Published version:
doi.org/10.1016/j.ins.2013.06.042

Minimizing the disclosure risk of semantic correlations in document sanitization

David Sánchez [1], Montserrat Batet, Alexandre Viejo

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

Abstract

Text sanitization is crucial to enable privacy-preserving declassification of confidential documents. Moreover, considering the advent of new information sharing technologies that enable the daily publication of thousands of textual documents, automatic and semi-automatic sanitization methods are needed. Even though several of these methods have been proposed, most of them detect and sanitize sensitive terms (e.g., people names, addresses, diseases, etc.) independently, neglecting the importance of semantic correlations. From the attacker's perspective, semantic correlations can be exploited to disclose a sanitized term from the presence of one or several non-sanitized words. To tackle this problem, this paper presents a general-purpose method that, by taking the output of a standard sanitization mechanism, analyses, detects and proposes for sanitization those semantically correlated terms that represent a plausible disclosure risk for the already sanitized ones. Our method relies on an information-theoretic formulation of disclosure risk which is able to adapt its behavior to the criterion of the initial sanitizer. The evaluation, carried on over a collection of real documents, shows that semantic correlations represent a real privacy threat in prior sanitized documents, and that our method is able to detect them effectively. As a result, the disclosure risk of the sanitized output is significantly reduced with respect to standard sanitization mechanisms.

Keywords: privacy, document sanitization, semantic correlation, information theory.

[1] Corresponding author. Address: Departament d'Enginyeria Informàtica i Matemàtiques. Universitat Rovira i Virgili. Avda. Països Catalans, 26. 43007. Tarragona. Spain
Tel.: +034 977 559657; Fax: +034 977 559710;
E-mail: david.sanchez@urv.cat.

1. Introduction

Declassification and publication of documents are of great interest in information-intensive environments such as medical or business decision-making and research. For example, declassification of medical records has played a significant role in improving the detection, verification and monitoring of new diagnostic examinations and treatment methodologies [16]. Regarding the analysis of commercial documents, it has been applied to support different business needs related to data sharing and integration [6].

Even though document declassification provides important advantages, it also brings some relevant concerns that require attention. In many cases, documents contain sensitive information related to individuals (e.g., identifiable data, personal information like diseases or economic status, etc.) or companies (e.g., sale operations, commercial partners, etc.). Hence, in the context of the Information Society in which thousands of documents are made public daily and digital information can be copied and re-distributed easily, untrusted parties might have access to large quantities of sensitive data [8, 20].

To tackle this problem, *document sanitization* pursuits to identify and remove sensitive pieces of information from documents before being declassified, so that the risk of re-identification of individuals and/or of revealing confidential information can be minimized. At the same time, non-sensitive terms are left in clear for so that the analytical utility of the sanitized output can be retained up to a degree. In the most general case, on which this paper focuses, input documents consist of raw and unstructured texts. Within this area, sanitization has traditionally been done manually, applying certain rules or guidelines [21], detailing the correct procedures to ensure irreversible suppression or distortion of sensitive parts in physical and electronic documents. Nevertheless, manual mechanisms are time-consuming [12], prone to disclosure risks [6] and they do not scale as the volume of data increases [7]. To minimize these problems, researchers have proposed automatic or semi-automatic sanitization methods. Some schemes like [7] use databases of entities (e.g., persons, products, diseases, etc.) to discover sensitive terms. Other proposals [39] find patterns of identifying information such as name, location, etc. There are also mechanisms [1] that use trained classifiers to discover apparently sensitive terms. Finally, other methods [31, 32] rely on the Information Theory to assess the degree of sensitiveness of terms according to the *amount of information* they provide.

All the above approaches analyze and detect sensitive terms independently. That is, the sensitivity of a set of textual terms is evaluated considering them independent variables. For example, in an approach like [7], the term *Acquired Immunodeficiency Syndrome* may be contained in a list of sensitive diseases that should be sanitized prior to publication, whereas terms like *influenza* and *blood transfusion*, which are less critical and more generic terms, may not. However, all these terms are semantically correlated when appearing in the same context. The fact that a non-sanitized term or a combination of them is highly correlated with a sanitized one may enable the re-identification of the latter by semantic inference. Since most textual terms appearing in a discourse (e.g., a sentence, a paragraph or a document) are semantically correlated [2], not only the sensitivity of individual terms, but also of term combinations, should be considered during the sanitization process in order to minimize disclosure risk.

The prevention of disclosure derived from the combination of several *individually* non-sensitive terms has been extensively tackled in the Statistical Disclosure Control (SDC) area [11, 17-19]. For structured databases, a distinction is made *a priori* between identifying, quasi-identifying and confidential attributes. *Identifiers* (e.g., social security numbers) are directly removed from the dataset, whereas *quasi-identifiers* (i.e., those attributes whose value combinations may unequivocally identify an individual, such as *job+birth place+age*) are masked according to a privacy model. *k*-anonymity [38] is one of the most popular ones, stating that each anonymized record should be indistinguishable from, at least, *k-1* other ones with respect to their quasi-identifiers. More general approaches deal with set-valued data, instead of relational databases. They protect transactional data (e.g., query logs), assuming that any item in the set is a potential quasi-identifier. A specific privacy model, k^m-anonymity [40], aims at masking set-valued data so that any adversary who knows, at most, *m* items of an individual, could not distinguish him from, at least, *k-1* other individuals.

In the above scenarios, quasi-identifiers are defined a priori, either by selecting a specific subset of attributes in the case of relational databases, or by considering all items of transactional data. The sanitization of unstructured text defines a more challenging scenario, since no structure can be assumed for input documents. In the most general case, given the context of a discourse containing at least one sensitive term, any combination of non-sanitized words with any cardinality may potentially disclose the former if terms are semantically correlated. A straightforward solution will systematically sanitize terms co-occurring with the sensitive one, destroying the utility of the sanitized output. A more appropriate solution should look for only those combinations of terms for which a strong semantic correlation exists with regard to the sensitive one. Hence, the sanitization can be focused solely on those terms that represent a feasible disclosure risk.

This paper proposes a general solution to minimize the disclosure risk derived from the presence of semantically correlated terms. Our method starts form the output of a standard sanitization mechanism in which textual terms have been analyzed independently, such as [1, 7, 31, 32, 39]. Then, relying on the foundations of the Information Theory to mathematically formulate the correlation between sensitive and non-sensitive term tuples, it quantifies the risk of re-identifying a sanitized term from one or several non-sanitized ones. As a result of this assessment, those terms that represent a high disclosure risk are automatically proposed for sanitization. In addition, the paper also describes the technical details of a fully automatic algorithm that enables the application of the theoretical framework in a practical scenario. Its implementation has been tested with the sanitized output provided by a widely used sanitization mechanism based on *Named-Entity recognition* [15]. Evaluation results show that, on the one hand, the sanitization output of a method which analyzes textual terms independently can be easily negated by exploiting term correlations and that, on the other hand, our method is able to accurately detect those correlated non-sanitized terms that represent a disclosure risk.

The rest of the paper is organized as follows. Section 2 describes and discusses related works in document sanitization, focusing on semi-automatic and automatics methods. Section 3 presents and formalizes the theoretical framework used to detect semantically correlated terms. Section 4 describes the algorithm and the technical details that enable a practical application of the theoretical framework. Section 5 presents the evaluation results obtained from a collection of real documents. The final section contains the conclusions and proposes several lines of future research.

2. Related work

As introduced above, manual sanitization has been usually applied to textual documents. This approach requires a human expert who applies certain standard guidelines that detail the correct procedures to sanitize sensitive entities [21]. As a result, manual sanitization is time-consuming and, hence, it is not scalable in the current context where there is a large volume of documents to be declassified daily. In order to speed up this task, the industry has provided certain solutions designed to highlight some of the sensitive elements and, in this way, ease the work of the human expert who has the final decision about erasing or keeping them. As an example of this, Adobe Acrobat Professional provides a semi-automatic sanitization process [22], which recognizes certain sensitive entities (e.g., email addresses, dates, etc.) using some standard patterns. Nevertheless, semi-automatic mechanisms still require the interaction of human

experts and, hence, they practically suffer from the same problems as manual ones. To address this issue, both academic and industrial researchers have proposed automatic sanitization mechanisms.

The first automatic proposals were based on the application of specific patterns to identify sensitive elements. Following this approach, the Scrub system [39] can detect general information, such as names or locations. Schemes like [42] and [13] use more specific patterns to remove sensitive terms from medical records. These patterns are designed according to the HIPAA "Safe Harbor" rules that specifies 18 data elements (called PHI: Protected Health Information) which must be eliminated from clinical data in order to anonymize a clinical text [10]. Nevertheless, due to the use of very specific patterns, these schemes can be hardly generalized to other document types and application domains.

As an alternative to manually specified patterns, several authors have relied on trained classifiers that recognize sensitive entities. Authors in [6] present a tool that focuses on the sanitization of documents directly linked to certain companies. The data to be detected include words and phrases that reveal the company the document belongs. A *Naive Bayes classifier* trained with tagged examples is used to recognize them, so that solely those topics that have been previously tagged are detected. More general systems [1, 6], based on trained classifiers, assume that Named Entities (NE) are sensitive by definition. A trained Named Entity Recognition (NER) package, such as the Stanford NER [15], is used to automatically recognize entities belonging to general categories such as person, organization and location names. Other general methods [31, 32] assume that sensitive terms are those that, due to their specificity, provide too much information. Hence, by quantifying the information content of textual terms, sensitive terms are detected and proposed for sanitization.

A common limitation of the above-described methods is that the sensitivity of textual terms is evaluated independently. As discussed in the introduction, this may produce disclosure if terms that are semantically correlated with a sanitized one are left in clear form.

Works considering term correlations in the sanitization context are scarce. In [6], authors propose a sanitization method that relies on a database of sensitive entities (e.g., persons and product names, diseases, etc.). Each entity is associated with a context which contains a set of terms related to that entity (e.g., the context of a disease could include symptoms, treatments, risk factors, etc.). Through the use of this information, the proposal detects terms to sanitize in a straightforward manner, by looking for sensitive entities and their contexts in the input database. In essence, this method is very similar to a manual approach, since all entities and their contexts

must be specified by hand a priori. Hence, it could be only applied to documents with very constrained scopes, since it suffers from the same scalability and generality problems of manual approaches.

Finally, authors in [2] demonstrate that sanitized terms could be re-identified given the presence of non-sanitized words in the same context. To do so, they rely on a contingency table that reflects the degree of co-occurrence of all possible term pairs. In the most extreme case, in which a pair of terms always co-occur, the sanitization can be completely negated. Moreover, using a background taxonomy that is coherent to the contingency table, in the sense that the co-occurrences monotonically increase as terms are generalized in the taxonomy, they demonstrate that it is also possible to negate sanitizations based on the presence of generalized terms in the same context. Even though the availability of such comprehensive and detailed contingency table and the associated taxonomy is unrealistic in a practical and general setting, the work is valuable since it demonstrates, from a theoretical perspective, that document sanitization without considering term correlations might be ineffective.

3. Measuring the disclosure risk of semantic correlations

In order to minimize the disclosure risk of sanitized terms caused by the presence of other non-sanitized but semantically correlated ones, two tasks should be performed. First, the degree of correlation between sanitized and non-sanitized term should be measured. A high correlation will state a feasible disclosure. Then, those non-sanitized terms for which the degree of correlation is high enough should be sanitized additionally (e.g., removed). In this manner, the disclosure risk of the sanitized output is decreased while its utility is maintained as much as possible, since non-correlated or slightly-correlated terms are left in clear form.

Our approach is designed as a complementary step to any sanitization mechanism that detects and removes sensitive terms individually, such as most of those described in section 2. We formalize this scenario as follows:

- D: the original document to sanitize, which can be viewed as an ordered sequence of textual terms (i.e., words or phrases).
- ζ: a sanitization mechanism that detects sensitive words/phrases independently (e.g., NER-based methods [1, 6]). We assume that these terms will be removed prior to publication.
- $D'=\zeta(D)$: the output of the initial sanitization mechanism ζ, in which sensitive terms have been individually identified, but not removed yet.

- $C_i \subseteq D'$: each of the textual contexts contained in D', that is, the length of a semantically related discourse. A context may be a sentence, a paragraph or the whole document.
- $s_{ij} \in C_i$: each of the terms contained in C_i that have been detected as sensitive by the sanitization mechanism ζ. Hence, given a context C_i, we define the initial set of sensitive terms as $S_i=\{s_{i1}, s_{i2}, \ldots, s_{in}\}$.
- $q_{ik} \in C_i$: each of the textual terms in C_i that have not been detected as sensitive by ζ. Since these may cause disclosure of sanitized terms due to semantic correlation, we consider them as potential *quasi-identifiers*. As above, we define the set of non-sensitive terms in C_i as $Q_i=\{q_{i1}, q_{i2}, \ldots, q_{im}\}$.
- Ψ: the implementation of our approach which, taking D' as input (i.e., $\Psi(D')$), detects those terms in Q_i that represent a disclosure risk for sanitized terms in S_i.
- $sq_{ik} \in Q_i$: a term in Q_i that has been assessed by our approach Ψ as a *quasi-identifier* (i.e., a term that may cause disclosure). The set of quasi-identifiers detected for a certain C_i is defined as $QS_i=\{qs_{i1}, qs_{i2}, \ldots, qs_{il}\}$. The ordered set $UQS=\{QS_{i1}, QS_{i2}, \ldots, QS_{ix}\}$ for a given D' represents the output of Ψ. Again, we assume that all terms in UQS will be removed prior to publication.

In a nutshell, Ψ acts as follows. Given a context C_i with a set of sanitized (S_i) and non-sanitized (Q_i) terms, for each s_{ij} in S_i, Ψ evaluates the *disclosure risk* caused by those q_{ik} in Q_i. The output is a set of quasi-identifiers QS_i detected for each C_i. Note that, the longer the context C_i is with regards to the document D', the larger the sets C_i and Q_i are; this results in a larger number of assessments. Usually, textual contexts for correlation analyses are defined as sentences or paragraphs [25, 43], even though some works focus solely on adjacent words [2, 36], whereas others consider complete documents [41].

In the following sections, we present the theoretical framework in which our approach relies and which, based on the Information Theory, evaluates term correlations and measures their disclosure risk.

3.1. An information theoretic assessment of disclosure risk

In order to evaluate the *disclosure risk* that a potential quasi-identifier q_{ik} may cause with regard to a sanitized term s_{ij}, we rely on the following premises.

First, the information given by any textual term can be expressed in terms of Information Theory as its Information Content (IC). Given a sensitive term s_{ij}, $IC(s_{ij})$ can be computed as the inverse of its probability of appearance in a corpus (log_2 base states that the unit of IC is in bits):

$$IC(s_{ij}) = -\log_2 p(s_{ij}) \quad (1)$$

Hence, frequently appearing terms provide less information that scarcer ones.

By using this formulation, we can conclude that s_{ij} has been sanitized by a mechanism ζ because it reveals an amount of sensitive information, which can be quantified as $IC(s_{ij})$.

Then, we hypothesize that the *disclosure risk* derived from the presence of a potential quasi-identifier q_{ik} with regard to a sanitized term s_{ij} when both appear in the same context C_i, can be measured according to the *amount of information* that q_{ik}, which remains in clear form, reveals about s_{ij}. In terms of Information Theory, this corresponds to their *Mutual Information* (MI) that constitutes a measure of correlation between variables. The instantiation of MI for two specific outcomes (i.e., textual terms, in this case) results in the well-known *Point-wise Mutual Information* (PMI), which quantifies the difference between the probability of term co-occurrence given their join distribution and their marginal distributions [9]:

$$PMI(s_{ij}; q_{ik}) = \log_2 \frac{p(s_{ij}, q_{ik})}{p(s_{ij})\, p(q_{ik})} \quad (2)$$

PMI has been successfully applied in the past to evaluate different types of semantic correlations such as word collocation [3, 34], synonymy [41], taxonomic subsumption and similarity [29, 30, 43], and a variety of non-taxonomic relationships [25, 35, 37]. Notice that, in the context of document sanitization, a *high* correlation between term pairs may enable disclosure regardless of the type of correlation (i.e., taxonomic or non-taxonomic).

Numerically, when s_{ij} and q_{ik} are independent, that is, when they co-occur in a textual context by chance, then $PMI(s_{ij};q_{ik})=0$. In this case, we can conclude that the presence of q_{ik} does not provide any particular evidence of s_{ij} and, hence, disclosure risk is null. On the contrary, if s_{ij} and q_{ik} are perfectly associated and always co-occur, either for an occurrence of s_{ij} and/or an occurrence of q_{ik}, PMI has the following value:

$$PMI(s_{ij}; q_{ik}) = \begin{cases} -\log_2(p(s_{ij})) = IC(s_{ij}) & if\ p(s_{ij}, q_{ik}) = p(q_{ij}) \\ -\log_2(p(q_{ik})) = IC(q_{ik}) & if\ p(s_{ij}, q_{ik}) = p(s_{ik}) \end{cases} \quad (3)$$

Particularly, if $p(s_{ij},q_{ik})=p(q_{ik})$, that is, whenever q_{ik} occurs, s_{ij} also occurs, we can then conclude that the presence of q_{ik} in a document completely discloses s_{ij}. Thus, according to the above formula, disclosure risk of s_{ij} is maximum when $PMI(s_{ij};q_{ik})=IC(s_{ij})$.

Finally, PMI will result in negative values if s_{ij} and q_{ik} are exclusive, providing a minimal value (i.e., $-\infty$) when they never co-occur. It is important to note that words are rarely exclusive since most of them are semantically correlated up to a degree, even though only some of them are highly correlated [2]. In fact, low or even null co-occurrence between word pairs is usually attributed to data sparseness of probability calculus (i.e., the fact that not enough data is available to extract reliable conclusions from their analysis), rather than to a real exclusiveness [30].

In terms of Information Content, PMI also fulfills the following relationship:

$$PMI(s_{ij};q_{ik}) = \log_2 \frac{p(s_{ij},q_{ik})}{p(s_{ij})p(q_{ik})} = \log_2 \frac{p(s_{ij} \mid q_{ik})}{p(s_{ij})} = \log_2 \frac{\frac{p(s_{ij},q_{ik})}{p(q_{ik})}}{p(s_{ij})} = \quad , \qquad (4)$$
$$= \log_2 \frac{p(s_{ij},q_{ik})}{p(q_{ik})} - \log_2 p(s_{ij}) = IC(s_{ij}) - IC(s_{ij} \mid q_{ik})$$

given that the *conditional information content* of s_{ij} given the presence of q_{ik} ($IC(s_{ij}|q_{ik})$) is computed as follows:

$$IC(s_{ij} \mid q_{ik}) = -\log_2 p(s_{ij} \mid q_{ik}) = -\log_2 \frac{p(s_{ij},q_{ik})}{p(q_{ik})} \qquad (5)$$

Hence, it turns out that $PMI(s_{ij};q_{ik})$ measures how much q_{ik} tells us about s_{ij}, which is assumed to be removed prior to publication. That is, $PMI(s_{ij};q_{ik})$ is a measure of how much sensitive information we gain from s_{ij} (i.e., from $IC(s_{ij})$), conditioned to the presence of q_{ik} (i.e., $IC(s_{ij}|q_{ik})$).

Given the above arguments and numerical bounds, we propose computing the disclosure risk (DR) of s_{ij} with regard to q_{ik} as their PMI value:

$$DR(s_{ij};q_{ik}) = PMI(s_{ij};q_{ik}) \qquad (6)$$

Negative or zero values (which are unlikely for word pairs, as discussed above) will result in no disclosure risk, whereas positive values will indicate an increasing risk up to a maximum of $min(IC(s_{ij}),IC(q_{ik}))$.

3.2. Setting up the disclosure risk threshold

As discussed above, since most words are correlated in some degree, PMI and, hence, DR will usually be positive. Considering that only some words are highly correlated, to avoid an excessive sanitization that may destroy the utility of the output, we should define the value (i.e., threshold) of DR above which a feasible re-identification may arise. According to this value, only those q_{ik} that result in a high enough DR with respect to a sanitized term s_{ij} will be considered as quasi-identifiers (qs_i) and will be proposed for further sanitization.

In order to define the *disclosure risk threshold* (t_{DR}) in a generic and adaptive manner, we propose computing it according to the behavior of the initial sanitization mechanism ζ. Different ζ mechanisms applied to the same document may result in different sanitization outputs, each one reflecting the sanitization needs of a specific scenario. Hence, it makes sense that the sanitization behavior of our method, which is guided by t_{DR}, would also be coherent with ζ.

In order to achieve this goal, we assume that, regardless of the concrete sanitization technique, ζ aims at detecting those terms that provide *too much sensitive information* according to a criterion (either fixed or manually defined) [31, 32]. For example, methods based on NER assume that Named Entities should be systematically sanitized from text because, in general, they provide more information (due to their higher concreteness) than regular words [1]. Other works specifically state that those terms that provide more information than a user-defined value should be sanitized [31, 32]. In all cases, given the output D' of ζ, the set of terms in S_i reflect the notion of what is sensitive, that is, what is too informative according to ζ. Particularly, the *least informative* s_{ij} (i.e., s_{min}) for all S_i reflects the lower bound of the sanitization mechanism/criterion ζ, above which sensitive terms were found to be too informative. Considering the above arguments, we propose using this lower bound, which can be automatically computed given the sanitization output of ζ, as our threshold t_{DR}.

Formally, for every s_{ij} in every S_i in D', we look for the one that provides minimal information, that is, minimal *Information Content*. This IC value will serve as t_{DR} for our proposal:

$$t_{DR} = IC(s_{min}) = \min_{\forall S_i \text{ from } D'} \left(\min_{\forall s_{ij} \in S_i} (IC(s_{ij})) \right) \quad (7)$$

This value can be coherently compared with DR (i.e., PMI) values, since both represent *amounts of information*. In fact, as shown in eq. (3), $PMI(s_{min};s_{min})=IC(s_{min})$. Hence, according to eq. (4), any q_{ik} for which $PMI(s_{ij};q_{ik})$ is equal or above, $IC(s_{min})$ will state that the amount of information provided by q_{ik} with regard to s_{ij} is above the sanitization criterion of ζ. Thus, in coherency with ζ, q_{ik} should be considered a quasi-identifier (*QS*) to be sanitized. Formally, given a set S_i of sanitized terms and t_{DR} computed as in eq. (7), QS_i is obtained as follows:

$$QS_i = \bigcup_{\forall s_{ij} \in S_i} \left\{ q_{ik} \in Q_i \mid DR(s_{ij}; q_{ik}) \geq t_{DR} \right\} \tag{8}$$

Notice that we define one t_{DR} for all S_i in *D'*, since we assume that the sanitization criterion of ζ is the same for all *D'*. In this manner, since the t_{DR} calculus will consider more evidences, that is, more sanitized terms, it will better capture the behavior of ζ.

3.3. Generalizing the disclosure risk assessment

With the above formulation, correlations between *one* sensitive term s_{ij} and *one* non-sensitive term q_{ik} can be measured. However, as stated in the introduction for the case of SDC in relational databases and set-valued data, disclosure may appear by the combination of *several* non-sanitized quasi-identifiers. In this case, any combination of terms in Q_i, with a cardinality up to $|Q_i|$ should be considered as a possible disclosure threat for each s_{ij} in S_i.

To consider this scenario, the above approach can be generalized to evaluate the disclosure risk of a subset of Q_i (i.e., $\{q_{i1},q_{i2},\ldots,q_{ip}\} \subseteq Q_i$) with respect to a s_{ij}. It is important to note that this should still evaluate a *binary* correlation between two elements: *one* sensitive term s_{ij} and *several* non-sensitive ones $\{q_{i1}, q_{i2},\ldots,q_{ip}\}$. Hence, PMI calculus can be formulated as follows:

$$\begin{aligned} PMI(s_{ij};\{q_{i1},q_{i2},...,q_{ip}\}) &= \log_2 \frac{p(s_{ij},q_{i1},q_{i2},...,q_{ip})}{p(s_{ij})\,p(q_{i1},q_{i2},...,q_{ip})} = \\ &= \log_2 \frac{p(s_{ij} \mid \{q_{i1},q_{i2},...,q_{ip}\})}{p(s_{ij})} = IC(s_{ij}) - IC(s_{ij} \mid \{q_{i1},q_{i2},...,q_{ip}\}) \end{aligned} \tag{9}$$

As in section 3.1, with this formulation, it turns that $PMI(s_{ij};\{q_{i1},q_{i2},\ldots,q_{ip}\})$ measures the information gain from s_{ij} (i.e., $IC(s_{ij})$) conditioned, in this case, to the presence of the set $\{q_{i1},q_{i2},\ldots,q_{ip}\}$ (i.e., $IC(s_{ij}|\{q_{i1},q_{i2},\ldots,q_{ip}\})$.

Consequently, DR can be formulated as follows:

$$DR(s_{ij};\{q_{i1},q_{i2},...,q_{ip}\}) = PMI(s_{ij};\{q_{i1},q_{i2},...,q_{ip}\}) \quad (10)$$

Notice that this is different from a multivariate generalization of PMI, in which each term is treated as an independent variable. In this last situation, a multivariate PMI, such as *specific correlation* or *specific interaction information* [5], would measure the amount of information shared or given by the set of terms, rather than the information gain of one term conditioned to the presence of the other ones, as in eq. (10).

In the most general case, for a given Q_i with a cardinality $|Q_i|=n$ in the same context as s_{ij}, the set of distinct tuples of potential quasi-identifiers created from the combination of up to n elements in Q_i will be the *power set* of Q_i, excluding the empty set, that is $P_{\geq 1}(Q_i)$. This contains the result of combining k elements from Q_i, with any order, for k in $[1..n]$.

4. A practical method for the sanitization of semantic correlations

The application of the theoretical framework presented above in a practical setting is not trivial. On the one hand, the probability calculus should consider the co-occurrence of a potentially large number of terms (i.e., $p(s_{ij},q_{i1},q_{i2},\ldots,q_{ip}\})$). As more terms are added to this calculus, data sparseness increases because it is more difficult to retrieve evidences of such scarce combinations of words [30]. On the other hand, the number of possible combinations of terms in Q_i could be quite large, especially for wide contexts. This would require a considerable number of probability assessments.

In the following subsections, we detail the practical implementation of the theoretical framework, discussing issues related to the probability calculus and the textual processing. As a result, a general algorithm, which focuses on providing a feasible and efficient implementation, is proposed.

4.1. Probability calculus

The framework presented in section 3 extensively relies on IC and PMI calculus that, in turn, depends on term occurrence/co-occurrence probabilities. Thus, the way in which probabilities are computed is crucial to ensure: i) the generality of the approach, in the sense that it can be applied to heterogeneous documents of different domains referring to terms/entities with

different degrees of concreteness, and ii) the quality of the results, that is, the accurate assessment of term correlations.

In the past, some authors [24] have used tagged textual data as corpora for probability calculus, so that term frequencies can be obtained unambiguously. Although this strategy provided accurate results when applied to general terms [23], the cost of manually compiling and tagging text produced relatively small corpora with limited coverage. Hence, data sparseness usually appears when computing probabilities for concrete or domain-specific terms (e.g., rare diseases), NEs (e.g., person or organization names) or newly coined terms or neologisms (e.g., smartphone) [30], which are usually the most sensitive ones from the sanitization point-of-view. Moreover, data sparseness is especially prone to appear when looking for term co-occurrences from which we assess disclosure risk, rather than individual occurrences.

The compilation of a corpus with a good coverage of those kinds of problematic terms is challenging due to its potential size and the need to be continuously updated to consider evolving domains and dynamic terms such as NEs. In this sense, the Web stands out as the largest general-purpose corpus covering almost any possible up-to-date term. In fact, it has been argued that the Web is so large and heterogeneous that it represents the current distribution of terms at a social scale [4]. An additional advantage of the Web is the availability of Web Search Engines (WSEs) that directly provide web-scale page counts for queried terms, avoiding the necessity of analyzing such enormous repository. Note that, even though page counts represent a subestimation of total number of term appearances, since a term may appear several times in a document, considering the size of the Web, this approximation is representative enough [4]. In fact, many authors have already exploited page counts of specific queries to compute term probabilities in a variety of tasks [33, 34, 41, 43, 44].

Given the above arguments, in order to minimize data sparseness of probability calculus and to enable a general and unconstrained assessment regardless of the document's domain, we compute the probability of a term *t* from its relative number of *page counts* as provided by a WSE:

$$p(t) = \frac{page\_counts("t")}{total\_webs} \tag{11}$$

where *page_counts*("*t*") is the number of web sites indexed by a WSE for the query *t* within double quotes (""), and *total_webs* is the total amount of web sites indexed by the WSE.

Likewise, probabilities of a set of terms $\{t_1,\ldots,t_n\}$ can be computed according to the number of web pages in which they co-occur, as follows:

$$p(t_1;...;t_n) = \frac{page\_counts("t_1" \; AND \; ... \; AND \; "t_n")}{total\_webs} \tag{12}$$

4.2. Extracting terms from textual documents

The framework introduced in section 3 requires the extraction of contexts (C_i) and sanitized and non-sanitized terms (S_i and Q_i) from a given document *D'*. By *context*, we refer to a sentence, paragraph or a whole document. By *term*, we refer to concepts (e.g., acquired immunodeficiency syndrome) or NEs (e.g., person names). These are found in text as *nouns* and *noun phrases* (NPs), which are the basic units of a discourse with rich semantic content.

In order to extract contexts and NPs, we use a set of natural language-processing tools[2] to perform: i) *sentence detection*, ii) *tokenization*, so that individual words are detected, including contraction separation, iii) *part-of-speech tagging* (*POS*) of individual tokens and iv) *syntactic parsing* of POS tagged tokens, obtaining sets of words tagged as verbal (VPs), prepositional (PPs) and nominal phrases (NPs). On the one hand, sentences or groups of sentences are used to bind the context of the analysis according to a desired criterion. On the other hand, only NPs are considered during the extraction of terms to analyze. Note that a NP is considered a sanitized (S_i) if one or several of its words have been detected as sensitive by ζ. Other NPs found in the same context are considered non-sanitized terms to analyze (Q_i).

Moreover, since terms will be evaluated according to their appearance probabilities in the web when querying them in a WSE (as detailed in section 4.1), in order to focus the correlation analysis on the information provided by the conceptualization behind each NP, we also remove *stop words* from sanitized and non-sanitized terms. Stop words define a finite list of domain independent words including determinants, prepositions or adverbs that can be removed from a NP without altering its conceptualization (e.g., *a blood transfusion -> blood transfusion*). In this manner we avoid altering the probability calculus, since the presence of a stop word may constraint the number of results provided by the WSE. For example, in a WSE like Bing[3], the query *"a blood transfusion"* results in a page count of 802.000, while *"blood transfusion"* results in 3.500.000, even though both refer to the same concept.

[2] OpenNLP, Apache Software Foundation. Available at: http://opennlp.apache.org [accessed: November 22th, 2012}
[3] http://www.bing.com/ [accessed: January 3th, 2013]

The above process is illustrated in Fig. 1 and Fig. 2. The former shows the output of an initial sanitization mechanism $D'=\zeta(D)$ for a given text.

> The patient suffers from **acquired immunodeficiency syndrome** because of a blood transfusion. He was diagnosed when his **immune system** responded poorly to influenza.

Fig. 1. Sample text outputted by ζ. Terms in **boldface** are those proposed for sanitization by ζ.

Fig. 2 shows the output of the linguistic analysis described above, in which NPs corresponding to terms in S_i and in Q_i have been detected.

> [*NP* The patient] suffers from [*NP* **acquired immunodeficiency syndrome**] because of [*NP* a blood transfusion]. [*NP* He] was diagnosed when [*NP* his **immune system**] responded poorly to [*NP* influenza].

Fig. 2. Noun Phrases (NP) detected from sample text.

Considering the whole text as context (C_i) and removing stop words from NPs, we obtain S_i={*acquired immunodeficiency syndrome, immune system*} and Q_i={*patient, blood transfusion, influenza*}.

4.3. Algorithm

In this section, we detail the practical implementation of the approach Ψ presented in section 3. In the following, an iterative algorithm is presented so that, taking $D'=\zeta(D)$ as input, it automatically detects semantically correlated quasi-identifiers.

Algorithm 1. Detection of quasi-identifiers

```
Input:  D'=ζ(D) //sanitized document
Output: UQS     // quasi-identifiers identified for all contexts in D'

1    t_DR=compute_threshold(D'); //According to eq. (7)
2    for each (C_i ⊆ D') do
3        S_i=getSanitizedTerms(C_i);  //Extract NPs as in section 4.2
4        Q_i=getNon-sanitizedTerms(C_i); //Extract NPs as in section 4.2
5        for each (s_ij ∈ S_i) do
6            QS_i=empty; //Quasi-identifiers for the context C_i
7            k=1;
8            while (k ≤ |Q_i|) do
9                X=create_combinations(Q_i, k);
10               while (!empty(X)) do
11                   x_p=obtain_combination(X); //x_p is a k-subset of Q_i
12                   remove(x_p, X);
13                   if (DR(s_ij; x_p) > t_DR) then //As in eq. (10)
14                       put(x_p, QS_i);
15                       remove_terms(x_p, Q_i);
16                       remove_combinations_with_terms(x_p, X);
17                   end if
18               end for
19               k++;
20           end while
21       end for
22       put(QS_i, UQS);
23   end for
24   return UQS;
```

The algorithm starts by computing the disclosure risk threshold (t_{DR}) as a function of the IC of those terms already sanitized in *D'* (line 1, recall to section 3.2 for concrete details). Note that this threshold is fixed for the whole *D'*, since it represents the behavior of ζ, which is also fixed for *D*. Table 1 illustrates this process for the sample text introduced in section 4.2; in that case, t_{DR}=9.86, which corresponds to the IC computed from Bing for the term *immune system*.

Table 1. IC figures for sanitized terms according to ζ and IC values from Bing with *total_webs*=11 billions. **Bold** tuple represents the threshold t_{DR}.

Sanitized terms	*IC*
acquired immunodeficiency syndrome	14.03
immune system	**9.86**

Next, the algorithm individually analyses each context C_i in *D'*, looking for sanitized (S_i) and non-sanitized (Q_i) terms (lines 2-4, as described in section 4.2). It then assesses which terms in Q_i reveal too much information for each s_{ij} in S_i by computing the disclosure risk. In order to do so, for each s_{ij} in S_i, the set of *combinations X* of *k* terms from Q_i, with k=[1…|Q_i|], is obtained (lines 5-9). Then, the disclosure risk (DR) of s_{ij} with respect to each combination of *k* terms (i.e., x_p={q_{i1},…,q_{ik}}) is computed as in eq. (10). If this value is higher than the threshold t_{DR}, these {q_{i1},…,q_{ik}} terms are considered quasi-identifiers and are added to the vector QS_i (lines 13-14). Since these terms are removed from Q_i (line 15), and combinations in *X* involving one or several of them are also removed (line 16), they will not be considered in subsequent combinations/analyses. Table 2 illustrates the DR calculus for each s_{ij} with combinations from Q_i with cardinality *k=1*. Notice that stop words appearing in the sample text (i.e., *The*, *a*, *He*, *his*) have been removed from extracted NPs and that none of them resulted in a DR higher than the threshold.

Table 2. Disclosure risk computed from Bing for each sanitized term and each combination of terms in Q_i with cardinality *k=1*.

Sanitized terms (s_{ij})	*Potential quasi-identifiers (from Q_i)*	*DR*
acquired immunodeficiency syndrome	influenza	7.63
	blood transfusion	8.75
	patient	5.86
immune system	influenza	6.28
	blood transfusion	7.16
	patient	5.30

The process is repeated until the cardinality of the combinations (*k*) is higher than the number of elements remaining in Q_i (line 8). This condition is reached either because of the increment of *k* at each iteration (line 19), or because of the removal of elements in Q_i when those are detected as quasi-identifiers (line 15). Table 3 shows the DR calculus for combinations from Q_i with cardinality *k=2*. Notice how, in this case, the combination *influenza AND blood transfusion* produces a DR higher than the threshold for the sanitized term *acquired immunodeficiency syndrome*. Since the former terms will be removed from Q_i, no further iterations will be performed. As a result, *influenza* and *blood transfusion* will be proposed for sanitization (i.e., removal), whereas *patient* will remain in clear form (see the final output in Fig. 3).

Table 3. Disclosure risk computed from Bing for each sanitized term and each combination of term in Q_i with cardinality $k=2$. **Bold** tuple represents a set of detected quasi-identifiers.

Sanitized terms (s_{ij})	*Potential quasi-identifiers (from Q_i)*	*DR*
acquired immunodeficiency syndrome	**Influenza AND blood transfusion**	**3.00**
	Influenza AND patient	2.76
	blood transfusion AND patient	2.88
immune system	Influenza AND blood transfusion	2.48
	Influenza AND patient	2.20
	blood transfusion AND patient	2.38

The patient suffers from **acquired immunodeficiency syndrome** because of a **blood transfusion**. He was diagnosed when his **immune system** responded poorly to **influenza**.

Fig. 3. The final sanitized result after $\Psi(\zeta(D))$. Terms in **boldface** are those proposed for final sanitization.

Each context C_i results in a vector of quasi-identifiers Q_i. The ordered set of all Q_i (i.e., *UQS*), which represents the terms in *D'* that may cause disclosure and should be sanitized, is the output of the algorithm (line 24).

It is worth noting two design aspects that help to improve the efficiency of the algorithm and also to retain the utility of the sanitized output:

- On the one hand, the algorithm avoids evaluating again a term or a combination of terms that have been already detected as quasi-identifiers for a given C_i. As shown in line 4, the set of non-sanitized terms Q_i is global for all the s_{ij} in S_i. Given that terms detected as quasi-identifiers are removed from Q_i in line 15, those will not be considered in the next iterations for C_i, either during the creation of subsequent combinations (line 9) or in the analysis of the next s_{ij} in C_i (line 5). In addition, once a combination x_p has been detected as quasi-identifier, other combinations in X involving one or several terms from x_p are also removed (line 16). This makes sense, since the sanitization (i.e., removal) of a quasi-identifier, or a set of them, would negate the re-identification against the same or another s_{ij} in the case in which they were highly correlated, even when combined with additional terms. In this manner, we decrease the number of unnecessary DR calculations since a lower amount of term combinations will be needed to be evaluated as the cardinalities of X and/or Q_i decrease (considered in lines 10 and 8, respectively).

- On the other hand, the order in which combinations of non-sensitive terms are evaluated is significant to avoid excessive sanitizations and, hence, to retain the utility of the sanitized output. Concretely, the algorithm avoids starting the analysis with large sets of quasi-identifiers $\{q_{i1},\ldots,q_{ik}\}$ since, in case of a high DR value, we cannot discern which term(s) in $\{q_{i1},\ldots,q_{ik}\}$ are highly correlated with s_{ij} and which ones are irrelevant. Thus, to avoid detecting false quasi-identifiers when non-correlated or slightly correlated terms are combined with other highly correlated terms, we start the analysis with individual terms (i.e., $k=1$, as shown in line 7). Since in this initial iteration those highly correlated terms will be removed from Q_i (line 15), they will not influence the analysis of term combinations with a higher cardinality (i.e., k>1, line 19). Hence, for a given k, we can be sure that the combinations of terms to be analyzed include only those that have not be found to be highly correlated when analyzing them individually or in smaller sets, but which, when considered jointly, may still incur in disclosure. Notice, however, that the order in which combinations of the *same* cardinality are evaluated may affect the set of terms to be removed, it does not negatively influence disclosure.

5. Evaluation

In this section, we test the accuracy of the proposed method Ψ, implemented as detailed in section 4. Since the detection of NEs is the most usual and less constrained approach for document sanitization [1, 44], we used the state-of-the-art *Stanford Named Entity Recognizer* (SNER) [15] as input sanitizer ζ for our tests. ζ directly applies SNER over the input document D, which evaluates textual terms individually. Any term detected as NE (*persons*, *locations* and *organizations* are supported) is proposed for sanitization. As formalized in section 3, its output $D'=\zeta(D)$ is the input for our method, $\Psi(D')$.

Evaluation data consists of *real* texts containing highly sensitive information. Taking and extending those texts already used in [31] and [32], the evaluation dataset corresponds to Wikipedia English articles of a set of entities of different domains. Articles describe *persons*, *organizations* and *locations* in order to offer a favorable scenario for NER-based sanitizers. Moreover, for each entity type, different kind or articles are considered. Some of them refer to Anglo-Saxon entities, so that most terms and NEs appearing in text would be expressed in English, easing the detection for English-trained NE recognizers such as SNER [15]. Others correspond to Spanish (but well-known) entities that, even though their descriptions are written in English, could include NEs expressed by non-translatable Spanish words or localisms. In this manner, we can evaluate the degree of language dependency of ζ and the adaptability of Ψ.

Evaluated entities are listed in Table 4, together with the term detected as sensitive by ζ which, according to the criterion detailed in section 3.2, defines the threshold t_{DR} for our method. The threshold value is computed as in eq. (7). To compute term probabilities, we used Bing as WSE, fixing the total amount of indexed web sites in 11 billions[4].

Table 4. Evaluated Wikipedia articles with associated threshold terms according to ζ and t_{DR} values computed from Bing.

Wikipedia Article	*Threshold term*	t_{DR}
Steve Wozniak	Steve Jobs	10.38
Steven Spielberg	Spielberg	10.30
Tom Cruise	Magnolia	10.04
Arnold Schwarzenegger	California	4.86
Sylvester Stallone	Stallone	10.52
Audrey Hepburn	London	5.03
Antoni Gaudi	Spain	4.88
Antonio Banderas	Antonio Banderas	11.13
Javier Bardem	Boca	7.25
Jordi Mollà	United States	4.43
Dreamworks	LLC	5.13
Microsoft	United States	4.43
Apple	United States	4.43
Aston Martin	England	5.68
Volkswagen	Audi	7.79
Port Aventura	Europe	4.74
Yellowstone	North America	6.86
Barcelona	Europe	4.74
Tarragona	Spain	4.88
Salou	Spain	4.88

5.1. Evaluating quasi-identifiers

Evaluation has been carried out by requesting two human experts to select and agree on which terms or term combinations (i.e., words or NPs, including NEs) that have not been detected as sensitive by ζ, would feasibly reveal any of the terms that have been tagged for sanitization. Hereinafter, we will refer to this set of terms as *Human_UQS*. By comparing *Human_UQS* with

[4] http://www.worldwidewebsize.com/ [last accessed: January 3th, 2013]

the output of our method, that is *UQS*, as detailed in section 3, the performance of our proposal has been quantified according to *precision*, *recall* and *F-measure*.

Precision (eq. (13)) is calculated as the percentage between the number of quasi-identifiers detected by Ψ (i.e., *UQS*) that have also been selected by the human experts in *Human_UQS*, and the total amount of automatically detected terms (i.e., |*UQS*|). The higher the precision, the lower the amount of incorrectly detected quasi-identifiers and, hence, the better the utility of the output.

$$Precision = \frac{|UQS \cap Human\_UQS|}{|UQS|} \times 100 \tag{13}$$

Recall (eq. (14)) is calculated as the ratio between the number of terms in *UQS* that are also in *Human_UQS*, and the total amount of terms in *Human_UQS*. Recall indicates how many quasi-identifiers have been detected. The higher the recall, the more private the sanitized output is.

$$Recall = \frac{|UQS \cap Human\_UQS|}{|Human\_UQS|} \times 100 \tag{14}$$

Finally, *F-measure* (eq. (15)) quantifies the harmonic mean of recall and precision, summarizing the accuracy of the sanitization when the same weight is given to the output's utility (i.e., precision) and privacy (i.e., recall).

$$F-measure = \frac{2 \times Recall \times Precision}{Recall + Precision} \tag{15}$$

Table 5 shows evaluation results for the set of *UQS* for each article. Two different tests have been performed modifying the notion of *context*. In the first one, the whole Wikipedia description has been defined as the context (i.e., C_x=*D'*), that is, all terms in *D'* are analyzed as a whole to evaluate their degree of semantic correlation. This makes sense since Wikipedia descriptions are tight and focused discourses. In the second test, the *context* has been set to individual *sentences*, so that only tuples of terms appearing within the same sentence are evaluated.

Table 5. Evaluation of quasi-identifiers (*UQS*) detected by Ψ for different Wikipedia articles setting the *context* for analysis to complete documents (C_x=*D'*) and to individual sentences (C_x=*sentence*).

Wikipedia Article	*Precision*		*Recall*		*F-Measure*	
	C_x=*D'*	C_x=*sentence*	C_x=*D'*	C_x=*sentence*	C_x=*D'*	C_x=*sentence*
Steve Wozniak	100%	100%	75%	25%	85,71%	40%
Steven Spielberg	71,42%	100%	55,5%	18,75%	62,5%	31,57%
Tom Cruise	100%	100%	82,35%	28,57%	90,32%	44,44%
Arnold Schwarzenegger	68,75%	84,61%	100%	91,66%	81,48%	88%
Sylvester Stallone	50%	66,66%	66,66%	40%	57,14%	50%
Audrey Hepburn	66,66%	81,48%	100%	81,48%	80%	81,48%
Antoni Gaudi	51,28%	72,72%	90,90%	38,09%	65,57%	50%
Antonio Banderas	80%	100%	57,14%	40%	66,66%	57,14%
Javier Bardem	77,77%	94,73%	100%	94,73%	87,5%	94,73%
Jordi Mollà	53,84%	87,5%	100%	70%	70%	77,77%
Dreamworks	35%	57,14%	100%	44,44%	51,85%	50%
Microsoft	48,57%	65,21%	100%	83,33%	65,38%	73,17%
Apple	58,33%	75%	84%	56,25%	68,85%	64,28%
Aston Martin	55,55%	55,55%	100%	83,33%	71,42%	66,66%
Volkswagen	80%	80%	66,66%	66,66%	72,72%	72,72%
Port Aventura	57,89%	64,28%	100%	56,25%	73,33%	60%
Yellowstone	71,42%	88,88%	90,90%	66,66%	80%	76,19%
Barcelona	31,88%	51,28%	88%	48,78%	46,80%	50%
Tarragona	50%	100%	100%	25%	66,66%	40%
Salou	52,94%	71,42%	100%	62,5%	69,23%	66,66%
Average	*63,06%*	*79,82%*	*87,86%*	*56,07%*	*70,65%*	*61,74%*

Results show significant differences between precision and recall. When using the whole document as context, precision falls to an average of 63% in comparison with the almost 80% achieved, in average, when correlations are evaluated within the context of individual sentences. Indeed, as contexts become larger, the degree of semantic ambiguity increases as term tuples to evaluate appear in more distant positions within the texts. Since the probabilistic assessment does not apply any kind of semantic disambiguation, which would require additional knowledge sources and/or more complex queries, this hampers the accuracy of the results. On the contrary, semantic ambiguity is more unlikely to appear within the same sentence, since it defines a tighter discourse.

Recall, on the other hand, behaves inversely: the analysis of the document as a whole produces recall figures of more than 87% in average, with half of the entities achieving a perfect 100%, whereas the analysis of individual sentences results in an average value below 56%. This shows that semantic correlations may appear between terms located in different sentences within the same document. The fact that some entities (i.e., *Steve Wozniak, Steven Spielberg, Tom Cruise, Tarragona*) show recall values below 30% when analyzed sentence by sentence states that the scope of this kind of analysis is insufficient to provide a robust sanitization. This makes sense since most sentences of Wikipedia articles describe the same underlying entity and, thus, their terms are likely to be highly correlated.

Notice that, even though a high precision is desirable to retain the utility of the output, recall usually plays a more important role in document sanitization. Low recall implies that a number of terms that enable a feasible disclosure of sensitive terms will appear in the sanitized output. Since disclosing of a sole sensitive term may negate the sanitization [2], a high recall figure is more important, in most scenarios, than precision. The above results also illustrate the potential problems of correlation-oriented methods focused on very narrow contexts, such as consecutive term pairs [2]. In any case, contexts (i.e., sentences, paragraphs, documents) should be specified for concrete documents, according to their structure and discourse and also according to the sanitization needs (i.e., utility preservation or low disclosure risk).

Through the analysis of individual results, we observe notorious differences. For some entities (e.g., *Steve Wozniak, Tom Cruise*) perfect precisions are achieved, whereas for others (e.g., *Barcelona, Microsoft*), figures are below 50%. This variability is closely related to the threshold value t_{DR} used to guide the analysis of Ψ (see Table 4). We observe that those entities with a very general threshold (e.g., *Europe* for *Barcelona*, *United States* for *Microsoft*) tend to produce looser results than those with more concrete thresholds (e.g., *Magnolia* for *Tom Cruise* or *Steve Jobs* for *Steve Wozniak*). Notice that a general threshold term (i.e., with a low IC) results in a low t_{DR} value and, according to eq. (8), in a strict criterion for selecting quasi-identifiers. Since many of these quasi-identifiers will reveal a low amount of sensitive information (i.e., low disclosure) their sanitization will negatively affect the utility of the output. The large differences between t_{DR} values are caused by the sanitization criterion of ζ, from which t_{DR} is computed as detailed in section 3.2. Since ζ is based on NE recognition, it systematically proposes for sanitization NEs regardless of the fact that they indeed reveal identifying and/or confidential information. Certainly, NEs are usually more specific than normal words because they refer to individuals rather than to concepts. However, many NEs are so general (e.g., continent or country names) and, hence, they provide such a low amount of information (i.e., IC), that they hardly pose a risk. Since these general NEs are precisely those that define the sanitization

threshold t_{DR}, their detection forces our method to implement a strict sanitization, resulting in many terms considered as quasi-identifiers. This illustrates the adaptability of Ψ with regard to the –even though imperfect- sanitization proposed by ζ.

5.2. Evaluating the sanitized output

In order to quantify the contribution of our method Ψ over the basic sanitization mechanism ζ, we also evaluated and compared the output of ζ alone against the one provided when applying Ψ after ζ. In this case, human experts were requested to identify individual terms and term combinations in *D* that should be sanitized to avoid disclosing identifying or confidential information for each entity. This set is compared against the set of terms tagged for sanitization by ζ alone and against those identified by the combination of $\Psi(\zeta)$ to compute precision, recall and F-measure as in eq. (13), (14) and (15). Table 6 depicts evaluation results for this scenario, when considering the whole document as the *context* analysis for Ψ.

Evaluation figures show that the application of Ψ after the sanitization performed by ζ positively affects the output. Even though the global precision lowers from an average of 89% to around 75%, recall almost doubles from 49% to over 93%, with 10 cases of perfect recall, which reflect a much reduced disclosure risk. Considering the sanitization behavior of ζ, two conclusions can be extracted. First, not only NEs appearing in text may reveal sensitive information, but also NPs referring to concrete terms [31, 32]. The latter ones are indirectly detected by our method through their degree of correlation with a sanitized NE. Second, semantic correlations between terms represent a tangible privacy threat since their co-occurrence would reveal, in many occasions, sensitive information and/or enable disclosing already sanitized entities. This compromises the privacy of the outputs of a sanitizer evaluating terms individually [2].

Moreover, for the case of ζ, we noticed that the SNER failed in several occasions in detecting real and sensitive NEs. This was especially relevant when dealing with Spanish entities (e.g., *Tarragona*, *Port Aventura*, etc.), for which even the entity name was omitted, and also when analyzing actor biographies, for which movie titles remained undetected, since they do not match with any of the predefined NE categories (i.e., persons, locations, organizations). This shows the limitations of trained classifiers, whose training data may not be enough to offer a general-purpose cross-domain and language-independent solution. In comparison, our method focuses on the most informative terms, which are assessed as those that appear less frequently in the Web. Thus, *recall*, as shown in Table 6, is less affected by data sparseness. The counterpart

is that data sparseness may appear when evaluating terms with complex syntactic constructions that tend to produce a relatively low page count when queried in a WSE. As a result, some syntactically complex terms referring to non-so specific conceptualizations may be detected as quasi-identifiers due to their apparent informativeness, producing a lower *precision*, as shown in Table 6. In any case, the improvement in recall more than compensates the lowered precision, resulting in an F-measure that is 20% higher in average.

Table 6. Evaluation of the sanitized output of ζ and of the combination of $\Psi(\zeta)$.

Wikipedia Article	*Precision*		*Recall*		*F-measure*	
	ζ	$\Psi(\zeta)$	ζ	$\Psi(\zeta)$	ζ	$\Psi(\zeta)$
Steve Wozniak	100%	100%	50%	87,5%	66,66%	93,33%
Steven Spielberg	100%	81,81%	30,76%	69,23%	47,05%	75%
Tom Cruise	100%	100%	29,16%	87,5%	45,16%	93,33%
Arnold Schwarzenegger	100%	77,27%	35,29%	100%	52,17%	87,17%
Sylvester Stallone	100%	81,81%	70%	90%	82,35%	85,71%
Audrey Hepburn	85,71%	73,68%	42,85%	100%	57,14%	84,84%
Antoni Gaudi	90%	59,18%	29,03%	93,54%	43,90%	72,5%
Antonio Banderas	100%	90%	41,66%	75%	58,82%	81,81%
Javier Bardem	94,11%	85,71%	53,33%	100%	68,08%	92,30%
Jordi Mollà	80%	67,85%	63,15%	100%	70,58%	80,85%
Dreamworks	90%	62,5%	72%	100%	80%	76,92%
Microsoft	78,57%	57,14%	39,28%	100%	52,38%	72,72%
Apple	62,5%	59,09%	16,66%	86,66%	26,31%	70,27%
Aston Martin	87,5%	76%	73,68%	100%	80%	86,36%
Volkswagen	100%	91,66%	53,84%	84,61%	70%	88%
Port Aventura	89,47%	73,68%	60,71%	100%	72,34%	84,84%
Yellowstone	71,42%	71,42%	47,61%	95,23%	57,14%	81,63%
Barcelona	62,5%	41,58%	44,44%	93,33%	51,94%	57,53%
Tarragona	100%	80%	75%	100%	85,71%	88,88%
Salou	90%	66,66%	50%	100%	64,28%	80%
Average	*89,09%*	*74,85%*	*48,92%*	*93,13%*	*61,60%*	*81,70%*

6. Conclusions

As discussed in section 2, most sanitization methods evaluate textual terms independently, neglecting the privacy threat caused by semantic correlations between sanitized and non-sanitized terms. As stated in [2], the presence of highly correlated terms may be exploited by an attacker to negate the sanitization of a sensitive one.

To tackle this issue, in this paper, we presented a general purpose method to minimize the disclosure risk of the outputs provided by sanitization methods dealing with textual terms individually. Our proposal relies on an information theoretical formulation of the disclosure risk inherent to term correlations to detect additional terms to sanitize. This formulation is general, in the sense that it can be applied to any text regardless of its discourse structure, and it can be used to detect correlations of any cardinality. Moreover, the detection criterion (i.e., threshold) is automatically adapted to the behavior of the mechanism used to sanitize individual terms. The implementation of our proposal, exploiting the Web as the source to evaluate term correlations, and its evaluation using a set of real texts, show a much improved sanitization recall over usual mechanisms based on NE recognition. As a result, our approach effectively helped to reduce the disclosure risk of the sanitized output.

As future work, some aspects may be considered to improve the utility of the sanitized document while maintaining its privacy. First, as stated in section 3, our method assumes that individually sanitized terms will be removed prior to publication as in [7, 42], that is, they will provide zero information. However, some sanitizers [1, 2] opt to replace sensitive terms by less informative versions, such as conceptual generalizations (e.g., AIDS->disease). Even though this strategy helps to better retain document's utility, it may also increase the disclosure risk, since an attacker could exploit the information provided by the generalized term, which is non-zero, in addition to that provided by correlated terms. To tackle this problem, our quantification of disclosure risk should be reformulated, so that it also considers the additional information, and hence the increased risk, provided by term generalizations. Also related to the improvement of document's utility, we may also opt to replace correlated terms by appropriate generalizations, instead of removing them. To do so, two aspects should be considered. First, a knowledge base modeling taxonomic relationships should be exploited to extract generalizations. General-purpose ontologies/taxonomies, such as WordNet [14], can be considered. Then, the informativeness of term generalizations towards their correlated terms should also be measured so that we can be sure that the selected generalization fulfills the privacy criterion t_{DR}. Finally, the accuracy and generality of the Web-based IC calculus may be compared with other IC computation models exploiting more structured knowledge sources [26-28].

Acknowledgements and disclaimer

Authors are solely responsible for the views expressed in this paper, which do not necessarily reflect the position of UNESCO nor commit that organization. This work was partly supported by the European Commission under FP7 project Inter-Trust, by the Spanish Ministry of Science and Innovation (through projects eAEGIS TSI2007-65406-C03-01, ICWT TIN2012-32757, CO-PRIVACY TIN2011-27076-C03-01, ARES-CONSOLIDER INGENIO 2010 CSD2007-00004, Audit Transparency Voting Process IPT-430000-2010-31 and BallotNext IPT-2012-0603-430000), by the Spanish Ministry of Industry, Commerce and Tourism (through projects eVerification2 TSI-020100-2011-39 and SeCloud TSI-020302-2010-153) and by the Government of Catalonia (under grant 2009 SGR 1135).

References

[1] D. Abril, G. Navarro-Arribas, V. Torra, On the declassification of confidential documents, in: Modeling Decision for Artificial Intelligence. 8th International Conference, MDAI 2011, Springer, 2011, pp. 235–246.

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

[3] G. Bouma, Normalized (Pointwise) Mutual Information in Collocation Extraction, in: E.d.C.S. Chiarcos (Ed.) Biennial GSCL Conference 2009, Gunter Narr Verlag, Tübingen, Germany, 2009, pp. 31–40.

[4] R.L. Cilibrasi, P.M.B. Vitányi, The Google Similarity Distance, IEEE Transactions on Knowledge and Data Engineering, 19 (2006) 370-383.

[5] T.V.d. Cruys, Two Multivariate Generalizations of Pointwise Mutual Information, in: Workshop on Distributional Semantics and Compositionality (DiSCo'2011), Association for Computational Linguistics, Portland, Oregon, USA, 2011, pp. 16-20.

[6] C. Cumby, R. Ghani, A machine learning based system for semiautomatically redacting documents, in: Twenty-Third Conference on Innovative Applications of Artificial Intelligence, San Francisco, California, USA, 2011, pp. 1628–1635.

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

[8] D. Chen, H. Zhao, Data security and privacy protection issues in cloud computing, in: 2012 International Conference on Computer Science and Electronics Engineering (ICCSEE'12), Hangzhou, China, 2012, pp. 647–651.

[9] K.W. Church, P. Hanks, Word association norms, mutual information, and lexicography, Computational Linguistics, 16 (1990) 22-29.

[10] Department of Health and Human Services, The health insurance portability and accountability act of 1996, in, 2000.
[11] J. Domingo-Ferrer, A Survey of Inference Control Methods for Privacy-Preserving Data Mining, in: C.C. Aggarwal, P.S. Yu (Eds.) Privacy-Preserving Data Mining, Springer, 2008, pp. 53-80.
[12] D. Dorr, W. Phillips, S. Phansalkar, S. Sims, J. Hurdle, Assessing the difficulty and time cost of de-identification in clinical narratives, Methods of Information in Medicine, 45 (2006) 246–252.
[13] M. Douglass, G. Cliffford, A. Reisner, W. Long, G. Moody, R. Mark, De-identification algorithm for free-text nursing notes, in: Computers in Cardiology, 2005, pp. 331–334.
[14] C. Fellbaum, WordNet: An Electronic Lexical Database, MIT Press, Cambridge, Massachusetts, 1998.
[15] J. Finkel, T. Grenager, C. Manning, Incorporating non-local information into information extraction systems by gibbs sampling, in: 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05), Ann Arbor, USA, 2005, pp. 363–370.
[16] W. Jiang, M. Murugesan, C. Clifton, L. Si, t-plausibility: Semantic preserving text sanitization, in: International Conference on Computational Science and Engineering (CSE'09), Vancouver, Canada, 2009, pp. 68–75.
[17] S. Martínez, D. Sánchez, A. Valls, Semantic adaptive microaggregation of categorical microdata, Computers & Security, 31 (2012) 653-672.
[18] S. Martínez, D. Sánchez, A. Valls, A semantic framework to protect the privacy of electronic health records with non-numerical attributes, Journal of Biomedical Informatics, 46 (2013) 294-303.
[19] S. Martínez, D. Sánchez, A. Valls, M. Batet, Privacy protection of textual attributes through a semantic-based masking method, Information Fusion, 13 (2012) 304-314.
[20] R. Mishra, S. Dash, D. Mishra, A. Tripathy, A privacy preserving repository for securing data across the cloud, in: 3rd International Conference on Electronics Computer Technology (ICECT'11), 2011, pp. 6–10.
[21] National Security Agency, Redacting with confidence: How to safely publish sanitized reports converted from word to pdf, in, 2005.
[22] National Security Agency, Redaction of pdf files using Adobe Acrobat Professional X, in, 2011.
[23] P. Resnik, Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language, Journal of Artificial Intelligence Research, 11 (1999) 95-130.
[24] P. Resnik, Using Information Content to Evalutate Semantic Similarity in a Taxonomy, in: C.S. Mellish (Ed.) 14th International Joint Conference on Artificial Intelligence, IJCAI 1995, Morgan Kaufmann Publishers Inc., Montreal, Quebec, Canada, 1995, pp. 448-453.
[25] D. Sánchez, A methodology to learn ontological attributes from the Web, Data & Knowledge Engineering 69 (2010) 573-597.
[26] D. Sánchez, M. Batet, A New Model to Compute the Information Content of Concepts from Taxonomic Knowledge, International Journal on Semantic Web and Information Systems, 8 (2012) 34-50.
[27] D. Sánchez, M. Batet, Semantic similarity estimation in the biomedical domain: An ontology-based information-theoretic perspective Journal of Biomedical Informatics, 44 (2011) 749-759.

[28] D. Sánchez, M. Batet, D. Isern, Ontology-based Information Content computation, Knowledge-based Systems, 24 (2011) 297-303.
[29] D. Sánchez, M. Batet, A. Valls, Web-Based Semantic Similarity: An Evaluation in the Biomedical Domain, International Journal of Software and Informatics, 4 (2010) 39-52.
[30] D. Sánchez, M. Batet, A. Valls, K. Gibert, Ontology-driven web-based semantic similarity, Journal of Intelligent Information Systems, 35 (2010) 383-413.
[31] D. Sánchez, M. Batet, A. Viejo, Automatic general-purpose sanitization of textual documents, IEEE Transactions on Information Forensics and Security, (2013) (in press).
[32] D. Sánchez, M. Batet, A. Viejo, Detecting sensitive information from textual documents: an information-theoretic approach, in: Modeling Decisions for Artificial Intelligence. 9th International Conference, MDAI 2012, Springer, 2012, pp. 173-184
[33] D. Sánchez, J. Castellà-Roca, A. Viejo, Knowledge-based scheme to create privacy-preserving but semantically-related queries for web search engines, Information Sciences, 218 (2013) 17-30.
[34] D. Sánchez, D. Isern, Automatic extraction of acronym definitions from the Web, Applied Intelligence, 34 (2011) 311-327.
[35] D. Sánchez, A. Moreno, Learning non-taxonomic relationships from web documents for domain ontology construction, Data & Knowledge Engineering, 63 (2008) 600-623.
[36] D. Sánchez, A. Moreno, Pattern-based automatic taxonomy learning from the Web, AI Communications, 21 (2008) 27-48.
[37] D. Sánchez, A. Moreno, L.D. Vasto-Terrientes, Learning relation axioms from text: An automatic Web-based approach, Expert Systems with Applications, 39 (2012) 5792-5805.
[38] L. Sweeney, k-anonymity: a model for protecting privacy, International Journal Uncertainty Fuzziness Knowledge-Based Systems, 10 (2002) 557-570.
[39] L. Sweeney, Replacing personally-identifying information in medical records, the scrub system, in: 1996 American Medical Informatics Association Annual Fall Symposium, Washington, DC, USA, 1996, pp. 333–337.
[40] M. Terrovitis, N. Mamoulis, P. Kalnis, Privacy-preserving anonymization of set-valued data, in: Proceedings of the VLDB Endowment, 2008, pp. 115-125.
[41] P.D. Turney, Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL, in: L. De Raedt, P. Flach (Eds.) 12th European Conference on Machine Learning, ECML 2001, Springer-Verlag, Freiburg, Germany, 2001, pp. 491-502.
[42] A. Tveit, O. Edsberg, T.B. Rost, A. Faxvaag, O. Nytro, T. Nordgard, M.T. Ranang, A. Grimsmo, Anonymization of general practioner medical records, in: second HelsIT Conference, 2004.
[43] C. Vicient, D. Sánchez, A. Moreno, An automatic approach for ontology-based feature extraction from heterogeneous textual resources, Engineering Applications of Artificial Intelligence, 26 (2013) 1092-1106.
[44] A. Viejo, D. Sánchez, J. Castellà-Roca, Preventing automatic user profiling in Web 2.0 applications, Knowledge-Based Systems, 36 (2012) 191-205.