Engineering Applications of Artificial Intelligence 59 (2017) 23–34

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

Toward sensitive document release with privacy guarantees

David Sánchez[a], Montserrat Batet[b,*]

[a] UNESCO Chair in Data Privacy, Department of Computer Science and Mathematics, Universitat Rovira i Virgili, Avda. Països Catalans, 26, 43007 Tarragona, Spain

[b] Internet Interdisciplinary Institute (IN3), Universitat Oberta de Catalunya, Parc Mediterrani de la Tecnologia, Av. Carl Friedrich Gauss, 5, 08860 Castelldefels, Spain

ARTICLE INFO

Keywords:
Document redaction
Sanitization
Semantics
Ontologies
Privacy

ABSTRACT

Privacy has become a serious concern for modern Information Societies. The sensitive nature of much of the data that are daily exchanged or released to untrusted parties requires that responsible organizations undertake appropriate privacy protection measures. Nowadays, much of these data are texts (e.g., emails, messages posted in social media, healthcare outcomes, etc.) that, because of their unstructured and semantic nature, constitute a challenge for automatic data protection methods. In fact, textual documents are usually protected manually, in a process known as document *redaction* or *sanitization*. To do so, human experts *identify* sensitive terms (i.e., terms that may reveal identities and/or confidential information) and *protect* them accordingly (e.g., via removal or, preferably, generalization). To relieve experts from this burdensome task, in a previous work we introduced the theoretical basis of *C-sanitization*, an inherently semantic privacy model that provides the basis to the development of automatic document redaction/sanitization algorithms and offers clear and a priori privacy guarantees on data protection; even though its potential benefits *C-sanitization* still presents some limitations when applied to practice (mainly regarding flexibility, efficiency and accuracy). In this paper, we propose a new more flexible model, named (*C, g*(*C*))-*sanitization*, which enables an intuitive configuration of the trade-off between the desired level of protection (i.e., controlled information disclosure) and the preservation of the utility of the protected data (i.e., amount of semantics to be preserved). Moreover, we also present a set of technical solutions and algorithms that provide an efficient and scalable implementation of the model and improve its practical accuracy, as we also illustrate through empirical experiments.

1. Introduction

Information Technologies have paved the way for global scale data sharing. Nowadays, companies, governments and subjects exchange and release large amounts of electronic data on daily basis. However, in many occasions, these data refer to personal features of individuals (e.g., identities, preferences, opinions, salaries, diagnoses, etc.), thus causing a serious privacy threat. To prevent this threat, appropriate data protection measures should be undertaken by responsible parties in order to fulfill with current legislations on data privacy (Department of Health and Human Services, 2000; The European Parliament and the Council of the European Union, 2016).

Because of the enormous amount of data to be managed and the burden and cost of manual data protection (Bier et al., 2009), many automated methods have been proposed in recent years under the umbrella of *Statistical Disclosure Control* (SDC) (Hundepool et al., 2013). These methods aim at masking input data in a way that either *identity* or *confidential attribute* disclosure are minimized. The former deals with the protection of information that can re-identify an individual (e.g., a social security number or unique combinations of several attributes, such as the age, job and address), and it is usually referred to as *anonymization*, whereas the latter deals with the protection of *confidential* data (e.g., salaries or diagnosis). To do so, protection methods remove, distort or coarse input data while balancing the trade-off between privacy and data utility: the more exhaustive the data protection is, the higher the privacy but the less useful the protected data becomes as a result of the applied distortion, and vice-versa. In addition to data protection methods, the computer science community has proposed formal *privacy models* (Drechsler, 2011), within the area of *Privacy-Preserving Data Publishing* (PPDP) (Fung et al., 2010) and *Data Mining* (PPDM) (Lin et al., 2016a, 2016b). In comparison to the ad-hoc masking of SDC methods, in which the level of protection is empirically evaluated *a posteriori* for a specific dataset (Drechsler, 2011), privacy models attain a *predefined* notion of privacy and offer *a priori* privacy guarantees over the protected data (e.g., a probability of re-identification (Samarati, 2001; Samarati and Sweeney, 1998)). This provides a clearer picture on the level of protection that is

* Corresponding author.
E-mail addresses: david.sanchez@urv.cat (D. Sánchez), mbatetsa@uoc.edu (M. Batet).

http://dx.doi.org/10.1016/j.engappai.2016.12.013
Received 10 August 2016; Received in revised form 8 November 2016; Accepted 13 December 2016

applied to the data, regardless the features or distribution of a specific dataset. Moreover, privacy models provide a *de facto* standard to develop privacy-preserving tools, which can be objectively compared by fixing the desired privacy level in advance.

So far, most privacy models and protection mechanisms have focused on structured statistical databases (Domingo-Ferrer et al., 2016), which present a regular structure (i.e., records refer to individuals that are described by a set of usually uni-valued attributes) and mostly contain numerical data. Privacy models such the well-known *k*-anonymity notion relied on such regularities to define privacy guarantees: a data base is said to be *k*-anonymous if any record is indistinguishable with regard to the attributes that may identify an individual from, at least, *k-1* other records (Samarati, 2001; Samarati and Sweeney, 1998).

However, many of the (sensitive) data that is exchanged in current data sharing scenarios is textual and unstructured (e.g., messages posted in social media, e-mails, medical reports, etc.). In comparison with structured databases, plain textual data protection entails additional challenges:

- Due to their lack of structure, we cannot pre-classify input data according to identifying and/or confidential attributes, as most data protection mechanisms do (Domingo-Ferrer et al., 2016); in fact, for plain text, any combination of textual terms of any cardinality may produce disclosure.
- In comparison with the usually numerical attributes found in structured databases, plain textual data cannot be compared and transformed by means of standard arithmetical operators. In fact, since textual documents are interpreted by data producers and consumers (and also potential attackers) according to the meaning of their contents, linguistic tools and semantic analyses are needed to properly protect them (Torra, 2011).

Because of the above challenges, the protection of plain textual documents has not received enough attention in the current literature (Anandan et al., 2012; Chow et al., 2008; Sánchez and Batet, 2016). As we discuss in the next section, most of the current methods and privacy models for textual data protection are naïve, unintuitive, require from a significant intervention of human experts and/or limit the protection to predefined types of textual entities.

1.1. Background on plain textual data protection

Traditionally, plain textual data protection has been performed manually, in a process by which several experts detect and mask terms that may disclose identities and/or confidential information, either directly (e.g., names, SS numbers, sensitive diseases, etc.) or by means of *semantic inferences* (e.g., treatments or drugs that may reveal sensitive diseases, readings that may suggest political preferences or habits that can be related to religion or sexual orientations) (Gordon, 2013). In this context, data semantics are crucial because they define the way by which humans (sanitizers, data analysts and also potential attackers) understand and manage textual data.

In general, plain textual data protection consists of two main tasks: i) identify textual terms that may disclose sensitive information according to a privacy criterion (e.g., names, addresses, authorship, personal features, etc.); and ii) mask these terms to minimize disclosure by means of an appropriate protection mechanism (e.g., removal, generalization, etc.). The community refers to the act of removing or blacking-out sensitive terms as *redaction*, whereas *sanitization* usually consists in coarsening them via generalization (e.g., *AIDS* can be replaced by a less detailed generalization such as *disease*) (Bier et al., 2009). The latter approach, which we use in this paper, better preserves the utility of the output.

To relieve human experts from the burden of manual sanitization, the research community has proposed mechanisms to tackle specific data protection needs. On the one hand, we can find works that aim at inferring sensitive information, such as the authorship of a resource (e.g., documents, emails, source code, etc.) (Koppel et al., 2013) or the profile of the author (e.g., gender) (Rangel et al., 2016); on the other hand, other works aim at preventing disclosure by masking the data that may disclose that authorship (Adimoolam et al., 2009; Almishari et al., 2014). In the healthcare context, we can find ad-hoc data protection approaches that focus on detecting *protected health information* (PHI, such as ages, e-mails, locations, dates or social security numbers) (Meystre et al., 2010), which are data that, according to the HIPAA "Safe Harbor" rules, must be eliminated before releasing electronic healthcare records to third parties. Most of these application-specific approaches exploit the regularities of the lexico-syntactic regularities of the entities to be detected (e.g., use of capitalizations for proper names, structure of dates or e-mails, etc.) to define patterns or employ machine learning techniques such as trained classifiers. However, the applicability of these methods is limited to the use case they consider, and they do not offer robust guarantees against disclosure outside the entities in which they focus.

General-purpose privacy solutions for plain text are scarce and they only focus on the protection of sensitive terms, which are assumed to be manually identified beforehand. We can find two privacy models that reformulate the notion of *k*-anonymity for documents rather than data bases: *K*-safety (Chakaravarthy et al., 2008) and *K*-confusability (Cumby and Ghani, 2011). Both approaches assume the availability of a large and homogenous collection of documents, and require each sensitive entity mentioned in each document of the collection to be indistinguishable from, at least, K-1 other entities in the collection. To do so, terms are generalized (so that they become less diverse and, hence, indistinguishable) in groups of *K* documents. However, documents cannot be sanitized individually and, due to the need to generalize terms to a common abstraction, data semantics will be hampered if the contents of the collection are not perfectly homogenous.

In (Anandan et al., 2012), a privacy model named *t*-plausibility that also relies on the generalization of manually identified sensitive terms was presented. A document is said to fulfill *t*-plausibility if, at least, *t* different *plausible* documents can be derived from the protected document by specializing sanitized entities; that is, the protected document generalizes, at least, *t* documents obtained by combining specializations of the sanitized terms. Even though this approach allows sanitizing documents individually, it is noted that setting the *t*-plausibility level is not intuitive and that one can hardly predict the results of a given *t*, because they would depend on the document size, the number of sensitive entities and the number of available generalizations and specializations.

To tackle the limitations of the above-described solutions, in (Sánchez and Batet, 2016) we presented an inherently semantic privacy model for textual data: *C-sanitization*. Its goal is to mimic and, hence, automatize the analysis of semantic inferences that human experts perform for document sanitization. Informally, the disclosure risk caused by semantic inferences is assessed by answering to this question: does a term or a combination of terms in a document to be released allow to univocally inferring and, thus, disclosing a sensitive entity defined in *C*? According to such vision, the privacy guarantees offered by the model state that a *C-sanitized* document should not contain any term that, individually or in aggregate, univocally reveals the semantics of the sensitive entities stated in *C*. In accordance with current privacy legislations, *C* may contain the entities that legal frameworks define as sensitive, such as religious and political topics or certain diseases (Terry and Francis, 2007). For example, an *AIDS-sanitized* medical record should not contain terms that enable a univocal inference of AIDS, such as HIV or closely related symptoms or treatments.

In (Sánchez and Batet, 2016), *C-sanitization* is formalized according to the following elements: (1) D: the document to be protected. (2) *C*: the set of sensitive entities that should be protected from univocal disclosure in *D* (e.g., *C* could be a set of sensitive diseases or religious or political topics and *D* a medical record or a message to be posted in a social network). (3) *T*: whatever group of terms of any cardinality occurring in *D* that could be used by an attacker to unambiguously infer any of the sensitive entities in *C* (e.g., if *C* is a sensitive disease, *T*