



Sixth Framework Programme  
Information Society Technologies  
Future Emerging Technologies



Project Acronym:

**GeoPKDD**

Project full title:

**Geographic Privacy-aware Knowledge Discovery and Delivery**

Project Number: FP6-014915

Instrument: Specific Target Research Project

Project Deliverable D5.2

**PRO report on current privacy regulations and societal requirements**

Date of preparation: 10/06/2006

Revision: final

Operative commencement date of Contract: 01/12/2005

Project Coordinator: Fosca Giannotti

Project institute coordinator: Knowledge Discovery and Delivery-LAB / ISTI-CNR

**Programme Name:** .....IST  
**Project Number:**.....014915  
**Project Title:** .....GEOPKDD

**Document Number:** .....D5.2  
**Work-Package:**.....WP5  
**Document Type:** .....Deliverable  
**Contractual Date of Delivery:** .....01/06/2006  
**Actual Date of Delivery:** .....10/06/2006  
**Title of Document:** .....PRO report on current privacy regulations and societal requirements

**Author(s):** .....

**Dissemination Level** .....Public

#### GeoPKDD consortium participants

Partic. Role	Partic. N.	Participant name	Short name	Country
CO	1	<b>KDD Lab.</b> joint research group of <b>ISTI-CNR</b> , Istituto di Scienza e Tecnologie dell'Informazione, Pisa. <a href="http://www-kdd.isti.cnr.it/">http://www-kdd.isti.cnr.it/</a> and <b>Univ. Pisa</b> , Dept. of Computer Science <a href="http://www.di.unipi.it">http://www.di.unipi.it</a>	<b>KDDLAB</b>	I
CR	2	<b>Univ. of Hasselt</b> , Theoretical Computer Science Group. <a href="http://alpha.uhasselt.be/research/groups/theocomp/">http://alpha.uhasselt.be/research/groups/theocomp/</a>	<b>HASSELT</b>	B
CR	3	<b>EPFL - Ecole Polytechnique Fédérale de Lausanne</b> , Lab. DB, Lausanne. <a href="http://bdwww.epfl.ch/e/">http://bdwww.epfl.ch/e/</a> and <b>University of Milan - Computer Science and Communication Department</b> <a href="http://www.dico.unimi.it/">http://www.dico.unimi.it/</a>	<b>EPFL</b>	CH
CR	4	<b>Fraunhofer Institute for Autonomous Intelligent Systems</b> , Sankt Augustin. <a href="http://www.ais.fraunhofer.de/">http://www.ais.fraunhofer.de/</a>	<b>FAIS</b>	D
CR	5	<b>Wageningen UR</b> , Centre for GeoInformation. <a href="http://cgi.girs.wageningen-ur.nl/">http://cgi.girs.wageningen-ur.nl/</a>	<b>WUR</b>	NL
CR	6	<b>Research Academic Computer Technology Institute</b> , Research and Development Division. <a href="http://www.cti.gr/">http://www.cti.gr/</a> and <b>Univ. Piraeus</b> , Dept. of Informatics <a href="http://www.unipi.gr">http://www.unipi.gr</a>	<b>CTI</b>	GR
CR	7	<b>Sabanci University</b> , Faculty of Engineering and Natural Sciences. <a href="http://www.sabanciuniv.edu/">http://www.sabanciuniv.edu/</a>	<b>UNISAB</b>	TK
CR	8	<b>WIND Telecomunicazioni SpA</b> , Direzione Reti Wind Progetti Finanziati & Technology Scouting. <a href="http://www.wind.it">http://www.wind.it</a>	<b>WIND</b>	I

Deliverable 5.2:  
PRO report on current privacy regulations and  
societal requirements

GeoPKDD Workpackage 5

December 2006

# Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
<b>2</b>	<b>Privacy Regulations</b>	<b>9</b>
2.1	Aspects and Models of Privacy Protection . . . . .	9
2.2	The Evolution of Data Protection . . . . .	11
2.2.1	The European Union Data Protection Directives . . . . .	11
2.2.2	The APEC Privacy Initiative . . . . .	13
2.2.3	Data Heavens and the Safe Harbor Arrangement . . . . .	13
2.3	Privacy Constraints in the GeoPKDD Context . . . . .	15
<b>3</b>	<b>Privacy Preserving Data Analysis</b>	<b>18</b>
<b>4</b>	<b>The Role of the Observatory</b>	<b>21</b>
<b>5</b>	<b>Conclusions</b>	<b>23</b>

# Chapter 1

## Introduction

Information & Communication Technologies - the ICTs – pervade every aspects of our lives, especially after the advent of mobile and wireless technologies. Some of the most evident examples are mobile phones and wireless communication, web browsing and e-mailing over the Internet, credit cards and point-of-sale e-transactions, e-banking, electronic administrative transactions and health records, shopping transactions with loyalty cards, and so on.

In all the above cases, our everyday actions leave (or can leave) digital *traces* into the information systems of the organizations that provide services through the ICTs. Examples of these traces are the positioning log data that record mobile phone location at each time, the web log data that record web pages or packet requests with associated client and server identities, transaction records with associated purchase location and time, sequences of progress states in an administrative process, and so on. These traces are indeed necessary for the correct delivery of services: the wireless network needs to know each mobile user’s position at each moment in time, in order to dispatch calls. Analogously, Internet needs to know IP addresses to dispatch requests from source to destination computers; automated teller machines need to authenticate their users to authorize a withdrawal; retail shops need to identify their loyal customers to assign them bonuses.

These traces can be either *forgotten*, as soon as they are no longer needed for service delivery, or instead *stored* – and the capability of storing larger and larger masses of data grows everyday, at increasingly cheaper costs. But why should we store traces? While traditional information systems manage *business-oriented* information, such as sales, customers, billing-related

records; elementary traces carry finer grained *process-oriented* information about how a complex organization *works*. This is the real reason why traces are worth being remembered: because they may hide a wealth of *knowledge* about the processes which govern the life of complex economical or social systems.

As a remarkable example, the wireless phone networks gather highly informative traces about the human activities in a territory, essentially due to two factors: pervasiveness and positioning accuracy. The number of mobile phone users worldwide was estimated at 1.5 billions in 2005, with regions, such as Italy, where number of mobile phones and number of inhabitants are very close; in other regions, especially developing countries, the numbers are still increasing at a high speed. Moreover, as explained in Chapter 3 of the forthcoming GeoPKDD book [?], the location technologies currently used by wireless carrier operators are capable of providing a good estimate of user location, and better localization is expected in the near future, due to the integration of various positioning technologies (GPS-equipped mobile devices, Wi-Fi and Bluetooth for indoor positioning, and so on).

A scenario of opportunities and threats opens up, provided that some forms of useful knowledge can be discovered from the traces left behind by mobile users in the information systems of wireless networks. Knowledge, in itself, is neither good nor bad: obviously, it can be used for either virtuous or evil purposes. What knowledge to be searched from digital traces? For what purposes? And, more importantly, which eyes to look at these traces with?

- The malicious eyes of the *Spy* – or the detective – interested in discovering the punctual knowledge about the behaviour of an individual person (or small groups) for surveillance purposes.
- The benevolent eyes of the *Historian* – or the archaeologist – interested in discovering the collective knowledge about the behaviour of whole communities, for the purpose of analysis, of understanding the dynamics of these communities, the way they live.

Before the ICT revolution, there was little chance of overlapping the time of criminal investigation and the time of historical investigation: the *Spy* looks for traces in the present, while the *Historian* looks for traces in the past. We are faced today, for the first time in history, with the concrete possibility of pursuing an *archaeology of the present*: discovering from the digital traces

of our mobile activity the knowledge that makes ourselves comprehend timely and precisely the way we live, the way we use today our time and our land. Thus, it is becoming possible, in principle, to understand how to live better by learning from our recent history, i.e., from the traces left behind us yesterday, or a few moments ago, recorded in the information systems and analyzed to produce usable, timely and reliable knowledge.

Concretely, from the traces of our mobile phones it is possible to reconstruct our mobile behaviour, the way we move – at a level of groups or communities, not individuals, and this knowledge may enable us to improve decision-making in mobility-related issues:

- the way we plan traffic and public mobility systems in metropolitan areas;
- the way we plan physical communication networks, such as new roads or railways;
- the way we localize new services in our towns;
- the way we forecast traffic-related phenomena;
- the way we organize postal and logistics systems;
- the way we can avoid repeating mistakes that emerge from the freshly analyzed movement behaviour;
- the way we can timely detect changes that occur in the movement behaviour.

In a few words, we may construct from the knowledge about our own mobile behaviour many concrete services of public interest.

There's a little path from opportunities to threats: we are aware that, at the basis of this positivist's scenario, lies a flaw of potentially dramatic impact, namely the fact that the donors of the mobility data are ourselves the citizens, and making this data publicly available for the mentioned purposes would put at risk our own privacy, our natural right to keep secret the places we visit, the places we live or work at, the people we meet - all in all, the way we live as an individual. In other words, the personal mobility data, as gathered by the wireless networks, are extremely sensitive information; their disclosure may represent a brutal violation of the privacy protection

rights, established in increasingly more laws and regulations internationally (see Section 2).

However, a (rather naïve) positivist Scientist may observe that, for the above mentioned analytical purposes, knowing the exact identity of individuals is not needed: contrarily to the Spy, that for surveillance purposes needs to know the personal movement data, the Historian does not need to know identities: anonymized trajectories are enough to reconstruct aggregate movement behaviour, pertaining to groups of people. Is this reasoning correct? Can we conclude that the Historian runs no risks, while working for the public interest, to inadvertently put in jeopardy the privacy of the individuals?

Unfortunately not: hiding identities is not enough. In certain cases, it is possible to reconstruct the exact identities from the released data, even when identities have been removed and replaced by pseudonyms. A well-known example by Latanya Sweeney [9], is a study on the US 1990 Census data, that showed that 87% of US citizens (216 millions out of 248) is uniquely identifiable given three pieces of information: ZIP code of birthplace, birth date, and gender. A combination of attributes like these three, erroneously believed to hide the real identities, is called a *quasi-identifier*. Clearly, the release of punctual data containing quasi-identifiers is not safe. Nonetheless, it is possible to legally acquire (from publicly available sources or by legal purchase) datasets that include these quasi-identifiers, and draw worrying conclusions. In Sweeney's experiment, a first database consisted of medical record information from the US National Association of the Health Data Organizations, made available to research institutes because believed anonymous, containing the attributes ZIP code, birth date, sex and various clinical information (date, exam or visit, diagnosis, treatment, ). The second database consisted of the poll list of voters of the county of Cambridge, MA (available for sale for political campaign purposes), containing the attributes ZIP code, birth date, sex, Name, Surname, Address, and so on. By joining the data in the two tables on the quasi-identifiers, it is possible in many cases to link the sensitive medical information (e.g., the diagnosis) to an individual. In the specific experiment, Sweeney considered the governor of Massachusetts: only 6 persons had his birth date in the joined table, only 3 of those were men, and only 1 had his own ZIP code! As a consequence, the medical records of the governor were uniquely identified from legally accessible sources!

So, removing identities does not suffice, at least if quasi identifiers are used as pseudonyms. This motivates the study of techniques, briefly discussed in

Section 3, that try to remove the danger of quasi-identifiers while preserving, as far as possible, the usefulness of data for analysis. But even a brute force solution, like replacing identities or quasi-identifiers with totally unintelligible codes may not work, when the data to be disclosed are movement data, such as personal trajectories. An example, discussed for instance in [4, 6] is the following: a totally anonymous trajectory which occurs periodically every working day from location A in the suburbs to location B downtown during the morning rush hours and in the reverse direction from B to A in the evening rush hours can be linked to the persons who live in A and work in B; therefore, if locations A and B are known at a sufficiently fine granularity, it is possible to identify specific persons and unveil their daily routes. This example shows how, in mobility data, positioning in space and time may act as a powerful quasi identifier.

Other interesting examples of *re-identification* are given by Bradley Malin and Latanya Sweeney [7, 8]. In these works the term *re-identification* refers to correctly relating seemingly anonymous data to explicitly identifying information (such as the name or address) of the person who is the subject of those data. While re-identification has usually been associated with data released from a single data holder, they show how an individual could be related to a trail of seemingly anonymous and homogenous data left across different locations. Successful re-identifications are reported for DNA sequences left by hospital patients and for IP addresses left by online consumers.

Of course, other alternatives may be followed, such as adding randomized noise to the data in such a way that punctual data are obfuscated while it is still possible to extract useful aggregated knowledge from them: but clearly the picture is not as easy as imagined.

Now, the Scientist might argue that, in the end, it is not needed to disclose the data: the Historian only may be given access to the data, as a trusted civil servant, in order to produce the knowledge (mobility patterns, models, rules) that is then disclosed for the public utility. In this view, only aggregated information is divulged, while source data are kept secret. In the terms of Chapter 2 of the forthcoming GeoPKDD book [?], only synoptic questions are allowed, not elementary ones, such as “*who was near the leaning tower yesterday between 3 and 4 a.m.?*”. Since aggregated information concerns large groups of individuals, we are tempted to conclude that its disclosure is safe. Once again, this reasoning is flawed, although in a rather subtle sense. The reason, as explained in [3], is that from rules with high support

(i.e., concerning many individuals) it is sometimes possible to deduce new rules with very limited support, capable of identifying precisely one or few individuals together with some of their sensitive attributes. As an example, assume that the following rule can be mined from the source data:

$$\begin{aligned} &Age = 27 \quad \wedge \quad ZIP = 45254 \quad \wedge \quad Diagnosis = HIV \rightarrow \\ &NativeCountry = USA \quad [sup = 758, conf = 99.8\%] \end{aligned}$$

Apparently, this is a safe rule, which tells us that 99.8% of 27 year old people from a given geographic area that have been diagnosed an HIV infection, are born in the US. From this rule, however, we can derive that only the 0.2% of the rule population of 758 persons are 27 years old, live in the given area, has contracted HIV and is not born in the US; a simple calculation shows that there is only one person with these combination of characteristics. Now, the triple Age, ZIP code and Native Country is a quasi-identifier, and it is well possible that we can find in the county demographic list that there only one 27-year old person in the given area who is not born in the US. This is clearly a privacy violation, as such person can be easily identified.

This discussion should have brought evidence that protecting privacy when disclosing information is not trivial: anonymization and aggregation do not necessarily put ourselves on the safe side from attacks to privacy. Clearly, for the very same reason the problem is scientifically attractive - besides socially relevant. As often happens in science, the problem is to find an optimal trade-off between two conflicting goals: from one side, we would like to have precise, fine-grained knowledge about mobility, which is useful for the analytic eyes of the Historian; from the other side, we would like to have imprecise, coarse-grained knowledge about mobility, which puts us in repair from the eyes of the Spy. It is interesting that the same conflict - essentially between opportunities and risks - can be either read as a mathematical problem or as a social (or ethical, or legal) challenge. Indeed, the privacy issues related to ICTs are unlikely to be solved by exclusively technological means: quoting Rakesh Agrawal, one of the first researchers to address privacy issues in data management, any real solution to privacy problems can only be achieved through an alliance of technology, legal regulations and social norms [1]. It is exactly with this observation in mind that we created, within the GeoPKDD project, an observatory on privacy regulations, whose aims are described in Section 4. Before getting to this, we briefly overview the current international situation on the legal front - in terms of laws and regulations about protection of personal data (Section 2) - and on the technical front

- in terms of privacy-preserving technologies in data management and data mining (Section 3), although a more detailed discussion on this latter theme is to be found in Chapters 8 and 11 of the forthcoming GeoPKDD book [?].

# Chapter 2

## Privacy Regulations

Privacy is a fundamental human right which underpins dignity and other values such as freedom of speech and association. Of all the human rights, privacy is perhaps the most difficult to define since it varies widely according to the context and environment. In many countries, the concept has been fused with data protection, which interprets privacy in terms of management of personal information. Privacy is recognized around the world in one way or the other. More specifically, it is protected in the Universal Declaration of Human Rights, the International Covenant on Civil and Political Rights, and in many other international and regional human rights treaties. Nearly every country in the world includes a right of privacy in its constitution. At a minimum, these provisions include rights of inviolability of the home and secrecy of communications. Most recently written constitutions include specific rights to access and control one's personal information. In many countries, international agreements that recognize privacy rights have been adopted into law <sup>1</sup>.

### 2.1 Aspects and Models of Privacy Protection

From among the various aspects that privacy can be considered, two are important to our study:

---

<sup>1</sup>Part of the material presented in this Section is borrowed from the "*Overview of Privacy*" of the 2005 edition of the Privacy and Human Rights Report ([www.privacyinternational.org](http://www.privacyinternational.org)).

*Information privacy*, which involves the establishment of rules governing the collection and handling of personal data such as credit information, and medical and government records. It is also known as “data protection”, and

*Privacy of communications*, which covers the security and privacy of mail, telephones, e-mail and other forms of communication.

There are also four major models for privacy protection that can be used either independently or simultaneously. In the countries that protect privacy most effectively, all of the models are used together to ensure privacy protection. The models are:

*Comprehensive Laws* These are general laws that govern the collection, use and dissemination of personal information by both the public and private sectors. An oversight body then ensures compliance. This is the preferred model of most countries adopting data protection laws and was adopted by the European Union to ensure compliance with its data protection regime. A variation of these laws, which is described as a “co-regulatory model” was adopted in Canada and Australia. Under this approach, industry develops rules for the protection of privacy that are enforced by the industry and overseen by the privacy agency.

*Sectoral Laws* Some countries, such as the United States, have avoided enacting general data protection rules in favor of specific sectoral laws governing for example, video rental records and financial privacy. In such cases, enforcement is achieved through a range of mechanisms. A major drawback with this approach is that it requires that new legislation be introduced with each new technology so protections frequently lag behind. The lack of legal protections for individual’s privacy on the Internet in the United States is a striking example of its limitations. There is also the problem of a lack of an oversight agency. In many countries, sectoral laws are used to complement comprehensive legislation by providing more detailed protections for certain categories of information, such as telecommunications, police files, or consumer credit records.

*Self-regulation* Data protection can also be achieved, at least in theory, through various forms of self-regulation, in which companies and industry bodies establish codes of practice and engage in self-policing. However,

in many countries, especially in the United States, these efforts have been disappointing, with little evidence that the aims of the codes are regularly fulfilled. Adequacy and enforcement are the major problem with these approaches. Industry codes in many countries have tended to provide only weak protections and lack enforcement.

*Technologies of Privacy* With the recent development of commercially available technology-based systems, privacy protection has also moved to into the hands of individual users. Users of the Internet and of physical applications can employ a range of programs and systems that provide varying degrees of privacy and security of communications. These include encryption, anonymous remailers, proxy servers and digital cash.

## **2.2 The Evolution of Data Protection**

Interest in the right of privacy increased in the 1960's and 1970's with the advent of information technology. The surveillance potential of powerful computer systems prompted demands for specific rules governing the collection and handling of personal information. The genesis of modern legislation in this area can be traced back to the first data protection law in the world enacted in the Land of Hesse in Germany in 1970. This was followed by national laws in Sweden (1973), the United States (1974), Germany (1977) and France (1978). Two crucial international instruments evolved from these laws. The Council of Europe's 1981 Convention for the Protection of Individuals with regard to the Automatic Processing of Personal Data and the Organization for Economic Cooperation and Development (OECD) Guidelines Governing the Protection of Privacy and Transborder Data Flows of Personal Data, set out specific rules covering the handling of electronic data. These rules describe personal information as data that are afforded protection at every step from collection to storage and dissemination.

### **2.2.1 The European Union Data Protection Directives**

In 1995, the European Union enacted the Data Protection Directive in order to harmonize member states' laws in providing consistent levels of protections for citizens and ensuring the free flow of personal data within the European Union. The directive sets a baseline common level of privacy that not only

reinforces current data protection law, but also establishes a range of new rights. It applies to the processing of personal information in electronic and manual files. A key concept in the European data protection model is “enforceability”. Data subjects have rights established in explicit rules. Every European Union country has a data protection commissioner or agency that enforces the rules. It is expected that the countries with which Europe does business will need to provide a similar level of oversight.

The basic principles established by the Directive are: the right to know where the data originated; the right to have inaccurate data rectified; a right of recourse in the event of unlawful processing; and the right to withhold permission to use data in some circumstances. For example, individuals have the right to opt-out free of charge from being sent direct marketing material. The Directive contains strengthened protections over the use of sensitive personal data relating, for example, to health, sex life, or religious or philosophical beliefs.

The 1995 Directive imposes an obligation on member states to ensure that the personal information relating to European citizens has the same level of protection when it is exported to, and processed in, countries outside the European Union. This requirement has resulted in growing pressure outside Europe for the passage of privacy laws. Those countries that refuse to adopt adequate privacy laws may find themselves unable to conduct certain types of information flows with Europe, particularly if they involve sensitive data. In 1997, the European Union supplemented the 1995 directive by introducing the Telecommunications Privacy Directive. This Directive established specific protections covering telephone, digital television, mobile networks and other telecommunications systems. It imposed wide-ranging obligations on carriers and service providers to ensure the privacy of user’s communications, including Internet-related activities. In July 2000, the European Commission issued a proposal for a new directive on privacy in the electronic communications sector to strengthen privacy rights for individuals by extending the protections that were already in place. During the process, however, the Council of Ministers began to push for the inclusion of data retention provisions, requiring Internet Service Providers and telecommunications operators to store logs of all telephone calls, emails, faxes, and Internet activity for law enforcement purposes.

Following the events of September 11, however, the political climate changed and the Parliament came under increasing pressure from member states to adopt the Council’s proposal for data retention and finally reached a

deal in favor of the Council's position. On June 25, 2002 the European Union Council adopted the new Privacy and Electronic Communications Directive as voted in the Parliament. Under the terms of the new Directive, member states may now pass laws mandating the retention of the traffic and location data of all communications taking place over mobile phones, SMS, landline telephones, faxes, emails, chatrooms, the Internet, or any other electronic communication device.

### **2.2.2 The APEC Privacy Initiative**

The 21 APEC (Asia-Pacific Economic Cooperation) economies commenced development in 2003 of an Asia-Pacific privacy standard, which can be considered as the most significant international privacy initiative since the European Union's Data Protection Directive of the mid-1990's. In February 2003, Australia put forward a proposal for the development of APEC Privacy Principles, using the 20-year old OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data (1980) as a starting point. A Privacy Sub Group was set up comprising Australia, Canada, China, Hong Kong, Japan, Korea, Malaysia, New Zealand, Thailand, and the United States. In March 2004, Version 9 of the APEC Privacy Principles was released as a public consultation draft.

The positive side of the APEC privacy initiative is that it has the potential to encourage the development of stronger privacy laws in those APEC economies that at present provide little privacy protection, and to help find a regional balance between the protection of privacy and the economic benefits of trade involving personal data. The negative side is that it also presents considerable potential dangers to long-term regional privacy protection, if it becomes a means by which the APEC economies accept a second-rate standard. Criticisms of the APEC Principles emphasize that they do not even meet the 20 year-old OECD standard, whereas they should include some significant strengthening where the OECD guidelines are now too weak.

### **2.2.3 Data Heavens and the Safe Harbor Arrangement**

The ease with which electronic data flows across borders leads to a concern that data protection laws could be circumvented by simply transferring personal information to third countries, where the national law of the country of origin does not apply. This data could then be processed in those countries,

frequently called “data heavens,” without limitations. For this reason, most data protection laws include restrictions on the transfer of information to third countries unless the information is protected in the destination country. This requirement has resulted in growing pressure outside Europe for the passage of strong data protection laws. Determination of a third country’s system for protecting privacy is made by the European Commission based on the principle that the level of protection in the receiving country must be “adequate” rather than “equivalent”. Another possible way to protect the privacy of information transferred to countries that do not provide “adequate protection” is to rely on a private contract containing standard data protection contractual clauses. This kind of contract would bind the data processor to respect the fair information practices such as the right to notice, consent, and access.

Although the Commission never issues a formal opinion on the adequacy of privacy protection in the United States, there were serious doubts whether the United States’ sectoral and self-regulatory approach to privacy protection would pass the adequacy standard set out in the Directive. The European Union commissioned two prominent United States law professors, who wrote a detailed report on the state of United States privacy protections and pointed out the many gaps in United States protection. The United States strongly lobbied the European Union and its member countries to find that United States system adequate. In 1998, the United States began negotiating a “Safe Harbor” agreement with the European Union in order to ensure the continued transborder flows of data. The idea of “Safe Harbor” was that United States companies would voluntarily self-certify to adhere to a set of privacy principles worked out by the United States Department of Commerce and the Internal Market Directorate of the European Commission. These companies would then have a presumption on adequacy and they could continue to receive personal data from the European Union. On July 26, 2000, the Commission approved the agreement. The Commission did, however, promise to re-open negotiations on the arrangement if the remedies available to European citizens proved inadequate.

Privacy advocates and consumer groups both in the United States and Europe are highly critical of the European Commission’s decision to approve the agreement, which they say will fail to provide European citizens with adequate protection for their personal data. The agreement rests on a self-regulatory system whereby companies merely promise not to violate their declared privacy practices. There is little enforcement or systematic review of

compliance and there is no individual right to appeal or right to compensation at the time of self-certification.

## 2.3 Privacy Constraints in the GeoPKDD Context

In the context of the European project GeoPKDD, the main reference to privacy regulations is directive 95/46/EC of the European Parliament and the Council, approved October, 24 1995. The directive provides both a number of definitions that are applicable to privacy preserving data publishing and mining and a number of open questions, left in the directive to further legislation of national countries, that may provide open ground for possible contributions of the project. In the sequel we recall some of such definitions and questions, and we try to relate them to the goals of the project.

Two are the basic definitions established by the directive that are important for knowledge discovery. Definition (a) states that:

“personal data” shall mean any information relating an identified or identifiable natural person (“data subject”); an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity;

Definition (b) states that:

“processing of personal data” (“processing”) shall mean any operation or set of operations which is performed upon personal data, whether or not by automatic means, such as collection, recording, organization, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, blocking, erasure or destruction.

The first general statement to take into account is premise (2):

Data-processing systems are designed to serve man; they must, whatever the nationality or residence of natural persons, respect

their fundamental rights and freedoms, notably the right to privacy, and contribute to economic and social progress, trade expansion and the well-being of individuals

This general statement establishes the two main constraints also for knowledge discovery techniques, i.e. the respect of freedom and its use for *good* purposes, i.e. the “economic and social progress, trade expansion and the well-being of individuals”.

Premise 26 establishes the general concept of *identifiable* and *anonymous* which are central to knowledge discovery. Such concepts have already been paid attention, and they have received more stringent definitions. However more research is still much needed in order to provide a firm technical ground for these concepts. Premise 26 reads as follows:

The principle of protection must apply to any information concerning an identified or identifiable person; whereas, to determine whether a person is identifiable, account should be taken of all means likely reasonably to be used either by the controller or by any other person to identify the said person; whereas the principles of protection shall not apply to data rendered anonymous in such a way that the data subject is no longer identifiable; whereas codes of conduct within the meaning of article 27 may be a useful instrument for providing guidance as to the ways in which data may be rendered anonymous and retained in a form in which identification of the data subject is no longer possible.

The last part of this article specifies that the directive may be integrated by member states by “codes of conduct”. For example, the Italian Authority for Privacy protection has issued a code of conduct for journalist. An extremely ambitious aim for our project could be the suggestion of definitions for privacy preserving data-mining to be incorporated in a code of conduct for data analysts.

It is worth noting that the directive shows clear awareness to favor the “good” use of data analysis, implicitly including also data mining. In fact, Premise 29 reads:

The further processing of personal data for historical, statistical or scientific purposes is not generally to be considered incompatible with the purposes for which the data have previously been collected provided that member states furnish suitable safeguards;

whereas these safeguards must in particular rule out the use of the data in support of measures or decisions regarding any particular individual.

It is very interesting to note that the directive implicitly refers to a concept close to *k-anonymity* in premise 40:

It is not necessary to impose this obligation<sup>2</sup> [...] if would involve disproportionate efforts, which could be the case where processing is for historical, statistic or scientific purposes; whereas in this regard the number of data subjects, the age of the data, and any compensatory measures adopted may be taken into consideration.

---

<sup>2</sup>of communicating to the owner of data the use of it

## Chapter 3

# Privacy Preserving Data Analysis

Data Mining is the process of unearthing new knowledge and information from data collected in the course of people conducting daily businesses or other customary actions of personal nature, like paying a bill in the restaurant by using a credit card, or by using a navigator system in an automobile to show the driver the way on a trip to a foreign country. Although data mining as a process can be either good or bad to the public (it depends on who is using the tools and for what purpose), the assumed willingness of the people to offer their personal information to a company (buying food from a supermarket by using a loyalty card, filling in a questionnaire related to the promotion campaign of a new product) for mining purposes most of the time has a payback for the people who want to participate into this. Although it is not always the benefits the main reason that people are giving out (or are letting others to collect) their personal information, it seems that most of the times it is a “one-way” street. For example, if people for making their purchases faster by using the Internet, are obliged to sacrifice some of their confidential information for registering to some m-commerce site, it is the case that to not provide their personal information is equivalent with not being able to use such a service, which sometimes is not even an option if we think of the hectic way of today’s life.

Local, National and International law (where ever this is applicable) and various regulatory directives are coming into place in order to protect the misuse of peoples’ personal information collected either by a public or private organization in the context of our previous discussion. Such legislative actions

which have been described earlier are asking from the data steward facility to respect the private texture of personal information and not to use this data for other purposes except the ones for which the data was collected, as well as to inform the subject about different usage of the data, unless there is a court order that calls for their collection.

In order to adhere to the existing legislation, scientists are investigating new methodologies for personal and corporate data de-identification, in such a way, that data mining will not impose any threat to the privacy of subjects participating in similar studies, or to the confidentiality of corporate secrets of competitors. Therefore, *Privacy preserving data mining*, i.e., the study of data mining side-effects on privacy, is capturing a growing attention from researchers and administrators across a large number of application domains [2, 10, 5]. This is made evident by the fact that major companies, including IBM, Microsoft, and Yahoo, are allocating significant resources to study this problem. Recent techniques which have been proposed to serve this purpose include the masking of raw data by adding noise, or the swapping of values, or the aggregation of neighboring values, cryptographic techniques for the secure sharing of private data in a collaborative environment as well as the hiding of sensitive knowledge from the data. An even more recent trend in the area is the investigation of various techniques for the privacy preserving data integration or else the so called privacy preserving file interconnection.

The privacy preserving data integration area is picking up on the new privacy threats created not only by the collection of recent forms of data like location and mobile data as well as trajectories and such but also by the unprecedented power of database systems to interconnect and link or match information originated from different data repositories putting at risk both the privacy provisions of individual data collections and the existing privacy regulations safeguarding contemporary private data.

Despite such efforts, and many important research results in the last years, there are still many open issues that deserve further investigation. One of today's critical challenges is that, despite increasing interest in privacy from academia, corporations, and government agencies, there remains a lack of technology transfer in privacy preserving data mining technologies. This problem stems from different facts: firstly, privacy concerns and data mining endeavors vary across application domains, and it is not straightforward how to generalize technical solutions from specific applications to principles; secondly, there exists an evident and obvious communication gap between scientists that develop theories and technical solutions, and the lawyers that

define the regulations regarding privacy issues in data collection and analysis.

We believe that real solutions to the challenges posed by the applications, as those ones studied within GeoPKDD project, can only be achieved through a combination of technical tools, legal regulations and social norms. On one side, a regulatory context poses challenges and constraints for novel technical solutions; in turn, new solutions might provide feedback and opportunities towards better norms and regulations. To implement this optimistic synergy, it is needed a more frequent and fruitful cooperation between the scientists and the lawyers: both sides need to be constantly aware of the progress developed by the opposite side.

By contemplating the current situation, we are convinced that there is a lot of work that need to be done in order to address a host of different circumstances ranging from new forms of data (personal, location, biometric and other) and new forms of data collection processes (point of sales, mobile commerce, stream and sensor data) to a wealth of different data mining techniques in use today for inferring knowledge and information of various forms. There are various initiatives that have been shaped up in response of these issues including the formation of independent bodies and data protection authorities whose daily task is to take action for securing private data against misuse and misconduct, the funding of research for moving the current state of the art to a broader dissemination of all kinds of information for the good of society while at the same time no sensitive information is jeopardized, and the passing of new bill amendments with respect to privacy protection and confidentiality that guarantee the smooth transition of our world from a closed to a widely open society that ensures the peoples' rights and penalizes any misbehavior.

## Chapter 4

# The Role of the Observatory

In the context of the GeoPKDD project, while we investigate the technical advances needed to embed privacy into the data mining tools, we have activated a *privacy regulation observatory*, aiming at involving the representatives of the national and European privacy authorities, as well as non-governmental privacy-related associations. The observatory will be aimed at harmonizing the activity in the project with existing regulations, which may emerge from the privacy-preserving methods developed within the project. The activities of the observatory will be to create and maintain relationships with the EC authority and with at least some of the national authorities of the countries of the partners of the consortium.

Such relationships will be aimed at implementing correctly the regulations into our methods and tools and, more ambitiously, to provide refinements of the technical regulations about privacy preserving analysis methods for future revisions of the regulations themselves.

The first steps in this direction have been the set up of regular relationships with the Italian Authority for Privacy (Autorità Garante della Privacy). Italy implemented the main European directive, Directive 95/46, in 1996 by law no. 675/96. The Authority analyzes many cases every year, and establishes sanctions when they find that the rules are violated. The directives, both the European one and the national implementations, are, as usual for directives, very declarative and qualitative.

Another important aspect is the interaction with all those organizations that recognize the need for location privacy standards. For example, Geo-priv (see the website), which is an IETF (Internet Engineering Task Force) working group aimed at examining some of the risks associated with location-

based services. Such a project has proposed several requirements for location privacy, including limited identifiability and customizable rules for controlling how data flows. Another example is Privacy International (PI), a human rights group formed in 1990 as a watchdog on surveillance by governments and corporations. A member of our project is in the advisory board of Privacy International. Dissemination will include an annual GeoPKDD workshop devoted to presenting achievements and to act as an international forum for spatio-temporal privacy-preserving data mining.

In summary, the aim of the GeoPKDD privacy observatory is to act for the Authorities as technical consultant in the field of privacy preserving data mining, as much as a mechanical engineer can help the judge in evaluating the speed of a car involved in an accident.

# Chapter 5

## Conclusions

Virtuous intentions are not enough to protect people's privacy: we have learnt in this Chapter that looking at the human mobility data (the traces) with the benevolent eye of the Historian does not prevent the possibility of privacy breaches when disseminating apparently harmless information. But, as the Scientist becomes less naïve and more aware of these threats, increasingly sophisticated privacy-preserving techniques are being studied. Their aim is to guarantee anonymity by means of controlled transformation of data and/or patterns - little distortion that avoids the undesired side-effect on privacy while preserving the possibility of discovering useful knowledge. A fascinating array of problems thus emerged, from the point of view of computer scientists and mathematicians, which already stimulated the production of important ideas and tools. Hopefully, in the near future, it will be possible to reach a win-win situation: obtaining the advantages of collective mobility knowledge without divulging inadvertently any individual mobility knowledge. This results, if achieved, may have an impact on laws and jurisprudence, as well as on the social acceptance and dissemination of ubiquitous technologies. We believe that this research effort must be tackled in a multi-disciplinary way, in such a way that the opportunities and risks may be shared by social analysts, jurists, policy makers, concerned citizens.

Indeed, there is a fourth character in this play, besides the Spy, the Historian, and the (positivist) Scientist. Let's call it the Politician, meaning with this term, broadly, both the policy makers and the public opinion. There is a list of provocative questions for the Politician: is the opportunity scenario advocated in this Chapter understood? Did the Politician fully comprehend that in the mines of mobility data that accumulate in the network operators'

databases lie veins of knowledge of inestimable value for the public community? Is there a chance that the Politician look at the privacy problem not only from the viewpoint of the Spy, but also from the viewpoint of the Historian? In these last years, there is a comprehensible increasing fear for a weakening of privacy rights that comes from the ICTs, and consequently there has been an increasing attention on media and regulations, at least in Europe, which put forward the defence from the eyes of the Spy - the defence of the rights to individual privacy obviously comes first. However, we would like to put at the Politician attention that the defence of the right of the Historian to discover socially useful knowledge is also becoming a (new!) right of the community as a whole. Our traces accumulate as an effect of our own living: why should they be simply given as a present to the service providers companies that, at most, would analyze such traces for commercial purposes? Traces are produced by the people, and therefore should be considered a public good, to be analyzed for the purpose of public interest in a safe, controlled (anonymous) way.

This play has not an end, for the moment. The best we can do so far is to express a wish: let us exercise the right to defend ourselves from the Spy together with the right to know what the Historian can tell us. To achieve this goal, we have to put the Politician and the Scientist at work together to invent regulations, methods and technologies that protect us from the risks without giving up to the opportunities.

# Bibliography

- [1] AGRAWAL, R. Privacy and data mining. In *The 15th European Conference on Machine Learning (ECML) and the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD)* (Pisa, Italy, September 2004). Invited Talk.
- [2] AGRAWAL, R., AND SRIKANT, R. Privacy-preserving data mining. In *Proceedings of the 2000 ACM SIGMOD on Management of Data*.
- [3] ATZORI, M., BONCHI, F., GIANNOTTI, F., AND PEDRESCHI, D.  $k$ -anonymous patterns. In *Proceedings of 9th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'05), Porto, Portugal, 2005*.
- [4] BETTINI, C., WANG, X. S., AND JAJODIA, S. Protecting privacy against location-based personal identification. In *Secure Data Management, Second VLDB Workshop, SDM 2005, Trondheim, Norway, September 2-3, 2005, Proceedings* (2005), vol. 3674 of *Lecture Notes in Computer Science*, Springer.
- [5] CLIFTON, C., KANTARCIOGLU, M., AND VAIDYA, J. Defining privacy for data mining. In *Natural Science Foundation Workshop on Next Generation Data Mining* (2002).
- [6] HOH, B., AND GRUTESER, M. Location privacy through path confusion. In *IEEE/CreateNet Intl. Conference on Security and Privacy for Emerging Areas in Communication Networks (SecureComm), Athens, Greece 2005*.
- [7] MALIN, B. Betrayed by my shadow: learning data identity via trail matching. *Journal of Privacy Technology*, 20050609001 (2005).

- [8] MALIN, B., SWEENEY, L., AND NEWTON, E. Trail re-identification: learning who you are from where you have been. Tech. rep., Data Privacy Laboratory, LIDAP-WP12, Carnegie Mellon University: Pittsburgh, PA, March 2003.
- [9] SWEENEY, L. Uniqueness of simple demographics in the u.s. population. Tech. rep., CMU, 2000.
- [10] VERYKIOS, V. S., BERTINO, E., FOVINO, I. N., PROVENZA, L. P., SAYGIN, Y., AND THEODORIDIS, Y. State-of-the-art in privacy preserving data mining. *SIGMOD Rec.* 33, 1 (2004), 50–57.