

On Privacy-Preservation of Text and Sparse Binary Data with Sketches

Charu C. Aggarwal*

Philip S. Yu†

Abstract

In recent years, privacy preserving data mining has become very important because of the proliferation of large amounts of data on the internet. Many data sets are inherently high dimensional, which are challenging to different privacy preservation algorithms. However, some domains of such data sets also have some special properties which make the use of sketch based techniques particularly useful. In this paper, we present a new method for privacy preserving data mining of text and binary data with the use of a sketch based approach. The special properties of such data sets which are exploited are that of *sparsity*; according to this property, only a small percentage of the attributes have non-zero values. We formalize an anonymity model for the sketch based approach, and utilize it in order to construct sketch based privacy preserving representations of the original data. This representation allows accurate computation of a number of important data mining primitives such as the dot product. Therefore, it can be used for a variety of data mining algorithms such as clustering and classification. We illustrate the effectiveness of our approach on a number of real and synthetic data sets. We show that the accuracy of data mining algorithms is preserved by the transformation even in the presence of increasing data dimensionality.

Keywords: Privacy, Sketches

1 Introduction

Privacy preserving data mining has been an important problem in recent years because of many new kinds of technology which facilitate the collection of different kinds of data. Such large collections of data have lead to an increasing need to develop methods to protect the privacy of the underlying data records. As a result, a considerable amount of research has been focussed on the problem in recent years [4, 6, 21]. However, most of these techniques are designed for the case of quantitative and categorical data. In this paper, we will discuss new methods for text and binary data with the use of a sketch-based approach. Our algorithms will work effectively for any kind of data which is sparse, and has very few attributes with non-null values.

The problem of privacy preserving data mining has been studied in the context of a number of different approaches in the literature. Some important methods for privacy preserving data mining include those of perturbation and k -anonymity. These are as follows:

- In perturbation, we add noise to the original data. Reconstructed distributions are then used to perform the privacy preserving data mining [4, 6, 11]. The perturbation method has the advantage that it can be performed at data collection time. This is because the behavior of the other records does not need to be used during the perturbation process.
- In k -anonymity, we transform the data such that a given record cannot be distinguished from at least k other records in the data [17, 21]. While this approach provides greater privacy in the presence of public information, it uses the distribution of other records, and needs the presence of all records on a single trusted server.

Both techniques are useful for different scenarios of privacy, though both techniques work well for lower dimensional data, and are not very effective for the high dimensional case [2]. This is because of the following reasons:

- In the high dimensional case, the concept of locality becomes ill-defined. Since the concept of k -anonymity depends deeply upon locality, it is no longer possible to anonymize the data without losing an unacceptable amount of information. Furthermore, with an increasing number of attributes the problem of k -anonymity becomes increasingly difficult. Since it has been shown that this problem is NP-hard [17], it also becomes impractical to anonymize the data.
- In the method of perturbation [6], it is possible to compute maximum likelihood estimates [3] for a record matching the public database. With increasing dimensionality, these estimates become increasingly accurate, and therefore privacy is lost.

We note that the problem of high dimensionality is a fundamental one for privacy purposes, and it cannot

*IBM T. J. Watson Research Center, charu@us.ibm.com

†IBM T. J. Watson Research Center, psyu@us.ibm.com

be solved by using more effective models and algorithms. Nevertheless, many high dimensional data sets have special structure which can be exploited in order to obtain more effective solutions. In this paper, we will study the domain of text and market basket data sets which are high dimensional, but are also *sparse*. By sparsity, we refer to the property that most of the attributes are zero, and only a few attributes take on non-zero values. For this special case, it is possible to design algorithms which work effectively under a variety of circumstances. Specifically, we will use the method of sketches in order to perform the anonymization. These methods work well since the final sketch based representation is defined only by the non-zero elements in the record. Therefore, they retain their effectiveness for the high dimensional case, as long as the number of non-zero elements is small.

Some recent work discusses techniques for privacy via pseudo-random sketches [18]. The method in [18] is designed specifically the problem of query resolution in quantitative data sets. However, the focus of our work is very different and it is designed as a solution to the high dimensional case in which standard methods such as k -anonymity and perturbation do not work very effectively. Furthermore, our primary focus is in the domain of text and categorical data, and in the context of general data mining problems. We will present sketch-based methods which are analogous to the different scenarios of perturbation and k -anonymity:

- We will present algorithms which rely on absolute levels of perturbation using the sketch based approach. These algorithms do not typically need to use other records in order to perform the anonymization. Therefore, as in [4], these methods can be performed at data collection time.
- We will present algorithms which are sensitive to local data distributions in order to perform the transformation. For example, outlier records require greater anonymization, since it can be easily distinguished from the other records in the data. These techniques use other records to perform the transformation, and therefore perturb the records in a way which is sensitive to the underlying data distribution.

We will use an adaptation of the AMS sketch [8] in order to perform the privacy preserving transformation. We will present algorithms to perform both kinds of anonymization. We will also apply these anonymized techniques to a number of different kinds of data sets and algorithms, and show that the perturbed data does retain its utility after anonymization. The sketch

based approach is also extremely efficient because of the simplicity in implementation.

This paper is organized as follows. The remainder of this section contains a discussion on related work on privacy preserving data mining. In the next section, we will discuss a model for sketch based anonymization. In section 3, we will discuss the implementation of the anonymization process, and its use for data mining algorithms. The experimental results are contained in section 4. Section 5 contains the conclusions and summary.

1.1 Related Work The privacy preserving data mining topic has been studied extensively in recent years because of recent focus on the topic of anonymization and privacy. A comprehensive survey of privacy methods may be found in [23]. The most widely researched classes of problems on the topic are as follows:

- **Randomization** The method of randomization adds additive or multiplicative noise to the data in order to hide the attribute values of the records [4, 6, 11, 20]. Modifications of data mining algorithms are applied to the randomized data in order to obtain the most effective results. Some examples of such data mining problems are those of classification [4, 6] and association rules.
- **Methods for k -anonymity:** A key point about privacy preserving data mining algorithms is that the underlying data can be combined with information available in public records in order to reveal sensitive information about the data. In k -anonymity based methods, the data is perturbed in a careful way so as to minimize the likelihood of discovery of identities of the underlying records [2, 9, 17, 21]. Other related techniques [1] include the use of pseudo-data for k -anonymization. In [9], a number of heuristics have been proposed for effective k -anonymization, though the problem of obtaining an optimal solution has been shown to be NP-hard in [17].
- **Distributed Privacy Preserving Data Mining:** In many applications, it is desirable for multiple entities to share records across horizontally or vertically partitioned data sets [12, 22]. In such cases, the individual entities may not wish to share the individual records, but they are interested in the aggregate mining results across the union of the data sets across these entities. In such cases, comprehensive protocols need to be designed for distributed privacy preserving data mining. A description of such methods may be found in [13].

- **Cryptographic Methods:** In cryptographic methods, we use functional transformations of the data which are difficult to invert without impractical computational or other costs. Many of these methods intersect with the distributed mining case, since they require the computation of functions with secure multi-party protocols. So the privacy of the data continues to be preserved from a practical point of view. A survey on cryptographic methods has been discussed in [19].

While the k -anonymity technique is generally quite effective, it has been shown to have some shortcomings in some special cases. For this reason, a technique called l -diversity was recently proposed in [16]. Some recent methods [14] also inject utility into the problem of privacy preserving data mining. These techniques try to design privacy preserving algorithms in such a way so as maximize the utility of the data from the point of view of the underlying mining algorithms.

In general, many of these techniques work quite well for the lower dimensional case, but are not specifically designed for high dimensional data. This is because of the curse of dimensionality which makes high dimensional data very difficult to handle. For the case of text data, a method has been proposed to perform the specific task of privacy preserving indexing of documents over a network [10]. This is different from the general problem of anonymizing a given data set for an arbitrary application which may be different from indexing. In the next section, we will discuss a sketch based anonymization algorithm for sparse data, which works well for the high dimensional case.

2 Sketch Based Anonymization of Sparse Data

In this paper, we define sparse data as one in which each record contains only a small number of non-zero values. Many domains of data such as text and transactional data satisfy this property. For example, a text document may contain only a few words, but may be drawn from a base lexicon of more than a hundred thousand words. The same is true of a market basket transaction drawn from items selling in a supermarket.

Before describing the algorithms further, we will first introduce some notations and definitions. We will assume that the database \mathcal{D} contains N records, each of which contains d dimensions. Each record \bar{X} in \mathcal{D} is denoted by $\bar{X} = (x_1 \dots x_d)$. In this case, we assume that $x_i \neq 0$ for only l different values of i , where $l \ll d$. Furthermore, d is typically quite large and its magnitude may range in the thousands, whereas l is rarely larger than a few hundred.

The *a sketch* of the record $(x_1 \dots x_d)$ is defined by

the quantity s^j such that:

$$(2.1) \quad s^j = \sum_{i=1}^d x_i \cdot r_i^j$$

Here, the random variable r_i^j is drawn from $\{-1, +1\}$ with a mean of 0, and is generated from a pseudo-random number generator [8] which produces 4-wise independent values for the variable r_i^j . Different values of j provide different instantiations for the random variable, and therefore different components s^j of the sketch $S = (s^1 \dots s^r)$. In general, the record \bar{X} can be re-constructed only approximately from the sketch. This approximation provides the privacy for that record. The larger the number of components r , the better the re-construction, but the lower the privacy. We also note that the sketch value s^j is defined only by the non-zero components of the record. Therefore, the noise in the sketch representation is primarily governed by the number of such non-zero components. This helps in preserving the effectiveness of the sketch based approach for the purpose of distance computations, while preserving the privacy at the attribute level.

We note that the value of x_k can be reconstructed by using the sketch derivative $E^k = s^j \cdot r_k^j$. This can be shown using the pairwise independence of different values of r_i^j , the fact that the square of r_k^j is always 1, and the fact that $E[r_i^j] = 0$. Therefore, we have:

$$(2.2) \quad E[E^k] = x_k$$

A key issue here is the variance of the estimation of the different values of x_k . This is because this variance defines the level of accuracy of the sketch representation. We compute the variance of E^k as follows:

$$(2.3) \quad \text{var}(E^k) = \text{var}(s^j \cdot r_k^j)$$

$$(2.4) \quad = \text{var}\left(\sum_{i=1}^d (x_i \cdot r_i^j) \cdot r_k^j\right)$$

$$(2.5)$$

The above expression simplifies to the following:

$$\begin{aligned} & E\left[\left(\sum_{i=1}^d (x_i \cdot r_i^j) \cdot r_k^j\right)^2\right] - E\left[\sum_{i=1}^d (x_i \cdot r_i^j) \cdot r_k^j\right]^2 \\ & = E\left[\left(\sum_{i=1}^d (x_i \cdot r_i^j)\right)^2\right] - x_k^2 = \sum_{i=1}^d x_i^2 - x_k^2 \end{aligned}$$

We note that the above-mentioned variance is dependent only on the non-null attributes in the data. A key function for many data mining algorithms is that of dot product computation. This is because dot products can

be used to compute the distances between records. Let $S = (s^1 \dots s^r)$ be one set of sketches from a given record \bar{X} , and let $T = (t^1 \dots t^r)$ be another set of sketches from a different record \bar{Y} . Then the expected dot product of \bar{X} and \bar{Y} is given by the following:

$$(2.6) \quad E[\bar{X} \cdot \bar{Y}] = \sum_{j=1}^r s^j \cdot t^j / r$$

As in the previous case, it is useful to compute the variance of the dot product. First, we will compute the variance of each component $s^j \cdot t^j$. Therefore, we have:

$$\begin{aligned} \text{var}(s^j \cdot t^j) &= \\ &= \text{var}\left(\left(\sum_{i=1}^d (x_i \cdot r_i^j)\right) \cdot \left(\sum_{i=1}^d (y_i \cdot r_i^j)\right)\right) \\ &= \sum_{i=1}^d \sum_{l=1}^d x_i^2 \cdot y_l^2 - \left(\sum_{i=1}^d x_i \cdot y_i\right)^2 \end{aligned}$$

Since each value of j defines an independent instantiation of the sketch derivative, it is possible to reduce the variance by averaging the different sketch derivatives $s^j \cdot t^j$. Specifically, the variance can be reduced by a factor of r (and standard deviation by \sqrt{r}) by averaging the sketch derivative over r independent instantiations. Therefore, we have:

$$\begin{aligned} \text{var}\left(\sum_{j=1}^r s^j \cdot t^j / r\right) &= \\ &= \left(\sum_{i=1}^d \sum_{l=1}^d x_i^2 \cdot y_l^2 - \left(\sum_{i=1}^d x_i \cdot y_i\right)^2\right) / r \end{aligned}$$

By varying the number of components in the different sketch derivatives, it is possible to increase or decrease the level of anonymity. In general, the anonymous representation will comprise a sketch for each record in the data. However, the number of components for each sketch can vary across different records, and is carefully controlled so as to provide a uniform measure of anonymity across different records. We note that in order to compute functions of two or more records, we need to use the minimum number of sketch components from the set of multiple records.

In general, let us assume that the database \mathcal{D} contains N records which are denoted by $\bar{X}^1 \dots \bar{X}^N$. We assume that the number of sketch components are defined by $m_1 \dots m_N$. In order to decide how the number of such sketch components are determined, we first need to define the privacy level. Specifically, we define the concept of δ -anonymity:

DEFINITION 1. A sketch-based randomization with r components is defined to be δ -anonymous, if the variance of the reconstruction of each attribute is larger than δ , when a total of r sketch components is used. Therefore we have:

$$(2.7) \quad \left(\sum_{i=1}^d x_i^2 - x_k^2\right) / r \geq \delta \quad \forall k \in \{1 \dots d\}$$

We note that for a given value of δ , it may not always be possible to construct a δ -anonymous representation. For example, let us consider the particular case in which attribute x_i is zero, except for one attribute which takes on a value less than $\gamma \ll \delta$. In such a case, a δ -anonymous representation for the record does not exist. In general, if the use of $r = 1$ provides a variance which is less than δ , then that record needs to be suppressed. Therefore, the suppression condition for a record is as follows:

DEFINITION 2. A record $\bar{X} = (x_1 \dots x_d)$ is suppressed for δ -anonymity, when the following condition is satisfied:

$$(2.8) \quad \left(\sum_{i=1}^d x_i^2 - x_k^2\right) / r < \delta \text{ for some } k \in \{1 \dots d\}$$

We note that record suppression is necessary for many anonymity-based approaches. One advantage of the δ -anonymity method is that since it does not depend on the behavior of the other records in the data, it can actually be performed at *data collection time*. One disadvantage of the δ -anonymity based definition is that it treats all records evenly irrespective of the behavior of the other records in its locality. In general, outlier records containing unique words should have a much larger anonymity requirement than records which are drawn from pre-defined clusters. Therefore, we provide a second definition of privacy which uses the records in the neighborhood of a given record in order to define the anonymity level. The key idea behind this definition is that the variance of the distance calculations for the k -nearest neighbors are higher than the absolute distances to each of these neighbors. This ensures that it becomes extremely difficult to distinguish a record from its k -nearest neighbors even when partial information about some of the records is available. Therefore, we define the concept of k -variance based anonymization as follows:

DEFINITION 3. A data set \mathcal{D} is said to be k -variance based anonymized, if the following conditions hold true for any pair of records \bar{X}^i and \bar{X}^j , with the corresponding sketches \bar{S}^i and \bar{S}^j :

Algorithm *DeltaAnonymity*(\overline{X} , AnonLevel: δ);
begin
 Find smallest $r \geq 1$ such that
 $(\sum_{i=1}^d x_i^2 - x_k^2)/r \geq \delta \quad \forall k \in \{1 \dots d\}$
if no such r exists suppress \overline{X}
else create sketch of \overline{X} with
 $\lfloor r \rfloor$ sketch components;
end

Figure 1: Algorithm for δ -anonymity

- \overline{X}^j is not among the k -nearest neighbors of \overline{X}^i .
- \overline{X}^j is among the k -nearest neighbors of \overline{X}^i , and the sketch based estimation of $\overline{X}^j \cdot \overline{X}^i$ has standard deviation which is at least equal to $\sqrt{(|\overline{X}^j| \cdot |\overline{X}^i|) - |\overline{X}^j \cdot \overline{X}^i|}$.

The above definition ensures that it becomes much more difficult to distinguish a record from its k -nearest neighbors, since the standard deviation of the similarity calculations is larger than the difference between the similarity to the k th neighbor and the maximum possible similarity value. For practical applications, the records may be normalized so that the each value of $|\overline{X}^i|$ is 1 unit. We note that in order to perform the sketch-based estimation of $\overline{X}^j \cdot \overline{X}^i$, we need to use only the first $\min\{m(i), m(j)\}$ components of \overline{S}^i and \overline{S}^j . The average of pairwise multiplication of corresponding sketch components provides a variance which is equal to $1/\min\{m(i), m(j)\}$ of the variance computed in Equation 2.7. As in the previous case, it may not always be possible to satisfy this condition for a given record \overline{X}^i . In such a case, the record \overline{X}^i needs to be suppressed.

3 Algorithms for Anonymization

In this section, we will discuss and present methods for sketch based anonymization. We will first present methods for δ -anonymity, and then present methods for k -variance based anonymization.

3.1 Algorithms for δ -Anonymity In this subsection, we will present a number of algorithms for δ -anonymity of the underlying data. In Figure 1, we have presented algorithms for δ -anonymity. The overall algorithm is extremely simple, and uses the results of Definition 1. The algorithm for δ -anonymity simply tries to find the smallest value of r which is at least 1, and which satisfies the conditions of Definition 1. One nice property of this method is that it is extremely simple, and can be easily implemented at *data collection time*,

Algorithm *Kvariance*(Database: \mathcal{D} , AnonyLevel: k);
begin
 Determine k -nearest neighbor distance to
 each record in \mathcal{D} ;
 $\{m_1 \dots m_N\} = \{\infty \dots \infty\}$;
 Determine sort index $\mathcal{SI}_{\mathcal{D}}$ for k -nearest
 neighbor distance in \mathcal{D} (largest first);
while *not(termination_criterion)* **do**
for each record $\overline{X} \in \mathcal{D}$
 in order of $\mathcal{SI}_{\mathcal{D}}$ **do**
begin
 Let q be the index of \overline{X} in database \mathcal{D} ;
 Choose minimum integral value of $m(q)$ such that
 for each \overline{X}_l among the k -nearest neighbors
 of \overline{X} , the use of $\min\{m(q), m(l)\}$
 sketch components satisfies the conditions of
 Definition 3 for \overline{X}_l with respect to \overline{X} ;
end;
 Construct the sketches of the N records with the use
 of $\{m(1) \dots m(N)\}$ sketch components respectively;
 We will need to suppress all records for which the
 value of $m(i)$ is 0;
end

Figure 2: Algorithm for k -variance based anonymity

since it does not use the behavior of the other records in the data. On the flip side, the use of the behavior of other records can help in improving the quality of privacy preservation. In the next section, we will discuss algorithms for k -variance based anonymity. This technique uses the distribution of the records in the data, and therefore provides greater privacy to outlier records which can be more easily distinguished from the rest of the data.

3.2 Algorithm for k -variance based anonymity

In this section, we will discuss algorithms for k -variance based anonymity. In the k -variance based anonymization method, our goal is to choose a number of sketch components so that k -variance based anonymity is preserved for each record. An important question here is to choose the number of sketch components appropriately for each record so that the key condition for k -variance based anonymity is satisfied. One possible approach is use an iterative technique so that the number of sketch components for each record are determined one by one. An important question here is the order in which the records should be processed in order to perform the anonymization.

In general, outlier records present a greater challenge from a privacy point of view. Therefore, these

records need to be processed first during the anonymization process. The overall process of anonymization is illustrated in Figure 2. The key to creating the sketch is determining the number of sketch components for the different records which are denoted by $\{m(1) \dots m(N)\}$. Once the number of components for each record have been determined, the sketches can be created in a straightforward manner. We start off by setting $\{m(1) \dots m(N)\}$ to $\{\infty \dots \infty\}$. The algorithm uses an iterative approach in which a pass through the entire database is performed in each iteration in order to set the values of $\{m(1) \dots m(N)\}$. We refer to each such iteration as a *major iteration*. In each pass through the database, we further refine the values of $\{m(1) \dots m(N)\}$. This successive refinement continues in each iteration, until a termination criterion determines that the change in the values of $\{m(1) \dots m(N)\}$ from one iteration to another is not very significant. We will discuss more on this issue slightly later.

In each major iteration, we process the records in the data in a certain sort order $\mathcal{ST}_{\mathcal{D}}$ which is in reverse order of the k -nearest neighbor distance of different records. Each such iteration of processing a record is referred to as a *minor iteration*. This order is determined during an initial pre-processing phase on the data. When we process a record $\overline{X}_q \in \mathcal{D}$, we set the number of sketch components $m(q)$ so that for each of the k -nearest neighbors of \overline{X}_q , the conditions of Definition 3 are satisfied. We note that while computing the conditions for Definition 3 for a record \overline{X}_l which lies within the k -nearest neighbors of \overline{X}_q , we only need to use the first $\min\{m(l), m(q)\}$ sketch components of both records. We note that since the values of the $m(\cdot)$ vector are retained from one iteration to the next, this helps in gradual convergence of these values. The termination criterion is to determine the total change in the values of the $m(\cdot)$ vector from one iteration to the next. If this change is below a certain threshold, then we terminate and construct the sketches for the different records. We note that at the end of the process, some of the values of $m(i)$ may be 0. The corresponding records may need to be suppressed.

3.3 Application of Sketch Based Representation to Data Mining Algorithms The sketch based representation can be easily used for a variety of data mining algorithms. This is because primitive computations such as the dot product are extremely easy with the use of the sketch based representation. In this section, we enumerate the basic primitive operations which can be computed with the sketch based representation, and the data mining algorithms which can be implemented with the use of these primitive operations.

1. **Determination of Original Attribute Values:** The technique can be used for determination of the original attribute values from the sketch. We have already discussed the process of determination of original attribute values in section 2.
2. **Determination of dot product:** This was also discussed in section 2. The dot product is an extremely important operation for a variety of proximity based computations. Many data mining algorithms such as clustering and classification can be achieved almost exclusively with the use of proximity based computations.
3. **Determination of Sparse Transaction Length:** In many applications such as text and transaction data, it is desirable to find the sum-squared length $\|\overline{X}\|^2$ of the document (or transaction) \overline{X} for the purposes of normalization during mining.
4. **Determination of Euclidean distance:** The euclidean distance between two records \overline{X} and \overline{Y} can be computed as follows:

$$(3.9) \quad \|\overline{X} - \overline{Y}\|^2 = \|\overline{X}\|^2 + \|\overline{Y}\|^2 - 2 \cdot \overline{X} \cdot \overline{Y}$$
 Since we have already discussed how to estimate all quantities on the right hand side, the quantity of the left hand side can be estimated as well.
5. **Centroid of a set of records:** The centroid of a set of records can be expressed as the average of all the records in the set.

The above primitives can be used for the purpose of a number of data mining algorithms. The important data mining algorithms which can be used for these purposes are as follows:

- **Clustering:** Many distance based clustering algorithms can be easily implemented with the use of the above primitives. For example, the K -means clustering algorithm requires the repeated computation of centroids in conjunction with the computation of distances of data points from these centroids.
- **Classification:** Many classifiers such as nearest neighbor methods use distance based computations in order to perform the classification. The above mentioned primitives can be easily used in order to perform the computations.
- **Reconstruction Based Algorithms:** Since the attributes can be approximately reconstructed, any general data mining algorithm can be applied to

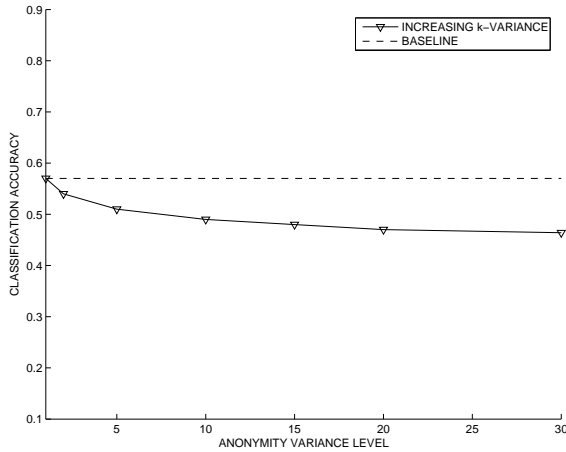


Figure 3: Classification Accuracy with Increasing Anonymity Level (Text17)

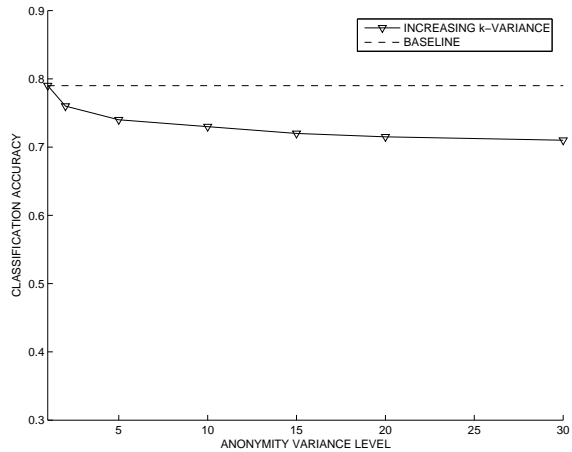


Figure 5: Classification Accuracy with Increasing Anonymity Level (C20.I6.D200k)

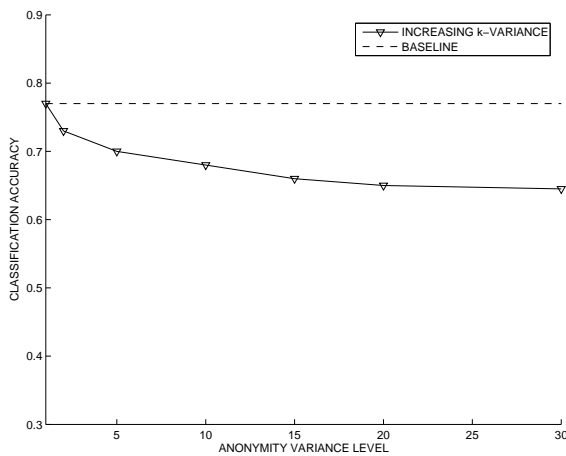


Figure 4: Classification Accuracy with Increasing Anonymity Level (C20.I4.D200K)

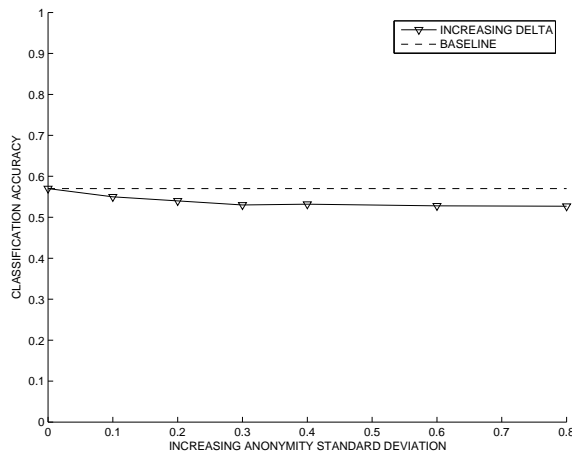


Figure 6: Classification Accuracy with increasing δ (Text17)

the problem with the use of the reconstructed attributes. This is a slightly less desirable option, since direct approximation of primitives is usually more accurate than computations from the reconstructed attributes.

4 Experimental Results

In this section, we will discuss the results of the sketch based method for privacy preserving data mining. We will test our method on a number of text and market basket data sets. We will test the effectiveness of our technique over the classification problem.

For the purpose of testing, we used a number of text and market basket data sets. The text data sets used include the *Yahoo!* taxonomy data set which contained 163,000 documents from a 1996 scan of the *Yahoo!*

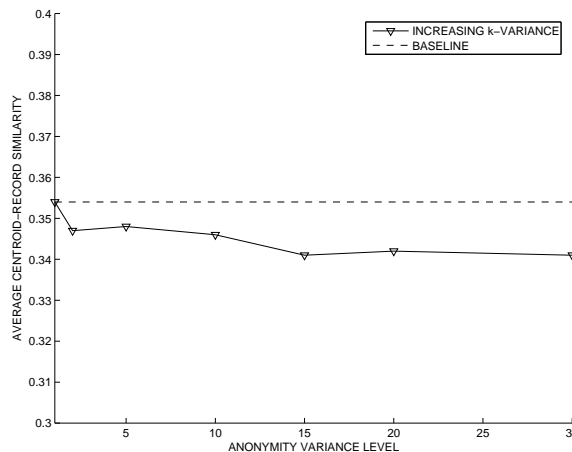


Figure 7: Clustering Effectiveness with Increasing Anonymity Level (Text17)

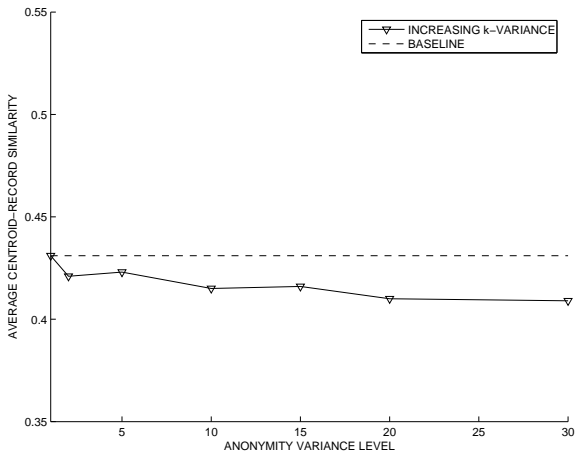


Figure 8: Clustering Effectiveness with Increasing Anonymity Level (C20.I4.D200k)

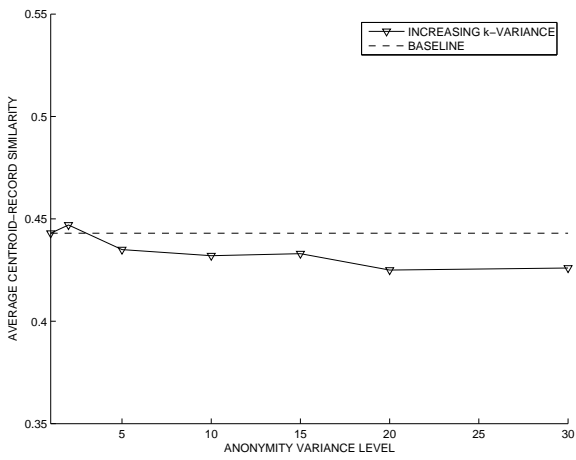


Figure 9: Clustering Effectiveness with Increasing Anonymity Level (C20.I6.D200k)

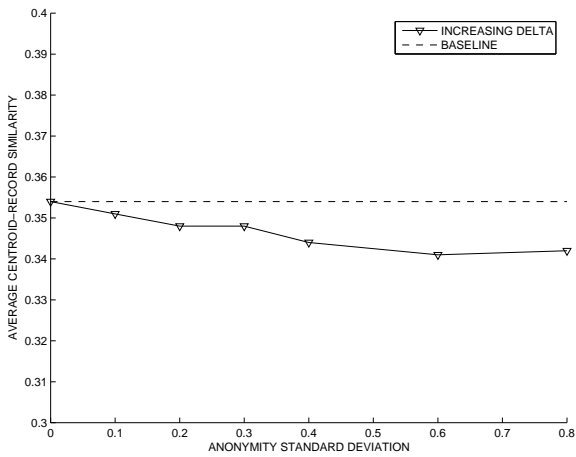


Figure 10: Clustering Effectiveness with increasing δ (Text17)

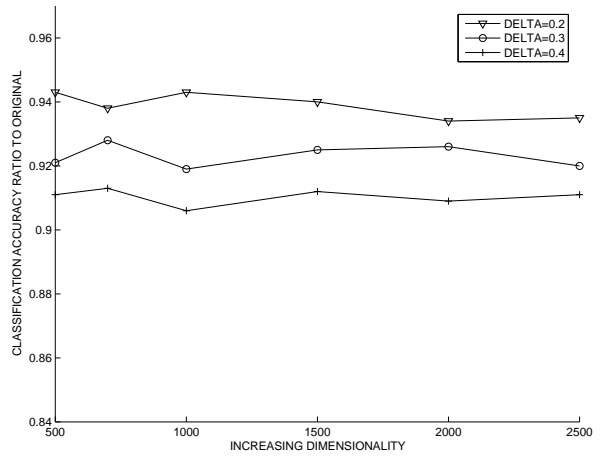


Figure 11: Effectiveness of Classification Method with Data Dimensionality

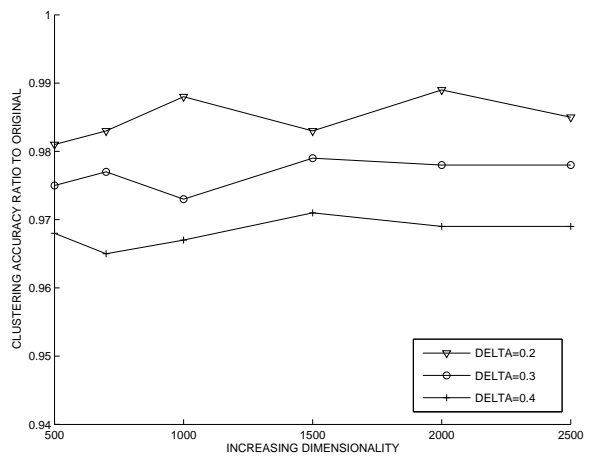


Figure 12: Effectiveness of Clustering with Increasing Data Dimensionality

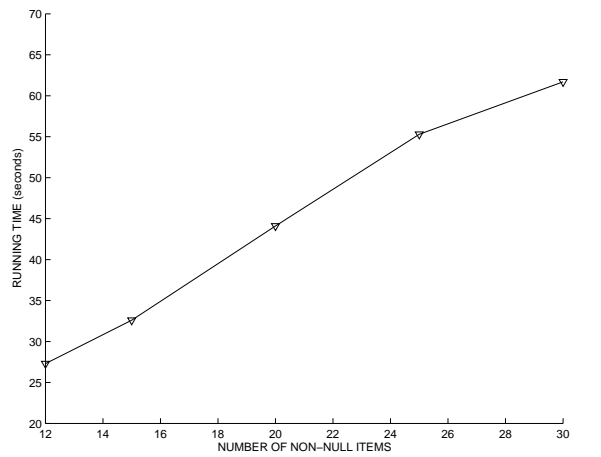


Figure 13: Increasing Efficiency with Data Dimensionality

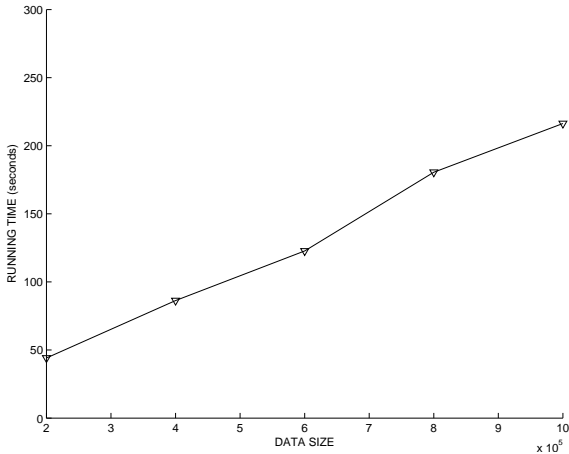


Figure 14: Increasing Efficiency with Data Size)

taxonomy. The data contained documents from the *Yahoo!* categories, which we truncated to 17 high level categories of the taxonomy. We will refer to this data set as Text17. The market basket data sets were generated using the *Apriori* data generator [7]. For the case of the *Apriori* data generator, we used a number of different data sets. In particular, we used two instances of the data set T20.I4.D100K to create¹ the two class data set C20.I4.D200K. Similarly, we created the data set C20.I6.D200K with the use of two instances of the data set T20.I6.D100K. Next, we will present the results of using the technique for the classification problem.

4.1 Classification Problem In this section, we will present the results of using the technique for the classification problem. The training data and test data were separately anonymized. The classification accuracy was tested using a bootstrapping procedure in which the ratio of the training to the test data was 2:1. In each case, we tested the k -variance based anonymity procedure for varying values of k . We note that some of the test data needed to be suppressed as a result of the procedure. This was a source of inaccuracy since the suppressed data was labeled at random in accordance with a-priori class distribution probabilities. In Figure 3, we have illustrated the classification accuracy on the text data set. It is clear that with increasing value of k , the classification accuracy was reduced. On the same graph, we have also illustrated the classification accuracy when the original data set was used with the use of a horizontal baseline. However, on an overall basis the accuracy reduced only slightly by about 5 – 10%. In Figures 4 and 5, we have illustrated the results for

¹The notations used here are from the paper [5].

the market basket data sets. The results on these data sets show the same trend as the text data set. In each case, one of the trends was that the accuracy initially reduced quickly, but then leveled out beyond a certain anonymity level.

We also tested the method with increasing value of δ when the δ -anonymity procedure was used. The results for the text data set are illustrated. The trend is quite similar to the case of the k -variance based anonymization method. As in the previous case, the rate of degradation of the results levels off with increasing value of δ . These results show that the technique is quite robust to increasing values of the parameter δ .

4.2 Clustering Problem We also tested the technique for the case of the clustering problem. In the case of the clustering problem, we implemented a version of the k -means algorithm, in which the value of k was chosen to be 10. The average similarity value of each point to its cluster centroid was reported as the quality measure for the clustering process. The average similarity value was computed in terms of the cosine distance, and therefore larger values imply greater similarity. While the data mining algorithm was applied on the sketch based representation, the original representation of the records were used for presenting the true quality of the results with respect to the original data. As before, records which were suppressed contributed to quality deterioration, since they were assigned a-posteriori to random clusters in proportion to the number of points in the different clusters. Such assignments increased with greater distortion, and therefore reduced the quality of the clustering process.

In Figures 7, 8, and 9, we have illustrated the effectiveness of the clustering problem with increasing variance level for the anonymization process. On the X-axis, we have illustrated the variance level of the anonymization, whereas on the Y-axis, we have illustrated the average similarity level of each record to its assigned cluster centroid. Since the computed values are *similarity values* as opposed to *distance values*, high numbers imply greater similarity. We have also illustrated a horizontal baseline which indicates the similarity value at the point of no distortion. The trends in this case are quite similar to the classification application, though they are a little bit more noisy because of the difference in the clustering process across different runs of the randomized k-means algorithm. In each case, the results degrade initially, but level off after a while.

In Figure 10, we have illustrated the effectiveness of the clustering technique with increasing value of the parameter δ . As in the previous cases, the clustering

accuracy initially reduces with increasing value of δ , but then it levels off when the value of δ is increased beyond 1. These results show that the anonymization method is quite robust to increase in the noise level. In the next section, we will discuss the computational results associated with the use of the sketch based privacy preservation method.

4.3 Effectiveness with Increasing Data Dimensionality We tested the effectiveness of the method with increasing data dimensionality. In order to generate data sets with increasing dimensionality, we used market basket data sets as generated in [5]. except we varied the base number of items from which the data set was generated. Thus, the data set Q20.I6.D200K.G1200 refers to the same data set as C20.I6.D200K, except that we used 1200 base items instead of 1000 items in order to generate the data sets. The data set C20.I6.D200K corresponds to Q20.I6.D200K.G1000. Thus, we varying the value of x in Q20.I6.D200K.G x , it is possible to measure the variation in effectiveness with increasing data dimensionality.

In Figure 11, we have illustrated the variation in effectiveness with increasing data dimensionality. We fixed the values of δ at 0.2, 0.3, 0.4, in order to plot the behavior at varying levels of perturbation. The base dimensionality is illustrated on the X-axis. On the Y-axis, we have illustrated the ratio of the accuracy on the transformed data set to the accuracy on the original data set. This ratio is always less than 1, and a higher value implies a lower degradation. It is clear that the technique does not degrade much with increasing data dimensionality, and the behavior does not vary much for different levels of perturbation. Similar results are observed for the case of the clustering problem in Figure 12. As in the previous case, we fixed the value of δ at 0.2, 0.3, and 0.4 in order to obtain the corresponding results. In this case, the value on the Y-axis illustrates the ratio of the average clustering similarity value (as in Figures 7, 8, 9, and 10) of the transformed data to the original data. As in the previous case, the method is extremely robust to increasing data dimensionality. Thus, this technique is very effective for the high dimensional case, as long as the data continues to be sparse.

4.4 Computational Results In this section, we will present the computational results illustrating the efficiency of the method. In particular, we tested the k -variance based anonymity method, since this algorithm was more computationally challenging of the two methods. Since the running time of the count based sketch method depended upon the number of items with non-zero counts, we tested the efficiency with

increasing number of items with non-zero counts. In Figure 13, we have illustrated the running time of the technique on a data set containing 200,000 records with increasing dimensionality of the market basket data set. The increasing dimensionality of the market basket data set was obtained by taking projections of increasing dimensionality of the data set Cx.I6.D200K. On the X-axis, we have used the number of non-zero items x , whereas on the Y-axis, we have illustrated the running time for the method. It is clear that the running time increases linearly with the number of non-zero items. This is because of the nature of the computations of the sketch-based method which require a constant number of computations for each non-zero item in the data.

Similarly, the increasing data set size was obtained by creating the data set C20.I6.1M, which contains 1 million records, and using samples of varying sizes of this data set. In Figure 14, we have illustrated the running times of the data set for with increasing size of the data set. The X-axis contains the size of the data set, whereas the Y-axis contains the running time in seconds. It is clear that in which case, the running time increases linearly with increasing data size. This is because of the use of a partition based method whose complexity increases linearly with database size. For both the scalability cases, we note that the running times are small on an absolute basis, and are of the order of a few seconds. Thus, the technique retains its efficiency with increasing database size.

5 Conclusions and Summary

In this paper, we presented a method for sketch based privacy preserving mining of text and market basket data streams. These data domains are especially difficult for traditional privacy-preserving methods because of their high dimensionality. Therefore, we utilize the sketch based technique in which the variance level is sensitive to the amount of noise in the data. This technique is extremely effective for high dimensional data sets, as long as the data is sparse. The sketch based method provides excellent privacy while allowing effective reconstruction of many aggregate distance measures. This is useful in computation of a number of data mining primitives which can be leveraged for a number of algorithms. We also tested the effectiveness of the method on a number of text and market basket data sets. The results show that the technique is accurate, maintains robustness with increasing dimensionality, and is also efficient to implement in practice.

References

- [1] C. C. Aggarwal, and P. S. Yu. *A Condensation Based*

- Approach to Privacy Preserving Data Mining*. Proceedings of the EDBT Conference, pp. 183–199, 2004.
- [2] C. C. Aggarwal. *On k -anonymity and the curse of dimensionality*. VLDB Conference, 2005.
- [3] C. C. Aggarwal. *On Randomization, Public Information, and the Curse of Dimensionality*, ICDE Conference, 2007.
- [4] R. Agrawal, and R. Srikant. *Privacy-Preserving Data Mining*. Proceedings of the ACM SIGMOD Conference, pp. 439–450, 2000.
- [5] R. Agrawal, and R. Srikant. *Fast Algorithms for Mining Association rules in Large Databases*. Proceedings of the VLDB Conference, pp. 487–499, 1994.
- [6] D. Agrawal, and C. C. Aggarwal. *On the Design and Quantification of Privacy Preserving Data Mining Algorithms*. Proceedings of the ACM PODS Conference, pp. 247–255, 2001.
- [7] R. Agrawal, and R. Srikant. *Fast Algorithms for Mining Association Rules in Large Databases*. VLDB Conference, 1994.
- [8] N. Alon, Y. Matias, and M. Szegedy. *The Space Complexity of Approximating the Frequency Moments*. ACM Symposium on Theory of Computing, pp. 20–29, 1996.
- [9] R. J. Bayardo, and R. Agrawal. *Data Privacy through Optimal k -Anonymization*. Proceedings of the ICDE Conference, pp. 217–228, 2005.
- [10] M. Bawa, R. J. Bayardo Jr. and R. Agrawal. *Privacy Preserving Indexing of Documents on the Network*. VLDB Conference, 2003.
- [11] A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke. *Privacy Preserving Mining Of Association Rules*. Proceedings of the ACM KDD Conference, pp. 217–228, 2002.
- [12] M. Kantarcioglu, and C. Clifton. *Privacy Preserving Data Mining of Association Rules on Horizontally Partitioned Data*, Transactions on Knowledge and Data Engineering, 16(9): pp. 1026-1037, 2004.
- [13] H. Kargupta, K. Liu, and J. Ryan. *Privacy Sensitive Distributed Data Mining from Multi-party Data*, Proceedings of the first NSF/NIJ Symposium on Intelligence and Security Informatics, pp. 336-342, 2003.
- [14] D. Kifer, and J. Gehrke. *Injecting utility into anonymized datasets*. SIGMOD Conference, pp. 217–228, 2006.
- [15] K. LeFevre, D. DeWitt, and R. Ramakrishnan. *Mondrian Multidimensional K -Anonymity*. ICDE Conference, 25, 2006.
- [16] A. Machanavajjhala A., J. Gehrke, D. Kifer, and M. Venkatasubramaniam. *l -Diversity: Privacy Beyond k -Anonymity*. ICDE, 2006.
- [17] A. Meyerson, and R. Williams. *On the Complexity of optimal k -anonymity*. Proceedings of the ACM PODS Conference, pp. 223–228, 2004.
- [18] N. Mishra, and M. Sandler. *Privacy via Pseudo-random sketches*. ACM PODS Conference, 2006.
- [19] B. Pinkas. *Cryptographic Techniques for Privacy-Preserving Data Mining*, SIGKDD Explorations, 2003.
- [20] S. Rizvi, and J. Haritsa. *Maintaining data privacy in association rule mining*. VLDB Conference, 2002.
- [21] P. Samarati, and L. Sweeney. *Protecting Privacy when Disclosing Information: k -Anonymity and its Enforcement Through Generalization and Suppression*. Proceedings of the IEEE Symposium on Research in Security and Privacy, May 1998.
- [22] J. Vaidya, and C. Clifton. *Privacy Preserving K -Means Clustering over Vertically Partitioned Data*. ACM KDD Conference, 2003.
- [23] V. Verykios, E. Bertino, I. Fovino, L. Provenza, Y. Saygin, and Y. Theodoridis. *State-of-the-art in privacy preserving data mining*. SIGMOD Record 33(1): pp. 50-57, 2004.