Sensitive-aware Privacy Index for Sanitized Text Data

Claudio Carpineto and Giovanni Romano

Fondazione Ugo Bordoni, Rome, Italy
{carpinet,romano}@fub.it

Abstract. Although a number of sanitization methods for text databases have been recently proposed, there has been so far little work on general ways to measure what can be learned about the user from the sanitized database relative to what can be learned from the original database. In this paper we propose a new privacy index, termed Sensitive-aware Privacy Index (SPI), that extends the common approach of comparing the global information content of the two databases by taking into account the relative importance of the single documents in each database. This is achieved through a form of weighted Jensen Shannon divergence, in which the weights reflect the document's sensitivity, as determined by an ad hoc classifier. Using search queries as an illustration, we show that SPI provides more reliable and consistent indications than sensitive-unaware information theoretic measures, both under a generic anonymization technique as well as with privacy-controlled query logs.

1 Introduction

The amount of text data formed by ‘documents’ associated with users (e.g. social network posts, search queries, medical notes, tweets, reviews, email messages) has been growing exponentially in recent years, giving rise to new opportunities and challenges. While the value of analyzing these data is widely recognized, their publication raises privacy concerns, both in terms of identity disclosure (i.e., when an attacker is able to match a document in a database to an individual) and attribute disclosure (i.e., when an attacker is able to find the sensitive documents, with or without reidentification). Even though we remove explicit identifiers (such as personal name, social security number, address), quasi identifiers (such as zip code, gender, birthdate), and directly-sensitive items (such as marriage status, national origin, salary, religion, sexual orientation, diseases) from a user's documents by using natural language processing techniques, it may still be possible to infer some of these attributes by combining other, seemingly irrelevant parts of the documents with external databases [11] [8].

To better protect the user's privacy while at the same time preserving the utility of the shared data, a number of sanitization methods for text data have been made available,[1] often as an extension of earlier privacy models for struc-

IIR 2018, May 28-30, 2018, Rome, Italy. Copyright held by the author(s).

[1] In this paper, by sanitization we mean both the protection of identity (usually referred to as anonymization) and of private information.

tured data. Among others, k-anonymity [1], differential privacy [12], and user clustering [9] . The availability of several sanitization strategies, each enforcing certain privacy guarantees for a specific set of parameters and type of output, raises the question of their evaluation and comparison.

Previous work in the microdata field has focused on global ways to to measure the changes in information content that follow data sanitization, without making specific assumptions about an attacker. Several information theoretic measures have been proposed, including mutual information [2] and Kullback-Leibler divergence [6]. Although a straightforward application of this approach to text data is possible [14], we argue that it does not seem very appropriate because the two domains are fundamentally different. In structured databases, the quasi-identifiers and sensitive attributes are known a priori, so that we can neglect the other attributes. In text databases, we have documents instead of attributes and each document may be sensitive or quasi-identifier. This suggests that we should be able to estimate how the single documents may affect the user privacy besides considering the changes in their probability distributions.

2 Sensitive-aware Privacy Index

Given a database X containing text documents associated with a set of users N, and a sanitized version of X denoted by Y, let $X_u = \{d_{u,1}, d_{u,2}, ..., d_{u,j}\}$ and $Y_u = \{d_{u,1}, d_{u,2}, ..., d_{u,k}\}$ be the set of documents associated with user u in, respectively, X and Y. We want to measure some difference between the set of a user's documents before and after sanitization, relating such a difference to the gain of privacy.

One natural starting point [6] is to compute the Kullback-Leibler divergence (KLD) of Y_u from X_u:

$$\mathrm{KLD}(X_u||Y_u) = \sum_d \left[P(d|X_u) \ \cdot \ log \frac{P(d|X_u)}{P(d|Y_u)} \right] \tag{1}$$

Intuitively, the larger the KLD value, the more difficult it is to find useful information to break the user's privacy. However, the use of KLD in our scenario does not come without problems. In order to compute Expression 1, we need to estimate $P(d|X_u)$ and $P(d|Y_u)$; i.e., the probability of the document d given the original and sanitized datasets. This problem is made difficult by the fact that there may be documents in X_u not present in Y_u (e.g., due to document suppression), as well as documents in Y_u not present in X_u (e.g., due to document perturbation). In particular, we cannot set $P(d|Y_u), d \in X_u, d \notin Y_u$, to zero, because $\mathrm{KLD}(X_u||Y_u)$ is not defined in this (very common) case.

To overcome this difficulty, rather than applying some smoothing procedure to KLD, we use the Jensen-Shannon divergence (JSD) between X_u and Y_u:

$$\mathrm{JSD}(X_u||Y_u) \ = \ \frac{1}{2}\,\mathrm{KLD}(X_u||M_u) + \frac{1}{2}\,\mathrm{KLD}(Y_u||M_u) \tag{2}$$

where $M_u = \frac{1}{2}(X_u + Y_u)$. Unlike KLD, JSD is always defined and is bounded by 1.

We next observe that in Equations 1 and 2 any document is treated in the same manner, whereas some documents are clearly more important than others for user identification or disclosure of confidential information. We assume that it is possible to estimate the sensitivity of a document d automatically by an ad-hoc classifier, and denote by σ_d the class membership probability of the document.

We can now use σ_d to weight the contribution made by the single documents to KLD (and JSD). A released document with a large σ_d affects privacy negatively (i.e., the chances of privacy breaks increase), which means that the divergence should become smaller (compared to the value obtained releasing a document with a lower σ_d). The contribution of a single document should thus be inversely related to its σ_d value. We set the weight of d (w_d) to 1 minus the probability that d belongs to the sensitivity class: $w_d = 1 - \sigma_d$. The weighted KLD (WKLD) is given by:

$$\text{WKLD}(w_d\,; X_u || Y_u) = \sum_d \left[w_d \cdot P(d|X_u) \ \cdot \ log \frac{P(d|X_u)}{P(d|Y_u)} \right] \tag{3}$$

The formula to compute the Sensitive-aware Privacy Index (SPI) for X and Y is finally obtained by averaging the weighted JSD (WJSD) over the set of users:

$$\text{SPI}(X, Y) = \frac{1}{N} \sum_u \text{WJSD}(w_d\,; X_u || Y_u) =$$

$$= \frac{1}{N} \sum_u \left[\frac{1}{2}\,\text{WKLD}(w_d\,; X_u || M_u) + \frac{1}{2}\,\text{WKLD}(w_d\,; Y_u || M_u) \right]$$

Note that when all the weights w_d are equal to 1 (i.e., if the probability that any document belongs to the sensitive class is equal to zero), SPI coincides with JSD. In general, the value of SPI will be smaller than JSD. Like JSD, SPI is bounded by 0 and 1. In particular, if X = Y, then SPI = 0.

3 Experiments with query logs

For our experiments, we consider search query logs, a specific but important type of text data that has been the focus of much privacy research in the last ten years [5]. In all our experiments we used a subset of the well known AOL dataset. It contained the queries associated with 10,000 'heavy' users, who were randomly selected among those who entered more than 44 queries; i.e., the average number of queries per AOL user. In this way we removed the users with very short profiles, who are not very interesting from the point of view of privacy. The number of random users (i.e., 10,000) was decided by experimenting with increasing samples, until the results stabilized.

3.1 Training and testing SPI's sensitivity weights

In order to compute SPI for the query log data, we need to find σ_d; i.e., the probability that any given search query is sensitive. As search queries are usually very short and do not contain repetitions, we trained a naive Bayes classifier using a bag of words model with binary features. As a training set, we used a subset of AOL queries that were manually labeled as sensitive or not-sensitive, first introduced in [4]. The classifier achieved 78% 10 fold cross validation accuracy on the labeled data set. When we ran the classifier on the heavy AOL users, we found that about half of the queries were labeled as sensitive. Compared to an earlier experiment on the whole population of AOL users using a small training set [16], we achieved better classification accuracy and found fewer sensitive queries.

3.2 Evaluating SPI under k-anonymity

To evaluate the performance of SPI on sanitized databases, we generated a number of query logs under k-anonymization [1], a simple privacy policy which requires that for each query there are at least other k-1 equal queries entered by distinct users. Using k-anonymization, the level of privacy protection can be increased by choosing larger values of k, which results in the suppression of rare as well as relatively frequent queries, with only the most common queries being released.[2] We let k vary from 1 to 1000, thus progressively strengthening the privacy requirements, and generated the corresponding released logs for the heavy AOL users. Then, for each released log, we computed four measures: (i) SPI, (ii) JSD (i.e., the unweighted version of SPI), (iii) the number of released queries (also known as impressions), and (iv) the Profile Exposure Level (PEL), one of the few earlier privacy measures for text data of which we are aware, presented in [7] and [14]. The definition of PEL is the following:

$$\mathrm{PEL} = \frac{\mathrm{I}(X,Y)}{\mathrm{H}(X)} \cdot 100; \quad \mathrm{I}(X,Y) = \sum_{x,y} p(x|y) \cdot p(y) \cdot log \frac{p(x|y)}{p(x)} \tag{4}$$

The ratio between mutual information I and entropy H (known in statistics as the uncertainty coefficient) gives a measure of the information that Y provides about X, normalized with respect to the information of X. To estimate $p(x)$, $p(y)$, and $p(x|y)$ in Equation 4 we used the method described by the authors.

In Figure 1 we show how the four measures varied as a function of k, for the heavy AOL users. The function JSD monotonically grew as k increased, because larger values of k are associated with smaller subsets of released queries, thus increasing the divergence from the original log. We checked that the percentage of impressions, the value of k, and JSD were all highly correlated with one another, with pairwise Pearson's correlation coefficients greater than 0.85 (in absolute value). Althoug JSD seems a more powerful privacy index than impressions and k, in the particular setting of our experiment they provided similar indications.

[2] Note that this behavior is not specific to k-anonymization; it can be observed in most privacy-protection methods, including differential privacy and user clustering.

The function SPI exhibited a different pattern and was weakly correlated with the former measures. While it generally grew as k increased, it also showed some notable oscillations pointing to problems with the privacy content of the queries suppressed at step k. In fact, unlike JSD, SPI may decrease when we remove some documents from the released database. This happens if we remove documents with high weights (i.e., with low sensitivity) while keeping documents with low weights (i.e., with high sensitivity). Intuitively, in this case the privacy decreases because it may be easier to find harmful documents in the released database. Figure 1 also shows that the values of SPI were always lower than the corresponding values of JSD, consistent with the observations made in the preceding section.

Turning to the behavior of PEL, we see that it remained nearly stable despite the large variations in the size of released logs. In general, it slightly decreased as k grew, but it occasionally increased. It can be proved that the latter phenomenon happens when we remove queries that are more frequent in the user population than those released. In practice, this situation may be common. Think of a user frequently entering some unpopular query of interest and less frequently a popular query: using most sanitization techniques, the less popular query will have more chances of being removed because it may be more harmful for the user privacy, but this will instead cause an increase of the level of exposure of the user according to PEL.

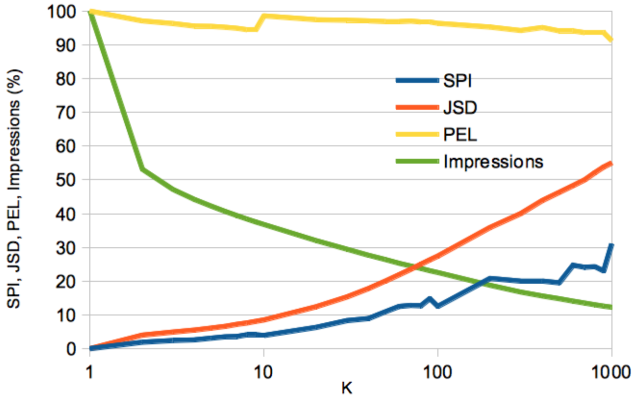

Fig. 1. Sensitive-aware Privacy Index (SPI) versus JSD, PEL, and impressions as a function of the degree of anonymity k, for heavy AOL users. The x axis is logarithmic.

3.3 Evaluating SPI using privacy-specific data

The second round of experiments leverages controlled privacy-related features. We focused on queries containing person or place names or private information (denoted as PPP queries) because this kind of information may be useful to break the user privacy. In order to extract such queries from the original search log we used three lists of proper nouns available on the web; one of about 150,000 surnames, one of 5,000 female and male names, and one of 200,000 populated places. We also collected a vocabulary of sensitive words from a search engine using the Google sensitive ad categories as seed queries and extracting the most informative terms from the search results. The search log associated with the heavy AOL users was partitioned in two groups of queries, those matching our lists of words (i.e., the PPP queries, covering about 50% of the sample) and those that did not (i.e., the remaining no-PPP queries). Then we considered two query selection strategies: releasing only PPP queries and releasing no PPP queries. For each strategy, we generated increasingly larger supersets of queries, letting the number of released queries vary from 10% to 50% of the size of the original search log. We also used a fourth query selection strategy, namely k-anonymization. You may think of it as reverse engineering of the chart in Figure 1. For each value of impressions (in the range from 10% to 50%), we found the corresponding value of k and computed the set of queries associated with it. Finally, for each strategy we computed the SPI values for the sets of released queries having the desired sizes. The results are shown in Figure 2.

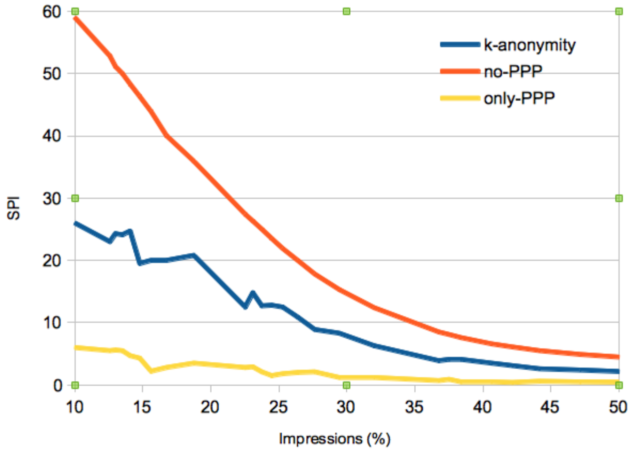

Fig. 2. Sensitive-aware Privacy Index as a function of the percentage of released impressions for three query selection strategies: releasing only queries with person and place names and private information (only-PPP), releasing no-PPP queries, and using k-anonymity.

The no-PPP strategy was a clear winner. Our simple method of removing queries containing profane words, even without considering lexical variations and context, was better than using k-anonymization. This finding confirms that k-anonymization, in general, does not guarantee a good level of privacy protection because there may be relatively popular sensitive queries that are released even for high values of k. The only-PPP strategy was clearly recognized as a bottom line. Our privacy evaluation measure seemed thus to provide reliable and consistent indications.

We also computed the analogous values of PEL. Its behavior was affected neither by the size of released log nor by the type of query selection strategy, thus confirming that PEL is not very suitable as a privacy evaluation measure, at least when the released search log is a strict subset of the original search log. PEL is limited not only by its inability to find and score sensitive queries, but also by the difficulty of estimating the joint (or conditional) probabilities of queries in original and released search logs in Equation 4.

4 Discussion and limitations

One question raised by this approach is the scope of applicability of SPI. We have estimated the probabilities in Equation 1 using frequency counts at the document level (i.e., search queries) for each user. In this way, in addition to requiring that the association between documents and users should be preserved in the sanitized database (which holds for most sanitization methods), we have implicitly assumed that many original documents were left unchanged by sanitization. This assumption is met by various sanitization methods, not only by the k-anonymity family but also by methods based on classification [13] and clustering [14], and by recent forms of differential privacy [10]. However, there are other privacy policies where this assumption would not always hold, e.g., due to systematic document perturbation [8] or generalization [9]. In such cases, it seems that SPI can still be applied provided that we use more flexible methods to compare a user before and after sanitization; e.g., by partial matching at the document level or by exact matching at the word level. This is left for future work.

Another practical difficulty concerns the paucity of annotated natural language datasets for training the classifier of document sensitivity for the type of text data of interest. These datasets are difficult to acquire, although there are recent works that automate this process to some extent; e.g., for Quora posts [15].

A final issue concerns the attack model. As SPI is intended to evaluate the level of protection offered by distinct privacy models, it does not make any specific assumption about the attacker. Its basic tenet is that any sanitization method for text data will result in the suppression or modification of leaked sensitive documents, and it measures the extent to which this has been achieved. This is a generic assumption that holds for most privacy models, usually implicitly but also explicitly; e.g., when an attacker's background knowledge is modeled in terms of machine learning [13] [8], or information retrieval techniques [3].

5 Conclusions

We introduced SPI, a novel privacy index for text data that extends the classical information theoretic approach to include the sensitivity of single documents. First experiments with query logs suggest that its indications are more reliable and consistent than those provided by existing methods. Future research directions include the generalizability of SPI to other types of text data and sanitization methods.

References

1. E. Adar. User 4xxxxx9: Anonymizing query logs. In *WWW Workshop on Query Log Analysis*, 2007.
2. M. Bezzi. An information theoretic approach for privacy metrics. *TDP*, 3:199–215, 2010.
3. J. A. Biega, K. P. Gummadi, I. Mele, D. Milchevski, C. Tryfonopoulos, and G. Weikum. R-Susceptibility: An IR-Centric Approach to Assessing Privacy Risks for Users in Online Communities. In *SIGIR'16*, pages 365–374, 2016.
4. C. Carpineto and G. Romano. K_θ-Affinity Privacy: Releasing Infrequent Query Refinements Safely. *IP&M*, 51:74–88, 2015.
5. C. Carpineto and G. Romano. A Review of Ten Year Research on Query Log Privacy. In *Proceedings of the 7th Italian Information Retrieval Workshop*, 2016.
6. G. Cormode, C. Procopiuc, D. Srivastava, E. Shen, and T. Yu. Empirical privacy and empirical utility of anonymized data. In *29th ICDEW*, pages 77–82, 2013.
7. A. Erola, J. Castella-Roca, A. Viejo, and J. Mateo-Sanz. Exploiting social networks to provide privacy in personalized web search. *J. Syst. Software*, 84(10):1734–1745, 2012.
8. Yi Fang, Archana Godavarthy, and Haibing Lu. A Utility Maximization Framework for Privacy Preservation of User Generated Content. In *ICTIR'16*, pages 281–290, 2016.
9. Y. He and J. F. Naughton. Anonymization of Set Valued Data via Top Down, Local Generalization. In *VLDB'09*, pages 934–945, 2009.
10. Y. Hong, J. Vaidya, H. Lu, and M. Wu. Differentially private search log sanitization with optimal output utility. In *15th EDBT*, pages 50–61, 2012.
11. R. Jones, R. Kumar, B. Pang, and A. Tomkins. 'I know what you did last summer': query logs and user privacy. In *CIKM'07*, pages 909–914, 2007.
12. A. Korolova, K. Kenthapadi, N. Mishra, and A. Ntoulas. Releasing search queries and clicks privately. In *WWW'09*, pages 171–180, 2009.
13. B. Li, Y. Vorobeychik, M. Li, and B. Malin. Iterative Classification for Sanitizing Large-Scale Datasets. In *ICDM'15*, pages 841–846, 2015.
14. G. Navarro-Arribas, V. Torra, A. Erola, and J. Castella-Roca. User k-anonymity for privacy preserving data mining of query logs. *IP&M*, 48(3):476–487, 2012.
15. S. T. Peddinti, A. Korolova, E. Bursztein, and G. Sampemane. Cloak and Swagger: Understanding Data Sensitivity Through the Lens of User Anonymity. In *S&P'14*, pages 493–508, 2014.
16. S. T. Peddinti and N. Saxena. On the Privacy of Web Search Based on Query Obfuscation: A Case Study of TrackMeNot. In *PETS'10*, pages 19–37, 2010.