Semantic Disclosure Control: semantics meets data privacy

David Sánchez[a], Montserrat Batet[b]

*[a]Department of Computer Engineering and Mathematics,
UNESCO Chair in Data Privacy, Universitat Rovira i Virgili,
Av. Països Catalans 26, E-43007 Tarragona, Catalonia (Spain)*

*[b]Internet Interdisciplinary Institute (IN3),
Universitat Oberta de Catalunya,
Avda. Carl Friedrich Gauss, 5, 08860 Castelldefels, Barcelona, Catalonia (Spain)*

Abstract

Purpose – To overcome the limitations of purely statistical approaches to data protection, this paper proposes *Semantic Disclosure Control* (SeDC): an inherently semantic privacy protection paradigm that, by relying on state of the art semantic technologies, rethinks privacy and data protection in terms of the *meaning* of the data.

Design/methodology/approach – The need for data protection mechanisms able to manage data from a semantic perspective is discussed and the limitations of statistical approaches are highlighted. Then, SeDC is presented by detailing how it can be enforced to detect and protect sensitive data.

Findings – So far, data privacy has been tackled from a statistical perspective; that is, available solutions focus just on the distribution of the data values. This contrasts with the semantic way by which humans understand and manage (sensitive) data. As a result, current solutions present limitations both in preventing disclosure risks and in preserving the semantics (utility) of the protected data.

Practical implications – SeDC captures more general, realistic and intuitive notions of privacy and information disclosure than purely statistical methods. As a result, it is better suited to protect heterogenous and unstructured data, which are the most common in current data release scenarios. Moreover, SeDC preserves the semantics of the protected data better than statistical approaches, which is crucial when using protected data for research.

Social implications – Individuals are increasingly aware of the privacy threats that the uncontrolled collection and exploitation of their personal data may produce. In this respect, SeDC offers an intuitive notion of privacy protection that users can easily understand. It also naturally captures the (non-quantitative) privacy notions stated in current legislations on personal data protection.

Originality/value – On the contrary to statistical approaches to data protection, SeDC assesses disclosure risks and enforces data protection from a semantic perspective. As a result, it offers more general, intuitive, robust and utility-preserving protection of data, regardless their type and structure.

Keywords Privacy, Personal data protection, Semantics, Knowledge.

Introduction

In the current context of information societies, it is quite common to refer to electronic data as "the new oil" of the XXI century (Rotella, Apr 2, 2012). On the one hand, the analysis of personal data *fuels* many research efforts (e.g., the analysis of medical records is essential to improve healthcare delivery). On the other hand, predictive market analytics derive *value* from the huge amount of personal data being gathered; for example, the compilation, aggregation and exploitation of data (e.g., social media) related to millions of Internet users is a billionaire business in which Data Brokers are the main providers of data and services, which include identity verification, marketing products, personal profiling, etc. (U.S. Federal Trade Commission, 2014). Even though there is no question that those services are of great interest for companies and consumers, at the same time, the confidential nature of many of the compiled data (e.g., census data gathered from government sources, personal opinions and preferences posted in social networks, medical records, etc.) may pose privacy risks to the subjects whom data refers to.

In order to guarantee the fundamental right to privacy of the individuals (The European Parliament and the Council of the EU, 2016), responsible parties should undertake appropriate protection measures. *Data protection* or *sanitization* methods are used to remove, coarse or perturb sensitive information prior releasing sensitive data, which may either reveal confidential information (e.g., salaries, political or sexual orientations, etc.) and/or enable the re-identification of an individual (e.g., names, addresses, ages, jobs, etc.). Data protection should be done in a way that the protected outcomes are still analytically useful in a variety of tasks (e.g., research), which is the main motivation underlying to data releases.

Most data protection methods proposed in the literature within the areas of *Statistical Disclosure Control* (Hundepool et al., 2013) and *Privacy Preserving Data Publishing* (Fung et al., 2010) manage data from a purely statistical perspective. Because they just focus on the distribution of (mainly numerical) data rather than on their actual *meaning*, they can hardly cope with the privacy threats that arise in current data release scenarios (Sánchez and Batet, 2016), in which data are usually heterogeneous (numerical, categorical, textual), dynamic (continuously generated, transient and even unbounded) and unstructured (e.g., messages published in social networks, emails and documents exchanged between stakeholders, etc.). Moreover, the plain statistical analyses implemented by those approaches contrast with the *semantic* manner by which humans (either data producers, readers and even attackers) understand, manage, exploit and protect information.

So far, protection mechanisms considering data semantics up to some degree have been scarce, scattered and unconnected. To provide a common ground for such works, and thus, to pave the way for further research on this topic, this paper defines and discusses the main challenges underlying to an inherently semantic data protection paradigm, named *Semantic Disclosure Control* (SeDC). The aim of SeDC is to mimic, from a formal perspective, the reasoning by which humans, either potential attackers, data sanitizers or data analysts, respectively disclose, protect or extract conclusions from data, regardless of their type and structure. To do so, SeDC relies on the maturity of semantic technologies, such as semantic analyses, knowledge bases or semantic operators. With this, and with respect to statistical approaches to data protection, SeDC-based solutions have the potential to capture more general, realistic and intuitive notions of privacy and information disclosure, to enforce sound and utility-preserving data protection and to cope with more challenging data protection scenarios.

The remaining of the paper discusses the need of SeDC and the advantages it brings over statistical solutions. After that, it identifies the main challenges to be tackled, proposes suitable solutions for each of them by relying and characterizing the incipient works on semantic data protection and identifies the shortcoming and needs of the state of the art. Finally, the benefits of the new paradigm are illustrated through several application scenarios.

Limitations of statistical data protection

Traditionally, data protection has been performed manually, in a process by which human experts detect and prevent *information disclosure risks* (Bier et al., 2009), that is, the chance of discovering sensitive data (identities and/or confidential information) by means of *direct* or *indirect semantic inferences*. For example, SS numbers, salaries, or disease names, explicitly reveal sensitive data by *direct inference*; on the other hand, treatments or drugs that are (semantically) related to a sensitive disease, readings that may suggest political preferences, or personal habits that can be associated to religious or sexual orientations may reveal sensitive information via *indirect inference* (Chow et al., 2008).

Because manual protection efforts can hardly cope with the volume of electronic data being generated and the complexity of dealing with *indirect inferences*, researchers have proposed a variety of automatic protection mechanisms. On the one hand, many algorithms have been proposed in the area of *Statistical Disclosure Control* (SDC), which aim at minimizing the damage to the statistical properties of the original data (e.g., means, variances, correlations) while lowering enough the probability of disclosure. To do so, *data protection methods* transform the original data by performing, among others, data suppressions, sampling, noise addition, swapping, recoding or microaggregation (Hundepool et al., 2013). Information disclosure has also been tackled by the computer science community in the field of *Privacy-Preserving Data Publishing* (PPDP) (Fung et al., 2010); whereas SDC methods are concerned with the statistical validity of data but offer vague privacy guarantees, PPDP approaches usually rely on formal *privacy models* (e.g., k-anonymity, t-closeness, ε-differential privacy), which offer *ex ante* privacy guarantees defined in terms of *information distribution* (Domingo-Ferrer et al., 2016). By instantiating privacy models, users have a clearer understanding of the privacy protection regardless of the specific data set. The enforcement of privacy models usually relies on constrained versions of SDC methods; for example, data suppression, generalization or microaggregation are commonly applied to enforce k-anonymity (Samarati, 2001), whereas ε-differential privacy relies on noise addition (Dwork, 2006).

However, because all the solutions above assess the information disclosure and define privacy guarantees from distributional and statistical perspectives, they face limitations regarding both their practical feasibility and the utility of the protected outcomes:

- Data protection methods require an expert to manually state *all* the data pieces in each data set that may cause disclosure and should be protected. For example, in a structured database, the expert is required to specify which attributes can, individually or in aggregate, re-identify an individual (Samarati, 2001); whereas, in a textual document or a document collection, the expert should tag which textual terms may cause disclosure, either directly or via semantic inferences (Bier et al., 2009).
- Privacy models use quantitative privacy parameters to express privacy guarantees: with k-anonymity, k states the minimum probability of reidentification resulting from making records indistinguishable (Samarati, 2001); with t-closeness, t defines the difference between the distributions of confidential attributes in the original and protected data sets (Li and Li, 2007); and

with ε-differential privacy, ε states the probability that the protected outcome is insensitive to changes in one input record (Dwork, 2006). These abstract parameters (and the privacy guarantees they offer) are often difficult to understand by the end user (Anandan et al., 2012), and can hardly capture the qualitative and inherently semantic requirements of current legal frameworks (Department of Health and Human Services, 2000, The European Parliament and the Council of the EU, 2016), which focus on the content of the data, rather than on their distributional features. For example, the EU General Data Protection Regulation states that sensitive data are such that concern the subject's race, ethnicity, political opinion, religion, trade union membership, health status or sex life; many US Federal laws on medical data privacy state that medical records should be sanitized in order to remove information that may disclose diseases such as AIDS/HIV, sexually transmitted diseases (STDs), mental disorders or substance abuse.

- Because sensitive attributes are protected according to their distributions in a data set, the protection is enforced *homogenously* for *all* the records in the data set. This limits the application protection mechanisms to structured and uniform data collections, such as relational databases, in which individuals are described by means of a uniform set of univalued attributes (i.e., *microdata sets*) (Hundepool et al., 2013). Moreover, these mechanisms cannot cope with heterogeneous sensitivities and per individual protection needs (e.g., patients with *AIDS* as diagnosis will be more sensitive than those with *flu*), and cannot protect records individually and independently, which is needed if data are created and released as a stream (Sánchez and Viejo, 2017).
- Due to their strict mathematical roots, most methods only try to preserve the statistical features of numerical attributes (e.g., mean, variance). Even though some methods have also been applied to categorical attributes (e.g., race, job) (Domingo-Ferrer and Torra, 2005), most of them consider them "flat" categories that can only be deemed as identical or completely different; this approach neglects and poorly preserves the semantics (and, thus, the utility) of such type of data (Martínez et al., 2013). In fact, most of the information that are nowadays involved in data releases and are unstructured, heterogeneous and textual (U.S. Federal Trade Commission, 2014).

Semantic Disclosure Control

To overcome the limitations imposed by the lack of semantic background of current data protection methods, this paper proposes an inherently semantic privacy protection paradigm named *Semantic Disclosure Control* (SeDC), which rethinks data privacy in terms of their semantics. Because semantics define the *meaning* of data and this is precisely what humans exploit to understand, manage and transform data, SeDC opens the door to develop realistic solutions that: *i*) more accurately mimic human reasoning and better cope with the actual privacy threats of current data releases; *ii*) are more flexible because they are not limited to numerical, homogeneous and structured data, but to any kind of semantically rich data (e.g. structured or unstructured text or meaningful numbers); and *iii*) offer *semantic* privacy guarantees that are more intuitive for the users. By managing data according to their semantic content, SeDC also aims at offering a protection that better preserves data semantics and, thus, utility. SeDC also contextualizes and characterizes the common ground of recent works exploiting semantic technologies for privacy protection, which we discuss in the following sections.

Figure 1 shows a multilayered view of SeDC in comparison with statistical approaches. The first layer "SDC vs. SeDC" illustrates the leitmotiv of SeDC: it fills the gap that Statistical Disclosure Control cannot cover. Specifically, statistics are not enough to capture all the dimensions of data privacy, that is, the disclosure risk inherent to (direct or indirect) semantic inferences, and data utility understood as the preservation of the *meaning* of data rather than just their statistical features. Semantics, on the other hand, provide a complementary and wider coverage of data privacy because they allow managing and transforming data according to their meaning, as done by human beings. Note that this does not mean to drop statistics at all when performing semantic analyses. As illustrated by the overlapping between the dashed circles, statistical tools can measure some semantic dimensions such as the strength of a semantic relationship between terms as a function of their co-occurrence (Chow et al., 2008). Moreover, as illustrated by the area of privacy that semantics do not cover, statistics exclusively captures some features of the data (e.g., statistical validity of protected data) and support abstract numerical attributes that lack semantic content.

Take in Figure (1)

Figure 1. Multilayered view of *Semantic Disclosure Control* vs. statistical approaches (SDC): tools and supported data structures and types

To enforce SeDC, two main tasks are considered. First, as discussed above, current methods assume that the data that may cause disclosure and that should be protected are manually identified beforehand by a human expert. Since many disclosure inferences are the result of the semantic relationships between data pieces, by understanding data semantics and by mimicking the reasoning of human sanitizers, SeDC can automatize the assessment of disclosure risks. Second, semantically-grounded protection methods should be defined and applied over the data that may cause disclosure. In semantic terms, a suitable protection method should prevent risky semantic inferences while retaining the meaning (i.e., utility) of the data as much as possible. In the following, the challenges related to these tasks are discussed, and several semantic tools and methods suitable to tackle them (shown in the "Tools" layer in Figure 1) are identified.

Automatic assessment of disclosure risks

To assess disclosure risks from a semantic perspective, three main challenges are identified.

First, the study of the semantics underlying data disclosure, that is, to characterize the different kinds of disclosure that may happen according to the type of semantic relationship that links the data to be released with data available to/known by attackers to perform inferences. By relying on the knowledge engineering theory, it is possible to study the different kinds of semantic relationships that appear between entities, which can be either taxonomic (hyponymy, hypernymy, instance of, synonymy) and non-taxonomic (causality, meronymy, holonymy, etc.). Features associated to the subjects, such as properties (e.g., religion professed by a subject stated in her personal profile), or those associated to the relationship itself, such as logical axioms (i.e., reflexivity, transitivity, inverse, symmetry, etc.), should be considered as evidences that may ease semantic inferences. In the trivial case, the semantics of a potentially sensitive term, such as being *Protestant*, is completely disclosed by the presence of one of its taxonomical specializations, such as being *Methodist*. In most other cases, partial disclosure will occur, which will depend on the kind and number of semantic relationships and/or semantic features that relate the entities (e.g., a collection of symptoms could partially or completely disclose a disease). In (Sánchez and Batet, 2016) a formal study of the disclosure risk underlying to taxonomic and generic non-taxonomic relationships is presented. Additional work is required to characterize the risks brought by more specific semantic nuances, such as attributes and axioms associated to semantic relationships.

A second challenge consists in quantifying the *amount* of disclosure that actually occurs, as a function of the strength or closeness of the semantic relationship between the information that should be protected from disclosure (e.g., the sexuality or health status of an individual) and the data to be released (e.g., her medical record). With this, it is possible to design quantitative disclosure evaluation measures to automatically detect sensitive data and to evaluate up to which level such data should be protected. To measure the closeness of a semantic relationship, it is possible to use well-studied techniques proposed in the fields of knowledge acquisition and information extraction, such as semantic similarity measures (Batet and Sánchez, 2014), association rule learning (Chow et al., 2008) or information theoretic approaches (Anandan and Clifton, 2011). In fact, the latter have been recently applied to the protection of unstructured plain texts (Anandan and Clifton, 2011, Sánchez et al., 2013) by relying on the observation that the semantics encompassed by any entity (i.e., either a sensitive topic to be protected or a textual term appearing in a document that may cause disclosure of the former) can be quantified by the amount of information it provides, that is, its *Information Content* (IC). The IC is defined as the inverse of the probability of occurrence of the entity. Under the same premise, the amount of semantics that terms appearing in a document *disclose* about a sensitive entity or topic to be protected can be measured according to their overlap of information, that is their *Point-wise Mutual Information* (PMI), which is the ratio between their joint and marginal probabilities of occurrence.

For example, let assume that a hospital desires to release the medical record of a patient that contains the terms *radiotherapy* and *pain*, but does not want to disclose that the patient had *cancer*. Figure 2 represents the informativeness and the mutual information of such terms computed from their probability of (co-)occurrence in the web, which have been gathered from the hit counts provided by Google when querying the terms. Specifically, *IC*(*cancer*) (dashed circle) quantifies the sensitive semantics that should not be disclosed. Because *radiotherapy* and *cancer* share a large amount of information (shaded area in Figure 2, *PMI*(*radiotherapy*, *cancer*)≈*IC*(*cancer*) due to many cancers being treated with *radiotherapy*), it results that *radiotherapy* discloses most of the semantics of *cancer* and, thus, produces a large disclosure risk; on the other hand, the information overlap between *pain* and *cancer* is much lower and, because of this, *pain* is not risky (i.e., semantic inferences will be ambiguous since there are many diseases other than *cancer* that cause *pain*).

Take in Figure (2)

Figure 2. Circles: information/semantic content of *radiotherapy*, *cancer* and *pain*; filled areas: information overlap/semantic disclosure between *radiotherapy* and *cancer,* and between *pain* and *cancer*.

By relying on information theory, the former assessment of disclosure risks can be extended to measure the disclosure caused by *groups* of terms w.r.t. a sensitive topic (e.g., several symptoms or treatments of the same disease), and also to protect a source of information (e.g., a medical record) w.r.t. *several* non-independent sensitive topics (e.g., several sensitive diseases).

Finally, as a need inherent to the two former challenges, it will be also necessary to assess and formalize the knowledge that attackers may gather from the available data sources (e.g., published census data, publications in social networks, other sanitized sources, etc.) to perform semantic inferences and to disclose sensitive data. For example, in the information theoretic approach depicted above, this knowledge will capture the probability of (co-)occurrence of the entities of interest, which are the base to assess the informativeness of terms in which the disclosure assessment relies. In approaches based on semantic similarity measures, structured knowledge bases (e.g., ontologies) formally modeling the semantic relationships that are exploited during the similarity assessment should be employed instead.

Knowledge models can be automatically and dynamically derived from available sources by means of unsupervised information extraction or knowledge acquisition techniques to formalize the knowledge gathered by the attackers. By relying on knowledge models automatically acquired from updated sources, the following advantages arise: *i)* models are general and the knowledge they model hold for any data set framed in the domain that the model covers; *ii)* they can be continuously and automatically updated independently to the privacy protection process; and *iii)* because models are built from current sources, they contain an up-to-date representation of the social knowledge, thus offering a realistic tool to evaluate plausible inferences.

Together with the characterization and quantification of the semantics of disclosure, knowledge models provide the base to detect the actual disclosure risks in a specific scenario, and to seamlessly adapt and update the protection process as new knowledge/data are made available. As source for learning models suitable for data protection, several authors have used the Web (Chow et al., 2008, Sánchez and Batet, 2016), because it provides a reasonable and up-to-date approximation of the knowledge that an attacker may exploit to enable disclosure.

From a technical standpoint, the implementation of the above lines should rely on mature semantic tools and methods developed under the umbrella of the Semantic Web. On the one hand, because the data to be protected and the data/knowledge sources to be considered can be unstructured and textual, natural language processing -NLP- tools (e.g., tokenization, part-of-speech tagging, linguistic parsing, stemming, semantic disambiguation, etc.) are needed to process text, whereas knowledge acquisition methods (e.g., information extraction, semantic annotation, ontology learning, etc.) are required to build knowledge models. On the other hand, *ontologies* are especially suitable to build knowledge models, because they formalize the semantics of concepts and their semantic relationships; modern ontological languages like OWL are highly expressive and allow modeling a large spectrum of semantic relationships and features.

Semantic data protection

To design semantic data protection mechanisms coherent with the notion of SeDC, the following challenges should be tackled.

First, most data protection methods available in the literature (see Table 1) are meant for numerical data and, thus, rely on standard arithmetic operators (e.g., distance, variance, average, etc.) to analyze and protect data. However, when dealing with textual data (e.g., nominal attributes, plain text documents) and/or when data semantics should be considered during the protection process, which is the case of

SeDC, arithmetical operators are not applicable and/or do not capture the semantic features underlying to the data, as it is illustrated in the layer "Tools" in Figure 1.

The challenge here consists in the definition of a complete set of mathematically and semantically consistent operators that capture the semantic features of non-numerical data, and that enable to protect them according to both their meaning and data distributions. Specifically, semantically-grounded comparison operators (e.g., *similarity/distance*) can be defined by relying on the state of the art in ontology-based semantic similarity (Batet and Sánchez, 2014), which quantifies the resemblance between the meaning of concepts according to the structural/semantic features they share in an ontology that models them. Then, by relying on these semantically-coherent comparisons, it is possible to build the ranks and (partial) orders on non-ordinal data (Soria-Comas et al., 2014) that are employed in data protection methods to group or coarse/generalize similar data, among other tasks (Hundepool et al., 2013). Likewise, measures quantifying the centrality of concepts within a knowledge base (as proposed in (Domingo-Ferrer et al., 2013)) can be used to compute semantically-coherent *means*, *variances* or *covariances* of samples of non-numerical data. Specifically, as proposed in (Martínez et al., 2012c), the *semantic mean* of a sample of a nominal attribute can be defined as the concept from the attribute domain that minimizes the semantic distance to all the values in the sample; the *semantic variance* can be estimated from the average semantic distance of the values in the sample toward the *semantic mean*, that is, from their average semantic dispersion within the sample (Rodríguez-García et al., 2017); and the *semantic covariance* between attributes pairs (e.g., a disease and its treatment) can be measured by assessing whether the corresponding value pairs have similar degrees of semantic dispersion (Rodríguez-García et al., 2016).

With these semantically grounded operators, it is possible to characterize the analytical utility of the data from a semantic perspective and guide the data protection process towards preserving it. The mathematical consistency of the semantic operators (e.g., the fact that semantic distance measures satisfy the properties of a *metric*) is also crucial to manage heterogenous data sets (e.g., involving numerical and non-numerical attributes), so that the standard arithmetic operators and their semantic counterparts can be coherently integrated.

By using adaptations of arithmetical operators to the semantic domain, very recently, some data protection methods meant for numerical data have been applied to textual data so that data semantics are better preserved in the protected outcomes. Table 1 surveys most of the data protection mechanisms available in the literature (Hundepool et al., 2013) and identifies adaptations to the semantic domain that have been recently proposed.

Table 1. Data protection methods.

Method	*Data transformation*	*Comments*	*Adapted to the semantic domain*
Sampling	A non-exhaustive sample of a data set is published, so that inferences on specific individuals are no longer univocal.	Significantly reduces the size of data. Records left untouched in the sample present great	No adaptation is needed (data is left untouched).

		disclosure risk.	
Global recoding/generalization	Values are grouped and replaced by common generalizations.	Rely on hierarchical discretizations of attribute domains.	Usually ad-hoc taxonomies are considered for categorical data (Samarati, 2001, Terrovitis et al., 2008).
Top and bottom coding	Special case of global recording for variables that can be ranked.	Only applicable to ordinal attributes.	No.
Local suppression	Sparse values that may be univocally re-identified are suppressed.	High utility loss, especially for outlying values.	No adaptation is needed (no utility is preserved after suppression).
Noise-addition	Original values are distorted by adding a random noise magnitude taken from a specific density function. Different types: uncorrelated, correlated, based on linear or non-linear transformations.	Can be applied to individual records (e.g., when data is generated as a stream).	Adapted in (Rodríguez-García et al., 2017)
Microaggregation	Similar records are grouped together and made indistinguishable by replacing them by the group average. Can be univariate or multivariate and with fixed or variable group size.	Alternative to global recording that maintains the granularity of the data and is less affected by outlaying values	Adapted in (Domingo-Ferrer et al., 2013, Martínez et al., 2012a) to mono-valued databases and in (Batet et al., 2013) to set-valued data.
Rank swapping	Values are ranked and swapped within fixed intervals.	Good utility preservation due to original values kept untouched but decoupled from the original records.	No (it is difficult to build ranks for categorical attributes).
Data shuffling	Special type of swapping which preserves marginal distributions.	Only applicable to continuous or ordinal data.	No
Re-sampling	Creates several independent samples of the data set and averages the values.		Adapted in (Martínez et al., 2012b)

To guarantee a certain level of protection and, also, to balance the trade-off between protection and utility preservation, the former methods should be enforced in the context of a *privacy model*. Privacy models offer *ex ante* and user-settable privacy guarantees on the protected outcomes, regardless the specific data set to be protected. By instantiating a privacy model, users can also abstract from the internals of the data protection algorithm used to enforce the model. Again, privacy models available in the literature, such as *k*-anonymity or *ε*-differential privacy define privacy guarantees in an abstract and numerical way and, as a result, they are unintuitive and difficult to understand for practitioners and data controllers (Anandan et al., 2012). To tackle this issue, inherently semantic privacy models should be defined, in which privacy guarantees are expressed in terms of the actual semantics to be protected or preserved. To do so, such models should allow defining which semantics should be protected (and which preserved) by means of intuitive linguistic labels and qualitative criteria, which is coherent with the (also qualitative and semantic) privacy guidelines of current regulations on data protection (Department of Health and Human Services, 2000, The European Parliament and the Council of the EU, 2016). For example, given a patient medical record, the model could be instantiated to allow readers to know that the patient suffers from an uncertain cancer but guaranteeing the non-disclosure of the type of cancer the she suffers, which could be instantiated by using "cancer" as linguistic label. Because the model instantiation is expressed in terms of semantics (instead of the data distributions, such in *k*-anonymity), no assumptions on the structure and type of the data are needed, so that any kind of input (e.g., structured databases, free text documents, etc.) can be managed homogeneously. In practice, enforcing a semantic model would imply that the protected outcomes cannot disclose more semantics about the sensitive entities to be protected that those allowed by the model instantiation. This guarantee can be achieved by means of (semantic) data transformations such as those depicted in Table 1. Very few semantic privacy models have been proposed so far (see (Sánchez and Batet, 2016, Anandan et al., 2012, Chakaravarthy et al., 2008)); from these, the only model that (partially) fulfills the principles enounced above was presented in (Sánchez and Batet, 2016) and extended in (Sánchez and Batet, 2017), which focuses on the protection of unstructured textual documents.

Finally, to guide and evaluate the data protection process, new semantically-coherent utility measures shall be defined. On the one hand, standard and generic measures, such as the widely used sum of errors, can be adapted to quantify the semantic discordance between original and protected data by relying on semantic similarities (Domingo-Ferrer et al., 2013). On the other hand, inherently semantic metrics that evaluate the preservation of the *meaning* of the protected data can be defined by relying on the semantically-grounded operators discussed above (e.g., the preservation of the semantic mean/centroid states that the protected output maintains the scope of the discourse (Rodríguez-García et al., 2017)). The goal is to ascertain whether the conclusions extracted from the semantics provided by the protected output are similar to those extracted from the original data.

SeDC through data types and data structures

Under the umbrella of SeDC, data are analyzed and protected homogeneously, regardless of their type and structure; that is, semantically-grounded methods should seamlessly support structured (e.g., relational databases) or unstructured (e.g., free text documents) data sets containing heterogeneous data types. The underlying idea is that the semantics and, thus, the protection needs for a record in a relational database detailing, for example, the name, address and diseases of a subject, are equivalent to those of a natural language textual document containing the same information in an unstructured way. Regarding data types, it is important to note that, even though semantics is usually associated to textual sources, it is also relevant for meaningful numerical data (e.g., a magnitude quantifying the blood pressure can be related to hypertension, or a date as significant as 9/11 may be related to terrorism).

In comparison with statistical approaches, the semantic interpretation of data also fosters that a much wider spectrum of data types can be naturally supported by data protection algorithms (see the "Data types" layer in Figure 1). In this respect, SeDC has the potential to support structured databases with heterogeneous attributes, including numerical values -either pure (e.g., a cell phone number) or semantically-rich numerical attributes (e.g., blood pressure)-, categorical values (e.g., race or country of birth) and also free text (e.g., hobbies). Thanks to the mathematical coherency of the semantic operators used to semantically interpret and transform data, it is possible to integrate them with the standard arithmetic operators meant for strictly numerical data and, thus, deal with heterogeneous data sets. Moreover, set-valued data, which consist of lists of transactions associated to an individual (e.g., queries performed to a web search engine, items purchased in an online store, etc.), are also supported (Batet et al., 2013). These data sets are characterized by the variable cardinality of the transaction elements and by the fact that any combination of any cardinality of elements may enable disclosure. In this case, the semantically-oriented assessment of disclosure risks of SeDC helps to automatically detect such risky combinations. Tagged multimedia data, which may include tagged text (e.g., microblogging messages), images or videos (e.g., tagged photos or videos in a social network) are also supported because tags define the semantics inherent to the associated resource. Thus, by semantically analyzing tags, one can evaluate the disclosure risk of the resource and protect it accordingly. Finally, SeDC can also support free textual data, such as e-mails, clinical outcomes written in natural language, reports, etc. These are the most challenging resources, because NLP tools are needed to identify the pieces of information that encompass the semantics of the discourse (typically noun phrases and named entities), and to evaluate their disclosure risk and protect them accordingly.

Also, because the data protection envisioned by SeDC is driven by the semantics of the data rather than by their distribution in the data set, individual records, transactions and documents can be protected independently. In comparison with statistical approaches that protect static data sets monolithically and homogenously, SeDC offers more flexibility (because the protection of data pieces can be tailored to per individual privacy needs) and provides a natural solution to protect dynamic data streams (e.g., queries, e-mails or messages posted in a social network).

Application scenarios

SeDC-based solutions can cope with the privacy needs of heterogeneous data release scenarios. In the following, some of these applications are detailed, by highlighting their disclosure issues, data types and privacy requirements and depicting how SeDC can cope with them.

Let consider a hospital must send a protected patient record containing semi-structured heterogeneous medical data (e.g., diagnoses, admission details, visit outcomes and tagged multimedia resources) to an insurance company in response to a worker's compensation claim. According to current legislations on health data privacy (Department of Health and Human Services, 2000), *confidentiality* regarding sensitive diseases (e.g., AIDS, STDs, mental disorders, etc.) must be guaranteed in order to avoid potential discrimination. In this case, the diseases considered sensitive should be used to instantiate the SeDC-based privacy model. In this way, an information-theoretic algorithm enforcing the model will automatically detect the terms (e.g., specific diagnoses, but also combinations of symptoms, treatments, drugs, etc.) that may disclose the sensitive diseases due to the large mutual information they share. These "risky" terms will be then subjected to sanitization, for example, by replacing them by less detailed generalizations, i.e., terms disclosing less information on the sensitive diseases.

A social media platform user wishes to protect herself from disclosing potentially *discriminatory* data when she iteratively publishes textual messages (e.g., on Twitter or Facebook). Again, the SeDC-based privacy model can be instantiated according to the topics identified as sensitive in the EU General Data Protection Regulation (The European Parliament and the Council of the EU, 2016), such as religion, sexuality or race. Then, as detailed in the previous paragraph, the algorithm enforcing the model would be able, for each individual message, to automatically process its free text content, detect terms that may disclose to the sensitive topics, warn the user about the potential risk, and propose sanitization measures (e.g., removal or generalization of terms) before publishing the message.

A statistical office wants to release census information (i.e., structured microdata with both numerical and nominal attributes) of a set of individual respondents while ensuring their *anonymity* in front of the inferences that the information available in external resources may enable (The European Parliament and the Council of the EU, 2016). With the envisioned solutions, the plausibility of disclosive semantic inferences between attribute value combinations in the data set can be automatically assessed, thus avoiding the need of a human expert specifying them. Afterwards, attribute value combinations will be individually protected (e.g., via generalization) so that univocal inferences are no longer possible; moreover, the protection could be tailored at record level, instead of enforcing a homogenous protection for all the data set.

Web search engines and online stores collect the queries and the items purchased by their users, respectively. These data (i.e., transactional or *set-valued* data) may be monetized by selling them to Data Brokers (Pàmies-Estrems et al., 2016). To implement the privacy-by-design recommendations made by the U.S. Federal Trade Commission (U.S. Federal Trade Commission, 2014) either the data owners or the Data Brokers should implement the disclosure risk assessment mechanisms discussed above consistently with current legal frameworks. As a result, the data owners may detect and decide not to sell pieces of data that may re-identify the individual to whom the data refer to, whereas the Data Brokers may avoid incurring in potentially discriminatory inferences (e.g., the interest of a costumer in sugar free products may be related to diabetes).

Conclusions

This paper has discussed the benefits that an intrinsically semantic management of data brings to data protection, in comparison with pure numerical approaches. This new paradigm, which is named SeDC, has the potential to improve the generality, flexibility, intuitiveness, accuracy and data utility preservation of protection algorithms and privacy models and, also, to support a much broader and heterogeneous set of data types and application scenarios.

This paper has characterized and discussed the main challenges SeDC entails, i.e., studying and quantifying of the semantics underlying disclosure via semantic inferences, modeling the knowledge available to attackers, defining semantically-grounded and mathematically consistent data management and transformation operators, adapting numerical data protection methods to the semantic domain, designing semantically-grounded privacy models providing *ex ante* privacy guarantees and defining intrinsically semantic data utility measures.

Works on data protection that, so far, have considered data semantics up to some degree are scarce, scattered and unconnected. SeDC aims at providing a common ground for such works. Specifically, in this paper, such incipient works have been characterized and discussed in the context of the identified challenges. As a result, shortcoming and needs of the state of the art have been identified, thus paving the ground for future research on this topic.

From a social perspective, SeDC has the potential of making users aware and participant of the protection of their own data, which is something that has been largely neglected in statistical approaches. On the one hand, inherently semantic solutions are much intuitive to use and understand for the end users than statistical approaches. Moreover, they can naturally enforce the privacy guidelines contained in current legislations on personal data protection, which focus on the content of the data, rather than on their distributional features. On the other hand, disclosure risks can be automatically assessed on per-individual basis by studying the semantics underlying to the data, thus avoiding the need for human intervention and/or homogenous data protection. As a result, individually tailored data protection may be enforced in a variety of scenarios and data types that can be hardly coped by statistical approaches.

Acknowledgements and disclaimer

This work was partly supported by the European Commission (projects H2020-644024 “CLARUS” and H2020-700540 “CANVAS”) and by the Spanish Government (projects TIN2014-57364-C2-R “SmartGlacis”, TIN2015-70054-REDC “Red de excelencia Consolider ARES” and TIN2016-80250-R “Sec-MCloud”). The opinions expressed in this paper are those of the authors and do not necessarily reflect the views of UNESCO. M. Batet is supported by a Postdoctoral grant from Ministry of Economy and Competitiveness (MINECO) (FPDI-2013-16589).

References

Anandan, B. and Clifton, C. (2011), "Significance of term relationships on anonymization", in *IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology - Workshops*, Lyon, France, pp. 253–256.

Anandan, B., Clifton, C., Jiang, W., Murugesan, M., Pastrana-Camacho, P. and L.Si (2012), "t-plausibility: Generalizing words to desensitize text", *Transactions on Data Privacy,* Vol. 5, pp. 505-534.

Batet, M., Erola, A., Sánchez, D. and Castellà-Roca, J. (2013), "Utility preserving query log anonymization via semantic microaggregation", *Information Sciences,* Vol. 242, pp. 49-63.

Batet, M. and Sánchez, D. (2014), "Review on Semantic Similarity", *Encyclopedia of Information Science and Technology (3rd edition)*, IGI Global, pp. 7575-7583.

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 Magazine,* Vol. 7 No. 6, pp. 46-53.

Chakaravarthy, V. T., Gupta, H., Roy, P. and Mohania, M. K. (2008), "Efficient techniques for document sanitization", in *17th ACM Conference on Information and Knowledge Management (CIKM'08)*, Napa Valley, California, USA, pp. 843–852.

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 Rufian-Torrell, G. (2013), "Anonymization of nominal data based on semantic marginality", *Information Sciences,* Vol. 242, pp. 35-48.

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

Domingo-Ferrer, J. and Torra, V. (2005), "Ordinal, continuous and heterogeneous k-anonymity through microaggregation", *Data Mining and Knowledge Discovery,* Vol. 11 No. 2, pp. 195-212.

Dwork, C. (2006), "Differential Privacy", in *33rd International Colloquium ICALP*, Venice, Italy, pp. 1-12.

Fung, B. C. M., Wang, K., Chen, R. and Yu, P. S. (2010), "Privacy-preserving data publishing: A survey of recent developments", *ACM Computer Surverys,* Vol. 42 No. 4, p. 14.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E. S., Spicer, K. and Wolf, P. P. d. (2013), *Statistical Disclosure Control,* Wiley.

Li, N. and Li, T. (2007), "t-Closeness: Privacy Beyond k-Anonymity and l-Diversity", in *IEEE 23rd. International Conference on Data Engineering*, Istanbul, pp. 106-115.

Martínez, S., Sánchez, D. and Valls, A. (2012a), "Semantic adaptive microaggregation of categorical microdata", *Computers & Security,* Vol. 31 No. 5, pp. 653-672.

Martínez, S., Sánchez, D. and Valls, A. (2012b), "Towards k-Anonymous Non-numerical Data via Semantic Resampling", in *Information Processing and Management of Uncertainty (IPMU)*, Montpellier (France), pp. 519-528.

Martínez, S., Sánchez, D. and Valls, A. (2013), "A semantic framework to protect the privacy of electronic health records with non-numerical attributes", *Journal of Biomedical Informatics,* Vol. 46 No. 2, pp. 294-303.

Martínez, S., Valls, A. and Sánchez, D. (2012c), "Semantically-grounded construction of centroids for datasets with textual attributes", *Knowledge-Based Systems,* Vol. 35, pp. 160-172.

Pàmies-Estrems, D., Castellà-Roca, J. and Viejo, A. (2016), "Working at the Web Search Engine Side to Generate Privacy-Preserving User Profiles", *Expert Systems with Applications,* Vol. 64, pp. 523-535.

Rodríguez-García, M., Batet, M. and Sánchez, D. (2017), "A semantic framework for noise addition with nominal data", *Knowledge-Based Systems,* Vol. 122, pp. 103-118.

Rodríguez-García, M., Sánchez, D. and Batet, M. (2016), "Perturbative Data Protection of Multivariate Nominal Datasets", in *Privacy in Statistical Databases*, Dubrovnik (Croatia), Vol. 9867, pp. 94-106.

Rotella, P. (Apr 2, 2012), "Is Data The New Oil?", *Forbes*.

Samarati, P. (2001), "Protecting Respondents' Identities in Microdata Release", *IEEE Transactions on Knowledge and Data Engineering,* Vol. 13 No. 8, pp. 1010-1027.

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. and Batet, M. (2017), "Toward sensitive document release with privacy guarantees", *Engineering Applications of Artificial Intelligence,* Vol. 59, pp. 23–34.

Sánchez, D., Batet, M. and Viejo, A. (2013), "Automatic General-Purpose Sanitization of Textual Documents", *IEEE Trans. Information Forensics and Security,* Vol. 8 No. 6, pp. 853-862.

Sánchez, D. and Viejo, A. (2017), "Personalized privacy in open data sharing scenarios", *Online Information Review,* Vol. 41 No. 3, pp. 298-310.

Soria-Comas, J., Domingo-Ferrer, J., Sánchez, D. and Martínez, S. (2014), "Enhancing Data Utility in Differential Privacy via Microaggregation-based k-Anonymity", *VLDB Journal,* Vol. 23 No. 5, pp. 771-794.

Terrovitis, M., Mamoulis, N. and Kalnis, P. (2008), "Privacy-preserving anonymization of set-valued data", in *VLDB Endowment*, Vol. 1, pp. 115-125.

The European Parliament and the Council of the EU (2016), "General Data Protection Regulation (GDPR) (Regulation (EU) 2016/679)".

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