their own data *independently*). However, a much better utility/privacy trade-off is reached by *global publishing* algorithms, also called *centralized publishing* algorithms [137], which exploit the prior knowledge of the complete dataset to be sanitized (e.g., data distribution) in order to perform a smarter sanitization. While fully decentralized, our protocol must preserve the quality reached by global publishing.
- *Genericity:* different sanitized releases must be delivered to different recipients depending on the utility/privacy trade-off required by each one. The proposed protocol must

[40]In order to avoid ambiguity we use the term *protocol* to denote the distributed implementation of an algorithm.

therefore be agnostic with respect to the targeted PPDP algorithms and their associated sanitization models (e.g., k-Anonymity, ℓ-Diversity, Differential Privacy).
- *Scalability:* to be usable, the approach must scale up to nationwide sized datasets (millions of participants, each using a secure device).

As Figure 3.2 suggests, the required distributed PPDP protocol looks like a Secure Multi-party Computation (SMC) protocol. Participants (i.e., trusted cells) must jointly compute a function (i.e., the sanitization process) without revealing their input (i.e., the personal data they host), such that only the final result (i.e., the sanitized release) can be observed by the target recipient. However, no state of the art SMC protocol tackles this problem accurately (See 3.2).

The solution promoted in this paper matches the three aforementioned requirements by exploiting the hardware security of the client devices, so that participants in the PPDP protocol can henceforth trust each other. The emergence of secure devices has already motivated the re-examination of a number of traditional SMC problems (e.g., [99, 94, 82]), pushing back well-established limits.

The chapter is organized as follows. Section 3.2 discusses the related works by first presenting the targeted PPDP algorithms and the running examples, and second positioning the state of the art solutions with respect to the hypothesis of our study. Section 3.3 presents Met$_{\mathbb{A}}$P, our global approach to define a generic and scalable PPDP meta-protocol matching the characteristics of the targeted distributed architecture. Section 3.4 presents future research directions based on this work.

3.2 Related Works

This section presents the background knowledge needed to understand the paper. First, we introduce the two major PPDP approaches targeted by Met$_{\mathbb{A}}$P and describe the models and algorithms selected from both approaches, which we will adapt to the Asymmetric Architecture. Then we position Met$_{\mathbb{A}}$P with respect to the state-of-the-art works: we overview the secure and distributed implementations of PPDP algorithms as well as the cryptographic protocols based on secure devices.

Targeted PPDP Models and Algorithms

There are two major approaches to the global (centralized) publishing problem: the more recent *differential privacy approach* and the more traditional *partition-based approach*, which is currently adopted in many experimental contexts such as epidemiology.

The Differential Privacy Approach

The *differential privacy* approach, initiated in the interactive query answering setting [75], is receiving an increasing attention in global publishing (e.g., [137, 120]). Loosely speaking, a differentially private algorithm is a *randomized* function which is defined such that the presence or

absence of each participant's record in the initial dataset has only a quantifiably insignificant impact on the algorithm outputs. The main advantage set forward with the differential approach is that it gives quantifiable and provable guarantees on the information gained when viewing the release. In this article, we consider more specifically the (d, γ)-PRIVACY model [137] which is part of the differential privacy approach. Indeed, [137] show that the (d, γ)-PRIVACY model is equivalent to the ϵ-INDISTINGUISHABILITY model [75]. (d, γ)-PRIVACY models the adversarial knowledge as a probability distribution over the possible initial records and guarantees that after accessing the sanitized release, the adversarial knowledge about any record remains below a fixed bound. In our work, since we are interested in enforcing the (d, γ)-PRIVACY model, we needed to choose an algorithm. We have considered to this end the $\alpha\beta$-ALGORITHM, introduced by [137], due to its simplicity and efficiency.

The Partition-Based Approach

Roughly speaking, the *partition-based* approach splits the attributes of the dataset in two categories: the *quasi-identifier* part and the *sensitive* part. A quasi-identifier (denoted QID) is a set of attributes for which some records may exhibit a combination of unique values in the dataset, and consequently be identifying for the corresponding participants (e.g., {ZipCode, BirthDate, Gender}). The sensitive part (denoted SD) encompasses the attribute(s) whose association with participants must be made ambiguous (e.g., {Disease}). Several quasi-identifiers, as well as several sensitive attributes, may exist in a dataset; but for simplicity, most of the literature assumes the presence of a single quasi-identifier (possibly encompassing all of them) and a single sensitive attribute.

Partition-based approaches essentially apply a controlled degradation to the association between participants (represented in the dataset by their quasi-identifier(s)) and their sensitive attribute(s). The initial dataset is deterministically partitioned into groups of records, called *equivalence classes*, where quasi-identifier and sensitive values satisfy a chosen partition-based privacy model. The seminal k-ANONYMITY model [141] requires each class to contain at least k records indistinguishable with respect to their (possibly coarsened) quasi-identifier values so that each sensitive data be associated with at least k records. Successors of k-ANONYMITY (e.g., ℓ-DIVERSITY [117], t-CLOSENESS [113], ϵ-PRIVACY [116], m-INVARIENCE[154], m-CONFIDENTIALITY [152]) further constrain the distribution of sensitive data within each class, tackling different adversarial assumptions.

Many different algorithms implementing these models exist, such as the well known MONDRIAN algorithm [111] for k-ANONYMITY, and used as the basis of the implementation of many successor models. However, other example exist, such as the BUCKETISATION algorithm [153], used to implement ℓ-DIVERSITY which is not built on MONDRIAN.

Secure Multi-party Computation

The traditional *Secure Multi-party Computation* (*SMC* for short) approach neither considers secure devices nor is generic and scalable *simultaneously*. It's trust assumptions however are even more pessimistic than ours, since there is no element of trust in SMC. Therefore the

context of SMC is actually quite different from the *Trusted Cells* context. As we will show in this Chapter, the use of a trusted computing element will greatly improve the feasability of generic complex distributed computations. The difficulty is that, as in SMC, we introduce an untrusted element in the architecture, which has to be dealt with.

Generic Secure Multi-party Computation

Early works such as [158, 92], have investigated methods for translating *any* centralized function to a decentralized one that provides SMC security guarantees. The resulting generic translation method essentially consists in expressing the centralized function as a combinatorial circuit and distributing its execution on the set of participants. The resulting cost depends on the number of inter-participant interactions (whose execution is mainly based on oblivious transfer protocols), which in turn depends exponentially on the size of the input data, on the complexity of the initial function, and on the number of participants. Despite their unquestionable theoretical interest, these generic approaches cannot be used in practical global publishing scenarios where inputs are large, participants numerous and sanitization algorithms complex.

Secure Multi-party Computation for Global Publishing Algorithms

The non-practicality of the generic SMC approach has led to the investigation of SMC solutions specific to each problem. To the best of our knowledge, only a small number of works has focused on the global publishing problem [162, 100, 122, 161, 101, 121, 155], and in a way that severely limits their scope. First, they are not generic: their internals consist in cryptographic protocols specifically designed to enforce a given privacy model/algorithm (k-Anonymity or variants for most). Second, they do not fit the Asymmetric architecture. They make strong assumptions of availability or computing resources for participants, or reintroduce a central point of attack.

The authors of [162] aim at producing for a analyst and without any central trusted server a k-anonymous release of the union of data held by a set of participants. The participants and the analyst are honest-but-curious. Zhong *et al.* design a first solution letting the analyst decrypt only the records whose raw quasi-identifiers appear more than k times in the dataset (other records are lost). To avoid the possibly high data loss, a second solution lets the analyst execute the partition-based algorithm proposed in [119] over encrypted records. To this end, participants disclose to the analyst a controlled amount of information consisting of the following: for each pair of quasi-identifiers the number of attribute values that differ. Both solutions do not answer our objectives in that they are strongly tied to the k-Anonymity model (no genericity), they consider an honest-but-curious analyst (no malicious attacker), and they are designed for fully connected participants (unfit to the Asymmetric architecture).

The approach proposed in [155] is similar to the second solution proposed in [162]: the difference mainly lies in the partition-based algorithm used to produce equivalence classes, i.e., [88], which is adapted by disclosing to the miner the relative orders and distances between quasi-identifiers. The proposed protocol between the participants and the miner is based on calls to *homomorphic* and *private information retrieval* schemes. Similar shortcomings preclude this approach.

In [161], Zhong *et al.* relax the high availability requirement made by [162] on the complete set of participants and consider a malicious miner. However, the protocol proposed is still unable to meet our objectives. First a single point of failure (attack) is reintroduced by requiring that a specific participant, called the *helper*, be fully connected and never collude with the miner. The complete protocol is a two-party protocol between the helper (that has collected the participants' data in an encrypted form) and the miner. Second, the internals of the proposed protocols are designed for the k-ANONYMITY model: participants' data is encrypted such that the two-party protocol lets the analyst suppress the quasi-identifiers appearing less than k times.

In [101], the authors propose an adaptation of the MONDRIAN algorithm to produce a k-anonymous release of the union of the datasets. However, the proposed protocol first does not fit the ASYMMETRIC architecture (full availability requirement of the participants), second is strongly tied to MONDRIAN (it is a distributed adaptation of MONDRIAN's recursive splitting procedure), and third limits the attack model to honest-but-curious.

The approach proposed in [121] differs from [101] only in the underlying privacy model (a home-made variant of k-ANONYMITY) and algorithm (designed for enforcing the variant they propose). Both present similar shortcomings with respect to our context: hardcoding of the privacy algorithm into the internals of the protocol, a full availability requirement for the participants, honest-but-curious attack model.

Finally, the approaches proposed in [100] and [122] consider a context where data is vertically partitionned accross the set of participants. The resulting protocols consist in an iterative sanitization of the global dataset (consisting in the (virtual) join of the vertical partitions) through local proposals emitted by participants (each participant has the full set of records (projected on a subset of attributes)). In addition to the usual shortcomings (e.g., honest-but-curious attack model, assumption of few participants, full availability), the approaches do not fit a context where data is horizontally partitioned (each participant has only a subset of records and is consequently unable to reach local sanitization decisions). We note that [100] is able to translate usual centralized k-ANONYMITY algorithms into secure distributed two-party protocols: though the need to provide genericity was felt, genericity was limited here it to the k-ANONYMITY model and algorithms.

Cryptography based on Secure Devices

The interest in cryptographic protocols founded on secure devices is resurfacing. Indeed, many works revisit the traditional approaches and their results, based on the use of secure hardware. For example, [104, 93] circumvent the theoretical impossibility results of secure multi-party computation protocols by benefiting from the tangible trust provided by tokens; [99] benefit from tamper-resistant devices to decrease the communication cost of the generic SMC approach by equipping parties with such devices which are - locally - in charge of a part of the circuit evaluation; [98, 82] propose two-party secure set intersection protocols based on a physical exchange of (preloaded) secure devices. To the best of our knowledge, the only approach to centralized publishing over an architecture based on secure devices is our own work.

Server-based approaches

Powerful (and costly) secure hardware exists and could act as a trusted proxy for the server (here, the recipient) [38]. Such an assumption would greatly simplify the protocol: all participants could send their raw data to this trusted proxy, which would implement the complete PPDP protocol and deliver the final result to the recipient. However, this would reintroduce a centralized trusted party in the architecture, and thus a clear incentive to attack it. For the same reason, we rule out any solution based on a very small number of secure devices (e.g., powerful secure clients running the PPDP protocol). The solution we propose adopts the opposite approach: it primarily relies on a wide distribution of data storage and processing on low cost personal secure devices (the *trusted cells*) in order to get rid of the single point of vulnerability introduced by any central entity.

3.3 Approach and Scientific Results

In this Section, we provide a brief overview of our approach and main contributions :

1. A generic, abstract and scalable protocol, called $\text{MET}_{\mathbb{A}}\text{P}$, which can be used to implement a variety of PPDP algorithms on the Asymmetric Architecture, and a proof of its correctness. Other global computations will be able to build on the security primitives discussed here.
2. Security proofs against *honest-but-curious* and *Malicious* adversaries (see Section 2.3.1).
3. Probabilistic countermeasures to deter *Malicious* adversaries having breached at least one PDS.
4. An experimental validation using smart tokens of the applicability and scalability of this approach.

The genericity of the $\text{MET}_{\mathbb{A}}\text{P}$ protocol is achieved by defining a common execution sequence embracing the behaviour of traditional global publishing algorithms. Scalability is reached by subsequently adapting this sequence to an Asymmetric Architecture, taking advantage of its intrinsic parallel processing capabilities. The correctness and security properties of $\text{MET}_{\mathbb{A}}\text{P}$ are proved, as a main step to guarantee the correctness and security of *all* concrete publishing algorithms instantiated through $\text{MET}_{\mathbb{A}}\text{P}$, using the following threat model.

3.3.1 Building a generic and scalable protocol

Genericity

All global publishing algorithms we are aware of can be arranged in a succession of three phases as follows. The *collection phase* gathers the input dataset to be sanitized. In this phase, the publisher collects the information generated by each agreeing participant, until reaching the expected cardinality. Then, the *construction phase* analyzes the complete dataset and,

according to this, produces the *sanitization information* which will maximize the utility/privacy tradeoff in the final release. Finally, the *sanitization phase* produces this final release, sanitizing each collected record individually according to this sanitization information. Putting aside the performance issue (which may introduce significant differences between PPDP algorithms), organizing the processing in such a way gives prominence to the fact that the specific part of each PPDP algorithm (or in general the computing of *any* aggregation function using data stored in individual cells) breaks down to the computation of the sanitization information.

Scalability

The objective is to adapt these three phases to an Asymmetric Architecture, exploiting its parallel computing potential without sacrificing genericity. The collection phase is intrinsically parallel, each participating cell sending its owner's record to the recipient with no synchrony requirement. The sanitization phase also can be easily computed with independent parallelism since it operates at a record granularity[41]. To make this possible, during the construction phase and once the sanitization information has been computed, the recipient partitions the dataset such that (1) each partition is a self-contained unit of treatment in that it contains both a subset of records and the sanitization information required to sanitize them and (2) the partition size fits the secure devices limited resources. The construction phase itself is unfortunately much more difficult to parallelize because (1) it executes at the dataset granularity and (2) it depends on the instantiated PPDP algorithm. As discussed in Section 3.2, no SMC implementation of such process could be envisioned without hurting either the genericity or scalability objective and could tackle the Asymmetric hypothesis. To circumvent this difficulty, we delegate the computation of the construction process to the recipient which benefits from high computing resources and explain next how to make this correct and secure.

3.3.2 Correctness and Security of Met$_{\mathbb{A}}$P

Met$_{\mathbb{A}}$P is actually a skeleton which can be instantiated given a PPDP algorithm $\mathbb{A}$. If we show that Met$_{\mathbb{A}}$P is correct and secure, then any instantiation of Met$_{\mathbb{A}}$P would inherit these same properties. Unfortunately, part of this proof depends on the information disclosed to the recipient by each algorithm's instantiation. Our methodology was therefore to prove once and for all the correctness and security of the generic part of Met$_{\mathbb{A}}$P, so that the attention can focus on the specific part of each instantiation. To this end, we followed the same strategy as for assessing the correctness and security of SMC protocols. Informally speaking, this strategy consists of showing that the SMC implementation of a given protocol and the *ideal* implementation of this same protocol are equivalent in terms of output and information leakage. We call *ideal* an implementation of a protocol by a centralized trusted third party (i.e., where dysfunctions linked to distribution and all forms of attacks are precluded).

[41] The degree of parallelism is determined by the number of secure devices which connect together during the sanitization phase, every token being eligible to participate to this phase, including those which did not participate to the collection phase. In the extreme case where secure devices connect one after the other, the processing will simply be performed sequentially.

It is worth noting that this problem is not specific to MET$_{\mathbb{A}}$P but to any attempt to adapt a global publishing algorithm to an untrusted distributed context. Correctness and security aspects exacerbate the benefit of a generic protocol skeleton. Indeed, each *ad-hoc* adaptation of an algorithm would require re-starting the correctness and security analysis from scratch, identifying specific attacks, designing home-made counter-measures, and proving their security. MET$_{\mathbb{A}}$P only needs to be proved correct and secure once and for all. Whatever the instantiated algorithm, it will inherit all the correctness and security guarantees provided by MET$_{\mathbb{A}}$P for its generic part.

Correctness

We say that the instantiation of a global publishing algorithm using MET$_{\mathbb{A}}$P is *correct* if, whatever the input dataset, the distribution ensembles of its *outputs* in our setting (i.e. ASYMMETRIC ARCHITECTURE) are computationally indistinguishable [91] from the distribution ensembles of the original algorithm's outputs in an ideal setting (i.e. trusted third party). In other words, the two results are equivalent (though they may not be equal strictly speaking due to the non determinism of PPDP algorithms).

The proof of correctness relies on the demonstration that the result of each phase (collection, construction, sanitization) is similar to its centralized counterpart. Regarding security, both passive and active attacks must be considered at each phase. To prevent passive attacks, each record is initially encrypted by its hosting secure device - with a randomized encryption scheme - before being delivered to the recipient during the collection phase. Records remain encrypted until the very last step of MET$_{\mathbb{A}}$P. During this same collection phase, and to allow execution of the construction phase, an extra amount of information must be disclosed (i.e., sent un-ciphered) to the recipient by each secure device[42]. This information, called hereafter *construction information*, is algorithm dependent and must not increase the attacker's knowledge. Based on it, the recipient then analyzes the dataset, appends to each encrypted record its corresponding sanitization information (e.g. false tuples for differential privacy or equivalence classes for k-anonymization), and partitions them. During the sanitization phase, secure devices download partitions one by one, decrypt their content, sanitize each record accurately using the sanitization information and produce part of the final release. Theorem 1 states the correctness of PPDP algorithms instantiated using a MET$_{\mathbb{A}}$P protocol (several exist depending on the adversary considered, *honest-but-curious* or *malicious*), depicted in Figure 3.3 :

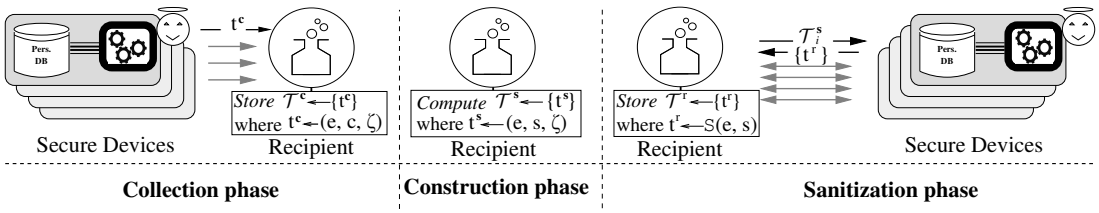

Figure 3.3: MET$_{\mathbb{A}}$P Execution Sequence

[42] A similar tradeoff occurs in the query outsourcing approach where an untrusted host must be able to issue queries over encrypted data (e.g., [95]).

1. During the collection phase, the recipient gathers the set of c-tuples $\mathcal{T}^c$ from the secure devices that connect. Each c-tuple contains the encrypted participant's record e together with its corresponding construction information c and security information ζ.

2. The construction phase consists in letting the publisher produce the set of s-tuples to be sanitized $\mathcal{T}^s$ based on the c-tuples collected, along with information on how to sanitize them s. The publisher then partitions $\mathcal{T}^s$ uniformly such that each partition $\mathcal{T}_i^s \in \mathcal{T}^s$ fits the limited resources of a single PDS.

3. Each secure device that connects during the sanitization phase downloads a self-contained partition $\mathcal{T}_i^s \in \mathcal{T}^s$, and uploads on the recipient the result of decrypting and sanitizing the s-tuples it contains (i.e. computing $r = S(e, s)$). The recipient then stores the resulting sanitized r-tuples in $\mathcal{T}^r$ and stops the sanitization phase once all the partitions have been sanitized.

Theorem 1 (MET$_{\mathbb{A}}$P Correctness)**.** Let π be an instance of a privacy algorithm $\mathbb{A}$ instantiated on an ASYMMETRIC ARCHITECTURE using MET$_{\mathbb{A}}$P. If the construction information is sufficient to perform the construction phase in a centralized setting, then π is *correct* and *terminates*.

Security

We say that the instantiation of a global publishing algorithm using MET$_{\mathbb{A}}$P is *secure* if the attacker learns nothing beyond its prior knowledge and the final release.

Honest-but-curious adversaries

The honest-but-curious recipient is a passive adversary. It aims at breaching the privacy of records by launching inferences on the information that it obtains from secure devices, explicitly or not (i.e. via inferences on information leaked by the execution sequence). Indeed, due to the "per-partition" organization of the correct but unprotected execution sequence, the recipient is able to trivially *link* the sanitized records corresponding to a given partition to the sanitization information corresponding to the same partition. Consequently, it may also link it to the corresponding construction information. Such *linking attacks* may lead to the full disclosure of records. We define the `Unlinkability` safety property to counter this attack, which informally states that the supporting server infrastructure must not be able to map cyphered tuples to (a subset of) sanitized tuples.

Theorem 2. Let $\pi_{(hc)}$ be the instance of a privacy algorithm $\mathbb{A}$ instantiated on an ASYMMETRIC ARCHITECTURE using MET$_{\mathbb{A}}$P$_{(hc)}$ (Algorithm 2 in [21]). If the construction information is sufficient to perform the construction phase in a centralized setting and does not independently provide any information to the recipient, then $\pi_{(hc)}$ correctly and securely computes $\mathbb{A}$ against a honest-but-curious recipient.

It is important to note that the instance of any algorithm under MET$_{\mathbb{A}}$P$_{(hc)}$ only requires to show that the construction information is harmless: if this is the case, Theorem 2 can consequently be used to prove that the global execution sequence is secure.

Malicious adversaries

A malicious attacker is not limited to passive attacks but may deviate arbitrarily from the protocol. The objectives of the malicious recipient are to conduct an attack that will increase its knowledge, while generating a correct result and remaining undetected by secure devices. It can attack the integrity of all data structures it can access in the course of the protocol (i.e. the collected dataset and the sanitization information) or try to tamper the execution sequence itself. This leads to three forms of active attacks: (1) Tampering the set of tuples to be sanitized, (2) Tampering the construction function and (3) tampering the execution sequence. For each of these attacks, we proposed safety properties (`Origin, Identifier Unicity, Mutual Exclusion, Membership, Invariance, Safe Construction Function` and `Execution Sequence Integrity`) to guarantee that tampering will be detected. As a result, safety properties force an active malicious adversary to be passive, a case covered by the *Honest-but-curious* adversary, and for which Theorem 2 provides a solution. Theorem 3 states the correctness and security of the *Malicious* protocol.

Theorem 3. Let $\pi_{(mal)}$ be the instance of the privacy algorithm $\mathbb{A}$ under $\text{MET}_{\mathbb{A}}\text{P}_{(mal)}$ (Algorithm 3 in [21]). If the information disclosed by the construction information does not independently provide any information to the recipient, and if `Execution Sequence Integrity` is enforced, then $\pi_{(mal)}$ correctly and securely computes $\mathbb{A}$ against a malicious recipient with weakly-malicious intents.

Malicious$_{Hard}$ attacks

In the $\text{MET}_{\mathbb{A}}\text{P}_{(mal)}$ protocol, if the adversarial recipient succeeds in breaking at least one secure device, it can unveil not only the devices' information but also its cryptographic keys which can in turn be used to decrypt the complete set of encrypted records. To limit the scope of such attacks, the traditional solution is to use clusters, each cluster using its own key, and organize the process so that the impact of compromising one key is divided by the number of clusters. We follow this approach and partition the devices into a set of n_c distinct *clusters*, *randomly* and *evenly*, such that the secure devices belonging to different clusters are equipped with distinct cryptographic keys. Breaking a secure device consequently brings the adversary the ability to decrypt the encrypted records from the device's cluster only. This security measure requires a few adjustments to $\text{MET}_{\mathbb{A}}\text{P}$ in order to handle the limited decryption abilities of secure devices. Moreover, in addition to decrypting data, the compromised device gains the capacity of encrypting and signing data (e.g., s-tuples, secure counts), allowing it to satisfy the `Origin` safety property. We protect $\text{MET}_{\mathbb{A}}\text{P}$ against this kind of forge actions by providing arbitrarily high probabilistic detection of compromised clusters (see [19]).

Using Met$_{\mathbb{A}}$P in practical cases

Proving the correctness and security of $\text{MET}_{\mathbb{A}}\text{P}$ is an important result of this research, stated by Theorems 1, 2, and 3, which are all detailed, along with their proofs or proof sketches in [21]. Considering this given, instantiating $\text{MET}_{\mathbb{A}}\text{P}$ using any global publishing algorithm and proving the correctness and security of the result leads to the following steps:

1. to characterize which *construction information* must be mandatorily disclosed to the recipient to allow it in turn to compute the *sanitization information*;
2. to prove that this *contruction information* cannot be source of passive attacks;
3. to characterize the test that must be performed by the secure devices in order to assess the legitimacy of the *sanitization information* produced by the recipient. This legitimacy is linked to the dataset and to the privacy guarantees enforced by the instantiated algorithm.

3.3.3 Experimental validation

We refer the reader interested to [21] for a detailed discussion of the implementation of MET$_{\mathbb{A}}$P on three different types of PPDP algorithms : the $\alpha\beta$-ALGORITHM, the BUCKETIZATION algorithm and the MONDRIAN algorithm. We can draw the following conclusions from the implementation and its performance analysis.

1. It is possible to securely implement a wide variety of global computations on the ASYMMETRIC ARCHITECTURE, illustrated by the safe computation of a variety of PPDP models, through very different algorithms. We propose general safety properties to this end.

2. The critical aspect of a global computation is the latency of the collection phase. This latency is indeed not bounded and highly application dependent. It is determined by the connection rate of participants and by the ratio of the whole population to be polled. However, in the case of PPDP, contrary to the manual approaches often used in practice, the latency is not linked to the size of the population.

3. Regarding internal time consumption, experiments show that MET$_{\mathbb{A}}$P can manage very large datasets (millions of tuples) with excellent performance due to its parallelism and to the hardware implementation of cryptographic operations (i.e. on a PDS). Hence, scalability is achieved.

3.4 Future Works

Deploying other types of computation on the Asymmetric Architecture

Our long term goal, as discussed in Chapter 2, is to provide a full relationnal query facility to data stored in large numbers of PDS. We have started working on this topic, and in particular the difficult task of computing agregate queries (`GROUP BY / HAVING`) using data stored in PDS, *without* leaking any information to the supporting server. This work is the topic of Cuong Quoc To's Ph.D. thesis, which started in october 2012. We have presented preliminary results in some workshops [145, 146, 147], involving computation of SQL aggregate queries with only local joins, and without considering any *Malicious* attacks. Future work involves investigating these two directions.

Non-trusted Hardware

In this piece of work, we have considered that the *Trusted Cells* are constructed using Secure Portable Tokens, which are highly secure, and in which one can place reasonable trust. However, the detection methods of $Malicious_{hard}$ adversaries, which have cracked a token, can be extended, so long as the number of hacked devices is smaller than the number of honest devices. Interesting work would involve distributed computations (such as PPDP or SQL queries) on lower trust elements, such as using computers of social network friends, or work colleagues. In particular, in this context, *Denial of Service* types of attacks are realistic, and should be covered. Protocols for automatic regulation and exclusion of malicious participants should also be included in such systems.

Chapter 4

Protecting Data outside the Cell : *Limited Data Collection*

"Experts often possess more data than judgement."

– *Colin Powell*

Summary:

In the Trusted Cell *context, users could be led to transmit personal documents to services* outside the trusted world*, such as social care, tax reduction, bank loans and many others. In real cases, hundreds of private data items are used to feed (more and more complex) decision processes. However, considering too much data not only leads to obvious privacy issues for the user, it can also entail important financial cost overheads for organizations. In this Chapter, we investigate a general privacy principle, called* Limited Data Collection *(LDC), which states that requested sets of personal data must be limited to the minimum necessary to achieve the purpose the user consents to, and formalize the underlying theoretical problem, termed* Minimum Exposure *(MinExp), and propose algorithms, some of which can be executed on secure tokens, to reduce the information transmitted to servers to the minimum necessary to correctly execute the service.*

Contents

This chapter presents the scientific results detailed in the following papers :

Contributions :

- [28, 29] our introductory papers, which present the concept of *Minimum Exposure* (MinExp), an implementation of the general privacy principle of *Limited Data Collection* (LDC), of which [30] is an extended version (in French).
- [26] which conducts an experimental study using multi-label classifiers on real datasets of the applicability of *MinExp* techniques.
- [23] which demonstrates the applicability of *MinExp* techniques using smart cards, which are even less powerful devices than Secure Portable Tokens.

This work was led in collaboration with Pr. Michalis Vazirgiannis of Athens University of Economics and Business (Greece), and the *Conseil Général des Yvelines*. Funding was provided by Pr. Vazirgiannis' Digiteo Chair, started in 2010, a partnership between University of Versailles, École Polytechnique, Telecom ParisTech and Exalead, and by the ANR KISS project (since 2011).

Prototype : The `MinExp-Card` prototype has been developed in partnership with Conseil Général des Yvelines and should be tested in the field in the comming year.

4.1 Context and Motivation

Data over-disclosure

DISCLOSING personal data when applying to services online is unavoidable. Indeed, data is most often legitimately collected to customize services to the specific situation of each applicant(e.g. social care, tax services, bank loans, etc). However, nearly half of the EU citizens report being asked for *more* information than necessary [78], thus falling victims to data *over-disclosure*. More than 70% of them are concerned by this issue, since their personal information will generally be checked and evaluated by employees and end up in a database for years (often as legal proof of the process). In a world full of trusted cells, focused on the protection of each individual's data, such a situation would be paradoxical : user's data would be better protected *inside* her trusted cell, but practically *all* her data would be sent to and stored by external service providers at some point. These service providers may be outside the trusted cell world, in which case we can make no assumption on the security of the data we send to them. We only acknowledge that they need *some* of our data, in order to process our request.

This privacy issue is well recognized, and privacy principles enacted worldwide [79, 131] promote *Limited Data Collection and Retention* (LDC), which state that data collected for a process should be limited to the *minimum data needed to provide the service*, and should not be kept for an excessive amount of time. However, it has been less emphasised that data over-disclosure also incurs important financial costs for the service providers themselves. Indeed, processing personal data often incurs manual operations, whose cost largely depends on the quantity of information processed. Collected data is checked, e.g., by comparing the applicants' declarations with copies of official documents or by crossing information with internal and external databases.

Furthermore, applicant's information must be retained for years (i.e., for the duration of the offer, for the legal delay required to prove the non discriminative nature of the offer), and organizations are held responsible in the case of data breaches. This is not minor issue. In 2011, the Open Security Foundation[43] reported over a thousand data loss incidents affecting more than a hundred million personal records. Every breach is a financial disaster for the companies in charge of the data. The 2012 Ponemon Institute study [134] on 49 US companies estimates the cost of data breaches at an average $194 per tuple, due to legal obligations to notify data owners and assist victims (e.g. cancelling their credit card if the number has been disclosed) and the impact of negative publicity. Security companies now provide online breach cost calculators[44] to draw attention to this new phenomenon.

Requirements to solve the data over-disclosure problem

In this work we have addressed the problem of data over-disclosure when applying to a service. Our objective is to provide data reduction techniques by respecting the following requirements:

[43] See http://www.datalossdb.org/reports
[44] See http://databreachcalculator.com.sapin.arvixe.com/

- **Minimum.** The devised techniques must aim at collecting *the least data possible in order to provide the service*[45]. This requirement corresponds to a strict understanding of the LDC principle as enacted in privacy laws. Note that it will also minimize the financial costs incurred for processing the data and in the (*likely*) event of a data breach.

- **Accountable.** Service providers are accountable for their decisions and must be able to verify input data items filled by users[46]. This may be anything from a simple "personalization of their web site experience", to receiving social benefits due to their severe medical status. The information provided (e.g. via online application forms) must thus be comprehensible enough to enable service providers to verify it, e.g., by checking digital signatures (when possible) or by checking conformity with internal databases or copies of other official documents. Applications and related decisions must also be retained by the service provider in order to be able to legally prove (later, if necessary) that their decision was taken in compliance with legal principles. In most domains like social care, tax returns, bank loans, insurance, applications are stored for many years (even for rejected applications) to attest non discrimination criteria or in the case of later disputes.

- **Broad-spectrum.** The proposed techniques must accommodate any kind of subsequent decision making technique. Decision making systems are based on classifiers transposed from human defined processes (e.g., built according to legislation or defined by experts) or built using data mining tools (e.g., decision trees, Support Vector Machines (SVM), etc.). Whatever their nature, classifiers can be *binary* (i.e., each application is associated to a yes/no decision), *multi-class* (i.e., each application is associated to a given class out of several possibilities) or *multi-label* (i.e., each application is associated with a set of labels [148]). Typically, multi-label classification is now used in many domains [148], such as social care, medical systems, tax administration (one label per tax exemption), bank loan proposals, etc.

- **Scalable.** In practice, banking applications or social care applications request up to several hundreds or even thousands of data items. The trend is upwards for decisions based on legislation (e.g. tax returns, social care benefits, etc.) since laws are endowed with more and more conditions and exemptions. When using data mining tools, the trend is also towards more complex classifiers to better calibrate decisions, leading to collect more (user) input data.

In this chapter, we overview the novel technique, called *Minimum Exposure* we proposed to tackle these requirements. A naïve solution is to move the decision process to the client side, and thus only provide final decisions (outputs) to the organization (without revealing the input personal data). This solution is however proscribed due to lack of verifiability (hurting the second requirement). Existing techniques transpose the limited data collection principle to computer systems [71, 8, 37]. However, they assume that useful data can be distinguished a priori for each application purpose. Under this assumption, the information transmitted by the user is obtained by constructing the union of all data items identified as potentially useful

[45]Note that minimizing the set of collected data is by nature orthogonal to any other security feature based on access control, cryptographic techniques, intrusion detection, etc.

[46]We call them *applicants* in this context, since they are applying for a tailored service.

for that purpose. However, the a priori assumption only holds for simple cases (e.g., when ordering online, the address of the customer is needed to deliver the purchased items). In the general case, the decision depends on the values of the data (as shown in Example 1). The assumption is thus invalid for complex decisions, leading the corresponding solutions to hurt the minimum requirement.

Example 1. Users declaring their revenue can benefit from a tax exemption in the following cases: having (i) an income under $30.000 and an age below 25, (ii) an income below $10.000, regardless of age, or (iii) a sufficient number of dependants (e.g. 2). Existing solution would request the union of potentially useful data items, namely [income, age, nb_dependants]. However, for a user with values u_1 = [income= $25.000, age=21, nb_dependants=1] the minimum data set would be [income, age], and for a user with u_2 =[income=$40.000, age=35, nb_dependants=2] it would be [nb_dependants].

The chapter is organized as follows. In Section 4.2 we present other approaches to implement general privacy principles, describe fields that deal with similar problems, and provide an overview of the concepts and tools used in this study. In Section 4.3 we provide an overview of the approach adopted and the results obtained in this study. We close this Chapter with Section 4.4 by providing directions for future works.

4.2 Related Works

Implementation of legal privacy principles

The transposition of legal privacy principles into privacy aware systems has fostered many studies. Emblematic examples include the P3P Platform for Privacy Preferences [71] and Hippocratic databases [8]. P3P highlights conflicting policies, but it offers no means to calibrate the data exposed by a user and achieve Limited Data Collection (LDC).

Many other policy languages have been proposed for different application scenarios, like EPAL [37], XACML [123] or WSPL [32], but to the best of our knowledge, no language has been introduced with LDC in mind.

The architecture of a Hippocratic database is based on ten guiding privacy principles including LDC. It addresses LDC by maintaining the set of attributes that are required for achieving each declared purpose. However, this solution assumes useful and useless data for a given purpose can be distinguished *at the time of the data collection*. As shown in Example 1 in Section 4.1, this assumption only holds for simple cases, but not in general decision making processes.

Fields with similar problems

Automated Trust Negotiation

Existing works closer to our problem are the areas of automated trust negotiation and credential based access control, where access decisions are based on the gradual confrontation of an

access control policy with a set of credentials. At each step, the minimal set of credentials, corresponding to a given credential request, is disclosed. Computing this set can be viewed as an application of Minimum Exposure, where the service objective is to be granted access, and the information to be transmitted is a set of credentials. Indeed, a few number of works including [35, 67, 159] even specifically address the minimization step using privacy metrics adapted to credentials. However, the problem and solutions are simpler than ours for two important reasons. First, the decision making processes that we consider are more complex than access control. Indeed, we consider multi-label decisions [148] (e.g. lower credit rate, longer duration, lower cost of insurance, larger portion of 0% loan, etc.) each one potentially impacting the final offer made to the applicant. Second, in our context, the decision making process requires by nature a huge amounts of personal data (hundreds to thousands) while in trust negotiation only a few credentials are considered (e.g. up to 35 in [35]). The problem we study is therefore much more general.

Privacy Preserving Data Mining

Works dealing with Privacy Preserving Data Mining (PPDM) also take a different direction than Minimum Exposure. Recent PPDM surveys [5, 81] refer neither to Minimum Exposure type problems nor to their legal foundation (i.e., the Limited Data Exposure principle). Unlike Minimum Exposure, PPDM techniques, such as recent developments in [112, 85, 120] which protect individual records with regards to the input of a data mining algorithm, turn original data into encrypted or randomly perturbed data, which becomes unverifiable. On the contrary, Minimum Exposure preserves the original data and its ability to be verified by a third party (a signature guarantees its integrity and origin). Another aspect of PPDM techniques is that they try to protect sensitive rules (i.e., the output of a data mining algorithm) by removing raw data [4, 150]. However, these techniques maximize the information retained in the output data set, so long as the private results remain secret, whereas the goal of Minimum Exposure is to minimize it. Note that this approach is in some sense orthogonal to Minimum Exposure. Indeed, the former (PPDM) would remove sensitive data upstream and the latter (MinExp) could minimize the remaining information, thereby achieving better privacy.

Novelty of the approach

Although well recognized by law, expected by citizens, and central to privacy aware data management, *Limited Data Collection* is not enforced, or is enforced in very simple cases. This study opens a new research direction, termed *Minimum Exposure* by formally defining the problem, studying its complexity and proposing solutions to implement it on PDS in practice, or more generally in any application wishing to enforce *Limited Data Collection*.

4.3 Approach and Scientific Results

In this Section, we provide a brief overview of our approach and our main contributions, which are :

1. A formal definition of the Minimum Exposure optimization problem.
2. The resolution and complexity analysis of this problem.
3. Algorithms and their experimental validation on central server and smart card to compute the solution to this problem, using real datasets and synthetic datasets.

4.3.1 Introducing the *MinExp* optimization problem

We consider that a user is defined by a set $Data_u$ of D distinct $(attribute, value)$ pairs, called *eq*-assertions, and noted as_i, stored in the *Trusted Cell*. We associate with them D Boolean variables, $b_1, \ldots, b_D$, where $b_i = true \Leftrightarrow as_i$ is published. The decision making process is modelled as $R = \{r_j\}$, a set of DNF formulas (each r_j is called a *collection rule* leading to a specific benefit j), called *Rule Set*. We note $E_R = \bigwedge_j r_j$ the conjunction of all the rules (i.e. the benefits the user will obtain) triggered by the user, called *rule set Boolean formula*. We consider a (linear) exposure function $\mathbb{E}$ which computes the privacy/cost risk associated to the publication of a subset of elements of $Data_u$ (e.g. a simple exposure function is to count the number of *eq*-assertions published).

n-exposure problem statement :

Definition 1. Given a rule set R, $Data_u = \{as_x\}$ a set of *eq*-assertions that uniquely prove R, a set of Boolean variables $B = \{b_1, \ldots, b_D\}$ such that $b_x = true \Leftrightarrow as_x$ is exposed, $E_R = \bigwedge_j (\bigvee_k (\bigwedge_m b_{j,k,m}))$, where $\forall j, k, m, b_{j,k,m} \in B$, the rule set Boolean formula associated to R, and an exposure function $\mathbb{E}$, $Data_u$ is n-exposable with regards to R if and only if there exists a truth assignment T_B of B such that $\mathbb{E}(T_B) \leq n$ and E_R is $true$.

Our work has in fact concentrated on the related *optimization problem*, whose goal is to minimize n, and that we call *the Boolean Minimum Exposure optimization problem*. Our approach has been to model the *MinExp* problem using a formalism that can easily be compared to *Min Weighted Sat*, and use complexity results on this problem, to quantify the difficulty and approximability of the *MinExp* problem. Expressed in a logical formalism, it is then easy to write an equivalent mathematical program that can be fed to a state-of-the-art exact solver. Since an exact solver is not always applicable, given the hardness of the problem, and the low powered devices that need to solve it, we also explored polynomial algorithms, although these algorithms have no guarantees of optimality. Nevertheless, experimental results on real datasets show that polynomial algorithms provide reasonable exposure reduction.

4.3.2 Complexity of the *MinExp* problem

We use complexity results on the classical *Min Weighted Sat* problem, from [40, 77, 10]. The *Min Weighted Sat* decision (resp. optimization) problem is defined in [10] as :

Definition 2. *Min Weighted Sat problem.*
Given an integer n, an instance $\{P_{j,k}\}$ of p Boolean variables, a Conjunctive Normal Form (CNF) formula $F = \bigwedge_j(\bigvee_k P_{j,k}))$ over $\{P_{j,k}\}$ and a positive weight function $w : \{P_{j,k}\} \to \mathbb{R}^+$, find a truth assignment T for $\{P_{j,k}\}$ that satisfies F such that $w(T) = \Sigma_{j,k} w(P_{j,k}) \times T(P_{j,k})$ is $\leq n$ (resp. is minimum).
Note : When the formula contains no negative variables, the problem is called *All Positive Min Weighted Sat* (APMWS).

It is well known [70, 103] that this problem is NP-Complete. We state the following Theorem 4, which has two direct consequences (Corrolaries 1 and 2 on the complexity of the *MinExp* problem.

Theorem 4. *All Positive Min Weighted SAT* (APMWS) decision (resp. optimization) problem is reducible to *n-exposure* decision (resp. *MinExp* optimization) problem.

Corollary 1. The n-exposure decision problem is NP-Complete.

Approximability

Given an instance I of an optimization problem, and a feasible solution S of I, we denote $m(I, S)$ the value of solution S, $opt(I)$ the value of an optimal solution of I and $W(I)$ the value of a worst solution of I. The traditional *approximation ratio* for a minimization problem is defined by :

$$AR(I, S) = \frac{m(I, S)}{opt(I)}$$

The *differential approximation* ratio of S is defined by :

$$DR(I, S) = \left| \frac{m(I, S) - W(S)}{opt(I) - W(I)} \right|$$

In this work, we used the following complexity classes :

- *APX* which is the class of NP-optimization problems that allow polynomial-time approximation algorithms with an approximation ratio bounded by a constant.
- *0-DAPX* is the class of NP-optimisation problems for which all polynomial approximation algorithms have a differential approximation ratio of 0.

Corollary 2. The *MinExp* optimization problem is NP-Hard, is not in *APX* and has a differential approximation ratio of 0-DAPX.

The proof of this corollary uses complexity results from [10, 40] and [77].

4.3.3 Solving the *MinExp* problem

Corollary 1 is a negative complexity result in the sense that it shows that the problem is difficult and that polynomial approximation algorithms will provide bad approximation guarantees in

the worst case. In [28], we explore the domain where it is possible to provide an exact resolution using a state of the art MINLP solver (COUENNE [52]). We show that using very simple and low complexity algorithms, it is experimentally possible to reach comparable results. We have indeed demonstrated in a prototype that these simple algorithms can be executed on a low power smart card [23].

Exact Resolution (AMPL Model)

We will formulate our problem in terms of a non-linear binary integer program, in order to use a state of the art *Binary Integer Program* (BIP) solver, generally termed *Mixed Integer Non-Linear Program* (MINLP) solver to compute the exact results of the *Minimum Exposure* problem. We have chosen the competitive and open source COUENNE solver [52] in this respect. Note that most solvers focus on *Mixed Integer Linear Program* solving, since non-linear programs are more difficult to solve. Future work involves linearizing our problem to use simpler solvers (such as the *Gnu Linear Programming Kit* (GLPK) [118]). To describe an instance of a *Minimum Exposure* problem, we use *AMPL* [83], an algebraic modelling language for optimization problems on discrete or continuous variables.

Producing an AMPL program is a direct transformation where each assertion corresponds to a Boolean variable, and in which we express, using AMPL, one non-linear constraint per collection rule

$$r_j : \Sigma_k \Pi_m b_{j,k,m} \geq 1$$

In the case of the *MinExp* problem, objective function $\mathbb{E}$ is simply a linear combination (e.g. sum) of these variables. The program is then fed to the COUENNE solver, that computes an exact solution (which always exists).

Approximate Solutions (Polynomial Time)

We need to revert to a polynomial time approximation in order to compute results for the instances of the problem that cannot be tackled within reasonable time by the solver. We use three algorithms: a naïve fully random algorithm called RAND*, a simulated annealing meta-heuristics based algorithm called SA*, and an algorithm called HME using a heuristic specially designed for the *MinExp* problem. These algorithms are non deterministic, therefore they can be run many times and the best solution kept. However, they produce their first result in linear or polynomial time, depending on the algorithm.

- RAND* is based on a random choice of rules and serves as a baseline, and has a complexity of $O(C \times \max_{atom}(atom.length) + D)$ where $\max_{atom}(atom.length)$ is the length of the longest atomic rule involved, C the number of rules and D the number of user assertions.

- SA* is based on the simulated annealing meta-heuristic [108] and has a complexity of $O(m + D)$, where m represents the number of cooling iterations, disregarding the initial phase that uses RAND* as initialization.

- HME is a specific heuristic, presented in [28]. Its complexity is $O(|R| \times d_C \times d_{QD} \times D^2)$, where $|R|$ is the number of collection rules, d_C is the number of atomic rules per collection rule and d_{QD} is the number of predicates per atomic rule.

The intuition behind the HME heuristic is to successively get rid of the assertions which require keeping the least number of other assertions (such that all benefits are preserved) among the remaining ones. This heuristic is particularly relevant when the number of atoms per collection rule is small. Our performance evaluation [28, 29] shows that HME is a better approximation than its random of meta heuristics guided counterparts.

Experimental Results

Algorithms, data, BIP model generator code, Multi-Label Classification algorithms, and smart card algorithms are available in open source at http://project.inria.fr/minexp/. We conducted two sets of experiments, the first on real datasets to show that the approach was feasible in practice, and the second on synthetic data (using the same topology as in real datasets), in order to demonstrate the scalability of the approach. The quality was measured by computing the reduction of the set of exposed assertions in the application form, called exposure reduction ratio, noted :

$$E_R(T_B) = 1 - \mathbb{E}(T_B)/|B|$$

Real datasets : current applicability of the technique The first two datasets used are called $ENRON$ (emails made public in the context of the ENRON scandal) and $MEDICAL$ (Cincinnati Children's Hospital Department of Radiology), and are publicly available from the MULAN website[47]. The third one is called $SOCIAL$, and was built with the help of the General Council of Yvelines District. We have developed a framework, presented in [26] that builds a multi-label classifier on these data sets. The resulting classifier is used as the Rule Set.

The main conclusions of our experimentation on real datasets are the following:

1. The exposure ratio gain is always important, above 40% in all our measures.
2. $RAND*$ performs relatively well considering that it is a random approximate algorithm.
3. $RAND*$ scales linearly with the number of labels and average number of predicates per atomic rule, ensuring overall scalability to any real world dataset.
4. $COUENNE$ gives, as expected, better results than $RAND*$.
5. For $COUENNE$, the execution time increases exponentially as the size and complexity of the problem grows (for $SOCIAL$, 1 hour in average was needed per application, and largest applications remained unsolved after 12 hours).

[47] http://mulan.sourceforge.net/datasets.html

Syntheric datasets : scalability of the techniques There are many parameters that must be taken into account in order to test the scalability of the system : number of collection rules, atomic rules per collection rule, number of distinct predicates/assertions, etc., but many are linked. We refer to [30, 31] for more details on the setup. The main idea is that the rule set and user application data is randomly generated, keeping a topology close to real cases.

We draw three main conclusions from these experiments.

1. The exposure reduction is important even with very simple algorithms (see RAND*), ranging from 30% to 80%, and is on average of 70% in the area of applicability of the exact solver. This means that on average only 30% of a user's data items is sent when using the *MinExp* approach compared to the traditional case.
2. The scope of the exact solution is limited, and therefore the use of approximation algorithms is unavoidable, even when using a powerful desktop.
3. HME provides the best results of the approximation algorithms, outperforming them by about 10%, and scales in polynomial time with regards to D.

Which algorithm for a *Trusted Cell* or smart card ?

*RAND** which provides rather satisfying results, could be used as a replacement of the optimal resolution, on a low powered and constrained device with low RAM. *RAND** gives the possibility of computing an approximate result in a bounded amount of time, without however having any formal guarantees on the quality of the approximation. We believe that experimental results show that the quality of this approximation is, in practice, quite acceptable. Therefore, as we have demonstrated in [23], the simple RAND* algorithm can be implemented in a low-cost smart card. In the case where the *Trusted Cell* has enough computing power (in particular RAM) to run a MINLP solver, then this can be used in some practical cases, roughly speaking when there are less than a couple hundred predicates.

4.4 Future Work

We conclude with three directions that we are currently exploring : (1) improving the performance of the exact resolutions by linearizing the Binary Program (work with B. Le Cun from University of Paris-Nanterre, PRiSM Lab), (2) taking into account background knowledge, and multi-releases, leading to a 3-value logic approach using lattices (work with Y. Loyer from University of Versailles) and (3) using entropy based metrics to capture the dependency between predicates (work with S. Gambs from University of Rennes I).

Problem Linearization

The problem with our current definition of a *MinExp* AMPL Program is that it involves non-linear constraints. In consequence, we use a non-linear mixed integer solver to process the problem. This is problematic because non-linear solvers are not as efficient as linear ones.

We are currently studying the linearization of the general *Minimum Exposure* problem. Once linearized, larger instances of the problem may be solvable, and maybe it will be realistic to solve some small instances on a Secure Portable Token itself. Porting simple linear solvers to a Token is an even longer term possible direction of this work (since solvers often require a lot of memory, devising an efficient implementation on low constrained devices is a challenge).

Attacks on minimized forms

The current framework does not take into account a certain number of attacks that can be conducted such as the minimality attack, which is based on the knowledge that a solution is minimal, and therefore infer the truth value of some predicates (see [106, 107]). We do not take into account either the fact that possible many different releases will be made of personal data in various contexts, and that malicious service providers could try to cross this information. We are currently investigating a new model of the reduction problem, using 3-valued logic lattices([114]) and logical programming stable models([115]) which helps identifying the set of unrevealed predicates that can be inferred from a reduced form, considering different kinds of background knowledge. Such a model is a primary requirement to quantify the information really exposed by applicants publishing application forms obtained which can then be plugged into existing reduction algorithms, such as those proposed in our current framework. This approach could also lay the foundations of a new family of reduction algorithms, which would be based by design on the computation of their own inference.

Entropy Based Metrics

One of the current limitations of this work is that the exposure function $\mathbb{E}$ must be linear. While introducing a non-linear objective function is not necessarily a problem when using a MINLP solver (we show this in [30]), the exposure function must be meaningful. We are currently investigating how to best model the quality of a solution. Using the same lattice model that we are introducing in order to quantify attacks on the form, we can devise simple metrics, such as computing the number of elements in a sub-lattice which reflect more correctly the entropy of a solution. Integrating such metrics is however difficult in practice, due to the size of the lattice, since testing individually all solutions is computationally unfeasible. Therefore we propose to investigate entropy based metrics which can be computed in reasonable time over a 3-valued logic lattice.

Chapter 5

Conclusion and Future Research Perspectives

"I am not young enough to know everything."

– *Oscar Wilde*

Summary: *This chapter wraps up the main matter of this document by presenting directions for my future research.*

Conclusion

THIS document has provided an overview of a privacy-centric approach to data management. This approach is based on *trust.* As we have seen, the originality of our work is that we currently consider a highly secure hardware element, the Secure Portable Token (SPT), to effectively enforce protocols and algorithms. Let us stress that the security of the approach stems both from the hardware security of the SPT, and from the intrinsic *distribution* of the private data and computing, and we have introduced this novel paradigm, called *Trusted Cells* in Chapter 2. In Chapter 3 we have shown that despite the SPT's low power and low connectivity, it is possible to build complex applications, such as Privacy-Preserving Data Publishing (PPDP) using the SPT as computing element of a *Trusted Cell*, while not sacrificing much in terms of efficiency, compared to an *untrusted* computation on a powerful infrastructure. Finally, in Chapter 4, we have studied the problem of *Limited Data Collection*, which appears once private data starts to leave the trusted environment. In each of these chapters, we have given several open issues, specifically related to the works that we have conducted. I will now conclude this document, by presenting more general lines of future research that I wish to conduct, in order to continue to methodically and consistently study *Privacy-Centric Data Management.* My research project builds on complementary aspects : theoretical foundations, technical innovation and public adoption of the techniques. I will now discuss some of the related important scientific challenges I wish to tackle : (1) devising a privacy reference architecture which will be used to develop privacy-centric applications with formal guarantees, (2) studying various trust elements used to implement a *Trusted Cell*, and their impact on protocols we have already devised, (3) extending these protocols to different data management techniques (e.g. SQL queries), and finally (4) interdisciplinary analysis of privacy applications.

A Privacy Reference Architecture

Existing information systems are not designed with privacy as a primary requirement. Usually privacy appears as an afterthought or an optional feature that one can choose to incorporate or not into the system, although some recent works, such as [68] are emerging to enhance UML to capture the requirements of privacy aware systems. Indeed, personal data is very often stored on servers which are claimed to be secure because strict access control policies are enforced, protecting sensitive data through encryption mechanisms and keeping track of every access in audit logs. We have argued in this document that this is not enough and that in fact data subjects completely lose control over their data as soon as it is collected. Privacy by design aims at breaking this vicious circle by putting privacy at the heart of the architecture design. The objective of a privacy preserving architecture should be to control data at its source (by the data subject, from the very moment of its creation), to exchange data in a restricted environment and to ensure that access and usage remains under the control of the subject. On the one hand, the *Trusted Cells* paradigm is a first step in the direction of proposing a complete privacy *reference architecture*, which would incorporate a conception methodology, and a formal validation, in which expressed privacy requirements (i.e. access control, usage control, accountability, etc.) could be checked. On the other hand, atomic components, or individual "bricks", already exist to construct such systems, based on PETs (Privacy Enhancing Technologies) but they are often

designed with a precise objective in mind (e.g. anonymous web browsing, privacy-preserving national identity cards, secure portable medical folder). Hence, they lack of interoperability and they are often difficult to integrate in open worlds composed of several distributed systems and actors.

I have started to investigate this issue in the CAPPRIS project. I am working (with Daniel Le Métayer and Philippe Pucheral) on designing a *high level reference architecture* incorporating privacy by design principles, and formal proofs of privacy protection in data flows through modules. We hope that this work will lead to a better understanding of architectural problems linked to privacy protection, better coverage and easier development of solutions and increased interoperability between PETs themselves and between PETs and other services. This interoperability should also favor the development and adoption of PETs and privacy by design in the future. Studying the implementation of a privacy-by-design application on different architectures is another important aspect of designing such systems.

Hardware and Software for a *Trusted Cell*

One of my long term objectives is to convince application developers to use the *Trusted Cells* paradigm, in order to propose new applications managing personal data with a privacy-centric approach.

Trusted Cell Core. A fully fledged application development environment must be proposed to developers. This environment must propose both data-centric primitives, such as a database engine, and privacy enforcement. Currently, SMIS has developed a fully fledged relational database engine, running on a SPT. The next step is to define and study a core set of privacy primitives, i.e. algorithms enforcing privacy parameters defined by the reference architecture. As our study on PPDP has taught is, the main difficulty is to find a small set of primitives that will cover a sufficiently large set of applications. Once this set is defined, a thorough study of the security of these primitives, and their combinations must be done.

Trust elements. A SPT has very high trust : it is running certified code on secure hardware. Trust is therefore *global*, since the user trusts her SPT and can also trust *other* SPTs. In fact this means that the user trusts the SPT constructor, assembler, and software developer. We believe that a better level of trust that a user can have in her own hardware will be reached if she can build her own SPT, and run open-source (thus verifiable) software. Organising and promoting such a project is clearly an exciting objective I would like to pursue.

One could argue that running open-source (thus freely modifiable) code could lead to lesser trust in *other* SPTs. Mechanisms such as using signed code, or managing social trust [63] and reputation [151], could be used when deciding which SPTs to trust during a global computation. Adapting recommendation system algorithms to global computations provides an interesting research challenge.

Computing on the Asymmetric Architecture

The Asymmetric Architecture introduces interesting problems of private execution of global operations. The results proposed in this document show that it is possible to run global computations on this architecture, such as PPDP. During this study, we have proposed several privacy primitives, and algorithms to enforce them. It is important to adapt many different computing paradigms to the Asymmetric Architecture. A ever longer term goal would be to find a generic methodology to convert any kind of central computation to a *Trusted Cell* execution.

Global Queries on Asymmetric Architecture Executing SQL queries is a natural extension, since each cell embeds a database engine. Computing basic `SELECT ...FROM ...WHERE...` queries can be done with the existing primitives defined in $\text{MET}_{\mathbb{A}}\text{P}$, and `GROUP BY` queries can also be dealt with using the same primitives. However, a more difficult aspect is the (private) join operation, on tables distributed among several cells. As stated in Chapter 3, I am currently co-supervising a Ph.D. thesis, that started last year, on the topic of SQL execution on the Asymmetric Architecture.

Private distributed computing e.g. Map/Reduce Another example of interesting computing paradigm to adapt would be a "private" Map/Reduce infrastructure, where both mappers and reducers should operate with specific privacy guarantees (e.g. differential privacy guarantees). Current solutions (see [6, 7] tutorials for more references) focus on querying encrypted data, while I would like to investigate an approach using large quantities of low powered trusted hardware. In this context, providing correct execution and protection against malicious participants is an open research problem.

XML management Current work in the SMIS team has been focused on relational databases. Semi-structured (XMl) data management would be an interesting extension to the core functionality of *Trusted Cells.* However, even the most lightweight XQuery engines (e.g. MXQuery[48]) needs more computing power than that available in an SPT. Devising a very low footprint XQuery (or XPath) engine would therefore prove very useful to improving the core capacities of the cells. As in the case of the relational engine, this is not simply an engineering problem : Data storage, access and query evaluation must be completely revisited given the specific low RAM and low power of the SPT.

Interdisciplinary Analysis of Privacy Applications

I have had many interactions and collaborations with two other disciplines : sociology, with the *WebStand* project and law, with the *Demotis* project. Similarly, the *Trusted Cells* approach has stemmed interest in economics, law and sociology. Indeed, although we have demonstrated technically the feasibility of a prototype system, we must now study with *other disciplines* the

[48] http://mxquery.org/

useability, applicability, legal and *societal impact* of the *Trusted Cells* paradigm. One example would be launching field experiments with real users, who would test an SPT implementation of a *Trusted Cell*, and privacy friendly applications, where currently only very intrusive smartphone applications are available. It is only by validating that our system can actually be used, the we will be convinced that it is possible to return control of private data to its legitimate owner. This is one of the goals of the Privacy Working Group of the *Institut de la Société Numérique*, that has launched this summer. I am co-chairing this Working Group with Fabrice Le Guel, an Economist from University of Paris-XI, and we are expecting collaborations through the means of dual Ph.D. students (one in computer science and one in economics) working on the topic of usage control via trusted cells, each with their discipline's point of view.

Appendix A

Ph.D. Students

I have co-supervised 3 Ph.D. students, and am currently supervising a fourth thesis.

Ivan Bedini [41]

Ivan Bedini defended his thesis, entitled *Deriving Ontologies automatically from XML Schemas applied to the B2B domain*, on January 15^{th} 2010. This Ph.D. was co-supervised at 50% with Pr. Georges Gardarin. It has lead to the following communications or publications : [42, 48, 47, 49, 50, 44, 46, 51] and to the *Janus* software [43], an automatic ontology builder (approx. $30K$ lines of code).

Ivan Bedini was senior researcher at Alcatel-Lucent, Bell Labs, Ireland from 2010 to 2013, and now senior researcher at Trento RISE the Italian EIT ICT Labs node.

Bogdan Butnaru [58]

Bogdan Butnaru defended his thesis, entitled *Optimizations of XQuery in Peer-to-Peer distributed databases*, on April 12^{th} 2012. This Ph.D. was co-supervised at 50% with Pr. Georges Gardarin. It has lead to the following communications or publications : [59, 59, 87, 61], the XQ2P software [62], a fully distributed, fully compliant XQuery processor (approx. $50K$ lines of code) and the P2PTester platform (approx. $20K$ lines of code).

Bogdan Butnaru is currently Project Leader at JAWS Software, Paris.

Tristan Allard [11]

Tristan Allard defended his thesis, entitled *Sanitizing Microdata Without Leak: a deventralized approach*, on December 17^{th} 2011. This Ph.D. was co-supervised at 50% with Pr. Philippe

Pucheral. It has lead to the following communications or publications : [1, 13, 12, 14, 15, 17, 16, 18, 19, 20, 21].

Tristan Allard is currently post-doc in the INRIA Zenith Team.

Cuong Quoc To

Cuong Quoc To is working since october 2012 on his Ph.D. thesis, co-supervised at 50% by Pr. Philippe Pucheral, on the topic of *Secure Global Computations on Personal Data Servers*. Preliminary results have been given in [145, 146, 147].

Appendix B

Curriculum Vitae

Dr. Benjamin NGUYEN — Curriculum Vitae – August 2013

Personal Information

Benjamin Quý Vinh NGUYEN MIHÉ
36 years old. French Citizen, British Citizen. Married, 1 child.

Contact Information

Projet SMIS – INRIA, Université de Versailles St-Quentin and CNRS
Laboratoire PRiSM, UVSQ, UMR 8144
45 av. des Etats-Unis
78035 Versailles Cedex
France

Mobile : +33-6-86-63-25-27
Landline : +33-1-39-25-40-49
Email : benjie.nguyen@gmail.com
Web : www.prism.uvsq.fr/~beng/

Current Situation

Associate Professor (Maître de Conférences, Classe Normale), Computer Science Section (CNU 27) at Université de Versailles St-Quentin (UVSQ) since September 2004. Member of the *Secured and Mobile Information Systems* (SMIS) joint-team (UVSQ, INRIA, CNRS) since January 2010.

Research Theme

My current research focuses on **Privacy & Security in Information Management Systems and Applications.** More specifically, I am interested in (a) methods to enforce existing privacy models using secure hardware devices, (b) design and implementation of large scale privacy-by-design personal information management applications, and (c) models to represent, quantify and enforce limited data collection.

Keywords

Databases, Personal Information, Privacy, Private Distributed Computation, Interdisciplinary applications of databases (Sociology, Law, Economics), Secure Hardware.

Past Research

Distributed XML (XQuery) Databases, Semantic Web and Sociological Applications.

Education

Ecole Normale Supérieure de Cachan (ENS-Cachan, French Top School for research careers), **Université de Paris-Sud** and **Institut National de Recherche en Informatique et Automatique** (INRIA)

Details

- **Doctor in Computer Science**, Université de Paris-Sud, December 2003
  Ph.D. dissertation : *Construction and Maintenance of a Web Warehouse*.
  Advisor : Serge Abiteboul, Senior Researcher (Directeur de Recherches) INRIA.
  Defended on December 17^{th} 2003, *with honours.*
- **Ecole Normale Supérieure de Cachan**, Physics Department, September 1996 - August 2000.
- **Magistère de Physique Fondamentale (Theoretical Physics)**, at Université de Paris Sud, September 1996 - August 2000.
  BA and MSc in Theoretical Physics, with honours.
  MSc in Computer Science, with distinction (First).
- **Classes préparatoires scientifiques** (2-3 year intensive program preparing the national competitive exam for entry to engineering schools) at Lycée Henri IV (Paris, France) – Option P' (September 1994- June 1996).
- **Baccalauréat C** (High School degree), with distinction (July 1994).

Scientific Career

- 2010-. . . : Associate Professor at Université de Versailles St-Quentin/INRIA/CNRS, PRiSM Laboratory, INRIA Team, Secured and Mobile Information Systems (SMIS), lead by Pr. Philippe Pucheral.
- 2004-2010 : Associate Professor at Université de Versailles St-Quentin/CNRS, PRiSM Laboratory, Data Integration and Management (DIM), lead by Pr. Georges Gardarin.
- 2003-2004 : Research and Teaching Assistant (ATER) at University of Paris-Sud (Team Gemo). Visiting researcher at Athens University of Economics and Business (Pr. Michalis Vazirgiannis).
- 2000-2003 : Ph.D. at INRIA-Rocquencourt Team Verso then INRIA-Futurs Team Gemo, lead by Serge Abiteboul, INRIA Senior Researcher.

Ph.D. Supervision

- To Quoc Cuong, on the topic "Secure Global Computations on Personal Data Servers", started on October the 1st, 2012, supervisor Pr. Philippe Pucheral. My supervision participation : 50%.
- Bogdan Butnaru, defended on April 12th, 2012, entitled "Optimizations of XQuery in peer-to-peer distributed XML Databases", supervisor Pr. Georges Gardarin. My supervision participation : 50%.
- Tristan Allard, defended on December 17th, 2011, entitled "Sanitizing Microdata Without Leak : a Decentralized Approach", supervisor Pr. Philippe Pucheral. My supervision participation : 50%.
- Ivan Bedini, defended on January 15th, 2010, entitled "Deriving ontologies automatically from XML Schemas applied to the B2B domain", supervisor Pr. Georges Gardarin. My supervision participation : 50%.

Research Projects

The *Agence Nationale de la Recherche* (ANR) is the French national research funding agency. The ACI projects are previous French national funding programs. The RNTL (Réseau National en Technologies Logicielles) projects are very large scale projects involving important software developments.

SMIS Team

- "Institut de la Société Numérique" (ISN) of Paris Saclay (Digital Society Institute), since July 2013 – **Co-coordinator** of the interdisciplinary Privacy Working Group (Computer Science, Economics, Law, Sociology).
- INRIA Project Lab "Collaborative Action on the Protection of Privacy Rights in the Information Society" (CAPPRIS), started January 2012. – **Leader** of the Joint Task "Privacy by Design Architecture".
- ANR "Keeping your Information Safe and Secure" (KISS), started December 2011. – **Contributor** to the Access and Usage Control task, the Distributed Services task, and the Demonstrator task.
- ANR "Définir, Evaluer et Modéliser les Technologies de l'Information de Santé" (DEMOTIS, *Defining, Analysing and Modelling Electronic Health Record Systems*) from January 2010 to January 2012. – **Leader** of the EHR data anonymization task. Contributions in the context of Tristan Allard's Ph.D. thesis.

DIM Team

- ANR "Une plateforme de gestion de données Web pour applications sociologiques" (WebStand, *Platform for Web Data Management for Sociological Applications*), from January 2005 to July 2009. – **Project Leader and Coordinator**. Interdisciplinary project with sociologists of of the Laboratoire d'Economie et Sociologie du Travail (LEST, *Economy and Industrial Sociology Laboratory*), Université d'Aix-Marseille on the use of XML databases to study and analyze interaction data, with an application to mailing lists. The project was distinguished by section 40 ("Politics, Authorities and Organisations") of the CNRS during the evaluation of LEST Lab in 2011.
- ANR-MDCO Project "Really Open and Simple Web Syndication" (ROSES) (2008-2011) – **Contributor** on P2P management of XML data, and in particular stream management and temporal series, in the context of Bogdan Butnaru's Ph.D. thesis.
- RNTL Project "WebContent" from May 2006 to December 2009 – **Member of the steering committee**, and **leader** of the work package "Centralized XML Store". **Contributor** to the architecture, and XML storage system.
- ACI-MD Project "Using XQuery to query the Semantic Web" (SemWeb) (July 2004-July 2007) – **Contributor** on Web Services, XQuery and Ontologies. This work was pursued during Ivan Bedini's Ph.D. thesis.
- ACI Project "Normes et Régulations des Politiques Publiques" (NPP, "Standards and Regulations of Public Policies") from 2004 to 2007 – **Leader** of the interdisciplinary work package with sociologists from the University de Paris-Dauphine on the profiling of information technology standard makers (W3C).

Verso / Gemo Teams

- "Xyleme" project : a large scale INRIA project which became a company (xyleme.com), on the management of XML data (2000-...) – **Leader** of the XML Monitoring task.
- European IST/FET Project "DB/GLOBE". Overall 3 months visits to Athens University of Economics and Business (2001-2004) – **Contributor** on distributed databases on the Web.
- RNTL Project "Entrepôts de Données Ouvert sur la Toile" (e.dot, *Web Data Warehouses*) – **Contributor** on the construction of a XML warehouse on food risk (January 2003 - December 2005).

International Collaborations

- Partner of the "Learning Techniques for Large and Evolving Networked Data" (LeTeVoNe) DIGITEO Chair of Pr. Michalis Vazirgiannis from Athens University of Economics and Business, since January 2010. – Participants are UVSQ, Ecole Polytechnique, Telecom ParisTech, and the Exalead company. My contributions revolve around the *Minimum Exposure* problem, i.e. privacy protection for individuals whose data is analyzed by data mining algorithms.
- "Bonus Qualité Internationale" (BQI) Project at UVSQ (with Pr. Alain Bui and Thu Ha Dao Thi) to organize a MSc level collaboration with many Vietnamese universities (2010-2011). Preparation of a common Masters program with the John Von Neumann Institute of the Vietnam National University of Ho Chi Minh City, and visits of researchers from the Ho Chi Minh University of Technology (HCMUT, ex. Ecole Polytechnique de Ho Chi Minh).
- Organization of the ERASMUS program with Athens University of Economics and Business (AUEB) and UVSQ since 2006.
- Participant in the European Project "DB-Globe" between University of Ioannina, Computer Technology Institute, University of Cyprus, Athens University of Economics and Business and INRIA in 2003/2004.

Administrative & Scientific Service

Université de Versailles St-Quentin

- Creator (in 2010) and Manager (since 2010) of the in-service training of high school teachers for the Terminale S (12^{th} grade) speciality "Informatique et Sciences du Numérique" (Computer Science) at UVSQ.
- Elected member of the Scientific Committee of the Sciences Faculty in 2007, reelected in 2011.
- Associate Director of the Computer Science department of UVSQ from September 2006 to September 2010.
- Manager of the Bachelors program in Computer Science from September 2008 to September 2010. Coordinator and writer (2009) of the "Dossier d'Habilitation" (Application File for the certification by the Ministry of Higher Education) of the Computer Science Bachelor's degree for 4 years.
- Manager of the Computer Science Masters program, since September 2011. Coordinator of the "Dossier d'Habilitation" in 2013.
- Co-representative of UVSQ (with Pr. Dominique Barth) in the Masters Commission of the Digicosme Labex since 2013.
- Technical manager of the Microsoft Developer Network Academic Alliance (MSDNAA) at the Computer Science Department since May 2005.

INRIA

- Member of the Post-doc and delegations committee in 2010 and 2011.

General Service

I have written many external reviews for top level international database conferences(e.g. SIGMOD, VLDB, ICDE or EDBT) and journals. I have also been chair or member of several program committees.

- Chair and Co-organiser with Nicolas Anciaux of the *4th Atelier Protection de la Vie Privée* (APVP), Les-Loges-en-Josas, 2013.
- Chair of the *29e Bases de Données Avancées (BDA)*, Demonstration Track, Nantes 2013.
- Co-organisor with Karine Zeitouni, Daniela Grigori, Stéphane Lopes and Tao Wan of the *2e Journées Francophones sur les Entrepôts de Données et l'Analyse en Ligne* (EDA), 2006

International Conference Program Committees

- International Conference on Very Large Databases (VLDB) 2012.
- IEEE International Conference on Data Engineering (ICDE) 2011.
- IEEE International Conference on Data Engineering Demonstrations track (ICDE-Demo) 2010.
- International Conference on Very Large Databases Demonstrations track (VLDB-Demo) 2009.
- European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD) 2011.
- International Conference on Computer Science, Applied Mathematics and Applications (ICCSAMA) 2013.
- International Conference on Advanced Computing and Applications (ACOMP) 2013.

Editorial Committee

- *Techniques et Sciences Informatiques*, Lavoisier Editions (since January 2012). Oldest French Computer Science journal (Founded in 1981).

National Conference PCs

- Member of the Atelier sur la Protection de la Vie Privée (APVP) steering committee, since 2012.
- Journées Bases de Données Avancées (BDA) 2009, 2010, 2011, 2012
- Journées Francophones sur les Entrepôts de Données et l'Analyse en Ligne (EDA) 2006 (co-organisor with Karine Zeitouni, Daniela Grigori, Stéphane Lopes and Tao Wan), 2007, 2008, 2009, 2010, 2011, 2013.

Journal Reviewer

- Communications of the ACM (CACM).
- Journal of Web Semantics (JWS).
- Very Large Databases Journal (VLDBJ).

- Information Systems (IS).
- Techniques et Sciences Informatiques (TSI) - French journal.
- Revue des Nouvelles Technologies de l'Information (RNTI) - French journal.
- Information, Interaction, Intelligence Journal (I3) - French journal.

Ph.D. Committees

- Johann Vincent, *Identité Numérique en Contexte Telecom*, Université de Caen Basse Normandie, June 6^{th}, 2013.
- Noor Malla, *Partitioning XML data, towards distributed and parallel management*, Université de Paris Sud, September 21^{th}, 2012.
- Bogdan Butnaru, *Optimizations of XQuery in peer-to-peer distributed XML databases*, Université de Versailles St-Quentin-en-Yvelines, April 12^{th}, 2012.
- Tristan Allard, *Sanitizing Microdata Without Leak : a Decentralized Approach*, Université de Versailles St-Quentin-en-Yvelines, December 12^{th}, 2011.
- Ivan Bedini, *Deriving ontologies automatically from XML Schemas applied to the B2B domain*, Université de Versailles St-Quentin-en-Yvelines, January 15^{th}, 2010.

Selection Committees & Expertise

Selection Committees are commissions in charge of hiring permanent researchers or post-doctoral researchers. All committees were for permanent researcher positions except Université de Paris-X Nanterre from 2006 to 2009.

- Université de Paris-VI (2012, 2013).
- Université de Paris Sud (2011).
- Université de Paris Dauphine (2011, 2013).
- Université de Versailles St-Quentin (2010, 2011).
- Université de Paris-X Nanterre (from 2006 to 2010).
- INSA Lyon (2012).
- ANR Expert(since 2009).
- Institut Mines-Telecom Expert (2013).

Other

- Member of the "Travaux d'Initiative Personnelle Encadrés" (TIPE) global jury for top french engineering schools from 2006 to 2010.

Technology Transfer, Dissemination

- XQuery on Peer-to-Peer (http ://cassiopee.prism.uvsq.fr/XQ2P/) project leader. XQ2P is a distributed XQuery database, 98.7% compliant to the official spec.
- *Advisory Commitee* Representative of UVSQ at the World Wide Web Consortium (W3C-Web Standardization Agency) from September 2008 to April 2011.
- W3C *XQuery* Working Group member from 2008 to 2011. Editor of the XQuery 3.0. Test Suite from 2010 to 2011.
- W3C *e-government* Interest Group Member (eGOV IG) from 2008 to 2009 and *Semantic Web Best Practises and Deployment* (SWBPD WG) Member from 2004 to 2006.
- Invited conference for a W3C *W3C Project Review* to present the *WebStand* Project to the W3C Director, April 2009.
- Xylème Project (XML Database System) : Design and implementation of the XML Pub/Sub system. Module commercialized by the start-up company in 2001.
- A Greek Patent (See bibliography).
- Many general public conferences (See bibliography : *Invited Conferences*).

MSc and *Grandes Ecoles* Students

- 4 MSc Student final thesis : T. Allard (COSY, 2007), N. Ababsa (COSY, 2008) H.T. Tran (Institut Francophone de l'Informatique, Viêt-Nam, 2011), Q.C. To (Hô Chi Minh City University of Technology, Viêt-Nam, 2012).
- 1 final research internship of Ecole Polytechnique : D. Boutara (2012).
- 2 final research internship of Ecole Nationale Supérieure des Mines de Nancy : L. Saint-Ghislain (2008), R. Vincent (2008).
- 2 research internship of Ecole Nationale Supérieure des Techniques Avancées (ENSTA) : W. Bezza (2012), M. Fazouane (2013).
- 1 "Projet Scientifique Collaboratif" (Collaborative Scientific Project) of 6 Ecole Polytechnique students : P-J. Bringer, V. Brillault, T. Chen, C. Dross, B. Gelineau, C. Zhu (2008/2009).
- 4 MSc first year student internships : D. Hadjout (UVSQ, 2007), A. Biane (UVSQ, 2010), M. Nasri (UVSQ, 2010), D. Maheswari (Indian Institute of Technology, Delhi, 2011).

Student Competitions

- Coach of the "Home Energy Optimizer" (HEO) Team of UVSQ and INP Grenoble, which reached the French Final (top 5 teams) in the Software Design category of the 2010 Imagine Cup 2010 (Microsoft).
- Co-supervision (with Béatrice Finance) of a group of 2 MSc students of UVSQ for the SIGMOD 2010 programming competition.

TEACHING / TRAINING

Since September 2004, I have been teaching at UVSQ for at least 192h per year, except in 2010/2011 where I was on sabbatical ("Congé pour Recherches et Conversion Thématique" or CRCT) at SMIS, 2011/2012 and 2012/2013 where I had a half year delegation at INRIA. I have also taught in various French top level engineering schools ("Grandes Ecoles") : the Ecole Polytechnique, Telecom-ParisTech, l'ESCP-EAP (Sup de Co. Paris, a top level business school), and also various professional training courses. For all courses listed next, unless specified, I directly elaborated the contents (lectures, exercises and labs). All course volumes are annual. Number of years I have given this course is also indicated.

In-service Training

- *PureXML Technology* days at IBM (2010). Introduction to XQuery and example applications of XML/XQuery in research projects (approx. 30 people).
- Creator (in 2010) and Manager (since 2010) of the in-service training of high school teachers for the Terminale S (12^{th} grade) speciality "Informatique et Sciences du Numérique" (Computer Science) at UVSQ. Joint elaboration of the contents with the "Délégation Académique à la formation des personnels de l'Education Nationale" (DAFPEN). Over 100 teachers have been trained since 2010. This training course spans 28 days over a total duration of 2 years, to get high school mathematics teachers up to speed on computer science.

Masters Level

- DBMS Internals (48h - 3×) : *Analyse et Conception de Systèmes d'Information Sûrs* Masters (UVSQ).
- XML Technology(48h - 3×) : *Analyse et Conception de Systèmes d'Information Sûrs* Masters (UVSQ).
- Java Programming (28h - 3×) : *Analyse et Conception de Systèmes d'Information Sûrs* Masters (UVSQ).
- Web Services Security (6h - 2×) : *Des Concepts aux Systèmes* research Masters (UVSQ).
- Database Technology (24h - 2×) : *Systèmes Informatiques en Réseau* Masters at Telecom-ParisTech, english course.
- XML/XSL/XQuery (24h - 2×) : final year students at Telecom Paris-Tech.
- Semantic Web (3h - 1×) : Masters at ESCP-EAP.

Bachelors level

- Basic Computer Science (54h - 5×) : 1st year students (UVSQ).
- Basic Computer Science (32h - 3×) : 1st year students (Univ. Paris Sud), *tutoring and lab work only.*
- Object Oriented Analysis, Conception and Programming (54h - 7×) : 2nd year students (UVSQ).
- Graph Theory (72h - 2×) : 3rd year students (UVSQ).
- Databases and Web Programming (54h - 8×) : 2nd year students (UVSQ)
- Using Database Tools in Social Science (36h - 3×) : 3rd year *social science* students L3 (UVSQ, Sociology and Humanities Faculty).
- Advanced Databases (24h - 1×) : 2nd year students at Ecole Polytechnique, *lab work only.*
- Algorithmics (36h - 3×) : 3rd year students (Univ. Paris Sud), *tutoring only.*
- Logics and Hardware (18h - 2×) : 3rd year students (Univ. Paris Sud), *lab work only.*

Lycée (High School)

- Physics teacher for 1 year for a class of *Première S* (Scientific 11th grade) at Lycée de la Vallée de Chevreuse, France (1996-1997).

Appendix C

Personal Bibliography

Overview

Top level conferences in the Database domain are equivalent to A or A* journals, with acceptation ratios under 15%, like VLDB, SIGMOD, ICDE, etc. where demonstration papers are also highly selective. I have also conducted interdisciplinary work with sociologists and jurists, and therefore published in their respective forums. In these disciplines, French national conferences are often regarded as of high difficulty and quality (e.g. Congrès de l'Association Française de Sociologique). Whenever information was available, I have indicated the CORE ranking, or acceptance ratio. All publications except those listed as invited papers have undergone a peer-review process.

Publication Type	**Total**
Books & Proceedings	2
Bookchapters	4
International Journals	7
French Journals	2
International Conference Papers and Demos	16
Posters in International Conferences	3
International Workshops	9
French Conferences or Workshops	17
Invited Conferences	10
Patents	1

Books or Proceedings

Jean-Pierre Archambault, Emmanuel Baccelli, Sylvie Boldo, Denis Bouhineau, Patrick Cégielski, Thomas Clausen, Gilles Dowek, Irène Guessarian, Stéphane Lopès, Laurent Mounier, Benjamin Nguyen, Franck Quessette, Anne Rasse, Brigitte Rozoy, Claude Timsit, Thierry Viéville, Jean-Marc Vincent. *Une introduction à la science informatique: Pour les enseignants de la discipline informatique au lycée*, Editions CRDP, ISBN13: 978286631188-9, 376p, 2011.

Daniela Grigori, Stephane Lopes, Benjamin Nguyen, Karine Zeitouni. *Entrepôts de données et Analyse en ligne, Actes de la conférence EDA'2006*, Numéro spécial de la Revue RNTI, Editions Cépaduès, 194p, 2006.

Bookchapters

Ivan Bedini, Benjamin Nguyen, Christopher Matheus, Peter F. Patel-Schneider, Aidan Boran, *Mining XML Schemas to Extract Conceptual Knowledge*, in *Semi-Automatic Ontology Development: Processes and Resources*, Maria Teresa Pazienza and Armando Stellato eds., IGI Global Publishing, ISBN13: 9781466601888, pp79-105, 2012.

Ivan Bedini, Georges Gardarin, Benjamin Nguyen, *Semantic Technologies and e-business*, in *Electronic Business Interoperability: Concepts, Opportunities and Challenges*, Ejub Kajan ed., IGI Global Publishing. ISBN13: 9781609604851, pp243-278, 2011.

Serge Abiteboul, Benjamin Nguyen, Gabriela Ruberg. *Building an Active Content Warehouse*, in *Processing and Managing Complex Data for Decision support*, Jérôme Darmont, Omar Boussaïd editors, IDEA Group Publishing. ISBN159140656-0, 2006.

Bernd Amann, Salima Benbernou, Benjamin Nguyen. *Web services: Technology issues and foundations*, in *Web Data Management Practices : Emerging Techniques and Technologies*, Athena Vakali and George Pallis editors, IDEA Group Publishing, ISBN159904228-2, 2006.

International Journals

Tristan Allard, Benjamin Nguyen, Philippe Pucheral, MET$_{\mathbb{A}}$P *: Revisiting Privacy-Preserving Data Publishing using Secure Devices*, to appear in *Distributed and Parallel Databases* (DAPD), 2013. (CORE : A).

Nicolas Anciaux, Danae Boutara, Benjamin Nguyen, Michalis Vazirgiannis, *Limiting Data Exposure in Multi-Label Classification Processes*, to appear in *Fundamenta Informaticae*, 2013. (CORE : B).

Nicolas Anciaux, Benjamin Nguyen, Michalis Vazirgiannis, *The Minimum Exposure Project : Limiting Data Collection in Online Forms*, in *ERCIM News*, vol 90, pp. 41-42, 2012.

Benjamin Nguyen, Antoine Vion, François-Xavier Dudouet, Dario Colazzo, Ioana Manolescu, Pierre Sennelart, *XML Content Warehousing : Improving Sociological Studies of Mailing Lists and Web Data*, in *Sociological Methodology Bulletin* (BMS), SAGE ed., vol 112(1), pp. 5-31, 2011. (AERES Political Science "A" Rank).

Iraklis Varlamis, Michalis Vazirgiannis, Maria Halkidi, Benjamin Nguyen, *THESUS, a Closer View on Web Content Management Enhanced with Link Semantics.* In *IEEE Transactions on Knowledge and Data Engineering*, vol. 16(6), pp. 685-700, 2004. (CORE : A).

Maria Halkidi, Benjamin Nguyen, Iraklis Varlamis, Michalis Vazirgiannis, *THESUS: Organizing Web document collections based on link semantics.* In *the VLDB Journal*, vol. 12(4), pp. 320-332, 2003. (CORE : A*).

Lucie Xylème[49], *Xyleme: A dynamic Warehouse for XML of the Web.* In *IEEE Data Engineering Bulletin*, vol. 24(2), pp. 40-47, 2001.

French Journals

Nicolas Anciaux, Benjamin Nguyen, Michalis Vazirgiannis, *Exposition Minimum de Données pour des Applications à Base de Classifieurs*, to appear in *Ingénierie des Systèmes d'Information* (ISI), 2013.

Benjamin Nguyen, Antoine Vion, François-Xavier Dudouet, Dario Colazzo, Ioana Manolescu, *WebStand, une plateforme de gestion de données Web pour applications sociologiques.* In *Revue de Technique et Science Informatiques* (TSI), numéro spécial sur L'informatique à l'Interface des Sciences Humaines et Sociales, vol. 29(8-9), pp. 1055-1080, 2010.

International Conferences & Demos

Nicolas Anciaux, Philippe Bonnet, Luc Bouganim, Benjamin Nguyen, Iulian Sandu Popa, Philippe Pucheral, *Trusted Cells: A Sea Change for Personal Data Services*, in 6^{th} *Biennial Conference on Innovative Database Research* (CIDR), 2013. (CORE : A).

Nicolas Anciaux, Walid Bezza, Benjamin Nguyen, Michalis Vazirgiannis, *MinExp-Card: Limiting Data Collection Using a Smart Card*, in 16^{th} *International Conference on Extending Database Technology* (EDBT), demonstration, pp753-756, 2013. (CORE : A).

Nicolas Anciaux, Benjamin Nguyen, Iulian Sandu-Popa, *Personal Data Management with Secure Hardware : the Advantage of Keeping your Data at Hand*, in *IEEE* 14^{th} *International Conference on Mobile Data Management* (MDM), 2*h Tutorial*, 2013.

Nicolas Anciaux, Benjamin Nguyen, Michalis Vazirgiannis, *Limiting Data Collection in Application Forms : A real-case application of a Founding Privacy Principle*, in *IEEE* 10^{th} *Annual Conference on Privacy, Security and Trust* (PST), 8p., 2012. (full paper acceptance 25%).

Ivan Bedini, Benjamin Nguyen, Christopher Matheus, Peter F. Patel-Schneider, Aidan Boran. *Transforming XML Schema to OWL Using Patterns*, in *IEEE* 5^{th} *International Conference on Semantic Computing* (ICSC), 8p., 2011. (acceptance 21%).

Tristan Allard, Benjamin Nguyen, Philippe Pucheral, *Safe Realization of the Generalization Privacy Mechanism*, in *IEEE* 9^{th} *Annual Conference on Privacy, Security and Trust* (PST), 8p., 2011. (**Best paper award.**, full paper acceptance 25%).

Tristan Allard, Benjamin Nguyen, Philippe Pucheral. *Sanitizing Microdata Without Leak: Combining Preventive and Curative Actions*, in *7th Information Security Practice and Experience Conference* (ISPEC), 10p. 2011. (CORE : B).

[49]Lucie Xyleme is a nickname for a large group of people who worked on the project: S. Abiteboul, V. Aguilera, S. Ailleret, B. Amann, F. Arambarri, S. Cluet, G. Cobena, G. Corona, G. Ferran, A. Galland, M. Hascoet, C-C. Kanne, B. Koechlin, D. Le Niniven, A. Marian, L. Mignet, G. Moerkotte, B. Nguyen, M. Preda, M-C. Rousset, M. Sebag, J-P. Sirot, P. Veltri, D. Vodislav, F. Watez and T. Westmann.

Tristan Allard, Nicolas Anciaux, Luc Bouganim, Yanli Guo, Lionel Le Folgoc, Benjamin Nguyen, Philippe Pucheral, Indrajit Ray, Indrakshi Ray, Shaoyi Yin. *Secure Personal Data Servers: a Vision Paper*, in 36^{th} *International Conference on Very Large Data Bases* (PVLDB), vol. 3(1), pp. 25-35, 2010. (CORE : A).

Ivan Bedini, Georges Gardarin, Benjamin Nguyen. *B2B Automatic Taxonomy Construction*, in 10^{th} *International Conference on Enterprise Information Systems* (ICEIS), pp.325-330, 2008. (CORE : C)

Ivan Bedini, Benjamin Nguyen, Georges Gardarin. *Janus: Automatic Ontology Builder from XSD Files.* In 17^{th} *International World Wide Web Conference* (WWW) Developer Track, 2008. (CORE : A).

Serge Abiteboul, Tristan Allard, Philippe Chatalic, Georges Gardarin, Anca Ghitescu, Francois Goasdoué, Ioana Manolescu, Benjamin Nguyen, Mohamed Ouazara, Aditya Somani, Nicolas Travers, Gabriel Vasile, Spyros Zoupanos. *WebContent: Efficient P2P Warehousing of Web Data*, in 34^{th} *International Conference on Very Large Data Bases* Demonstration Track (PVLDB), vol. 1(2), pp. 1428-1431, 2008. (CORE : A).

Bogdan Butnaru, Florin Dragan, Georges Gardarin, Ioana Manolescu, Benjamin Nguyen, Radu Pop, Nicoleta Preda, Laurent Yeh. *P2PTester: a tool for measuring P2P platform performance*, in *IEEE* 23^{rd} *International Conference on Data Engineering* (ICDE) Demonstration Track, pp. 1501-1502, 2007. (CORE : A).

François-Xavier Dudouet, Ioana Manolescu, Benjamin Nguyen, Pierre Senellart. *XML Warehousing Meets Sociology*, in *Proceedings of the IADIS International Conference on WWW/Internet* (ICWI), pp. 170-175, 2005. (acceptance 22%).

Serge Abiteboul, Vikas Bensal, Grégory Cobéna, Benjamin Nguyen and Antonella Poggi, *Model, Design and Construction of a Service-Oriented Web Warehouse*, in *7th European Conference on Digital Libraries* (ECDL), demonstration, Lecture Notes in Computer Science, Volume 2769, pp529, 2003. (acceptance 29%).

Benjamin Nguyen, Serge Abiteboul, Grégory Cobena, Mihaí Preda. *Monitoring XML Data on the Web*, in Proceedings of the ACM Special Interest Group on the Management of Data Conference (SIGMOD), vol. 30(2), pp. 437-448, 2001. (CORE : A).

Lucie Xylème[1], *A dynamic Warehouse for XML data of the Web*, in *International Database Engineering and Applications Symposium* (IDEAS), pp 3-7, 2001. (CORE : B).

Workshops & Posters

Nicolas Anciaux, Danae Boutara, Benjamin Nguyen, Michalis Vazirgiannis, *Limiting Data Exposure in Multi-Label Classification Processes*, in *International Workshop on Privacy-AwaRe Intelligent Systems* (PARIS), 2012.

Nicolas Anciaux, Benjamin Nguyen, Michalis Vazirgiannis, *Minimum Exposure*, in *Digiteo Workshop on Web Mining*, 2011.

Tristan Allard, Benjamin Nguyen, Philippe Pucheral. *Towards a Safe Realization of Privacy-Preserving Data Publishing Mechanisms.* In *12th International Conference on Mobile Data Management* Ph.D. Colloquium (MDM-PhD), 4p. 2011.

Benjamin Nguyen, Spyros Zoupanos. *The WebContent Store.* In *Atelier Sources Ouvertes et Services, Conférence en Reconnaissance de Formes et Intelligence Artificielle* (RFIA), 2010.

François-Xavier Dudouet, Benjamin Nguyen, Antoine Vion. *The governance of web standards. Economic struggles in the XML case.* In *Second International Workshop on Global Internet Governance: An Interdisciplinary Research Field in Construction*, 2009.

Benjamin Nguyen, Antoine Vion, François-Xavier Dudouet, Ioana Manolescu, Dario Colazzo, Pierre Senellart, *The WebStand Project.* In *WebSci'09 - The Web Science Overlay Journal* (Poster), 2009.

Benjamin Nguyen, Antoine Vion, François-Xavier Dudouet, Loïc Saint-Ghislain. *Applying an XML Warehouse to Social Network Analysis, Lessons from the WebStand Project*, in *W3C Workshop on the Future of Social Networking*, 5p. 2009.

Ivan Bedini, Georges Gardarin, Benjamin Nguyen. *Janus: Automatic Ontology Construction Tool. International Conference on Knowledge Engineering and Knowledge Management Knowledge Patterns* (EKAW), poster, 2008.

François-Xavier Dudouet, Benjamin Nguyen, Antoine Vion. *New web standards in the making: Transnational private governance and beyond.* In *Global Internet Governance Academic Network* (GigaNet) Third Annual Symposium, 29p. 2008.

Serge Abiteboul, Ioana Manolescu, Benjamin Nguyen, Nicoleta Preda. *A Test Platform for the INEX Heterogeneous Track.* In *Proceedings of the International Workshop of the Initiative for the Evaluation of XML retrieval* (INEX), 2004.

Benjamin Nguyen, Iraklis Varlamis, Maria Halkidi, Michalis Vazirgiannis, *Organization of Web Document Collections Based on Link Semantics*, in *7th European Conference on Digital Libraries* (ECDL), poster, Lecture Notes in Computer Science, Volume 2769, pp533, 2003.

Benjamin Nguyen, Serge Abiteboul, Grégory Cobéna, Laurent Mignet, *Query Subscription in an XML Webhouse*, in *DELOS Workshop: Information Seeking, Searching and Querying in Digital Libraries*, 2000.

Invited Conferences

Benjamin Nguyen, *Databases in the Classes Préparatoires aux Grandes Ecoles*, Luminy seminar for CPGE teachers, 2013.

Benjamin Nguyen, *Les études d'Informatique à l'Université*, Lycée Joliot-Curie (Nanterre), 2013.

Jean-Pierre Archambault, Laurent Bloch, Laurent Chéno, Benjamin Nguyen, *Logiciels libres, culture et enseignement de l'informatique*, Round Table at the *5th Open World Forum* (OWF), 2012.

Daniel Le Métayer, Benjamin Nguyen, *Le projet CAPPRIS*, dans *les mardis du CRIDS* (Université de Namur), 2012.

Benjamin Nguyen, Franck Quessette, *Training High School Teachers in Computer Science, a first experiment at University of Versailles*, in *Free Open Source Software Academia Conference (fOSSa)*, 2011.

Dominique Cardon, Guillaume Desgens-Pasanau, Benjamin Nguyen. *Le droit à l'oubli sur Internet est-il possible ?*, in *Conférence-débat au Café des techniques*, Musée des arts et métiers, 2011.

Tristan Allard, Benjamin Nguyen, *L'Anonymisation des données du DMP*, in *Conférence sur le Partage et secret de l'information de santé*, 2010.

François-Xavier Dudouet, Benjamin Nguyen, Antoine Vion, *New web standards in the making*, in *Technical Regulations of Internet workshop*, X-Telecom ParisTech (Orange chair), 2009.

Ivan Bedini, Benjamin Nguyen, Georges Gardarin, *Deriving Ontologies from XML Schema*, in *Entrepôts de Données et Analyse en Ligne* (EDA), 15p. 2008.

Antoine Vion, Benjamin Nguyen, *Outils informatiques pour la sociologie ; étude de la sociologie des normalisateurs du groupe XQuery du W3C*, in *Colloque Socio-Informatique Ecole des Hautes Etudes en Sciences Sociales (EHESS)*, 2007.

French Conferences

Nicolas Anciaux, Wallid Bezza, Benjamin Nguyen, Michalis Vazirgiannis, *MinExp-Card: Limiting Data Collection Using a Smart Card*, in 4^e *Atelier sur la Protection de la Vie Privée* (APVP), 2013.

Cuong-Quoc To, Benjamin Nguyen, Philippe Pucheral, *Privacy-Preserving SQL Query Execution on Distributed Data*, in 4^e *Atelier sur la Protection de la Vie Privée* (APVP), 2013.

Quoc-Cuong To, Benjamin Nguyen, Philippe Pucheral, *Secure global protocol for computing aggregate functions*, in *First Association of Vietnamese Scientists and Experts Doctoral Workshop* (AVSE Doctoral Workshop), 2012.

Nicolas Anciaux, Benjamin Nguyen, Michalis Vazirgiannis, *Cas d'usage d'un principe fondamental de protection de la vie privée*, in *Bases de Données Avancées* (BDA), 2012.

Nicolas Anciaux, Benjamin Nguyen, Michalis Vazirgiannis, *Limiting Data Collection in Application Forms*, in *Atelier Protection de la Vie Privée* (APVP), 2012.

Tristan Allard, Benjamin Nguyen, Philippe Pucheral. *Safe Anonymization of Data Hosted in Smart Tokens*, un *Bases de Données Avancées* , 2010. (acceptance 34%).

Pablo Andres Diaz, François-Xavier Dudouet, Jean-Christophe Graz, Benjamin Nguyen, Antoine Vion, *Gouverner la standardisation par les changements d'arène. Le cas du XML*, in Session Economie du politique et politique de l'économie, 9^e *Congrès de l'Association Française de Sciences Politiques* (AFSP), 36p. 2009.

Georges Gardarin, Benjamin Nguyen, Laurent Yeh, Karine Zeitouni, Bogdan Butnaru, Iulian Sandu-Popa, *Gestion efficace de séries temporelles en P2P: Application à l'analyse technique et l'étude des objets mobiles*, in *Bases de Données Avancées* (BDA), 2009. (acceptance 30%).

Bogdan Butnaru, Benjamin Nguyen, Georges Gardarin, Laurent Yeh, *XQ2P: Efficient XQuery P2P Time Series Processing*, in *Bases de Données Avancées* Demonstration Session, 2009.

Antoine Vion, François-Xavier Dudouet, Benjamin Nguyen, *Expériences de modélisation et de temporalisation des données Web en entrepôts XML*, in *Congrès de l'Association Française de Sociologie* (AFS), RT 20 Methods, 2009.

Dario Colazzo, François-Xavier Dudouet, Ioana Manolescu, Benjamin Nguyen, Antoine Vion, *Traiter des corpus d'information sur le Web. Vers de nouveaux usages informatiques de l'enquête*, in *Roundtable reflection on the methods of Political Science on both sides of the Atlantic*, 7^e *Congrès de l'Association Française de Sciences Politiques* (AFSP), 28 p. 2007.

Florin Dragan, Georges Gardarin, Benjamin Nguyen, Laurent Yeh, *On Indexing Multidimensional Values In A P2P Architecture*, in *Bases de Données Avancées* (BDA), 2006 (acceptance 29%).

Ivan Bedini, Fabrice Bourge, Benjamin Nguyen, *RepXML: Experimenting an ebXML Registry to Store Semantics and Content of Business Messages*, in *Bases de Données Avancées* (BDA) Demonstration Session, 2006.

Bogdan Butnaru, Florin Dragan, Georges Gardarin, Ioana Manolescu, Nicoleta Preda, Benjamin Nguyen, Radu Pop, Laurent Yeh, *P2PTester: testing P2P platform performance*, in *Bases de Données Avancées* (BDA) Demonstration Session, 2006.

Benjamin Nguyen, Iraklis Varlamis, Maria Halkidi, Michalis Vazirgiannis, *Construction de Classes de Documents Web*, in Journées Francophones de la Toile (JFT), 2003.

Benjamin Nguyen, Michalis Vazirgiannis, Iraklis Varlamis, Maria Halkidi, *Organising Web Documents in Thematic Subsets using an Ontology (THESUS)*, in *Journées AS Web Sémantique*, 2002.

Serge Abiteboul, Grégory Cobéna, Benjamin Nguyen, Antonella Poggi, *Construction and Maintenance of a Set of Pages of Interest (SPIN) using ActiveXML*, in *Bases de Données Avancées* (BDA), Evry, 2002.

Patents

Iraklis Varlamis, Michalis Vazirgiannis, Benjamin Nguyen, Maria Halkidi, *Method and system for the collection, description, management and manipulation of digital/web documents based on semantics from a thematic ontology.* Patent No.: 1004662. Greek Society of Industrial Property.

Bibliography

[1] Abiteboul, S., Allard, T., Chatalic, P., Gardarin, G., Ghitescu, A., Goasdoué, F., Manolescu, I., Nguyen, B., Ouazara, M., Somani, A., Travers, N., Vasile, G., and Zoupanos, S. WebContent: efficient P2P Warehousing of web data. *PVLDB 1*, 2 (2008), 1428–1431.

Cited on Page(s): 3, 6, 64.

[2] Abiteboul, S., Bensal, V., Cobéna, G., Nguyen, B., and Poggi, A. Model, design and construction of a service-oriented web warehouse. In *7th European Conference on Digital Libraries (ECDL), LNCS 2769* (2003).

Cited on Page(s): 6.

[3] Abiteboul, S., Nguyen, B., and Ruberg, G. *Complex Data for Decision Suppport*. IDEA Group Publishing, 2006, ch. Building an Active Content Warehouse.

Cited on Page(s): 9.

[4] Aggarwal, C. C., Pei, J., and Zhang, B. On Privacy Preservation against adversarial Data Mining. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)* (2006).

Cited on Page(s): 50.

[5] Aggarwal, C. C., and Yu, P. S. A general survey of privacy preserving data mining models and algorithms. *Advances in Database Systems 34* (2008).

Cited on Page(s): 50.

[6] Agrawal, D., El Abbadi, A., and Wang, S. Secure and privacy-preserving data services in the cloud: A data centric view. In *PVLDB* (2012), vol. 5, pp. 2028–2029.

Cited on Page(s): 60.

[7] Agrawal, D., El Abbadi, A., and Wang, S. Secure and privacy-preserving data services in the cloud. In *ICDE* (2013), pp. 1268–1271.

Cited on Page(s): 60.

[8] Agrawal, R., Kiernan, J., Srikant, R., and Xu, Y. Hippocratic databases. In *Proceedings of the 28th International Conference on Very Large Databases (VLDB)* (2002).

Cited on Page(s): 20, 48, 49.

[9] Agrawal, S., and Haritsa, J. R. A framework for high-accuracy privacy-preserving mining. In *Proceedings of the 21st International Conference on Data Engineering* (Washington, DC, USA, 2005), ICDE '05, IEEE Computer Society, pp. 193–204.

Cited on Page(s): 32.

[10] Alimonti, P., Ausiello, G., Giovaniello, L., and Protasi, M. On the Complexity of approximating weighted satisfiability problems. Tech. rep., Università di Roma, 1998.

Cited on Page(s): 51, 52.

[11] Allard, T. *Sanitizing microdata without leak : a decentralized approach.* PhD thesis, University of Versailles St-Quentin, 2011.

Cited on Page(s): 4, 8, 30, 63.

[12] Allard, T., Anciaux, N., Bouganim, L., Guo, Y., Folgoc, L. L., Nguyen, B., Pucheral, P., Ray, I., Ray, I., and Yin, S. Secure personal data servers: a vision paper. *PVLDB 3*, 1 (2010), 25–35.

Cited on Page(s): 4, 14, 21, 64.

[13] Allard, T., Anciaux, N., Bouganim, L., Pucheral, P., and Thion, R. Seamless access to healthcare folders with strong privacy guarantees. *International Journal of Healthcare Delivery Reform Initiatives (IJHDRI) 1*, 4 (2009), 82–107.

Cited on Page(s): 64.

[14] Allard, T., Anciaux, N., Bouganim, L., Pucheral, P., and Thion, R. Concilier ubiquité et sécurité des données médicales. *Les technologies de l'information au service des droits: opportunités, défis, limites 32* (2010), 173–219.

Cited on Page(s): 64.

[15] Allard, T., Anciaux, N., Bouganim, L., Pucheral, P., and Thion, R. Trustworthiness of pervasive healthcare folders. *Pervasive and Smart Technologies forHealthcare* (2010), 1–24.

Cited on Page(s): 64.

[16] Allard, T., and Nguyen, B. L'anonymisation des données du DMP. In *Conférence sur le partage et secret de l'information de santé* (2010).

Cited on Page(s): 64.

[17] Allard, T., Nguyen, B., and Pucheral, P. Safe anonymization of data hosted in smart tokens. In *Bases de Données Avancées (BDA)* (2010).

Cited on Page(s): 64.

[18] Allard, T., Nguyen, B., and Pucheral, P. Safe realization of the generalization privacy mechanism. In *PST* (2011), pp. 16–23. Best Paper Award.

Cited on Page(s): 4, 30, 64.

[19] Allard, T., Nguyen, B., and Pucheral, P. Sanitizing microdata without leak: Combining preventive and curative actions. In *ISPEC* (2011), pp. 333–342.

Cited on Page(s): 4, 30, 41, 64.

[20] Allard, T., Nguyen, B., and Pucheral, P. Towards a safe realization of privacy-preserving data publishing mechanisms. In *Mobile Data Management (2)* (2011), pp. 31–34.

Cited on Page(s): 30, 64.

[21] Allard, T., Nguyen, B., and Pucheral, P. MET$_{\mathbb{A}}$P : Revisiting Privacy-Preserving Data Publishing using Secure Devices. *Distributed and Parallel Databases (DAPD)* (to appear, 2013), 1–54.

Cited on Page(s): 4, 8, 30, 40, 41, 42, 64.

[22] Amann, B., Benbernou, S., and Nguyen, B. *Web Data Management Practices: Emerging Technologies*. IDEA Group Inc., 2006, ch. Web Services : Technology issues and foundations.

Cited on Page(s): 9.

[23] Anciaux, N., Bezza, W., Nguyen, B., and Vazirgiannis, M. Minexp-card: limiting data collection using a smart card. In *EDBT* (2013), pp. 753–756.

Cited on Page(s): 4, 7, 46, 53, 55.

[24] Anciaux, N., Bonnet, P., Bouganim, L., Nguyen, B., Popa, I. S., and Pucheral, P. Trusted cells: A sea change for personal data services. In *CIDR* (2013).

Cited on Page(s): 4, 7, 14.

[25] Anciaux, N., Bouganim, L., Pucheral, P., Guo, Y., Folgoc, L. L., and Yin, S. MILo-DB: a personal, secure and portable database machine. *Distributed and Parallel Databases (DAPD* (2013). to appear.

Cited on Page(s): 26.

[26] Anciaux, N., Boutara, D., Nguyen, B., and Vazirgiannis, M. Limiting Data Exposure in Multi-Label Classification Processes. *Fundamenta Informaticae* (2013), 1–19. to appear.

Cited on Page(s): 4, 46, 54.

[27] Anciaux, N., Nguyen, B., and Sandu-Popa, I. Personal Data Management with Secure Hardware : The Advantage of Keeping your Data at Hand. In *Mobile Data Management (2)* (2013), pp. 1–2.

Cited on Page(s): 14.

[28] Anciaux, N., Nguyen, B., and Vazirgiannis, M. Limiting data collection in application forms: A real-case application of a founding privacy principle. In *PST* (2012), pp. 59–66.

Cited on Page(s): 4, 46, 53, 54.

[29] Anciaux, N., Nguyen, B., and Vazirgiannis, M. The Minimum Exposure Project : Limiting Data Collection in Online Forms. *ERCIM News 90* (2012).

Cited on Page(s): 7, 46, 54.

[30] Anciaux, N., Nguyen, B., and Vazirgiannis, M. Exposition Minimum de Données pour des Applications à Base de Classifieurs (extended version). *Ingénierie des Systèmes d'Information 18*, 4 (2013).

Cited on Page(s): 46, 55, 56.

[31] Anciaux, N., Nguyen, B., and Vazirgiannis, M. Minimum Exposure in Classification Scenarios. Tech. rep., INRIA, 2013.

Cited on Page(s): 55.

[32] Anderson, A. An Introduction to the Web Services Policy Language (WSPL). In *Proceedings of the POLICY Workshop* (2004).

Cited on Page(s): 49.

[33] Aranda, C. B., Corby, O., Das, S., Feigenbaum, L., Gearon, P., Glimm, B., Harris, S., Hawke, S., Herman, I., Humfrey, N., Michaelis, N., Ogbuji, C., Perry, M., Passant, A., Polleres, A., Prud'hommeaux, E., Seaborne, A., and Williams, G. T. SPARQL 1.1 Overview. W3C Recommendation, 2013.

Cited on Page(s): 9.

[34] Archambault, J.-P., Baccelli, E., Boldo, S., Bouhineau, D., Cégielski, P., Clausen, T., Dowek, G., Guessarian, I., Lopès, S., Mounier, L., Nguyen, B., Quessette, F., Rasse, A., Rozoy, B., Timsit, C., Viéville, T., and Vincent, J.-M. *Une introduction à la science informatique: Pour les enseignants de la discipline informatique au lycée*. CRDP Eds., 2011.

Cited on Page(s): 11.

[35] Ardagna, C. A., de Capitani di Vimercati, S., Foresti, S., Paraboschi, S., and Samarati, P. Minimising Disclosure of Client Information in Credential-Based Interactions. *International Journal of Information Privacy, Security and Integrity 1*, 2-3 (2012).

Cited on Page(s): 50.

[36] Arduino. http://www.arduino.cc/.

Cited on Page(s): 21.

[37] Ashley, P., Hada, S., Karjoth, G., Powers, C., and Schunter, M. Enterprise privacy authorization language 1.2 (EPAL 1.2). W3C Member Submission, 2003.

Cited on Page(s): 48, 49.

[38] Bajaj, S., and Sion, R. Trusteddb: a trusted hardware based database with privacy and data confidentiality. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data* (2011), SIGMOD '11, pp. 205–216.

Cited on Page(s): 20, 37.

[39] Bao, J., Calvanese, D., Grau, B. C., Dzbor, M., Fokoue, A., Golbreich, C., Hawke, S., Herman, I., Hoekstra, R., Horrocks, I., Kendall, E., Krotzsch, M., Lutz, C., McGuinness, D. L., Motik, B., Pan, J., Parsia, B., Patel-Schneider, P. F., Rudolph, S., Ruttenberg, A., Sattler, U., Schneider, M., Smith, M., Wallace, E., Wu, Z., Zimmermann, A., Carroll, J., Hendler, J., and Kashyap, V. OWL 2 Web Ontology Language. W3C Recommendation, 2012.

Cited on Page(s): 5.

[40] Bazgan, C., and Paschos, V. T. Differential approximation for optimal satisfiability and related problems. *European Journal of Operational research 147*, 2 (2003).

Cited on Page(s): 51, 52.

[41] Bedini, I. *Deriving ontologies automatically from XML schemas applied to the B2B domain.* PhD thesis, University of Versailles St-Quentin, 2010.

Cited on Page(s): 3, 8, 63.

[42] Bedini, I., Bourge, F., and Nguyen, B. RepXML: Experimenting an ebXML registry to store semantics and content of business messages. Bases de Données Avancées, Demonstration, 2006.

Cited on Page(s): 63.

[43] Bedini, I., Gardarin, G., and Nguyen, B. Janus : Automatic Ontology Builder, 2010. available at : `http://bivan.free.fr/Janus/`.

Cited on Page(s): 63.

[44] Bedini, I., Gardarin, G., and Nguyen, B. *Electronic Business Interoperability: Concepts, Opportunities and Challenges.* IGI Global Publishing, 2011, ch. Semantic Technologies and e-business, pp. 243–278.

Cited on Page(s): 63.

[45] Bedini, I., Gardarin, G., and Nguyen, B. *Semantic Technologies and e-business.* IGI Global Publishing, 2011, pp. 243–278.

Cited on Page(s): 3.

[46] Bedini, I., Matheus, C. J., Patel-Schneider, P. F., Boran, A., and Nguyen, B. Transforming XML Schema to OWL Using Patterns. In *ICSC* (2011), pp. 102–109.

Cited on Page(s): 3, 5, 63.

[47] Bedini, I., Nguyen, B., and Gardarin, G. Automatic Ontology Builder from XSD Files. In *World Wide Web Conference* (2008). Developper Track.

Cited on Page(s): 3, 63.

[48] Bedini, I., Nguyen, B., and Gardarin, G. B2B Automatic Taxonomy Construction. In *ICEIS (3-2)* (2008), pp. 325–330.

Cited on Page(s): 3, 63.

[49] Bedini, I., Nguyen, B., and Gardarin, G. Deriving Ontologies from XML Schema. In *Entrepôts de Données et Analyse en Ligne (EDA)* (2008).

Cited on Page(s): 63.

[50] Bedini, I., Nguyen, B., and Gardarin, G. Janus: Automatic Ontology Construction Tool. In *16th International Conference on Knowledge Engineering, Knowledge Management and Knowledge Patterns (EKAW)* (2008). Poster.

Cited on Page(s): 63.

[51] Bedini, I., Nguyen, B., Matheus, C. J., Patel-Schneider, P. F., and Boran, A. *Semi-Automatic Ontology Development: Processes and Resources.* IGI Global Publishing, 2012, ch. Mining XML Schemas to Extract Conceptual Knowledge, pp. 79–2015.

Cited on Page(s): 63.

[52] Belotti, P., Lee, J., Liberti, L., Margot, F., and Wachter, A. Branching and bounds tightening techniques for non-convex MINLP. *Optimization Methods and Software 24*, 4-5 (2009).

Cited on Page(s): 53.

[53] Biron, P. V., and Malhotra, A. XML Schema Part 2: Datatypes. W3C Recommendation, 2004.

Cited on Page(s): 5.

[54] Boag, S., Chamberlain, D., Fernández, M. F., Florescu, D., Robie, J., and Siméon, J. XQuery 1.0: An XML Query Language. Tech. rep., W3C, 2007.

Cited on Page(s): 4.

[55] Boncz, P., Grust, T., van Keulen, M., Manegold, S., Rittinger, J., and Teubner, J. Monetdb/xquery: a fast xquery processor powered by a relational engine. In *Proceedings of the 2006 ACM SIGMOD international conference on Management of data* (2006), SIGMOD '06, pp. 479–490.

Cited on Page(s): 6.

[56] Boyd, D. Identity Production in a Networked Culture : Why Youth Heart MySpace. In *American Association for the Advancement of Sciences (AAAS)* (2006).

Cited on Page(s): 17.

[57] Bray, T., Paoli, J., and Sperberg-McQueen, C. Extensible Markup Language (XML) 1.0. Tech. rep., W3C, 1998.

Cited on Page(s): 4.

[58] Butnaru, B. *Optimizations of XQuery in Peer-to-Peer distributed environments.* PhD thesis, University of Versailles St-Quentin, 2012.

Cited on Page(s): 3, 7, 63.

[59] Butnaru, B., Dragan, F., Gardarin, G., Manolescu, I., Nguyen, B., Pop, R., Preda, N., and Yeh, L. P2PTester: Testing P2P Platform Performance. In *Bases de Données Avancées (BDA)* (2006). Demonstration.

Cited on Page(s): 63.

[60] Butnaru, B., Dragan, F., Gardarin, G., Manolescu, I., Nguyen, B., Pop, R., Preda, N., and Yeh, L. P2PTester: a tool for measuring P2P platform performance. In *ICDE* (2007), pp. 1501–1502.

Cited on Page(s): 3, 6.

[61] Butnaru, B., Gardarin, G., Nguyen, B., and Yeh, L. XQ2P: Efficient XQuery P2P Time Series Processing. In *Bases de Données Avancées (BDA)* (2009). Demonstration.

Cited on Page(s): 63.

[62] Butnaru, B., Nguyen, B., and Gardarin, G. XQuery on Peer-to-Peer (XQ2P), 2010. source code available under GPL, https://cassiopee.prism.uvsq.fr:8443/svn/XQ2P/ conformance tests results available at http://cassiopee.prism.uvsq.fr/XQ2P/.

Cited on Page(s): 3, 5, 7, 63.

[63] Carminati, B., Ferrari, E., and Girardi, J. Trust and share: Trusted information sharing in online social networks. In *IEEE 29th International Conference on Data Engineering (ICDE)* (2012), pp. 1281–1284.

Cited on Page(s): 59.

[64] Carminati, B., Ferrari, E., Heatherly, R., Kantarcioglu, M., and Thuraisingham, B. Semantic web-based social network access control. *Computers & Security 30*, 2-3 (2011), 108–115.

Cited on Page(s): 26.

[65] Carminati, B., Ferrari, E., and Perego, A. Enforcing access control in web-based social networks. *ACM Transactions on Information and System Security (TISSEC) 13* (2009).

Cited on Page(s): 26.

[66] Chen, B.-C., Kifer, D., LeFevre, K., and Machanavajjhala, A. Privacy-preserving data publishing. *Found. Trends in Databases 2*, 1-2 (January 2009), 1–167.

Cited on Page(s): 31.

[67] Chen, W., Clarke, L., Kurose, J., and Towsley, D. Optimizing cost-sensitive trust-negotiation protocols. In *Proceedings of the 24th Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM)* (2005).

Cited on Page(s): 50.

[68] Colombo, P., and Ferrari, E. Towards a modeling and analysis framework for privacy-aware systems. In *SocialCom/PASSAT* (2012), pp. 81–90.

Cited on Page(s): 58.

[69] Commission Nationale de l'Informatique et des Libertés. Rapport d'activité 2012. Tech. rep., CNIL, 2012. available online at : http://www.cnil.fr/fileadmin/documents/La_CNIL/publications/CNIL_RA2012_web.pdf.

Cited on Page(s): 17.

[70] Cook, S. A. The complexity of theorem-proving procedures. In *Proceedings of the 3rd Annual ACM Symposium on Theory of Computing* (1971).

Cited on Page(s): 52.

[71] Cranor, L., Langheinrich, M., Marchiori, M., Presler-Marshall, M., and Reagle, J. The Platform for Privacy Preferences 1.0 (P3P1.0) Specification. Tech. rep., W3C Recommendation, 2002.

Cited on Page(s): 48, 49.

[72] Dean, M., Schreiber, G., Bechhofer, S., van Harmelen, F., Hendler, J., Horrocks, I., McGuinness, D. L., Patel-Schneider, P. F., and Stein, L. A. OWL Web Ontology Language Reference. W3C Recommendation, 2004.

Cited on Page(s): 5.

[73] Dingledine, R., Mathewson, N., and Syverson, P. F. Tor: The second-generation onion router. In *USENIX Security Symposium* (2004), pp. 303–320.

Cited on Page(s): 21.

[74] Dudouet, F.-X., Manolescu, I., Nguyen, B., and Senellart, P. XML Warehousing Meets Sociology. In *In Proceedings of the IADIS International Conference on WWW/Internet (ICWI)* (2005).

Cited on Page(s): 3, 7.

[75] Dwork, C. Differential privacy. In *Proceeding of the 39th International Colloquium on Automata, Languages and Programming* (2006), vol. 4052 of *Lecture Notes in Computer Science*, Springer Berlin / Heidelberg, pp. 1–12.

Cited on Page(s): 33, 34.

[76] eMarketer. Email Marketing Benchmarks: Key Data, Trends and Metrics, 2012.

Cited on Page(s): 15.

[77] Escoffier, B., and Paschos, V. Differential approximation of min sat, max sat and related problems. *European Journal of Operational research 181*, 2 (2007).

Cited on Page(s): 51, 52.

[78] EU Commission. Attitudes on Data Protection and Electronic Identity in the European Union. *Eurobarometer Special Surveys 359* (2011).

Cited on Page(s): 47.

[79] European Parliament. Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data. *Official Journal of the European Union 23* (1995).

Cited on Page(s): 31, 47.

[80] Eurosmart. Smart USB Token. White paper, 2008.

Cited on Page(s): 21, 22.

[81] Evfimievski, A., and Grandison, T. *Privacy Preserving Data Mining.* IGI Global, 2009.

Cited on Page(s): 50.

[82] Fischlin, M., Pinkas, B., Sadeghi, A.-R., Schneider, T., and Visconti, I. Secure set intersection with untrusted hardware tokens. In *Proceedings of the 11th international conference on Topics in cryptology: CT-RSA 2011* (Berlin, Heidelberg, 2011), CT-RSA'11, Springer-Verlag, pp. 1–16.

Cited on Page(s): 33, 36.

[83] Fourer, R., Gay, D. M., and Kernighan, B. W. *AMPL : A Modeling Language for Mathematical Programming, second edition.* Duxbury Press, 2002.

Cited on Page(s): 53.

[84] FreedomBox Foundation. http://www.freedomboxfoundation.org/, 2013.

Cited on Page(s): 21.

[85] Friedman, A., and Schuster, A. Data mining with differential privacy. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge Discovery and Data mining* (2010).

Cited on Page(s): 50.

[86] Fung, B. C. M., Wang, K., Chen, R., and Yu, P. S. Privacy-preserving data publishing: A survey of recent developments. *ACM Comput. Surv. 42* (June 2010), 14:1–14:53.

Cited on Page(s): 31.

[87] Gardarin, G., Nguyen, B., Yeh, L., Zeitouni, K., Butnaru, B., and Sandu-Popa, I. Gestion efficace de séries temporelles en P2P: Application à l'analyse technique et l'étude des objets mobiles. In *Bases de Données Avancées (BDA)* (2009).

Cited on Page(s): 63.

[88] Ghinita, G., Karras, P., Kalnis, P., and Mamoulis, N. Fast data anonymization with low information loss. In *Proceedings of the 33rd international conference on Very large data bases* (2007), VLDB '07, VLDB Endowment, pp. 758–769.

Cited on Page(s): 35.

[89] Giesecke and Devrient. StarSign Crypto USB Token, 2012.

Cited on Page(s): 21.

[90] Global Platform Device Technology. Trusted Execution Environment Internal API Specification. Version 1.0., 2011.

Cited on Page(s): 18.

[91] Goldreich, O. Foundations of cryptography: a primer. *Found. Trends in Theoretical Computer Science 1*, 1 (2005), 1–116.

Cited on Page(s): 39.

[92] Goldreich, O., Micali, S., and Wigderson, A. How to play any mental game. In *Proceedings of the nineteenth annual ACM symposium on Theory of computing* (New York, NY, USA, 1987), STOC '87, ACM, pp. 218–229.

Cited on Page(s): 35.

[93] Goyal, V., Ishai, Y., Mahmoody, M., and Sahai, A. Interactive locking, zero-knowledge pcps, and unconditional cryptography. In *Advances in Cryptology - CRYPTO 2010*, T. Rabin, Ed., vol. 6223 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, 2010, pp. 173–190.

Cited on Page(s): 36.

[94] Goyal, V., Ishai, Y., Sahai, A., Venkatesan, R., and Wadia, A. Founding cryptography on tamper-proof hardware tokens. In *Theory of Cryptography*, D. Micciancio, Ed., vol. 5978 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, 2010, pp. 308–326.

Cited on Page(s): 33.

[95] Hacigümüş, H., Iyer, B., Li, C., and Mehrotra, S. Executing sql over encrypted data in the database-service-provider model. In *Proceedings of the 2002 ACM SIGMOD international conference on Management of data* (New York, NY, USA, 2002), SIGMOD '02, ACM, pp. 216–227.

Cited on Page(s): 39.

[96] Haddon, L., and Livingstrone, S. EU Kids Online: national perspectives. Tech. rep., The London School of Economics and Political Science, 2012.

Cited on Page(s): 17.

[97] Halkidi, M., Nguyen, B., Varlamis, I., and Vazirgiannis, M. THESUS: Organizing Web document collections based on link semantics. *VLDB J. 12*, 4 (2003), 320–332.

Cited on Page(s): 3.

[98] Hazay, C., and Lindell, Y. Constructions of truly practical secure protocols using standard smartcards. In *Proceedings of the 15th ACM conference on Computer and communications security* (New York, NY, USA, 2008), CCS '08, ACM, pp. 491–500.

Cited on Page(s): 36.

[99] Järvinen, K., Kolesnikov, V., Sadeghi, A.-R., and Schneider, T. Embedded sfe: offloading server and network using hardware tokens. In *Proceedings of the 14th international conference on Financial Cryptography and Data Security* (Berlin, Heidelberg, 2010), FC'10, Springer-Verlag, pp. 207–221.

Cited on Page(s): 33, 36.

[100] Jiang, W., and Clifton, C. A secure distributed framework for achieving k-anonymity. *The VLDB Journal 15* (2006), 316–333.

Cited on Page(s): 35, 36.

[101] Jurczyk, P., and Xiong, L. Distributed anonymization: Achieving privacy for both data subjects and data providers. In *IFIP WG 11.3 Working Conference on Data and Applications Security* (Berlin, Heidelberg, 2009), Springer-Verlag, pp. 191–207.

Cited on Page(s): 35, 36.

[102] Kaplan, B., and Harris-Salamone, K. D. Health IT Success and Failure: Recommendations from Literature and an AMIA Workshop. *Journal of the American Medical Informatics Association (JAMIA) 16*, 3 (2009), 291–299.

Cited on Page(s): 17.

[103] Karp, R. M. Reducibility among combinatorial problems. In *Complexity of Computer Computations* (1972), pp. 85–103.

Cited on Page(s): 52.

[104] Katz, J. Universally composable multi-party computation using tamper-proof hardware. In *Proceedings of the 26th annual international conference on Advances in Cryptology* (Berlin, Heidelberg, 2007), EUROCRYPT '07, Springer-Verlag, pp. 115–128.

Cited on Page(s): 36.

[105] Katzenbeisser, S., Kursawe, K., Preneel, B., and Sadeghi, A.-R. Privacy and Security in Smart Energy Grids (Dagstuhl Seminar 11511). *Dagstuhl Reports 1*, 12 (2012), 62–68.

Cited on Page(s): 19.

[106] Kifer, D. Attacks on privacy and definetti's theorem. In *Proceedings of the 35th SIGMOD international conference on Management of data* (New York, NY, USA, 2009), SIGMOD '09, ACM, pp. 127–138.

Cited on Page(s): 56.

[107] Kifer, D., and Machanavajjhala, A. No free lunch in data privacy. In *Proceedings of the 2011 international conference on Management of data* (New York, NY, USA, 2011), SIGMOD '11, ACM, pp. 193–204.

Cited on Page(s): 56.

[108] Kirkpatrick, S., Gelatt, C., and Vecchi, M. Optimization by Simulated Annealing. *Science 220*, 4598 (1983).

Cited on Page(s): 53.

[109] Kruk, S. R., Gronkowski, S., Gzella, A., Woroniecki, T., and Choi, H.-C. D-FOAF: Distributed Identity Management with Access Rights Delegation. *LNCS 4185* (2006), 140–154.

Cited on Page(s): 26.

[110] Lam, H., Fung, G., and Lee, W. A Novel Method to Construct Taxonomy Electrical Appliances Based on Load Signatures. *IEEE Transactions on Consumer Electronics 53*, 2 (2007), 653–660.

Cited on Page(s): 19.

[111] LeFevre, K., DeWitt, D. J., and Ramakrishnan, R. Mondrian multidimensional k-anonymity. In *Proceedings of the 22nd International Conference on Data Engineering* (Washington, DC, USA, 2006), ICDE '06, IEEE Computer Society, pp. 25–.

Cited on Page(s): 34.

[112] LeFevre, K., DeWitt, D. J., and Ramakrishnan, R. Workload-aware anonymization techniques for large-scale datasets. *ACM Transations on Database Systems (TODS) 33*, 3 (2008).

Cited on Page(s): 50.

[113] Li, N., Li, T., and Venkatasubramanian, S. t-closeness: Privacy beyond k-anonymity and l-diversity. In *Proceedings of the 23rd IEEE International Conference on Data Engineering* (april 2007), ICDE '07, pp. 106–115.

Cited on Page(s): 34.

[114] Loyer, Y., and Straccia, U. Any-world assumptions in logic programming. *Theoretical Computer Science 342*, 2â3 (2005), 351–381.

Cited on Page(s): 56.

[115] Loyer, Y., and Straccia, U. Epistemic foundation of stable model semantics. *Theory Pract. Log. Program. 6*, 4 (2006), 355–393.

Cited on Page(s): 56.

[116] Machanavajjhala, A., Gehrke, J., and Götz, M. Data publishing against realistic adversaries. *Proc. VLDB Endow. 2*, 1 (Aug. 2009), 790–801.

Cited on Page(s): 34.

[117] Machanavajjhala, A., Gehrke, J., Kifer, D., and Venkitasubramaniam, M. l-diversity: Privacy beyond k-anonymity. In *Proceedings of the 22nd International Conference on Data Engineering* (Washington, DC, USA, 2006), ICDE '06, IEEE Computer Society, pp. 24–.

Cited on Page(s): 34.

[118] Makhorin, A. O. GLPK: GNU Linear Programming Kit [computer software], 2000.

Cited on Page(s): 53.

[119] Meyerson, A., and Williams, R. On the complexity of optimal k-anonymity. In *Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems* (New York, NY, USA, 2004), PODS '04, ACM, pp. 223–228.

Cited on Page(s): 35.

[120] Mohammed, N., Chen, R., Fung, B. C. M., and Yu, P. S. Differentially private data release for data mining. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data mining (KDD)* (2011).

Cited on Page(s): 33, 50.

[121] Mohammed, N., Fung, B. C. M., Hung, P. C. K., and Lee, C.-K. Centralized and distributed anonymization for high-dimensional healthcare data. *ACM Trans. Knowl. Discov. Data 4* (October 2010), 18:1–18:33.

Cited on Page(s): 35, 36.

[122] Mohammed, N., Fung, B. C. M., Wang, K., and Hung, P. C. K. Privacy-preserving data mashup. In *Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology* (New York, NY, USA, 2009), EDBT '09, ACM, pp. 228–239.

Cited on Page(s): 35, 36.

[123] Moses, T. Extensible Access Control Markup Language (XACML) version 2.0. OASIS Standard, 2005.

Cited on Page(s): 49.

[124] Narayanan, A., Barocas, S., Toubiana, V., Nissenbaum, H., and Boneh, D. A Critical Look at Decentralized Personal Data Architectures. Tech. rep., University of Stanford, 2012.

Cited on Page(s): 20, 21.

[125] Nguyen, B. *Construction and Maintenance of a Web Warehouse*. PhD thesis, University of Paris XI, 2003.

Cited on Page(s): 3.

[126] Nguyen, B., Abiteboul, S., Cobena, G., and Preda, M. Monitoring XML data on the Web. In *Proceedings of the 2001 ACM SIGMOD international conference on Management of data* (New York, NY, USA, 2001), SIGMOD '01, ACM, pp. 437–448.

Cited on Page(s): 3, 6.

[127] Nguyen, B., Vion, A., Dudouet, F.-X., Colazzo, D., and Manolescu, I. WebStand, une plateforme de gestion de données Web pour applications sociologiques. *Revue de Technique et Science Informatiques (TSI) 29*, 8-9 (2010), 1055–1080.

Cited on Page(s): 3, 5.

[128] Nguyen, B., Vion, A., Dudouet, F.-X., Colazzo, D., Manolescu, I., and Senellart, P. XML Content Warehousing : Improving Sociological Studies of Mailing Lists and Web Data. *Sociological Methodology Bulletin 112*, 1 (2011), 5–31.

Cited on Page(s): 3, 4, 7.

[129] Nguyen, B., and Zoupanos, S. The webcontent store. In *Atelier Sources Ouvertes et Services, RFIA Conference* (2010).

Cited on Page(s): 6.

[130] Nissenbaum, H. *Privacy in context: Technology, policy, and the integrity of social life.* Stanford Law Books, 2010.

Cited on Page(s): 18.

[131] OECD. OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data. Tech. rep., OECD, 1980.

Cited on Page(s): 47.

[132] Park, J., and Sandhu, R. The $ucon_{ABC}$ usage control model. *ACM Trans. Inf. Syst. Secur. 7*, 1 (Feb. 2004), 128–174.

Cited on Page(s): 22, 27.

[133] Petronio, S. *Unpacking the paradoxes of privacy in CMC relationships: The challenges of blogging and relational communication on the internet.* Peter Lang Editions, 2011.

Cited on Page(s): 18.

[134] Ponemon Institute, LLC. 2011 Annual Study: U.S. Cost of a Data Breach. Tech. rep., Symantec, 2012.

Cited on Page(s): 47.

[135] Prud'hommeaux, E., and Seaborne, A. SPARQL Query Language for RDF, 2008.

Cited on Page(s): 9.

[136] Raspberry Pi. http://www.raspberrypi.org, 2013.

Cited on Page(s): 21.

[137] Rastogi, V., Suciu, D., and Hong, S. The boundary between privacy and utility in data publishing. In *Proceedings of the 33rd international conference on Very large data bases* (2007), VLDB '07, VLDB Endowment, pp. 531–542.

Cited on Page(s): 32, 33, 34.

[138] Robie, J., Chamberlain, D., Dyck, M., and Snelson, J. XQuery 3.0: An XML Query Language. Tech. rep., W3C, 2013.

Cited on Page(s): 4.

[139] *Responsabilité Sociale des Organismes Publics, Une approche responsable du capital humain* (2012), Ministère de l'Écologie, du Développement durable, des Transports et du Logement, Ministère du Travail, de l'Emploi et de la Santé, Ministère des Solidarités et de la Cohésion Sociale. available at : http://www.developpement-durable.gouv.fr/IMG/pdf/Colloque_RSO_-_9_janvier_2012_-_ACTES.pdf.

Cited on Page(s): 17.

[140] Schwartz, P. M. Property, privacy and personal data. *Harvard Law Review 117*, 7 (2004).

Cited on Page(s): 20.

[141] Sweeney, L. k-anonymity: a model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst. 10*, 5 (2002), 557–570.

Cited on Page(s): 34.

[142] The World Economic Forum. Personal Data: The emergence of a new asset class. Report, 2011.

Cited on Page(s): 18.

[143] The World Economic Forum. Rethinking Personal Data: Strengthening Trust. Industrial Report, 2012.

Cited on Page(s): 15.

[144] Thompson, H. S., Beech, D., Maloney, M., and Mendelsohn, N. XML Schema Part 1: Structures. W3C Recommendation, 2004.

Cited on Page(s): 5.

[145] To, Q.-C., Nguyen, B., and Pucheral, P. Secure global protocol for computing aggregate functions. AVSE Doctoral Workshop, 2012.

Cited on Page(s): 42, 64.

[146] To, Q.-C., Nguyen, B., and Pucheral, P. Privacy-Preserving SQL Query Execution on Distributed Data. Bases de Données Avancées, 2013.

Cited on Page(s): 31, 42, 64.

[147] To, Q.-C., Nguyen, B., and Pucheral, P. Privacy-Preserving SQL Query Execution on Distributed Data (demonstration). 4^e Atelier sur la Protection de la Vie Privée (APVP), 2013.

Cited on Page(s): 30, 42, 64.

[148] Tsoumakas, G., and Katakis, I. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining 3*, 3 (2007).

Cited on Page(s): 48, 50.

[149] Varlamis, I., Vazirgiannis, M., Halkidi, M., and Nguyen, B. THESUS, a Closer View on Web Content Management Enhanced with Link Semantics. *IEEE Trans. Knowl. Data Eng. 16*, 6 (2004), 685–700.

Cited on Page(s): 3, 6, 9.

[150] Verykios, V. S., Elagarmid, A. K., Bertino, E., Saygin, Y., and Dasseni, E. Association Rule Hiding. *Transactions on Knowledge and Data Engineering (TKDE) 16*, 4 (2004).

Cited on Page(s): 50.

[151] Wierzbicki, A. *Trust and Fairness in Open, Distributed Systems.* Springer, 2010.

Cited on Page(s): 59.

[152] Wong, R. C.-W., Fu, A. W.-C., Wang, K., and Pei, J. Anonymization-based attacks in privacy-preserving data publishing. *ACM Trans. Database Syst. 34*, 2 (July 2009), 8:1–8:46.

Cited on Page(s): 34.

[153] Xiao, X., and Tao, Y. Anatomy: simple and effective privacy preservation. In *Proceedings of the 32nd international conference on Very large data bases* (2006), VLDB '06, VLDB Endowment, pp. 139–150.

Cited on Page(s): 34.

[154] Xiao, X., and Tao, Y. M-invariance: towards privacy preserving re-publication of dynamic datasets. In *Proceedings of the 2007 ACM SIGMOD international conference on Management of data* (2007), SIGMOD '07, pp. 689–700.

Cited on Page(s): 34.

[155] Xue, M., Papadimitriou, P., Raïssi, C., Kalnis, P., and Pung, H. K. Distributed privacy preserving data collection. In *Proceedings of the 16th international conference on Database systems for advanced applications - Volume Part I* (Berlin, Heidelberg, 2011), DASFAA'11, Springer-Verlag, pp. 93–107.

Cited on Page(s): 35.

[156] Xyleme, L. A dynamic warehouse for XML Data of the Web. *IEEE Data Eng. Bull. 24*, 2 (2001), 40–47.

Cited on Page(s): 3.

[157] Xyleme Inc. http://www.xyleme.com/, 2003.

Cited on Page(s): 6.

[158] Yao, A. C. Protocols for secure computations. In *Proceedings of the 23rd Annual Symposium on Foundations of Computer Science* (Washington, DC, USA, 1982), SFCS '82, IEEE Computer Society, pp. 160–164.

Cited on Page(s): 35.

[159] Yao, D., Frikken, F. B., Atallah, M. J., and Tamassia, R. Private information: To reveal or not to reveal. *ACM Transactions on Information and System Security (TISSEC) 12*, 1 (2008).

Cited on Page(s): 50.

[160] Zhang, N., and Zhao, W. Distributed privacy preserving information sharing. In *Proceedings of the 31st international conference on Very large data bases* (2005), VLDB '05, VLDB Endowment, pp. 889–900.

Cited on Page(s): 25.

[161] Zhong, S., Yang, Z., and Chen, T. k-anonymous data collection. *Inf. Sci. 179* (August 2009), 2948–2963.

Cited on Page(s): 35.

[162] Zhong, S., Yang, Z., and Wright, R. N. Privacy-enhancing k-anonymization of customer data. In *Proceedings of the twenty-fourth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems* (New York, NY, USA, 2005), PODS '05, ACM, pp. 139–147.

Cited on Page(s): 35.