A Utility Maximization Framework for Privacy Preservation of User Generated Content

Yi Fang
Department of Computer Engineering
Santa Clara University
500 El Camino Real
Santa Clara, CA, 95053, USA
yfang@scu.edu

Archana Godavarthy
Department of Computer Engineering
Santa Clara University
500 El Camino Real
Santa Clara, CA, 95053, USA
agodavarthy@gmail.com

Haibing Lu
Operations Management and Information Systems
Santa Clara University
500 El Camino Real
Santa Clara, CA, 95053, USA
hlu@scu.edu

ABSTRACT

The prodigious amount of user-generated content continues to grow at an enormous rate. While it greatly facilitates the flow of information and ideas among people and communities, it may pose great threat to our individual privacy. In this paper, we demonstrate that the private traits of individuals can be inferred from user-generated content by using text classification techniques. Specifically, we study three private attributes on Twitter users: religion, political leaning, and marital status. The ground truth labels of the private traits can be readily collected from the Twitter bio field. Based on the tweets posted by the users and their corresponding bios, we show that text classification yields a high accuracy of identification of these personal attributes, which poses a great privacy risk on user-generated content.

We further propose a constrained utility maximization framework for preserving user privacy. The goal is to maximize the utility of data when modifying the user-generated content, while degrading the prediction performance of the adversary. The KL divergence is minimized between the prior knowledge about the private attribute and the posterior probability after seeing the user-generated data. Based on this proposed framework, we investigate several specific data sanitization operations for privacy preservation: add, delete, or replace words in the tweets. We derive the exact transformation of the data under each operation. The experiments demonstrate the effectiveness of the proposed framework.

Keywords

Privacy preservation; User generated content

1. INTRODUCTION

Social media has become an indispensable part of many lives. More and more people are using it to create their own content than ever before. According to [9], in 2014, 71% of online adults use Facebook and 28%, 28%, 26%, and 23% for the other social network services LinkedIn, Pinterest, Instagram and Twitter respectively. 52% of online adults now use two or more social media sites, more than half of all online adults 65 and older use Facebook, and some 36% of Twitter users visit the site daily. The prodigious amount of user-generated content continues to grow at an unprecedented rate.

While it greatly facilitates the flow of information and ideas among people and communities, it may pose great threat to our individual privacy. Personal attributes may be derived from analysis of user-generated content, which could yield valid conclusions that the individual would not want to disclose. For example, an anonymous Twitter user may occasionally post tweets, e.g. describing some cosmetic products, praising services at some convenience store, mentioning his/her terrible birthday party. The comment on cosmetic products may suggest that person is female. The convenience store frequently patronized by that user would imply that the user lives in the neighborhood of that store. The posting date of the birthday party comment would suggest his/her date of birth. It was reported in [32] that 87% population in the United States could be uniquely identified based only on 5-digit zip, gender, and date of birth. Thus, the three seemingly irrelevant comments would reveal the identity of the user. Besides identity/anonymity, user-generated content may reveal confidential information that users want to keep to themselves. A user may comment on books that he/she recently read. The books could indicate his/her political opinion. Comments on youth sports center and education quality of local school districts could suggest the commenter is a soccer-mom. Discussion on some medicine could reveal a person's medical history. Posts on some ethnic restaurant could infer national origin of that user. Purchasing an infant bed could suggest this user is an expecting parent. Activities on weekend could reveal the user's religion.

Although many sites offer privacy controls that enable users to restrict how their data is viewed by other users, content analysis of seemingly innocuous user-generated text may reveal much personal information about the users. In this paper, we demonstrate that the private traits of individuals can be inferred from user-generated content by text classification techniques. Specifically, we study three private attributes on Twitter users: religion, political leaning, and marital status. The ground truth labels of the private traits

ICTIR '16, September 12-16, 2016, Newark, DE, USA
© 2016 ACM. ISBN 978-1-4503-4497-5/16/09...$15.00
DOI: http://dx.doi.org/10.1145/2970398.2970417

can be readily collected from the Twitter bio field in an automatic fashion. Based on the tweets posted by the users and their corresponding bios, we show that a simple logistic regression model may yield a high accuracy of identification of these personal attributes, which poses a great privacy risk of user-generated content.

Privacy preservation on user-generated content has multiple challenges. First of all, user-generated content is typically in an unstructured form, i.e., text. Most of the existing privacy-preserving data publishing work has focused on structured data, which can be represented in a relational table that may include identifier (e.g., id) or quasi-identifier (e.g., zip code) attributes. Secondly, it is difficult to devise a good data sanitization strategy. The simplest solution would be deleting all user-generated data, which retains the perfect data privacy. Such a solution renders the data useless and hurts user experience the most. It is indeed against the spirit of the Internet and social media that is to share information. We have to find an effective data sanitization strategy to modify the data in some way such that an adversary cannot infer sensitive information while the sanitized data still carry semantic information and can be used for other legitimate uses. Possible data sanitization operations include deleting words, sentences, or whole records, replacing sensitive words with general words, publishing summary of data instead of complete data, and many others. Different operations would have different impacts on data utility and the amount of information that an adversary can infer. A unified private preservation strategy is needed for handling various data sanitization operations.

To tackle these challenges, we propose a utility maximization framework for preserving user privacy. The goal is to maximize the data utility when modifying the user-generated content, while degrading the prediction performance of the adversary as much as possible. The KL divergence is minimized between the prior knowledge about a private attribute and the posterior probability after seeing the user-generated data. In this way, the data reveals little or no information about the personal attribute. The main contributions of this paper can be summarized as follows:

- We demonstrate that private attributes of Twitter users can be inferred from tweets by using text classification techniques. We utilize the Twitter bio field to collect ground truth labels for training and testing. The experiments show that text classification would yield high accuracy of identification of personal attributes, which poses a great privacy risk on user-generated content.

- We propose a utility maximization framework with the constraint of minimizing the KL divergence between the prior and posterior probabilities of the personal attribute. This is a general framework applicable to various definitions of utilities (e.g., minimal modification, cost-sensitive, personal preference, etc.) and various data sanitization operations.

- Based on the proposed framework, we investigate several specific data sanitization operations for privacy preservation: add, delete, or replace words in the tweets, under the logistic regression model. We derive the exact closed-from data transformation for each operation.

- The experiments are conducted on three personal attributes of Twitter users: religion, political leaning, and marital status. The results demonstrate that the data transformations derived from the proposed framework can effectively degrade the prediction performance of adversaries especially when they are linear models.

2. RELATED WORK

Most existing privacy models are within the area of structured databases [12], also known as micro data, in which released data consist of records with attributes of individuals. A canonical example is census data. Sanitizing micro data by simply removing identifiers, e.g. social security number, cannot prevent privacy inference, because basic demographics can uniquely identify record owners. k-anonymity [33] is the privacy protection model ensuring that any record in the set must be indistinguishable from at least $k-1$ other records in the same set. Although k-anonymity helps to minimize identity disclosure, it may not protect against confidential attribute disclosure. This is the case of a group of k-anonymous records that share the same confidential value (e.g., patients being or not AIDS-positive). In such a case, even though an attacker would not be able to identify a particular individual, he can learn the individual's confidential attribute because all records have the same attribute value. To fix the issue, several privacy protection models are proposed such as differential privacy [10].

While there exist a number of privacy models for structured data, much less attention has been paid to unstructured data (e.g., plain text documents). There are several methods of detecting sensitive information in a text. To find barely semantic identifying data in textual documents (e.g., ID numbers, addresses, ages, dates, etc.), there exist automatic methods that exploit their regular structure in order to detect them by means of rules, patterns, or trained classifiers [21]. As stated in current legislation like HIPAA [22], identifiers should be directly removed/redacted in order to preserve the anonymity. Douglass et al. [8] presents schemes, which consist of techniques such as pattern matching, lexical matching, and heuristics, to identify protected identifiers from free-text nursing notes. Ruch et al. [25] uses a specialized medical semantic lexicon for finding personally identifying information in patient records. Atallah et al. [4] uses an ontological representation of a text document to find and remove sensitive sentences. Chakaravarthy et al. [6] assumes the existence of an external database containing demographic information. Abril et al. [1] use some named entity recognition techniques (e.g. Stanford Named Entity Recognizer [11]) to identify the entities of the documents that require protection. Named identity recognition techniques may not necessarily identify sensitive terms in a text. Staddon et al. [30] proposes a web-based inference detection method. It first extracts salient keywords from the private data, and then issues search queries for documents that match subsets of these keywords within the public web, and finally parses the documents returned by the search queries for keywords not present in the original private data. These additional keywords are used to estimate the likelihood of certain inferences. The idea of using a corpus, like the public web, to detect sensitive information is also found in [26, 27, 23]. Some prior work investigated the privacy issues on Twitter [13, 2].

To protect privacy in textual data, two common methods

Table 1: Bio examples on Twitter

happily married to my wonderful husband and best friend and mother to our wonderful sweet baby boy, Brady. life couldn't be any more sweeter
I love music, my kid's and i'm from Brooklyn, birthday is Feb,28.1978.
I have a brain and I'll use it any way I want. Meanwhile, I wife, mother, knit, read, cook, garden, and work.
Love God, husband, animals, camping, beading, crocheting and friends
married, daughter, christian, hard worker, animals, antiques, travel, cookout

are removing sensitive entities, referred to as redaction, and obfuscation by replacing sensitive pieces with appropriate generalizations (e.g., replacing AIDS by disease) that is also referred to as sanitization. One disadvantage of redaction is the loss of data utility. The other disadvantage is that the existence of blacked-out parts in the released document can raise awareness of the document's sensitivity to potential attackers [5]. It is easier to perform redaction than obfuscation. There are a few obfuscation methods for textual data. Some studies [14, 3] use less sensitive information to replace sensitive information, such as changing marijuana to drug. The Scrub system [31] finds and replaces patterns of identifying information such as name, location, Social Security Number, medical terms, age, date, etc. The replacement strategy is to change the identified word to another word of similar type, and it is not clear whether the semantics of the reports themselves reveal the individuals. Tveit et al. [34] provide a six-step anonymization scheme that finds and replaces identifying words in (Norwegian) patient records with pseudonyms. Saygin et al. [28] propose a two-phase scheme that employs both sanitization and anonymization. There are attempts of adapting the notion of k-anonymity in structured databases to unstructured data, such as k-safety [6], k-confusability ([7]), and k-plausibility [14, 3]. Li et al. [17] performed iterative classification to learn sensitive words by assuming the attacker would choose an optimal classifier from a set of classification models.

In the recent years, SIGIR has hosted workshops on Privacy Preserving IR (PIR) [29, 36]. It aims at exploring and understanding the privacy and security risks in information retrieval. The workshops cover privacy issues in various IR subfields including sentiment analysis [35], document summarization [20], passage retrieval [19], query log analysis [37], and medical records [15].

3. PRIVATE ATTRIBUTE IDENTIFICATION

Given a set of user-generated content for a particular user, we want to know whether we can automatically infer personal sensitive information from the data. In this section, we look at Twitter with one personal attribute that may be regarded as sensitive information for some people: marital status. We formulate the task of identification of marital status as a binary classification problem, i.e., whether a user is married or not given the tweets posted by the user. Thus, it can be viewed as a text classification problem with individual words as features. To obtain a set of training data with known marital status, we utilize the bio box on Twitter where users can give some information about themselves in fewer than 160 characters. Based on our observations, this bio box may contain some personal information such as marital status, age, location, occupation, etc. Table 1 shows some examples of bios that indicate marital status. Some of them also reveal birthday (2nd Tweet) and religion (4th and 5th tweets). The public tweets and user bios can be collected from Twitter Streaming API[1]. We use a list of marital status related keywords (e.g., "wife", "husband", "married", etc.) to match against the bios. We assume that the bios that contain any of these keywords indicate the user is married. Based on our preliminary study, this keyword-matching approach yields very high precision, probably due to the fact that the bios are in a limited length and lack variations in expressions. Section 5.1 presents the details about the data that we collected. We find that over 50% of Twitter users have filled out their bios and about 2% of the bios indicate marital status. Given the huge user base of Twitter, we would expect to obtain a large number of users with marital status by this automatic keyword-matching approach. Similarly, we can collect the negative labels (i,e. unmarried users) using keywords which reveal unmarried status (e.g., "single", "in a relationship", "unmarried", etc.).

After collecting both positive and negative labels along with the tweets posted by users, we can apply text categorization techniques to train a model and predict the marital status of a new user based on her posted tweets. The experiments in Section 6.1 demonstrate that just using conventional text classification models such as logistic regression or Support Vector Machine (SVM) would yield high accuracy of identification of personal attributes. These results would raise severe privacy concerns. The publicly available user-generated content constitutes a large knowledge base that allows an adversary to mine relations between content and personal attributes. Those learned relationships can then be used to infer the personal attributes of users who did not intentionally disclose such information and did not anticipate such disclosure either.

4. A UTILITY MAXIMIZATION FRAMEWORK

4.1 Privacy Model

A privacy model defines what type of privacy is guaranteed, which forms a basis on how privacy protection methods should be designed. In our setting, privacy is the true class (personal attribute) of a user, which might be inferred by some adversary/classifier. The goal of our privacy model is to hide the true user attribute value by degrading the classifier's performance. We argue that a classifier has degraded performance if it does not enhance one's prediction accuracy. In other words, a sanitized text is considered safe if its class prediction result by a classifier is the same as or close enough to random guessing based on the prior knowledge. To formalize the privacy problem, we denote $P(C|D)$ as the trained classifier, which outputs a probability distribution of predicted classes C. D is the original text and $\widetilde{D}$ is the sanitized text . We define the privacy as follows.

Definition 1 (Strong Privacy). *The sanitized document $\widetilde{D}$ is considered strongly private if releasing $\widetilde{D}$ does not improve one's chance in guessing the true class of D than without disclosing D.*

[1] https://dev.twitter.com/streaming/

The above definition is strong in the sense that releasing the sanitized document does not provide any information about the true class of D. If we know what kind of information are correlated with the true class of D, we can simply remove such information and only retain irrelevant information, or perturb D by adding noise to dilute the relevant features. In practice, we can often tolerate certain degree of predictability of the sanitized document as long as the prediction is not of high confidence. Intuitively, we aim to have the posterior probability $P(C|\widetilde{D})$ of the personal attribute C after seeing the sanitized document $\widetilde{D}$ close enough to the prior probability $P(C)$ of the class. In this way, the sanitized document $\widetilde{D}$ does not reveal much information about the attribute C. Mathematically, we can use the KL divergence $KL\big(P(C)||P(C|\widetilde{D})\big)$ to measure the closeness between the two probability distributions. Thus, we give another privacy definition.

Definition 2 (Weak Privacy). *The sanitized document $\widetilde{D}$ is considered private if $KL\big(P(C)||P(C|\widetilde{D})\big) \leq \epsilon$ where ϵ is the privacy tolerance threshold.*

4.2 Sanitization Operations

Sanitization of a document involves modifying the information that may help a classifier perform accurate prediction on the document. Common text sanitization operations can be divided into two categories: suppression and generalization. Suppression is widely used for protecting privacy for structured database. On the other hand, the generalization operation attempts to generalize the key terms so that they are not indicative any more. Different operations would have different impacts on user experience and usefulness of sanitized data. Some utility function can be built to quantify the effects of various types of operations. One can also allow the use of personalized perceptions on sanitized data utility to enable individualized sanitization decision.

4.3 Utility Maximization

While we know what to protect and the tools (operations) used for protection, we need to find out how to effectively use those tools and achieve a balance between privacy and utility. The level of privacy would depend on the ability of adversaries on inferring the true class of a document, while utility can be measured by how much a sanitized text $\widetilde{D}$ differs from the original text D. Mathematically, we propose the following constrained utility maximization objective:

$$\max U(\widetilde{D}, D) \tag{1}$$

$$s.t. \quad KL\big(P(C)||P(C|\widetilde{D})\big) \leq \epsilon \tag{2}$$

where $U(\widetilde{D}, D)$ measures the utility of the sanitized document $\widetilde{D}$ with respect to the original document D. For example, it can be defined as the total number of changes done on the original document D, i.e., $U(\widetilde{D}, D) = -(E + F + G)$ where E, F, G is the number of terms added, deleted, or replaced, respectively. The utility could also be cost-sensitive since some operations may be less desirable than others. For example, some users may not want any terms in the tweets to be deleted, while adding some words may be fine. Thus, a utility function can be defined as $U(\widetilde{D}, D) = -(w_e \times E + w_f \times F + w_g \times G)$ where w_e, w_f, and w_g are the weights for the corresponding operations. To achieve perfect privacy, the extreme solution is to remove the whole text, which is not a desired solution as no data utility is retained. A good privacy protection strategy should allow a trade-off between utility and privacy.

4.4 Privacy Preservation

If we specify in the proposed framework the utility function, prediction model, the privacy threshold, and the type of sanitization operation (e.g., Add/Delete), we can obtain the exact actions that we need to perform (e.g., what words to delete and how many of them to be deleted) by solving the optimization problem in Eqn.(1). In this section, we use logistic regression as the text classification model and show how to preserve user private attributes based on the proposed framework. In practice, it is infeasible to know the prediction model that adversary will use. However, the experiments in Section 6 show that the sanitization operations based on logistic regression can still substantially degrade the performance of other linear prediction models. This may be explained by the fact that the terms identified by logistic regression to be added/deleted/replaced are indicative and predictive of the personal attributes and thus modifying them would also largely affect the results of other classifiers.

We concatenate all the tweets posted by a given user u in a single document represented as a v-dimensional feature vector $u = (f_1, f_2, ..., f_v)$ where v is the vocabulary size. Here, we use TF-IDF weighting scheme $f_j = tf_j \times idf_j$ where tf_j is the term frequency of the j-th word in the vocabulary and idf_j is the inverse document frequency of the word. In logistic regression, the probability of being the positive personal attribute/class given the feature vector is modeled by:

$$P(C = 1|u) = \frac{1}{1 + \exp(-\sum_{j=1}^{v} \beta_j f_j - b)} \tag{3}$$

where β_j are b are the weight and bias parameters respectively, and are learned from the training data.

Let us assume strong privacy by setting the privacy threshold $\epsilon = 0$.

$$\begin{aligned}
& KL(P(C)||P(C|\widetilde{D})) = 0 \\
\Rightarrow \quad & \sum_C P(C) \log \frac{P(C)}{P(C|\widetilde{D})} = 0 \\
\Rightarrow \quad & P(C|\widetilde{D}) = P(C) \\
\Rightarrow \quad & \frac{1}{1 + \exp(-\sum_{j=1}^{v} \beta_j f_j - b)} = P(C) \\
\Rightarrow \quad & \sum_{j=1}^{v} \beta_j f_j + b = \log \frac{P(C)}{1 - P(C)} \qquad (4)
\end{aligned}$$

where $P(C)$ comes from the prior knowledge about the personal attribute. For example, based on the demographics of Twitter users, one can roughly estimate the percentage of married people, or just use $P(C = 1) = 0.5$ if we do not have any prior knowledge. For a trained model, the parameters β_j and b are constant in the equation. Thus, any reasonable data sanitization operation can be boiled down to change f_j so that Eqn.(4) is satisfied and consequently the user data $\widetilde{D}$ does not improve one's chance in guessing the true class C over the prior knowledge. For weak privacy with nonzero ϵ, we can similarly derive the condition to be satisfied but there may exist no simple and closed form as Eqn.(4). This is a general framework applicable to various definitions of utilities (e.g., minimal modification, cost-sensitive, personal preference, etc.), various sanitization operations, and various learning models. In the subsections below, we derive the specific actions for each sanitization operation based on the privacy condition in Eqn.(4).

4.4.1 Delete or Add Operation

Assume we want to delete x_k number of the k-th word in the vocabulary. f_k is the original TF-IDF feature value of word k. After the delete operation, Eqn.(4) must be satisfied. Thus, we have

$$\begin{aligned}
&\beta_k(f_k - x_k \times idf_k) + \sum_{j \neq k}^{v} \beta_j f_j + b = \log \frac{P(C)}{1-P(C)} \\
&\Rightarrow -\beta_k x_k idf_k + \sum_{j=1}^{v} \beta_j f_j + b = \log \frac{P(C)}{1-P(C)} \\
&\Rightarrow -\beta_k idf_k x_k = \log \frac{P(C)}{1-P(C)} - \sum_{j=1}^{v} \beta_j f_j - b \\
&\Rightarrow x_k = \frac{\sum_{j=1}^{v} \beta_j f_j + b - \log \frac{P(C)}{1-P(C)}}{\beta_k idf_k} \qquad (5)
\end{aligned}$$

Eqn.(5) specifies the exact action if we want to add or delete the word k. If $\sum_{j=1}^{v} \beta_j f_j + b \geq \log \frac{P(C)}{1-P(C)}$, x_k is positive and we will delete x_k occurrences of the word. If it is negative, we will add x_k occurrences of the word.

If the utility function is $U(\widetilde{D}, D) = -E$, i.e., minimizing the number of times the word to be added or deleted, we can solve the constrained optimization problem (Eqn.(1)) by calculating x_k for each word k and choose the word that has the smallest x_k, which indicates the minimum changes to the original document. The proposed framework is also applicable to Add/Delete multiple words. Similar formulas can be derived based on Eqn.(4) which specifies the condition when KL divergence is minimized to zero.

4.4.2 Replace Operation

If we want to replace the word k by the word s, we can derive the exact number of such replacements, denoted by y_{ks}, that are needed. Denote the original TF-IDF feature values of the word k and s by f_k and f_s, and the sanitized feature values by $\widetilde{f_k}$ and $\widetilde{f_s}$, respectively, and the corresponding learned weights by β_k and β_s. Based on Eqn.(4), we have

$$\begin{aligned}
&\beta_k \widetilde{f_k} + \beta_s \widetilde{f_s} + \sum_{j \neq k,s}^{v} \beta_j f_j + b = \log \frac{P(C)}{1-P(C)} \\
&\Rightarrow \beta_k(f_k - y_{ks} \times idf_k) + \beta_s(f_s + y_{ks} \times idf_s) \\
&= \log \frac{P(C)}{1-P(C)} - \sum_{j \neq k,s}^{v} \beta_j f_j - b \\
&\Rightarrow (\beta_s idf_s - \beta_k idf_k) y_{ks} = \log \frac{P(C)}{1-P(C)} - \sum_{j=1}^{v} \beta_j f_j - b \\
&\Rightarrow y_{ks} = \frac{\sum_{j=1}^{v} \beta_j f_j + b - \log \frac{P(C)}{1-P(C)}}{\beta_k idf_k - \beta_s idf_s} \qquad (6)
\end{aligned}$$

Based on Eqn.(6), if y_{ks} is a positive number, we will replace the word k by the word s in y_{ks} number of times. If y_{ks} is a negative number, we will replace the word s by the word k in y_{ks} number of occurrences.

4.4.3 More Sophisticated Operations

The above Add, Delete, and Replace are just simple operations to illustrate the usage of the proposed utility maximization framework. More sophisticated sanitization operations can be defined based on the general framework. For example, we may not want to replace word k by any word s. Otherwise, it may lead to grammatical errors or alter the meaning of the text. We can enforce constraints on the data transformation, e.g., requiring the word s has the same part-of-speech with the word k or it is a generalized concept of the work k (referred to as obfuscation as introduced in Section 2). The candidate words to replace can be chosen based on WordNet[2]. In addition, we can further add semantic preservation into the utility function so that the sanitized text retains the semantics of the original text as much as possible. Distributed representation of text such as Doc2Vec [16] can be used to measure the semantic similarity between the texts. This paper is just an initial step to demonstrate the proposed privacy preservation framework. We will study more sophisticated sanitization operations in the future work.

5. EXPERIMENTAL SETUP

5.1 Data Collection and Preprocessing

We used Twitter data as our testbed. We collected the user ids through Twitters Streaming API in the period of December, 2015 to March, 2016. and then identified the user profiles using Tweepy[3] get_user API. In the experiments, we focus on three categories, i.e., political leaning (Democratic vs Republican), religious affiliation (Christian vs non-Christian) and marital status (married vs unmarried). To collect the ground truth labels for users of different categories, we employ the keyword-matching approach as introduced in Section 3 to identify positive and negative users for each category. Table 2 contains the list of the keywords. The tweets were then collected for these users using Tweepy user_timeline() API. We only included the users who posted at least 100 tweets. Table 3 shows the data statistics. We preprocessed these tweets by removing stop words, punctuation, user ids, urls, and non-ascii characters. The top 10,000 terms with the highest TF-IDF values in each category were selected as word features. The whole data was randomly split into 80% for training and the rest 20% for testing.

We use standard evaluation metrics for classification including Precision, Recall, F-score, and Accuracy [18] to determine to what extent the model can correctly classify the instances. We evaluated the results using these metrics on the data before and after the sanitization transformations based on our proposed framework.

5.2 Research Questions

We performed an extensive set of experiments to address the following questions related to the proposed research:

- Can the users' private traits be inferred with high accuracy using off-the-shelf machine learning algorithms like logistic regression? (Section 6.1)
- Can these operations be applied in order to maximize the data utility? (Section 6.2.2)

[2] https://wordnet.princeton.edu/
[3] http://www.tweepy.org/

Table 2: Keywords for matching personal attributes in Twitter user bios

Christian
christian, jesus, christ, church, bible
Non-Christian
islam, hindu, buddhist, sikh, quran, jew, atheist
Democratic
democrat, liberal, left-leaning, progressives
Republican
republican, gop, conservative, right-leaning, pro-life,
pro-gun
Married
marriage, married, wife, husband, dad, mom, father, mother, parent, family
Unmarried
single, i am available, looking for a relationship, dating, boyfriend, girlfriend, bachelor

Table 3: Statistics of the testbed

	Religious	Political	Marital
Positive	1,001	801	3,500
Negative	745	772	3,400
Average # of tweets / user	2,831	2,834	2,861
Average # of words / user	41,672	43,573	40,839

- Given the risk of exposure of the private traits, what kind of measures can be taken to protect the privacy of the user? (Section 6.2.1)
- Can we monitor the user tweet stream to detect sensitive tweets in real time? Would suppressing selected tweets help protect the privacy? (Section 6.3)

6. EXPERIMENTS

In this section, we present both quantitative and qualitative experiments to address the above research questions in Section 5.2.

6.1 Privacy Risk Analysis

We show the effectiveness of a simple off-the-shelf machine learning algorithm like logistic regression in predicting the private traits of a user. With the ground truth data, we train the model and apply it to predict the class labels of the test users. Table 4 contains the prediction results. In general, the model yielded high accuracy on all the three categories with the best results on the religious category. Based on the results in Accuracy, at least more than 70% of the users in each category have their personal attributes correctly identified. Given the fact that we only used a linear model with a modest size of training data, these results raise a significant privacy concern on the user-generated content.

Table 4: Prediction results on the test users by logistic regression

	Precision	Recall	F-Score	Accuracy
Religious	0.938	0.829	0.880	0.835
Political	0.796	0.704	0.747	0.741
Marital	0.746	0.739	0.743	0.726

6.2 Privacy Protection

6.2.1 Utility Maximization

Given the privacy risk exposed in Section 6.1, we apply the proposed privacy preservation framework in Section 4, which is based on maximizing the utility of user data given a transformation. Here we define the utility as the negative amount of change we make on the user data. For example, deleting or adding m words will yield the utility of $-m$. In other words, the objective is to minimize the amount of change on the data. As shown in Section 4, in order to preserve user privacy, we minimize the KL divergence between prior knowledge about the personal attribute and the posterior probability as the constraint. The prior probabilities are assumed to be 0.5.

Specifically, we investigate the Add/Delete operation and the Replace operation introduced in Section 4.4.1 and 4.4.2. For each operation, we have two variants: *Random* and *Minimum*. For Add/Delete, *Random* is to randomly pick a term in the user tweets to add/delete and then solve Eqn.(5) to obtain the number of the term occurrences to add/delete. On the other hand, the *Minimum* operation is to pick the term that causes the minimum number of changes on the tweets, by solving Eqn.(5) for each vocabulary term and finding the minimum of x_k. Similarly *Random* and *Minimum* procedures are defined for the Replace operation based on Eqn.(6). In other words, the *Random* procedure only minimizes the KL divergence without considering data utility, while the *minimum* procedure applies the full proposed utility maximization framework.

Table 5 contains the results of the average changes in term frequency over all the users for different operations. As we can see, the average changes in term frequency for the *Minimum* procedure are significantly smaller than for the *Random* procedure across all the three categories. These results demonstrate that the data utility preserved by the proposed utility maximization framework is much higher under the *Minimum* procedure. The average changes for religious and political categories are comparable while they are much less for the marital category. This may be due to the different numbers of training instances in different categories. Moreover, we can find that the frequency change for the Replace operation is generally smaller than for Add/Delete, which may be explained by the involvement of two terms in the Replace operation while only one term in Add/Delete.

Table 5: Average changes in term frequency over the users under different sanitization operations for the three categories.

	Add/Delete Random	Add/Delete Minimum	Replace Random	Replace Minimum
Religious	32,825	85	8,834	39
Political	2,251	85	2,767	38
Marital	2,719	19	10,204	8

6.2.2 Privacy Preservation

Table 6 shows the evaluation metrics on various classifiers before (Pre) and after (Post) applying the proposed sanitization operations. In these experiments, we only look at the users whose attributes were correctly identified by the logistic regression model, since these users are at high privacy risk. As a result, the Pre-transformation metrics for the

Table 6: Performance of various classifiers before (Pre) and after (Post) applying the privacy preservation operations Add/Delete *Minimum* (A/D-Min) and Replace *Minimum* (Rep-Min)

Category	Classifier	Operation		Precision	Recall	F-Score	Accuracy
Religious	Logistic	Pre		1.0	1.0	1.0	1.0
		Post	A/D-Min	0.698	0.413	0.519	0.444
			Rep-Min	0.753	0.446	0.560	0.492
	Linear-SVM	Pre		0.987	0.949	0.968	0.954
		Post	A/D-Min	0.325	0.163	0.217	0.146
			Rep-Min	0.528	0.300	0.383	0.298
	KNN	Pre		0.914	0.889	0.901	0.859
		Post	A/D-Min	0.911	0.886	0.898	0.855
			Rep-Min	0.910	0.877	0.893	0.848
	Nonlinear-SVM	Pre		0.968	0.928	0.948	0.926
		Post	A/D-Min	0.959	0.916	0.937	0.911
			Rep-Min	0.959	0.916	0.937	0.911
Political	Logistic	Pre		1.0	1.0	1.0	1.0
		Post	A/D-Min	0.561	0.457	0.504	0.536
			Rep-Min	0.480	0.387	0.429	0.468
	Linear-SVM	Pre		0.924	0.945	0.934	0.932
		Post	A/D-Min	0.403	0.403	0.403	0.384
			Rep-Min	0.631	0.705	0.666	0.636
	KNN	Pre		0.774	0.744	0.758	0.756
		Post	A/D-Min	0.75	0.744	0.747	0.74
			Rep-Min	0.761	0.744	0.752	0.748
	Nonlinear-SVM	Pre		0.887	0.852	0.869	0.868
		Post	A/D-Min	0.755	0.790	0.772	0.76
			Rep-Min	0.863	0.837	0.850	0.848
Marital	Logistic	Pre		1.0	1.0	1.0	1.0
		Post	A/D-Min	0.582	0.538	0.559	0.539
			Rep-Min	0.610	0.601	0.606	0.574
	Linear-SVM	Pre		0.935	0.901	0.917	0.912
		Post	A/D-Min	0.562	0.528	0.545	0.520
			Rep-Min	0.631	0.705	0.666	0.636
	KNN	Pre		0.780	0.781	0.780	0.761
		Post	A/D-Min	0.771	0.776	0.773	0.753
			Rep-Min	0.776	0.776	0.776	0.756
	Nonlinear-SVM	Pre		0.809	0.833	0.821	0.803
		Post	A/D-Min	0.697	0.760	0.728	0.690
			Rep-Min	0.756	0.762	0.759	0.737

logistic regression model are all 1.0. We apply both linear and nonlinear classifiers including logistic regression (LR), linear SVM, nonlinear SVM (with quadratic kernel), and K-Nearest Neighborhood (KNN) (where $K = 20$) to test the effect of the proposed transformations. We utilize the Scikit-learn machine learning library [24] for these classifiers. We have a number of observations from the table.

First of all, the proposed sanitization operations reduce the predictive performance of all the classifiers across all the metrics. The transformations are much more effective on the linear classifiers (LR and Linear-SVM) than the nonlinear ones (nonlinear SVM and KNN) across all the categories. For logistic regression, the accuracy has reduced to be around 0.5 for all the three categories, which is close to random guessing. This is expected since the transformations are based on logistic regression. However, Linear-SVM also has much degraded predictive performance after the data transformations. These results show that even when the adversary uses a different or unknown prediction model, we may still be able to preserve user privacy based on the proposed sanitization operations.

Secondly, the nonlinear classifiers are affected to a much lesser extent, which may be due to the fact that the transformations are based on a linear model. A noticeable performance drop can be observed on Nonlinear-SVM in the marital category. This may be explained by the fact that this category has much more training data than the other two categories.

Thirdly, for the linear classifiers, the performance seems to be affected more by the Add/Delete operation than the Replace operation. However, as shown in Table 5, this comes at the expense of doing more changes to the user generated data. As we can see in the table, the Replace *Min* operation only caused half of the changes than the Add/Delete operation did. Therefore, the data utility is much higher in Replace *Min*. Considering the fact that the performance drops based on these two operations are quite similar, we may prefer the Replace *Min* operation.

The above observations demonstrate the effectiveness of our proposed framework in dampening the prediction accuracy and thus retaining the privacy of the users, especially when the adversary uses a linear model. In the future work, we will study nonlinear prediction models under the proposed framework and derive the corresponding data transformation operations.

6.2.3 Term Sanitization

To gain a further insight into the proposed utility maximization framework, we take a closer look at the specific terms that are selected by the proposed Add/Delete operation for data sanitization. Table 7 includes the probabilities on the personal attributes before (Pre) and after (Post) the Add/Delete *Minimum* operation for some example users. We can see that the terms picked by the proposed approach are quite indicative of the corresponding personal attributes. For example, the term *dnc2012* (i.e., 2012 Democratic National Convention) was identified for a user who is predicted to be democratic. A total of 19 *dnc2012* occurrences are deleted so that the Post transformation probability is decreased to 0.5, which was set as the prior probability on the attributes. When the model predicts a user to be Republican, the term *conservative* is deleted for 18 occurrences to keep the tweets less indicative. Similarly for the user predicted to be married, the proposed approach adds 10 occurrences of *bffs* to confuse the classifier. In sum, the terms identified by the proposed approach under logistic regression seem quite important and predictive of the user attributes. This explains why these operations also affect other prediction models, which makes privacy preservation possible even when the adversary model is unknown.

Table 7: Attribute probabilities before (Pre) and after (Post) the Add/Delete *Minimum* operation for some example users

User Attribute	Term	Freq change	Pre	Post
Democratic	*dnc2012*	19	[0.054,0.946]	[0.5,0.5]
Republican	*conservative*	18	[0.903,0.097]	[0.5,0.5]
Christian	*vote*	-21	[0.002,0.998]	[0.5,0.5]
Christian	*praying*	742	[0.000,1.000]	[0.5,0.5]
Married	*bffs*	-10	[0.004,0.996]	[0.5,0.5]
Married	*agreement*	18	[0.003,0.997]	[0.5,0.5]

6.2.4 ROC Curves

To further investigate the effect of the proposed operations for privacy preservation, we plot in Figure 1 the ROC (Receiver Operating Characteristic) curves of different classification results for users at high privacy risk (whose personal attributes have been correctly identified in Section 6.1) in three categories. Three classification models are compared including logistic regression (LR), linear SVM (LNR-SVM), and SVM with quadratic kernel (SVM-Poly). Pre and Post transformations are investigated. Two data sanitization operations are shown: Add/Delete (A) and Replace (R), with the *Minimum* procedure (minimizing the amount of changes). Figure 2 shows the similar ROC curves for the entire test users. We can see from the graphs that the Post ROC curve of logistic regression nearly lies on the random-prediction line, which indicates the predictions after the transformations are very poor. Moreover, Linear-SVM clearly responds to the transformations. It is worth noting that Linear-SVM shows different patterns for different categories. For the religious category, both Add/Delete and Replace operations move the predictions to the other side of the random-prediction line. For the political category, they are on either side of the line. For the marital category, both operations are mostly around the random-prediction line. On the other hand, the nonlinear models are not much affected by the data transformations. These observations are generally consistent in the two figures.

6.3 Monitoring Indicative Tweets

Tweets appear in a streaming fashion and sensitive tweets may emerge unexpectedly. We can develop a temporal based privacy preservation strategy by monitoring the real-time stream of user tweets. In the previous section, the methods were based on changing the frequency of words in tweets to disguise user private traits. These operations may change the syntactic structure of the sentences. This may not be as a serious issue as that in regular documents since tweets usually do not follow very strict syntactic rules. An alternative operation is to suppress an entire new tweet that might expose too much personal information when combined with the prior tweets. The prediction model can compute on the regular basis the KL divergence $KL\big(P(C)||P(C|\widetilde{D})\big)$ between the prior probability and posterior probability of the user attribute based on the currently observed tweets. If the KL divergence exceeds a certain threshold, it would generate an alarm for raising a privacy concern for the user.

We randomly select one user from each category for case studies. Figure 3 shows how the predicted probability of the correct attribute changes over time before and after suppressing the new tweets. As we can see for the religious category, there are spikes around the tweets related to Christianity, e.g., "*Love my CG! God is doing cool stuff!*" and "*8:36am and my day has already been made by a simple text from a friend who experienced God's grace in a real way. Thank you Jesus for grace.*" Here the terms like *Love*, *God*, *Jesus*, and *grace* are all related to Christianity. The prediction accuracy spiked around the appearance of these tweets. When we detect such tweets and suppress them, it lowers the prediction accuracy, and thus preserves privacy. Similarly in the political category, there were spikes at the beginning related to tweets such as *"RT @GKMTNtwits: MSM Misinforms/Polls/Reports“ outcomes. @AymanM ”...See Hillary “Liar” Poll/A Joke*" and "*@DWStweets needs to be stripped of the DNC Chair title then voted out of office aTRAITOR ;Pro-Israel first over USA*". The terms like *Hillary*, *DNC*, *Pro-Israel* are related to Democrats. When we suppress them, it degrades the prediction. There might be more tweets related to Democrats, but when we suppress the tweet early, it may cause future prediction ineffective, until the appearance of some strongly indicative tweets. For the marital category, although it started out with low prediction accuracy, it slowly gained accuracy around the tweets "*Children don't stress they only want to laugh - Children don't stress they just want to have fun and laugh...*" and "*10 Seconds That Will Change Your Life - 10 Seconds That Will Change Your Life. For most people we never...*". The terms like *Children*, *stress*, *fun*, and *laugh* are associated with marriage, although they are not as strong indicators as in the previous two cases. Therefore, the prediction accuracy is just slightly above 0.5 in this case. By conducting the proposed operation in an online fashion, we can monitor user tweets in real time and generate alarms so that users are aware of the privacy risk associated with their posts.

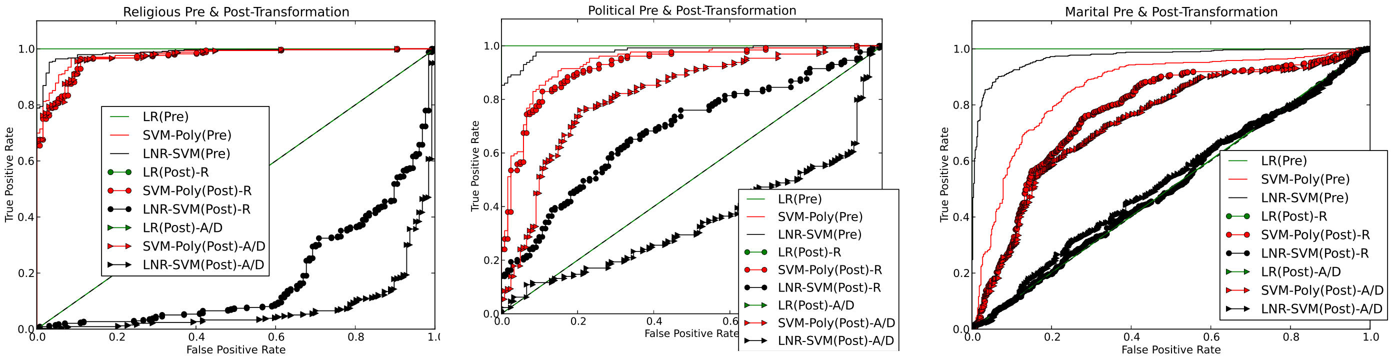


Figure 1: ROC curves of various classifiers for the test users who are at privacy risk

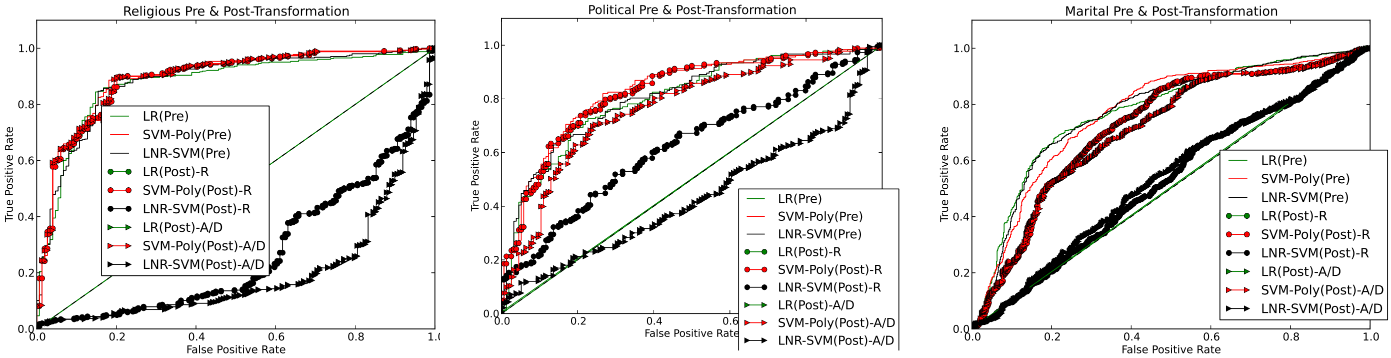
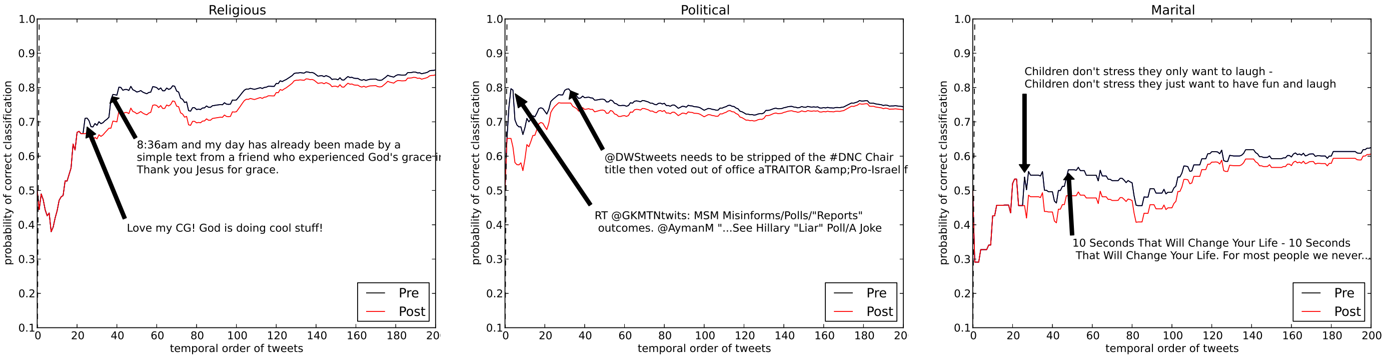


Figure 2: ROC curves of various classifiers for all the test users

Figure 3: Temporal detection of sensitive tweets



7. CONCLUSION AND FUTURE WORK

We demonstrate that private attributes of Twitter users can be accurately inferred by simple text categorization with the labels automatically extracted from the bio fields. We propose a general utility maximization framework to preserve user privacy while maximizing data utility. Specific data sanitization transformations are derived based on the framework.

This work is an initial step towards a promising research direction, as there exists few work on privacy preservation of user-generated content. First of all, the simple sanitization operations presented in this paper may lead to grammatical errors or hamper the meaning of the user tweets. In the future work, we plan to explore semantic preserving operations to produce meaningful tweets in practice. This may require to enforce constraints on the transformation and proper definition of utility function as discussed in Section 4.4.3. Secondly, we will make more relaxed assumptions about the adversary by allowing uncertainties, which can be be modeled by probability distributions over a wide range of models and features that the attacker might use. We can extend the proposed utility maximization framework with probabilistic inference. Last but not the least, we will conduct more comprehensive experiments to evaluate the proposed approach.

8. REFERENCES

[1] D. Abril, G. Navarro-Arribas, and V. Torra. On the declassification of confidential documents. In *MDAI*. 2011.

[2] H. Almuhimedi, S. Wilson, B. Liu, N. Sadeh, and A. Acquisti. Tweets are forever: a large-scale quantitative analysis of deleted tweets. In *CSCW*, 2013.

[3] B. Anandan, C. Clifton, W. Jiang, M. Murugesan, P. Pastrana-Camacho, and L. Si. t-plausibility: Generalizing words to desensitize text. *Transactions on Data Privacy*, 2012.

[4] M. J. Atallah, C. J. McDonough, V. Raskin, and S. Nirenburg. Natural language processing for information assurance and security: an overview and implementations. In *NSPW*, 2001.

[5] E. Bier, R. Chow, P. Golle, T. H. King, and J. Staddon. The rules of redaction: Identify, protect, review (and repeat). *W2SP*, 2009.

[6] V. T. Chakaravarthy, H. Gupta, P. Roy, and M. K. Mohania. Efficient techniques for document sanitization. In *CIKM*, 2008.

[7] C. M. Cumby and R. Ghani. A machine learning based system for semi-automatically redacting documents. In *IAAI*, 2011.

[8] M. Douglass, G. Cliffford, A. Reisner, W. Long, G. Moody, and R. Mark. De-identification algorithm for free-text nursing notes. In *CINC*, 2005.

[9] M. Duggan, N. B. Ellison, C. Lampe, A. Lenhart, and M. Madden. Social media update 2014. *Pew Research Center*, 2015.

[10] C. Dwork. Differential privacy. In *ICALP*, 2006.

[11] J. R. Finkel, T. Grenager, and C. Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *ACL*, 2005.

[12] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu. Privacy-preserving data publishing: A survey of recent developments. *ACM Computing Survey*, 2010.

[13] L. Humphreys, P. Gill, and B. Krishnamurthy. How much is too much? privacy issues on twitter. In *Conference of International Communication Association*, 2010.

[14] W. Jiang, M. Murugesan, C. Clifton, and L. Si. t-plausibility: Semantic preserving text sanitization. In *ICSE*, 2009.

[15] N. Kazantsev, D. Korolev, D. Torshin, and A. Mikhailova. An approach to automate health monitoring in compliance with personal privacy. In *Smart Health*. Springer, 2015.

[16] Q. V. Le and T. Mikolov. Distributed representations of sentences and documents. In *ICML*, 2014.

[17] B. Li, Y. Vorobeychik, M. Li, and B. Malin. Iterative classification for sanitizing large-scale datasets. In *ICDM*, 2015.

[18] C. D. Manning, P. Raghavan, H. Schütze, et al. *Introduction to information retrieval*. Cambridge university press, 2008.

[19] L. Marujo, J. Portêlo, D. M. De Matos, J. P. Neto, A. Gershman, J. Carbonell, I. Trancoso, and B. Raj. Privacy-preserving important passage retrieval. In *SIGIR Workshop on PIR*, 2014.

[20] L. Marujo, J. Portêlo, W. Ling, D. M. de Matos, J. P. Neto, A. Gershman, J. Carbonell, I. Trancoso, and B. Raj. Privacy-preserving multi-document summarization. In *SIGIR Workshop on PIR*, 2015.

[21] S. M. Meystre, F. J. Friedlin, B. R. South, S. Shen, and M. H. Samore. Automatic de-identification of textual documents in the electronic health record: a review of recent research. *BMC Medical Research Methodology*, 10, 2010.

[22] U. D. of Health, H. Services, et al. Summary of the hipaa privacy rule. *HHS*, 2003.

[23] S. T. Peddinti, A. Korolova, E. Bursztein, and G. Sampemane. Cloak and swagger: Understanding data sensitivity through the lens of user anonymity. In *IEEE Symposium on Security and Privacy*, 2014.

[24] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn: Machine learning in python. *JMLR*, 12, 2011.

[25] P. Ruch, R. H. Baud, A.-M. Rassinoux, P. Bouillon, and G. Robert. Medical document anonymization with a semantic lexicon. In *AMIA Symposium*, 2000.

[26] D. Sánchez, M. Batet, and A. Viejo. Detecting sensitive information from textual documents: an information-theoretic approach. In *MDAI*. 2012.

[27] D. Sánchez, M. Batet, and A. Viejo. Minimizing the disclosure risk of semantic correlations in document sanitization. *Information Sciences*, 2013.

[28] Y. Saygin, D. Hakkani-Tur, and G. Tur. Sanitization and anonymization of document repositories. *Web and information security*, 2005.

[29] L. Si and H. Yang. Pir 2014 the first international workshop on privacy-preserving ir: When information retrieval meets privacy and security. In *ACM SIGIR Forum*, 2014.

[30] J. Staddon, P. Golle, and B. Zimny. Web-based inference detection. In *USENIX Security*, 2007.

[31] L. Sweeney. Replacing personally-identifying information in medical records, the scrub system. In *AMIA*, 1996.

[32] L. Sweeney. Simple demographics often identify people uniquely. *Health*, 2000.

[33] L. Sweeney. k-anonymity: A model for protecting privacy. *IJUFKS*, 2002.

[34] A. Tveit, O. Edsberg, T. Rost, A. Faxvaag, O. Nytro, M. Nordgard, M. T. Ranang, and A. Grimsmo. Anonymization of general practioner medical records. In *HelsIT*, 2004.

[35] S. S. Woo and H. Manjunatha. Empirical data analysis on user privacy and sentiment in personal blogs. In *SIGIR Workshop on PIR*, 2015.

[36] H. Yang and I. Soboroff. Privacy-preserving ir 2015: When information retrieval meets privacy and security. In *SIGIR Workshop on PIR*, 2015.

[37] S. Zhang, H. Yang, and L. Singh. Applying epsilon-differential private query log releasing scheme to document retrieval. 2015.