Personalized privacy in open data sharing scenarios

David Sánchez and Alexandre Viejo

Universitat Rovira i Virgili
Department of Computer Engineering and Mathematics
UNESCO Chair in Data Privacy
Av. Països Catalans 26
E-43007 Tarragona, Spain

Abstract

Purpose – This work proposes a privacy-preserving paradigm for open data sharing based on the following foundations: i) subjects have unique privacy requirements; ii) personal data is usually published incrementally in different sources; and iii) privacy has a time-dependent element.

Design/methodology/approach – This study first discusses the privacy threats related to open data sharing. Next, these threats are tackled by proposing a new privacy-preserving paradigm. The main challenges related to the enforcement of the paradigm are discussed, and some suitable solutions are identified.

Findings – Classic privacy-preserving mechanisms are ineffective against observers constantly monitoring and aggregating pieces of personal data released through the Internet. Moreover, these methods do not consider individual privacy needs.

Research limitations/implications – This study characterizes the challenges to the tackled by the new paradigm and identifies some promising works, but further research proposing specific technical solutions is suggested.

Practical implications – This work provides a natural solution to dynamic and heterogeneous open data sharing scenarios that require user-controlled personalized privacy protection.

Social implications – There is an increasing social understanding of the privacy threats that the uncontrolled collection and exploitation of personal data may produce. The new paradigm allows subjects to be aware of the risks inherent to their data and to control their release.

Originality/value – Contrary to classic data protection mechanisms, the new proposal centers privacy protection on the individuals, and considers the privacy risks through the whole life cycle of the data release.

Keywords Data sharing, Privacy, Personalized data protection, Data brokers.

Paper type Conceptual paper

Introduction

Within the current context of Information Societies, it is possible and generally easy to find information about our identities, habits, interests or opinions in several electronic sources such as government

databases, commercial platforms or social networks. Moreover, this shared information tends to increase over time and usually remains available for long periods, so that the data that can be gathered from us grow and become more detailed with every new release. Because personal data are considered the fuel of the modern digital economy, their collection and exploitation have become an extremely lucrative business. For example, Data Brokers, who compile and analyze consumers' information to resell it or to provide business services, such as identity verification, marketing products or personal profiling (U.S. Federal Trade Commission, 2014), have reported annual revenues of more than 1 billion USD.

In parallel with the growth of personal data sharing, there is an increasing social understanding of the privacy threats that the uncontrolled collection and exploitation of these personal data may produce, which include discriminatory actions, unethical exploitation of data or phishing. In fact, the Microsoft's 2015 annual report "How Personal Technology is Changing our Lives" (Microsoft, 2015) reported that a 64% of Internet users from developed countries think that technology advances have a negative impact on their privacy, a value that increased five points from the 2014 results. In this respect, in May 2014, the U.S. Federal Trade Commission published a report about data collection and use practices of the most relevant Data Brokers (U.S. Federal Trade Commission, 2014) that put the spotlight on the privacy risks they pose to consumers. To mitigate these issues, the Commission suggested the implementation of mechanisms (i.e., legislations and technological solutions) that would enable consumers to access their data and give them the ability to opt out of having them shared for secondary use. In this respect, the Commission also highlighted that the sensitivity of some types of information may vary from one individual to another and, thus, consumers should be given the ability to control the uses of their data according to their own privacy needs. Finally, the Commission also recommended Data Brokers to practice privacy-by-design, which includes considering privacy issues at every stage of the data management: Data Brokers should collect only the data they need and implement measures to refrain from collecting intrinsically sensitive information (e.g., healthcare data, information about children and teens, etc.).

Due to the huge amount of data to be managed, privacy-by-design requires the existence of (automatic) privacy-preserving technologies that prevent the disclosure of sensitive information through the life cycle of data release. In the last years, the scientific community has designed many anonymization mechanism and models for ensuring privacy (Kieseberg et al., 2014). More specifically, statisticians and computer scientists have extensively considered this issue within the areas of "statistical disclosure control" (Hundepool et al., 2012) and "privacy-preserving data publishing" (Fung et al., 2010), respectively. Well-known scenarios and solutions are: the protection against identity disclosure of a statistical database in which each record details information of an individual (e.g., a census database) by making records indistinguishable from each other with respect to the (quasi-)identifying attributes (Samarati and Sweeney, 1998); or the "redaction" of declassified documents, by identifying and removing sensitive terms (e.g., names, addresses, social security numbers, certain diseases, political opinions, etc.) (Bier et al., 2009) in accordance with current legislations on data privacy (The European Parliament and the Council of the EU, 1995, Terry and Francis, 2007, Department of Health and Human Services, 2000). Nevertheless, in all these scenarios, data protection is carried out as a "homogenous", "static" and "independent" task without any involvement (or knowledge) of the individuals to whom data refer; that is, data sets (usually involving data of several subjects) are homogenously protected as monolithic elements

considering neither the past or future data releases related to the individuals nor their particular privacy needs. As a result, most of the available privacy-preserving mechanisms cannot avoid the disclosure risks that arise if a capable observer, such as a Data Broker, constantly monitors, links and aggregates the either clear or protected pieces of data of a subject that are released through the Internet. Moreover, since these methods apply the same kind of protection to *all* the entities in a given dataset, they cannot support heterogeneous per-individual privacy needs, which is crucial to increase the control of the users on their own data (U.S. Federal Trade Commission, 2014). In this respect, the EU Commission has identified as a key research topic for the upcoming years the development of privacy-preserving technologies that *"offer the basis for empowering the data subjects to understand and be informed of (and, where appropriate, control) the use of their personal data"* (European Commission, 2015).

One-to-one privacy

To address the need for per-individual privacy protection in dynamic open data sharing environments, we propose what we call *one-to-one privacy*, a privacy-preserving paradigm that builds on the following premises:

- Privacy expectations and data sensitivity are relative, and depend on our privacy requirements and also on the party with which our data will be shared (e.g., our relatives, our employer, a medical doctor, etc.).
- The possibility to compile, link and aggregate our data that get incrementally published in different platforms significantly expands the knowledge that third parties (e.g., government agencies, Data Brokers, cybercriminals, etc.) gain about us, and thus, increases the privacy threats. For example, our census data could be linked to a priori anonymous political opinions or sexual preferences, even though each piece of data was published independently in different sources and at different times. The disclosure risks related to data aggregation have been extensively considered in the literature (Hundepool et al., 2012), even though in a static non-incremental way, and mainly for structured data. In fact, the incremental protection of data sets is a really challenging issue (Byun et al., 2009, Cao et al., 2011), and many data protection mechanisms can hardly cope with it (Soria-Comas and Domingo-Ferrer, 2015).
- Privacy has a time-dependent element. On the one hand, it is possible to compile and aggregate data of an individual over time due to the indefinite storage of digital data. On the other hand, the privacy risks associated to data tend to diminish as time passes.

In coherence with these premises, *one-to-one privacy* empowers subjects to *control* the protection of their data and to be aware of the risks inherent to their publication. The data protection is thus tailored to the privacy needs of the individual to whom the data refer to, and considers the whole life cycle of the data releases through time. Specifically, each new piece of individual's data to be released, either published by her or by any other entity, is protected according to: i) the privacy requirements of the individual; and ii) the record of publications that refer to that subject. In this manner, we can ensure that the aggregation of all the clear or protected data about an individual that is available at a specific time fulfills her privacy needs.

In Figure 1 we depict the general components and workflow of our approach. The main element is the privacy protection infrastructure that represents the implementation of our paradigm in a specific setting (e.g., a social network, by a privacy-concerned Data Broker, within the Intranet of a company, etc.) and receives the individuals' data to be released as input. The infrastructure manages the privacy requirements of each individual and the data release record related to them, sorted by the time of publication. The output is a *one-to-one private* version of the input data, which ensures that the aggregation of these protected data plus the already published information satisfies the individual's privacy requirements; or, in other words, that a capable observer does not learn more information about the individual from the published data than what is allowed by her privacy requirements. In this sense, the publication record of each individual represents her actual aggregated level of exposure or information disclosure. To implement the necessary data protection, our approach relies on *risk management*, which is enforced by evaluating the accumulated privacy risk that the new data to be released plus the publication record of the subject produce with regard to the privacy requirements. As a result, risky data is detected and coherently protected prior to publication. The actual data protection can be implemented, for example, by removing some terms or attribute values or, more desirably, by replacing them by less specific data (e.g., AIDS→disease) so we can better preserve the analytical utility and/or the intelligibility of the protected data. When the protected data are finally released, they are also incorporated into the publication record of the subject with an absolute time stamp stating the publication time. In this manner, we keep updated the exposure level of each individual.

Let us illustrate this workflow with a basic example: a subject named *Alice,* who lives in "Madrid", states in the privacy requirements about her location that anything more specific than her country of living should not be disclosed. According to such requirement, if, at some point, *Alice* (or any other party within the domain of the privacy protection infrastructure) tries to share with the world that she lives in "Madrid", the privacy risk assessment mechanism will consider this term as sensitive; as data protection strategy, "Madrid" will be replaced by "Spain", which is the most informative generalization that is allowed by the privacy requirements of *Alice* regarding her location. Now, let us assume that, after releasing the (protected) piece of data "I live in Spain", *Alice* tries to share information about her usual whereabouts, which include repeated visits to the "National Archaeological Museum". The name of this museum, if analyzed independently, cannot unequivocally disclose *Alice*'s location because, for example, there is a "National Archaeological Museum" in Athens, Greece or in Sofia, Bulgaria; but if this data is aggregated with the [already published and privacy-compliant] location ("I live in Spain"), a knowledgeable observer can straightforwardly infer that *Alice* likely lives in "Madrid", which is the only city in Spain with such museum. Thus, the privacy risk assessment mechanism would consider the risk of the new data with regard to the [already protected] *Alice*'s publication record and, in order to avoid a violation of *Alice*'s privacy requirements, it will propose a safe generalization to replace "National Archaeological Museum", such as "archaeological museum".

One-to-one privacy protection brings some interesting benefits over classic paradigms focusing on data releases, such as those based on the well-known *k*-anonymity model (Samarati and Sweeney, 1998). Specifically, to enforce *k*-anonymity in a dataset, records (each one referring to an individual) are made indistinguishable in fixed groups of *k* with regard to the set of attributes that may identify the subjects (e.g., addresses, jobs, ages, etc.), and which are common to all of them; as a result, *all* the

records/individuals are protected homogenously: the *same* set of attributes are protected (e.g., generalized) in the *same* manner. This is problematic because either we consider the strictest privacy needs of any subject in the database, thus hampering the utility of the protected data because of the systematic overprotection, or we protect data from the point of view of the majority, thus compromising the protection of some individuals with stricter privacy needs. In contrast, with our approach, we can protect *some* attribute values of *some* individuals more than others, in coherence with their potentially heterogeneous privacy needs. By considering these heterogeneous needs, we have that the utility of the protected data is optimized, and the assessment of the individual privacy risks is more accurate. This allows us to optimize the trade-off between privacy protection and data utility preservation, which is the *raison d'être* of privacy-preserving data sharing (Hundepool et al., 2012).

In contrast, the methods developed under the framework of the more recent ε-differential privacy model (Dwork, 2006) may provide a more appropriate alternative to enforce our paradigm. Because differential privacy is based on adding an amount of noise to the data proportional to the domain of the attributes (rather than the actual values within the dataset), it more naturally supports per-individual protection, which could be tailored for different levels of protection. However, because of the very strict privacy guarantees defined by differential privacy (i.e., the protected data should be insensitive to modifications of one individual's data in the input) the utility of general-purpose differentially private data releases is still far from being usable (Sánchez et al., 2016); because of this, differential privacy is better suited for an interactive setting in which data consumers perform queries on a trusted data provider that holds the original data and provides differentially private answers.

The other cornerstone of our paradigm is the risk management that drives data protection through the whole life cycle, which incrementally assesses the privacy risks that arise from the aggregated publication record related to the individual. Most previous works assume that the data to be released is *static*, that is, it would not be updated in the future; moreover the protection is done with respect to an also statically defined background knowledge available to external entities (e.g. fixed sets of attributes), which may enable them to disclose sensitive data (Domingo-Ferrer et al., 2016). This approach cannot cope with the *incremental* nature of data sharing in open environments, in which data gets updated and external entities may have access to new knowledge. For example, in k-anonymous data releases, incremental data releases may render previous k-anonymous generalizations invalid (Byun et al., 2009); differential privacy still provides some protection in front of the incremental publication of correlated data via sequential composition, but the privacy guarantees diminish as new data get published (Soria-Comas and Domingo-Ferrer, 2015).

Because of the incremental nature of our approach, the more data released, the stricter the protection of new data would become in order to fulfill the privacy requirements. In the worst case, we could reach a state in which no additional publications were allowed by the requirements and the publication record. To avoid this problem, we can consider publication dates because, the nearer in time the data are, the more feasible the inferences will be (because they likely refer to the same events or features), and the more sensitive the data become (because they represent the individual's current state). Thus, in our approach, the time stamps associated to the entries in the publication records can be used to decide which of them

are really relevant to be considered when protecting the current publication. This can be implemented as a time frame for data aggregation during the assessment of disclosure risks.

Challenges and solutions

To implement the *one-to-one privacy* paradigm depicted above, there are several research and technical challenges that should be tackled in order to: i) manage individuals' requirements and publication records; ii) perform an accurate assessment of privacy risks; and iii) implement an appropriate protection of the data. In the following, we discuss them and propose some practical solutions.

First, we define the privacy requirements of an individual as i) the set of topics or data types (e.g., health data, sexuality, location data, dates, etc.) that she considers sensitive because they disclose her confidential information and/or her identify; and, for each one, ii) the maximum level of information she allows to be disclosed. Moreover, if the *one-to-one* privacy protection infrastructure is deployed in controlled scenarios in which users are identified and classified via an access control mechanism (e.g., contact types in a social network, types of employees in a company, etc.), these requirements can be also defined according to the type of party that would have access to the data (e.g., relatives, friends, employer, etc.). For example, recalling previous example, *Alice* may specify that anyone can know that she lives in "Spain", but that only her relatives can know she actually lives in "Madrid".

We also assume that each person may have a very unique opinion about what is sensitive or not according to her environment and personal circumstances. Thus, ideally, each individual should state her needs on the sensitive topics, so that subjects are actually empowered to *control* the protection of their privacy. If this is unfeasible, for example, because of the lack of awareness of users about privacy risks or because of the number of topics to consider, we can rely on the notions of privacy that are defined in current legislations. For example, the E.U. Data Protection Directive (The European Parliament and the Council of the EU, 1995) defines religion, political and sexual orientation or race as sensitive topics, several U.S. federal laws on medical data (Terry and Francis, 2007) provide lists of diseases that can be potentially discriminatory and should not be disclosed, and the HIPAA (Department of Health and Human Services, 2000) states sensitive and identifiable health-related elements that should be protected in medical records. In fact, the use of these legal frameworks to define the privacy requirements would be completely aligned with the legislative recommendations made by the U.S. Federal Trade Commission with regard to the exploitation of data gathered through the Internet (U.S. Federal Trade Commission, 2014).

Ascertaining how much information can be disclosed for each sensitive topic is another essential point of the privacy requirements. This would depend on who manages the privacy requirements, that is, which organization is in charge of implementing the privacy-protection infrastructure. For example, let us consider that a social network operator like Facebook is in charge of storing the users' privacy requirements and implement the *one-to-one* privacy protection prior data publication. Because a direct interaction between the users and the social network exists, the former can define their privacy settings by answering a questionnaire or by filling a form, in the same manner as they configure other aspects of their user accounts. We can also imagine the same process in other controlled scenarios in which "trusted

central entities" exist, such as companies, governmental organizations, etc. If there are no direct interaction between the user and the management entity, which is the case of Data Brokers for which users are generally unaware of them, the privacy protection infrastructure, deployed in the Data Broker for example, can still implement predefined privacy requirements suitable to the specific groups of people it deals with (e.g., users of a healthcare system, residents of a certain country, etc.) and to the applicable legal frameworks.

Once the privacy requirements have been defined, the system should be able to assess the privacy risks of the data to be published within the context of the publication record of the individual to whom the data refers to. This is a crucial core component of our approach and, given the amount of information that is generated daily, it should be automatic. Within the area of statistical disclosure control, privacy risk assessment relies on the structure of the data (e.g., the attribute schema in a SQL data base) and a set of manually defined rules that define the background knowledge assumed to be available to external entities to perform inferences; that is, which attributes should be protected because they can be used to disclose identities and/or confidential information) (Hundepool et al., 2012). However, because of the purely statistical ground of these methods, they cannot cope with unstructured and heterogeneous data sets (Sánchez and Batet, 2016). On the other hand, in the area of document redaction/sanitization, most of the methods exploit the linguistic regularities that characterize some types of sensitive information (e.g., names, addresses, dates, Social Security numbers, etc.) in order to automatically detect them in unstructured documents by means of manually defined patterns or trained classifiers (Meystre et al., 2010). This is neither versatile nor scalable enough to deal with unstructured sensitive data, such as the outcomes of a medical visit, political or sexual opinions posted in a social network, etc. To overcome these limitations, we need automatic mechanisms that, much like human sanitizers do (Bier et al., 2009), consider the *meaning* that the data to be published disclose. At this respect, recent approaches rely on the Information Theory as a mechanism to quantify the semantics disclosed by the data as a function of their informativeness (Sánchez and Batet, 2016, Chow et al., 2008, Anandan and Clifton, 2011). The idea is that those pieces of data, either being textual terms or semantically rich numerical values in a document or attribute values in a database, which disclose a large amount of information/semantics about a certain sensitive topic constitute a privacy threat. This inherently semantic approach perfectly fits with the needs of our paradigm because it is fully automatic, it is not restricted to specific data types and does not assume any regular structure on the data, and can be enforced with respect of the topics defined by each individual in her privacy requirements.

On the other hand, many of the methods available in the literature evaluate the privacy risks of pieces of data in an independent way and, as a result, they neglect the risks that arise from the semantic inferences enabled by an aggregated analysis of the data that are incrementally published. This is a realistic threat that has been considered by the scientific community in the past (Anandan and Clifton, 2011, Chow et al., 2008), and that is also highlighted in (U.S. Federal Trade Commission, 2014) due the fact that Data Brokers not only use raw individual's data, but they also aggregate and analyze such data to perform inferences and classify users; an example of the actual inferences that Data Brokers currently perform include classifying a person as "diabetic" because of her constant interest in sugar-free products and diets. To identify these risky inferences, we need mechanisms capable of evaluating the semantic relationships between terms and to measure the accumulated risk associated to their co-occurrence in one or several

publications. This is certainly challenging, because the aggregated risk cannot be computed by simply adding the risks of individual terms or publications, but by quantifying the semantics that the aggregated data discloses about a certain sensitive topic. For structured data bases, the notion of *quasi-identifying attributes* (Samarati and Sweeney, 1998), which are combinations of attributes that individually do not disclose sensitive information but that they do in aggregate, exactly captures this notion. However, in these approaches, quasi-identifying attributes are manually defined beforehand according to the knowledge that an expert assumes to be available for external attackers. In contrast, information theoretic approaches (Sánchez and Batet, 2016, Anandan and Clifton, 2011) are able to automatically detect risky combinations of terms according to their co-occurrence and mutual information toward a sensitive topic, which measure the amount of (sensitive) semantics that the former disclose from the latter. In our approach, the combinations of terms to evaluate would be the aggregation of those in the publication record of the individual (filtered by their publication time stamp that state their validity) and those in the new data to be published, and the sensitive topics would be those defined in the privacy requirements of the individual. To achieve accurate assessments of risky inferences, large data sources should be employed to estimate the (co-)occurrence of information; in this respect, the research community have extensively used the Web (Staddon et al., 2007, Chow et al., 2008, Sánchez et al., 2014), which, due to its size and heterogeneity, realistically captures the information distribution at a social scale (Cilibrasi and Vitányi, 2006).

Once we are able to automatically identify those pieces of data in the new publication that cause a privacy risk with respect to the privacy requirements, we need to implement a protection mechanism to avoid or mitigate enough that risk. Different protection approaches based on data transformations can be found in the literature (Hundepool et al., 2012). Many of them are based on creating groups/clusters of similar –risky- data (e.g., records of different individuals in a data base containing the same set of attributes), and making them indistinguishable by replacing them with a common value according to a privacy criterion such as *k*-anonymity (Samarati and Sweeney, 1998); because of this indistinguishability, disclosure inferences are not unequivocal anymore, thus lowering the privacy risk. Even though most of these approaches focus on "static" anonymization (i.e., input data are completely anonymized and released once), recently, some authors have proposed mechanisms to support incremental data releases/updates (Cao et al., 2011, Byun et al., 2009). However, data protection is still "homogenous" for all the records and can be only applied if: i) a set of individuals' data with the same structure is available; and ii) the privacy requirements are equal to all the individuals. Because in our approach we assume users with heterogeneous privacy needs, we require protection methods that are able to protect values according to a per-individual privacy criterion. Mechanisms that fit with this need are: data *suppression* (or *redaction*), which eliminates or blacks-out sensitive terms but severely hampers the utility of the data (Bier et al., 2009); independent *generalization* (or recoding), which replaces terms or values by less detailed versions (e.g., "AIDS"→"chronic disease", "19 years"→ "[15-25] years") to reduce the amount of disclosure while still retaining a subset of the data semantics and, thus, data utility (Sánchez et al., 2014); and *noise addition*, which distorts data by adding a constrained random error to the original value (for numerical data, e.g., "19 years" →(+2)→"21 years") (Domingo-Ferrer et al., 2004) or by randomly replacing terms by semantically similar ones (for textual data, e.g., "pneumonia"→"bronchitis") (Rodríguez-Garcia et al., 2015).

Finally, even though privacy-preserving data transformation methods can be applied to all scenarios, we can also rely on access control as an alternative in controlled environments in which users are authenticated, and the access to the resources is managed by a centralized entity that implements the privacy protection infrastructure (e.g., a social network operator, corporative managers, etc.). Our idea is to keep the data untouched and to allow or deny the access of a party to a specific resource according to the above-detailed analysis of risks of the resource and the privacy requirements defined by the data owner towards that type of party (e.g., my *relatives* can learn *anything* about my *health*). In this manner, authorized parties will gain full access to the data and, thus, perfect utility, whereas unauthorized entities will not learn anything of the individual.

Proof-of-concept case study: a privacy-enabled social network

In the following, we illustrate the possibilities of *one-to-one privacy* in one of the most used and privacy-challenging data sharing scenarios: a social network. On the one hand, a social network offers a natural solution to implement the *one-to-one privacy* infrastructure, since network operators can centralize, store and manage user requirements and publication records, and evaluate the privacy risks of new publications from the perspective of the user to whom the data refers to. On the other hand, most of the data published in social networks are unstructured plain texts, which are challenging to analyze and protect by standard statistical methods. Finally, social networks are a core source of personal data for Data Brokers (U.S. Federal Trade Commission, 2014); thus, in absence of legal frameworks that regulate the compilation and exploitation of personal data by third parties in such open environment, social network users can use our paradigm as a preventive data protection mechanism.

Specifically, by means of a proof-of-concept implementation of our approach in a simulated social network, we show how a network operator such as Facebook or Twitter can follow the privacy-by-design principle and implement *one-to-one privacy* to address the privacy concerns of their users (Microsoft, 2015) and empower users to control their privacy (European Commission, 2015) by i) allowing users to define their own privacy requirements and making them aware of the privacy risks inherent to their publications; ii) suggesting actions to mitigate those risks; and iii) enforcing data protection in an automatic and personalized way.

First, the social network operator should ask their users about their privacy requirements. This can be intuitively done once, by asking them about the maximum levels of disclosure they allow for a set of sensitive topics at the moment of creating the user accounts. For example, let us assume a user named *Bob*; Figure 2 shows a sample form with *Bob*'s personal choices in coherence with the topics defined as sensitive in current legislations on data privacy (The European Parliament and the Council of the EU, 1995, Terry and Francis, 2007, Department of Health and Human Services, 2000) (e.g., disorders such as *AIDS*, *STDs*, or *substance abuse* are stated as sensitive for medical records). With this, the social network operator stores *Bob*'s requirements and initializes his –empty- publication record.

Now, let us suppose that *Bob* submits for publication his first message: "I've got HIV". Figure 3a depicts a dialog implemented in the proof-of-concept social network that allows *Bob* to introduce his publication

and to check if it is compliant with his privacy requirements. By pressing the "Privacy Check" button, the privacy protection infrastructure implemented by the social network analyzes each term of the message with respect to *Bob*'s requirements. We have enforced this according to the information theoretic assessment of disclosure risks in (Sánchez and Batet, 2016), which we have introduced above. In a nutshell, the system iteratively evaluates the relative co-occurrence between each term in the message and each element in each sensitive topic in the privacy requirements; then, it checks whether their mutual information is larger than the information content of the level of disclosure allowed by *Bob* for the corresponding topic, because this would mean that the data to be published disclose more sensitive information that what the user have stated in his requirements. In this case, since the term "HIV" -in the message- and the term "AIDS" -included as a sensitive element in the *health* topic of the requirements- co-occur very often (e.g. by querying those terms in Google, more than 92% of the web sites talking about *HIV* explicitly mention *AIDS*), they have very large mutual information; moreover, since this mutual information is larger than the maximum level of disclosure allowed by *Bob* for the *health* topic (i.e., the information content of the term *chronic disorder* that is the maximum information that *Bob* wants to disclose), the system concludes that the term "HIV" should not be published in clear because it would likely disclose that *Bob* suffers from *AIDS*, thus violating his privacy requirements, and it warns the user about this risk.

Then, as a privacy-protection measure, the system suggests *Bob* that the term "HIV" in the message should be replaced (generalized) by "virus", whose mutual information with regard to "AIDS" is lower than the allowed level of disclosure, because there are thousands of viruses that do not cause *AIDS*. As done in (Sánchez and Batet, 2016), to measure term (co-)occurrences and to calculate the information content of terms and their mutual information, we used the Web as corpora and the hit count provided by a Web search engine as a proxy to gather these statistics; then, to retrieve term generalizations to be used as privacy-preserving replacements (e.g. "virus" for "HIV"), we used WordNet, which is a widely-used taxonomically structured lexical database. Finally, as shown in Figure 3b the protected message is published and the system stores it in *Bob*'s publication record. Note that, in this scenario, users would have the final decision about accepting or discarding any suggested term replacement in order to avoid any possible uncomfortable situation, and also to supervise the protection process that, as any automatic process, cannot achieve *perfect* accuracy (Holzinger, 2016, Kieseberg et al., 2016).

Continuing with the example, some days later, *Bob* is willing to publish the text: *"*I was transmitted because of a blood transfusion". The system now evaluates the mutual information between the sensitive elements in *Bob*'s requirements (such as "AIDS") and the aggregation (co-occurrence) of the terms in the new message ("blood transfusion") and those in *Bob*'s publication record ("virus"). Since there are many viral diseases other than *AIDS* that can be transmitted via blood contact, mutual information toward the sensitive elements is now smaller than the informativeness defined by the maximum level of disclosure in the *health* requirement, which indicates a higher degree of uncertainty in the inferences; hence, the disclosure assessment concludes that the new message is safe and can be published in its original form, as shown in Figure 3b.

Finally, some days later, *Bob* decides to publish the message *"*I was diagnosed last June, when my immune system responded poorly to influenza". As shown in Figure 3c, the system now aggregates the

terms in this new message with those two in the publication record (because they are close in time) and evaluates them against the privacy requirements. Now, because of the very univocal relationship between "virus", "blood transfusion" -in *Bob*'s publication record-, "immune system", "influenza" -in the new message- and "AIDS" -in *Bob*'s requirements-, mutual information figures are higher than allowed and the system concludes that the terms "immune system" and "influenza" are too risky to be published in clear, because they disclose more information than the fact that *Bob* suffers from a *chronic disease*; in fact, an observer can easily infer *AIDS* as such *chronic disease* according to those evidences. As shown in Figure 3c, the system warns the user about this risk and, as shown in Figure 3d, both terms are finally replaced by the less informative and safer generalizations "body" and "disease" (retrieved from WordNet), respectively.

Other applications and future directions

Our *one-to-one privacy* paradigm can also be applied in scenarios less controlled and centralized than social networks. For example, Data Brokers may adopt *one-to-one privacy* to fulfill the privacy-by-design recommendations stated by the U.S. Federal Trade Commission and consider the privacy issues at every stage of the product development (i.e., data collection, aggregation, analysis and usage) (U.S. Federal Trade Commission, 2014). To do that, Data Brokers should first consider the privacy requirements of the data owners prior to exploiting the data or, at least, they should incorporate generic requirements coherent with the applicable legislations. Then, they can rely on the automatic assessment of privacy risks to detect the potential semantic inferences enabled by the compiled data, and refrain from compiling or using sensitive data that may be used for eligibility determinations or for unlawful discriminatory purposes, as suggested by the Commission.

In a different context, the recent Snowden and Wikileaks scandals have made companies more conscious about the potential damage that insiders, who are gradually exposed to more sensitive data, may cause. To mitigate such threats, companies may enforce a privacy-driven access control mechanism by means of the *one-to-one privacy* paradigm. First, they can use accounting to monitor the accesses of employees to classified data; then, they can rely on the assessment of privacy risks to quantify the aggregated amount of sensitive data to which an employee has accessed. This last metric is also referred as *Misuseability Score* (Harel et al., 2012), and quantifies the level of harm that might be inflicted by an employee in the hypothetical case of data leakage. According to this score, central authorities may deploy dynamic access control policies and decide whether to grant the access to new content to specific employees, or may detect individuals with unusually high scores. In this scenario, privacy requirements that drive the risk assessment mechanism can be stated by the company with regard to the type of data to be accessed and the credentials of specific employees.

These and other data sharing scenarios found in modern societies are characterized by the high dynamicity and heterogeneity of the data, and by the need for implementing a user-controlled personalized privacy protection; under these circumstances, our privacy paradigm provides a natural solution. This work identified and contextualized recent works that can help to enforce our paradigm, and lays the foundation stone of our further research, which will focus on proposing technical solutions to the

identified challenges and developing practical implementations of our paradigm in realistic settings; moreover, we hope that it will motivate other researchers in advancing in the same direction.

References

Anandan, B. and Clifton, C. (2011), "Significance of term relationships on anonymization", *IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology Workshops*, Lyon, France.

Bier, E., Chow, R., P. Golle, T. H. King and Staddon, J. (2009), "The Rules of Redaction: identify, protect, review (and repeat)", *IEEE Security and Privacy,* Vol. 7 No. 6, pp. 46-53.

Byun, J.-W., Li, T., Bertino, E., Li, N. and Sohn, Y. (2009), "Privacy-preserving incremental data dissemination", *Journal of Computer Security,* Vol. 17 No. 1, pp. 43-68.

Cao, J., Carminati, B., Ferrari, E. and Tan, K.-L. (2011), "CASTLE: Continuously Anonymizing Data Streams", *IEEE Transactions on Dependable and Secure Computing,* Vol. 8 No. 3, pp. 337-352

Cilibrasi, R. L. and Vitányi, P. M. B. (2006), "The Google Similarity Distance", *IEEE Transactions on Knowledge and Data Engineering,* Vol. 19 No. 3, pp. 370-383.

Chow, R., Golle, P. and Staddon, J. (2008), "Detecting Privacy Leacks Using Corpus-based Association Rules", in *14th Conference on Knowledge Discovery and Data Mining*, Las Vegas, NV, pp. 893-901.

Department of Health and Human Services (2000), "The health insurance portability and accountability act of 1996".

Domingo-Ferrer, J., Sánchez, D. and Soria-Comas, J. (2016), *Database Anonymization: Privacy Models, Data Utility, and Microaggregation-based Inter-model Connections,* Morgan & Claypool Publishers.

Domingo-Ferrer, J., Sebé, F. and Castellà-Roca, J. (2004), "On the Security of Noise Addition for Privacy in Statistical Databases", *Privacy in Statistical Databases*, Springer, pp. 149-161.

Dwork, C. (2006), "Differential privacy", in *33rd International Colloquium on Automata, Languages and Programming*, pp. 1-12.

European Commission (2015), "Big data PPP: privacy-preserving big data technologies".

Fung, B., Wang, K., Chen, R. and Yu, P. S. (2010), "Privacy-preserving data publishing", *ACM Computing Surveys,* Vol. 42 No. 4.

Harel, A., Shabtai, A., Rokach, L. and Elovici, Y. (2012), "M-Score: A Misuseability Weight Measure", *IEEE Transactions on Dependable and Secure Computing,* Vol. 9 No. 3, pp. 414-428.

Holzinger, A. (2016), "Interactive machine learning for health informatics: when do we need the human-in-the-loop?", *Brain Informatics,* Vol. 3 No. 2, pp. 119-131.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E. S., Spicer, K. and Wolf, P.-P. d. (2012), *Statistical Disclosure Control,* Wiley.

Kieseberg, P., Malle, B., Frühwirt, P., Weippl, E. and Holzinger, A. (2016), "A tamper-proof audit and control system for the doctor in the loop", *Brain Informatics,* Vol. to appear.

Meystre, S. M., Friedlin, F. J., South, B. R., Shen, S. and Samore, M. H. (2010), "Automatic de-identification of textual documents in the electronic health record: a review of recent research", *BMC Medical Research Methodology,* Vol. 10 No. 70.

Microsoft (2015), "2nd Annual Report on How Personal Technology is Changing our Lives".

Rodríguez-Garcia, M., Batet, M. and Sánchez, D. (2015), "Semantic Noise: Privacy-protection of Nominal Microdata through Uncorrelated Noise Addition", *27th International Conference on Tools with Artificial Intelligence*, Vietri Sul Mare, Italy, IEEE Computer Society.

Samarati, P. and Sweeney, L. (1998), "Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression", *Technical Report*, SRI International

Sánchez, D. and Batet, M. (2016), "C-sanitized: a privacy model for document redaction and sanitization", *Journal of the Association for Information Science and Technology,* Vol. 67 No. 1, pp. 148-163.

Sánchez, D., Batet, M. and Viejo, A. (2014), "Utility-preserving sanitization of semantically correlated terms in textual documents", *Information Sciences,* Vol. 279, pp. 77-93.

Sánchez, D., Domingo-Ferrer, J., Martínez, S. and Soria-Comas, J. (2016), "Utility-preserving differentially private data releases via individual ranking microaggregation", *Information Fusion,* Vol. 30, pp. 1-14.

Soria-Comas, J. and Domingo-Ferrer, J. (2015), "Big data privacy: challanges to privacy, principles and models", *Data Science and Engineering,* Vol. 1 No. 1, pp. 1-8.

Staddon, J., Golle, P. and Zimmy, B. (2007), "Web-based inference detection", in *16th USENIX Security Symposium*, p. Article No. 6.

Terry, N. and Francis, L. (2007), "Ensuring the privacy and confidentiality of electronic health records", *University of Illinois Law Review,* Vol. 2007 No. 2, pp. 681-735.

The European Parliament and the Council of the EU (1995), "Data Protection Directive 95/46/EC".

U.S. Federal Trade Commission (2014), "Data Brokers, A Call for Transparency and Accountability".