On k-Anonymity and the Curse of Dimensionality

Charu C. Aggarwal

IBM T. J. Watson Research Center
Route 134 & Taconic State Parkway
Yorktown Heights, NY
USA
charu@us.ibm.com

Abstract

In recent years, the wide availability of personal data has made the problem of privacy preserving data mining an important one. A number of methods have recently been proposed for privacy preserving data mining of multidimensional data records. One of the methods for privacy preserving data mining is that of *anonymization*, in which a record is released only if it is indistinguishable from k other entities in the data. We note that methods such as k-anonymity are highly dependent upon spatial locality in order to effectively implement the technique in a statistically robust way. In high dimensional space the data becomes sparse, and the concept of spatial locality is no longer easy to define from an application point of view. In this paper, we view the k-anonymization problem from the perspective of inference attacks over all possible combinations of attributes. We show that when the data contains a large number of attributes which may be considered quasi-identifiers, it becomes difficult to anonymize the data without an unacceptably high amount of information loss. This is because an exponential number of combinations of dimensions can be used to make precise inference attacks, even when individual attributes are partially specified within a range. We provide an analysis of the effect of dimensionality on k-anonymity methods. We conclude that when a data set contains a large number of attributes which are open to inference attacks, we are faced with a choice of either completely suppressing most of the data or losing the desired level of anonymity. Thus, this paper shows that the curse of high dimensionality also applies to the problem of privacy preserving data mining.

Proceedings of the 31st VLDB Conference, Trondheim, Norway, 2005

1 Introduction

The privacy preserving data mining problem has gained considerable importance in recent years because of the vast amounts of personal data about individuals stored at different commercial vendors and organizations. In many cases, users are willing to divulge information about themselves only if the privacy of the data is guaranteed. Thus methods need to be proposed to mask the sensitive information in the records. This creates the natural challenge of mining the data in an effective way with a limited data representation. A variety of techniques [3, 4, 6, 7, 9, 10, 11, 12, 14] have recently been proposed both to represent and mine the data without loss of privacy. Some important techniques for privacy include methods such as perturbation [4], k-anonymity [14], condensation [1], and data hiding with conceptual reconstruction [3].

In this paper, we will analyze the k-anonymity approach [14] for the high dimensional case. The idea behind this class of approaches is that many of the fields in the data can be treated as *pseudo-identifiers* or *quasi-identifiers* which can be matched with publically known data in order to identify individuals. For example, a commercial database containing birthdates, gender and zip-codes can be matched with voter registration lists in order to identify the individuals precisely. Another related class of methods to deal with the issue of k-anonymity is the k-indistinguishability approach. The k-anonymity and k-indistinguishability approaches are briefly discussed below:

- In the k-anonymity approach [14], *generalization* techniques are applied in order to mask the exact values of attributes. For example, a quantitative attribute such as the age may only be specified to a range. This is referred to as *attribute generalization.* By defining a high enough level of generalization on each attribute, it is possible to guarantee k-anonymity. On the other hand, attribute generalization also leads to a loss of information.

- In the k-indistinguishability model [1], clustering techniques are used in order to construct indistinguishable groups of k records. The statistical characteristics of these clusters are used to generate pseudo-data which is used for data mining purposes. While such pseudo-data does not represent the true data records, it is useful for most modelling purposes, since it reflects the original distribution of the records. There are some advantages in the use of pseudo-data, in that it is more resistant to hacking, and it does not require any modification of the underlying data representation as in a generalization approach.

While the k-anonymity and k-indistinguishability model differ in the final anonymized data representation, they are similar in methodology and concept. For example, both the approaches are aimed to thwart the case where an inference driven user may use a combination of attributes in order to infer the identity of the individual record. Typical anonymization approaches assume that only a small number of fields which are available from public data are used as quasi-identifiers. These methods assume that these publically defined fields are well studied from a domain specific point of view and use generalizations on corresponding domain hierarchies of these small number of fields. These hierarchies are used to construct privacy preserving generalizations of the data set. While such solutions are useful for the case of small subsets of quasi-identifiers, they cannot be used effectively in the high dimensional case. In general, a quasi-identifier may not be derived from a public database, but may be any field which is partially or substantially known to any particular group or entity (such as an employer). In such cases, the number of combinations of dimensions available for inference attacks increases rapidly, and also makes the data more challenging for the privacy preservation process. In such cases, many attributes (eg. salary) continue to be sensitive, but also cannot be ruled out as quasi-identifiers. Such situations are quite likely in real life, since an adversary may also have personal knowledge of the target of interest. It is in fact quite likely that an adversary who is acquainted with a target of interest knows much more than is available from public information. Since one cannot make a-priori assumptions about what different adversaries know about the various records in the data set, it makes the problem of privacy preservation more challenging. In general, the distinction between sensitive attributes and quasi-identifiers becomes blurred when sensitive attributes are partially known to specific entities. In such cases, one has no choice but to include such attributes in the anonymization process. This paper investigates such real situations in which a large fraction of the attributes from the data are available for inference attacks.

We will see that inter-attribute combinations within a record have such a powerful revealing effect in the high dimensional case, that the amount of data required to preserve anonymity increases beyond most practical limits. While an earlier paper [1] has discussed the data mining advantages of preserving inter-attribute statistics, the results of this paper would seem to indicate that there are also some advantages in privacy preservation approaches which do *not* preserve inter-dimensional statistics (as in the perturbation model [4]).

This paper is organized as follows. In the next section, we will discuss some quantifications of information loss resulting from the anonymization process. We will analyze both axis-parallel and generic methods for the anonymization process. In section 3, we will illustrate some empirical results showing the effectiveness of different methods of privacy preserving data mining. Section 4 contains discussions and conclusions.

2 The Privacy Model

For ease in exposition, we will assume that any dimension in the database is a potentially identifying quasi-identifier. This assumption can be made without loss of generality, since we can restrict our analysis only to such identifying attributes. We will further assume the use of quantitative attributes. This assumption can also be made without loss of generality. The results can be easily extended to categorical data, since both the quantitative and categorical data domains can be represented in binary form.

We note that all anonymization techniques depend upon some notion of spatial locality in order to perform the generalization. This spatial locality is often defined in the form of a distance function as in [1]. However distance functions begin to show loss of intra-record distinctiveness in high dimensional space. It has been argued in [2, 8], that under certain reasonable assumptions on the data distribution, the distances of the nearest and farthest neighbors to a given target in high dimensional space is almost the same for a variety of data distributions and distance functions. In such a case, the concept of spatial locality becomes ill defined, since the contrasts between the distances to different data points do not exist. Generalization based approaches to privacy preserving data mining are deeply dependent upon spatial locality, since they use the ambiguity of different data points within a given spatial locality in order to preserve privacy. We will see that privacy preservation by anonymization becomes impractical in very high dimensional cases, since it leads to an unacceptable level of information loss.

In order to facilitate further discussion, we will es-

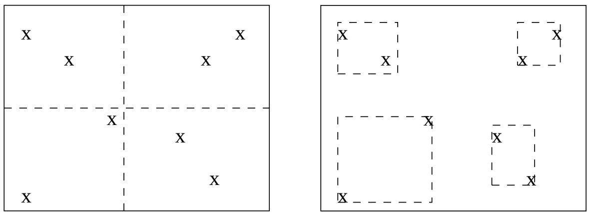
Notation	Definition
d	Dimensionality of the data space
N	Number of data points
$\mathcal{F}$	1-dimensional data distribution in $(0, 1)$
X_d	Data point from $\mathcal{F}^d$ with each coord. drawn from $\mathcal{F}$
$dist_d^k(x, y)$	Distance between $(x^1, \ldots x^d)$ and $(y^1, \ldots y^d)$ using L_k metric $= \sum_{i=1}^{d} [(x_1^i - x_2^i)^k]^{1/k}$
$\lVert \cdot \rVert_k$	Distance of a vector to the origin $(0, \ldots, 0)$ using the function $dist_d^k(\cdot, \cdot)$
$E[X]$, $var[X]$	Expected value and variance of a random variable X
$Y_d \rightarrow_p c$	A sequence of vectors $Y_1, \ldots, Y_d$ converges in probability to a constant vector c if: $\forall \epsilon > 0 \ lim_{d \rightarrow \infty} P[dist_d(Y_d, c) \leq \epsilon] = 1$

Table 1: Notations and Definitions

tablish certain notations and definitions. We assume that all points are distributed in the unit cube. In Table 1, we have introduced some notations and definitions, which we will use throughout this paper.

In Figure 1, we have illustrated two cases of generalization of data points into a range along each dimension. In Figure 1(a), 2-anonymization is achieved by simple discretization without much optimization. In Figure 1(b), more careful clustering methods are utilized to achieve 2-anonymity, so that the sizes of the bounding rectangles are reduced. The latter is also an example of optimized axis-parallel generalizations. It is not necessary to generalize using axis-parallel ranges only. Instead, the condensed statistics of arbitrary clusters can be used for the anonymization process [1]. In general, the problem of finding the optimal k-anonymous representation is known to be NP-hard [13]. Therefore, we will analyze both the methods of axis-parallel generalization and arbitrary clustering. We will show that the asymptotic information loss with increasing dimensionality is sufficiently high to make the privacy preservation process impractical.

First, let us consider the axis-parallel generalization approach, in which individual attribute values are replaced by a randomly chosen interval from which they are drawn. In order to analyze the behavior of anonymization approaches with increasing dimensionality, we consider the case of data in which individual dimensions are independent and identically distributed. The resulting bounds provide insight into the behavior of the anonymization process with increasing *implicit* dimensionality. We construct a bounding box around a target point $\overline{X_d}$ in order to generalize it. The value of the data point $\overline{X_d}$ in this grid cube is generalized to the corresponding partially specified range of this bounding box. For data point $\overline{X_d}$ to maintain k-anonymity, this bounding box must contain *at least* $(k-1)$ other points. First, we will consider the case when the generalization of each point uses a maximum fraction f of the data points along each of the d partially specified dimensions. Thus, data points which do not satisfy this condition may need to be *suppressed* [14]. It has been suggested [14], that suppression of a larger percentage of the data leads to an unacceptable aggregate change in the statistical characteristics of the data for mining purposes. In the following analysis, we will show the difficulty of preserving k-anonymity using the approach of partial range masking.



(a) (b)

Figure 1: Some Examples of Generalization for 2-Anonymity

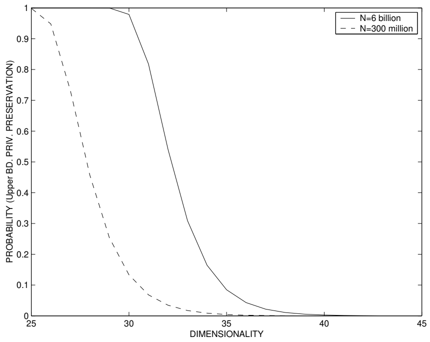


Figure 2: Upper Bound of 2-anonymity Probability in an Non-Empty Grid Cell

Lemma 1 *Let $\mathcal{D}$ be a set of N points drawn from the d-dimensional distribution $\mathcal{F}^d$ in which individual dimensions are independently distributed. Consider a randomly chosen grid cell, such that each partially masked dimension contains a fraction f of the total data points in the specified range. Then, the probability P^q of exactly q points in the cell is given by $\binom{N}{q} \cdot f^{q \cdot d} \cdot (1 - f^d)^{(N-q)}$.*

Proof: We note that the probability of a data point in a grid cell with range specificity of f along each of the d dimensions is given by $x = f^d$. Then, the probability that a given grid cube contains exactly q points is given by the binomial distribution with parameters N and x. Therefore, we can use the binomial distribution formula to define the corresponding probability P^q:

$$P^q = \binom{N}{q} \cdot x^q \cdot (1 - x)^{(N-q)} \tag{1}$$

A direct corollary of the above result is the following:

Corollary 1 *Let B_k be the event that the grid cell corresponding to the partially specified dimensions contains k or more data points. The corresponding probability $P(B_k)$ is given by:*

$$P(B_k) = \sum_{q=k}^{N} \binom{N}{q} \cdot f^{q \cdot d} \cdot (1 - f^d)^{(N-q)} \tag{2}$$

Proof: We note that $P(B_k) = \sum_{q=k}^{N} P^q$. By substituting $x = f^d$ from Equation 1, we get the corresponding result. ∎

We note that a set of partially specified dimensions violates the conditions of k-anonymity, when the corresponding set of partially specified ranges contain at least one data point, but less than k data points. Therefore, we need to find the conditional probability denoted by $P(B_k|B_1)$. The value of this conditional probability is defined by the Lemma below.

Lemma 2 *Let B_k be the event that the set of partially masked ranges contains at least k data points. Then the following result for the conditional probability $P(B_k|B_1)$ holds true:*

$$P(B_k|B_1) = \frac{\sum_{q=k}^{N} \binom{N}{q} \cdot f^{q \cdot d} \cdot (1 - f^d)^{(N-q)}}{\sum_{q=1}^{N} \binom{N}{q} \cdot f^{q \cdot d} \cdot (1 - f^d)^{(N-q)}} \tag{3}$$

Proof: We know from elementary probability theory that:

$$P(B_k|B_1) = P(B_k \cap B_1)/P(B_1) \tag{4}$$

However, the event B_k is a special case of B_1. This is because if a set of masked ranges contain at least k points, the corresponding set of ranges must also be non-empty. Therefore, we have:

$$P(B_k \cap B_1) = P(B_k) \tag{5}$$

Therefore, we have:

$$P(B_k|B_1) = P(B_k)/P(B_1) \tag{6}$$

By substituting for the value of $P(B_k)$ and $P(B_1)$ in Equation 2, we get the desired result. ∎

We note the following simple observation:

Observation 1 *For all $k > 2$, we have $P(B_k|B_1) \leq P(B_2|B_1)$.*

The above observation is true because the event B_k is subsumed by the event B_2 for any value of k larger than 2. Therefore, by finding an upper bound on $P(B_2|B_1)$, we can also find a upper bound on the probability that k-anonymity is achieved on a randomly chosen (non-empty) set of non-empty grid changes. Next, we observe the following:

$$P(B_2|B_1) = \frac{1 - N \cdot f^d \cdot (1 - f^d)^{(N-1)} - (1 - f^d)^N}{1 - (1 - f^d)^N} \tag{7}$$

The above observation can be easily verified by substituting the values of $k = 2$, $P(B_k)$ and $P(B_1)$ in Equation 3 of Lemma 2. We are simply expressing the events $P(B_2)$ and $P(B_1)$ in the complementary form[1] of the binomial expression. Next, we will show that the probability of achieving 2-anonymity in a non-empty grid cell is zero for the limiting case of high dimensionality. We formalize this result as follows:

Lemma 3 *The limiting probability for achieving 2-anonymity in a set of partially specified ranges, each containing a fraction $f < 1$ of the data points is zero. In other words, we have:*

$$lim_{d \to \infty} P(B_2|B_1) = 0 \tag{8}$$

Proof: By substituting $x = f^d$ in Equation 7, we get:

$$P(B_2|B_1) = 1 - \frac{N \cdot x \cdot (1 - x)^{N-1}}{1 - (1 - x)^N} \tag{9}$$

We note that as $d \to \infty$, we have $x \to 0$. This is because $f < 1$. Consequently, we get:

$$\lim_{d \to \infty} P(B_2|B_1) = 1 - \lim_{x \to 0} \frac{N \cdot x \cdot (1 - x)^{N-1}}{1 - (1 - x)^N} \tag{10}$$

Since both the numerator and denominator tend to zero in the limiting case, we can use L'Hopital's rule to differentiate the numerator and denominator. Therefore, we have:

$$P(B_2|B_1) = 1 - \lim_{x \to 0} \frac{N \cdot (1-x)^{(N-1)} - N \cdot x \cdot (1-x)^{(N-2)}}{N \cdot (1-x)^{(N-1)}}$$

It is easy to verify that this expression evaluates to zero. ∎

The following result follows directly:

Corollary 2 *The limiting probability for achieving k-anonymity in a non-empty set of masked ranges containing a fraction $f < 1$ of the data points is zero. In other words, we have:*

$$lim_{d \to \infty} P(B_k|B_1) = 0 \tag{11}$$

[1] Another way of deriving this would be to simply use the fact that the event of k or more data points occurring in the unit cube is the complementary event to that of less than k points in the unit cube.

This result follows because of our earlier observation that $P(B_k|B_1) \leq P(B_2|B_1)$. In order to derive a further practical understanding of this bound, let us consider some practical values of f. While it is clear that larger values of the population size (denoted by N) and f result in increased privacy, it is interesting to analyze some practical limts on these numbers. Therefore, we will set f and N to the largest practical values possible and calculate the variation of privacy probability with increasing dimensionality. Therefore we will set $f = 0.5$, and N to the values of $3 * 10^8$ and $6 * 10^9$. The latter two values represent the populations of the United States and the earth respectively. In Figure 2, we have plotted the 2-anonymity bound with increasing value of the dimensionality d. It is clear that even for modest values of the dimensionality between 25 and 35, the probability of achieving 2-anonymity within a non-empty grid cell fall off rapidly. Furthermore, we note that these are *upper bounds* for very liberally set values, and represent the probability of 2-anonymity preservation in *each non-empty cell*. In order for privacy to be preserved over the entire data set, the privacy of each non-empty cell must be preserved. Consequently, the overall probability for 2-anonymity preservation would be much lower than that predicted by Figure 2. We note that while these results are derived for uniformly distributed data, they conceptually represent the behavior of the privacy preservation process with increasing *implicit dimensionality* of the data set. In the empirical section, we will also illustrate the cases when correlations are present in the data and show that a very large fraction of the records would continue to violate the privacy requirements. This would require a large level of suppression.

In the previous discussion, we analyzed the privacy requirements for the case of randomly chosen masked attributes. Next, we will analyze the case where the masking can be performed in a more effective way with optimization techniques such as clustering. An example is the anonymization approach of [1] which uses the technique of multi-group cluster formation without the use of bounding rectangles. In the following discussion, we will try to find a lower bound on the information loss for achieving 2-anonymity using any kind of optimized group formation. We will show that in this case, the privacy preservation process requires an unacceptably high loss of information in order to satisfy the anonymity requirements.

We assume that a set S of k data points are merged together in one group for the purpose of condensation. Let $M(S)$ be the maximum euclidian distance between any pair of data points in this group. We note that larger values of $M(S)$ represent a greater loss of information, since the points within a group cannot be distinguished for the purposes of data mining. Similarly, let $M(\mathcal{D})$ represent the corresponding measure for the global database $\mathcal{D}$. This provides us a global base for the overall contrast between different data points. Then, we define the *relative condensation loss* $\mathcal{L}(S)$ for that group of k entities as follows:

Definition 1 *The relative condensation loss $\mathcal{L}(S)$ for the group S is defined as the following ratio:*

$$\mathcal{L}(S) = M(S)/M(\mathcal{D}) \tag{12}$$

Intuitively speaking, the above definition measures how much of the *relative contrast* between the data points (in a group) is lost with respect to the base contrast of the remaining data set. A value of $\mathcal{L}(S)$ which is close to one implies that most of the distinguishing information is lost as a result of the privacy preservation process. We further note that $\mathcal{L}(S)$ represents the *very minimum* level of information loss that any anonymization or condensation technique is likely to be achieve. This is because a particular algorithm for condensation or anonymization may use domain specific considerations [14], which are not always optimal from the information preservation perspective. In the following analysis, we will show how the value of $\mathcal{L}(S)$ is affected by the dimensionality d.

In order to provide a better understanding of the results, we will first analyze the behavior of a uniform distribution of $N = 3$ data points, and deal with the particular case of 2-anonymity. For ease in analysis, we will assume that one of these 3 points is the origin O_d, and the remaining two points are A_d and B_d which are uniformly distributed in the data cube. We also assume that the closest of the two points A_d and B_d need to be merged with O_d in order to preserve 2-anonymity of O_d. Later, we will generalize the results to the case of $N = n$ data points. Since the information loss $\mathcal{L}(\cdot)$ depends upon relative distances among data points, we will start by establishing some convergence results about the distances between A_d, B_d, and O_d in high dimensionality.

Lemma 4 *Let $\mathcal{F}^d$ be uniform distribution of $N = 2$ points. Let us assume that the closest of the 2 points to O_d is merged with O_d to preserve 2-anonymity of the underlying data. Let q_d be the Euclidean distance of O_d to the merged point, and let r_d be the distance of O_d to the remaining point. Then, we have: $lim_{d\to\infty} E\left[r_d - q_d\right] = C$, where C is some constant.*

Proof: Let A_d and B_d be the two points in a d dimensional data distribution such that each coordinate is independently drawn from the data distribution $\mathcal{F}$. Specifically $A_d = (P_1 \ldots P_d)$ and $B_d = (Q_1 \ldots Q_d)$ with P_i and Q_i being drawn from $\mathcal{F}$. Let $PA_d = \{\sum_{i=1}^{d}(P_i)^2\}^{1/2}$ be the distance of A_d to the origin O_d, and $PB_d = \{\sum_{i=1}^{d}(Q_i)^2\}^{1/2}$ the distance of B_d from O_d. The difference of distances is $PA_d - PB_d = \{\sum_{i=1}^{d}(P_i)^2\}^{1/2} - \{\sum_{i=1}^{d}(Q_i)^2\}^{1/2}$.

It can be shown [2] that the random variable $(P_i)^2$ has mean $\frac{1}{3}$ and standard deviation $\left(\frac{2}{3}\right)\sqrt{\left(\frac{1}{5}\right)}$. This means that $(PA_d)^2/d \rightarrow_p 1/3, \quad (PB_d)^k/d \rightarrow_p 1/3$ and therefore we have:

$$PA_d/d^{1/2} \rightarrow_p (1/3)^{1/2}, \quad PB_d/d^{1/2} \rightarrow_p (1/3)^{1/2} \tag{13}$$

We intend to show that $|PA_d - PB_d| \rightarrow_p C'''$ for some constant C'''. We can express $|PA_d - PB_d|$ in the following numerator/denominator form:

$$|PA_d - PB_d| = \frac{|(PA_d)^2 - (PB_d)^2|}{(PA_d) + (PB_d)} \tag{14}$$

Now, we will analyze the convergence behavior of the numerator and denominator individually. By dividing numerator and denominator on RHS by the same value, we get:

$$|PA_d - PB_d| = \frac{|((PA_d)^2 - (PB_d)^2)|/\sqrt{d}}{\frac{PA_d}{d^{1/2}} + \frac{PB_d}{d^{1/2}}} \tag{15}$$

Consequently, using Slutsky's theorem [3] and the results of Equation 13 we obtain

$$\left(\frac{PA_d}{d^{1/2}}\right) + \left(\frac{PB_d}{d^{1/2}}\right) \rightarrow_p 2/\sqrt{3} \tag{16}$$

Having characterized the convergence behavior of the denominator of the right hand side of Equation 15, let us now examine the behavior of the numerator: $|(PA_d)^2 - (PB_d)^2|/\sqrt{d} = |\sum_{i=1}^{d}((P_i)^2 - (Q_i)^2)|/\sqrt{d} = |\sum_{i=1}^{d} R_i|/\sqrt{d}$. Here R_i is the new random variable defined by $((P_i)^2 - (Q_i)^2)\ \forall i \in \{1, \ldots d\}$. This random variable has zero mean and standard deviation which is $\sqrt{2} \cdot \sigma$ where σ is the standard deviation of $(P_i)^2$. The sum of different values of R_i over d dimensions will converge to a normal distribution with mean 0 and standard deviation $\sqrt{2} \cdot \sigma \cdot \sqrt{d}$ because of the central limit theorem. Consequently, the mean average deviation of this distribution will be $C \cdot \sigma$ for some constant C. Therefore, we have:

$$\lim_{d\rightarrow\infty} E\left[\frac{|(PA_d)^2 - (PB_d)^2|}{\sqrt{d}}\right] \leq= C'' \tag{17}$$

Here C'' is a new constant defined by a product of the above mentioned constants. Since the denominator of Equation 15 shows probabilistic convergence to $2/\sqrt{3}$, we can combine the results of Equations 15, 16 and 17 to obtain the following result for some constant C'''=$C'' \cdot \sqrt{3}/2$.

$$\lim_{d\rightarrow\infty} E\left[|PA_d - PB_d|\right] = C''' \tag{18}$$

[2]This is because $E[P_i^2] = 1/3$ and $E[P_i^4] = 1/5$.

[3]**Slutsky's Theorem:** Let $Y_1 \ldots Y_d \ldots$ be a sequence of random vectors and $h(\cdot)$ be a continuous function. If $Y_d \rightarrow_p c$ then $h(Y_d) \rightarrow_p h(c)$.

■

We can easily generalize the result for a database of $N = n$ uniformly distributed points. The following corollary provides the result.

Corollary 3 *Let $\mathcal{F}^d$ be uniform distribution of $N = n$ points. Let us assume that the closest of the n points is merged with O_d to preserve 2-anonymity. Let q_d be the Euclidean distance of O_d to the merged point, and let r_d be the distance of the furthest point from O_d. Then, we have: $C''' \leq lim_{d\rightarrow\infty} E\left[r_d - q_d\right] \leq (n-1) \cdot C'''$, where C''' is some constant.*

Proof: This is because if L is the expected difference between the maximum and minimum of two randomly drawn points, then the same value for n points drawn from the same distribution must be in the range $(L, (n-1) \cdot L)$. ■

A further corollary of the above results is as follows:

Corollary 4 *Let $\mathcal{F}^d$ be uniform distribution of $N = n$ points. Let us assume that the closest of the n points is merged with O_d to preserve 2-anonymity. Let q_d be the Euclidean distance of O_d to the merged point, and let r_d be the distance of the furthest point from O_d. Then, we have: $lim_{d\rightarrow\infty} E\left[\frac{r_d - q_d}{r_d}\right] = 0$, where C''' is some constant.*

Proof: This result can be proved by showing that $r_d \rightarrow_p \sqrt{d}$. Note that the distance of each point to the origin in d-dimensional space increases at this rate. Combining the result with Corollary 3, we see that both the lower and upper bounds on the expression converge to 0. ■

Let S be the two point set represented by O_d and the closest point to O_d. We note that the information loss $M(S)/M(\mathcal{D})$ for 2-anonymity can be expressed[4] as $1 - E\left[\frac{r_d - q_d}{r_d}\right]$. It is easy to see that the value of the information loss converges to 1 in the limiting case in order to achieve 2-anonymity. We also note that the bounds for 2-anonymity also provide lower bounds for the general case of k-anonymity. Therefore, the following result holds:

Theorem 1 *For any set S of data points to achieve k-anonymity, the information loss on the set of points S must satisfy:*

$$lim_{d\rightarrow\infty} E[M(S)/M(\mathcal{D})] = 1 \tag{19}$$

Thus, these results show that with increasing dimensionality, all the discriminatory information in the data is lost in order to achieve k-anonymity. In the next section, we will experimentally examine the behavior of these privacy metrics over a variety of data domains and distributions.

[4]Here we are approximating $M(\mathcal{D})$ to r_d since the origin of the cube is probabilistically expected to be one of extreme corners among the maximum distance pair in the database.

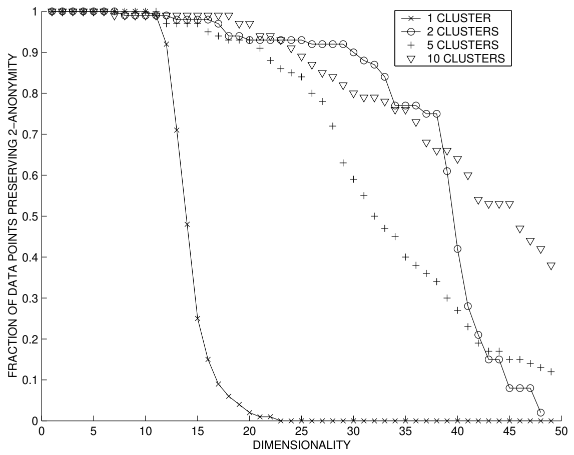


Figure 3: Fraction of Data Points Preserving 2-Anonymity with Data Dimensionality (Gaussian Clusters)

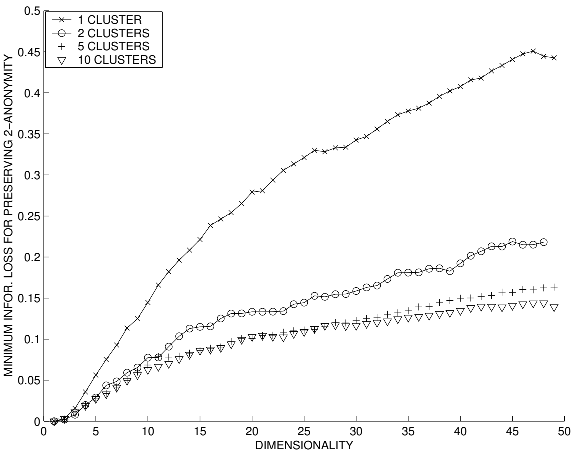


Figure 4: Minimum Information Loss for 2-Anonymity (Gaussian Clusters)

3 Experimental Analysis

In this section, we will provide some experimental analysis of the behavior of the different data sets. We will show that the behavior discussed earlier in this paper is exhibited over a variety of real and synthetic data sets. The synthetic data sets were generated as Gaussian clusters with randomly distributed centers in the unit cube. The radius along each dimension of each of the clusters was a random variable with a mean of 0.075 and standard deviation of 0.025. Thus, a given cluster could be elongated differently along different dimensions by varying the corresponding standard deviation. Each data set was generated with $N = 10000$ data points in a total of 50 dimensions. Finally, the data set was normalized such that the variance along each dimension was 1 unit. We generated the data sets with different numbers (1, 2, 5 and 10) of clusters in order to test the effectiveness of the method with data skew. A larger number of clusters lead to a greater

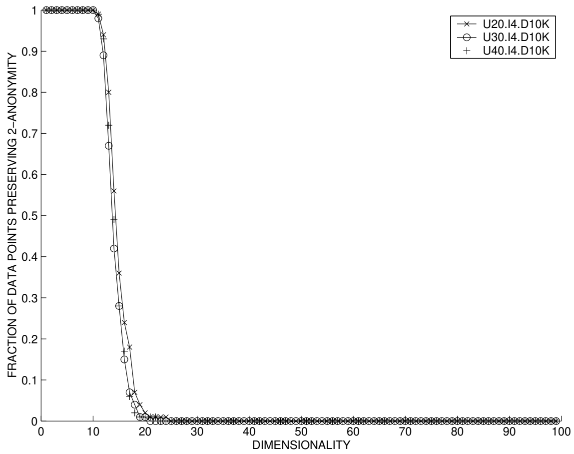


Figure 5: Fraction of Data Points Preserving 2-Anonymity with Data Dimensionality (Market Basket Data)

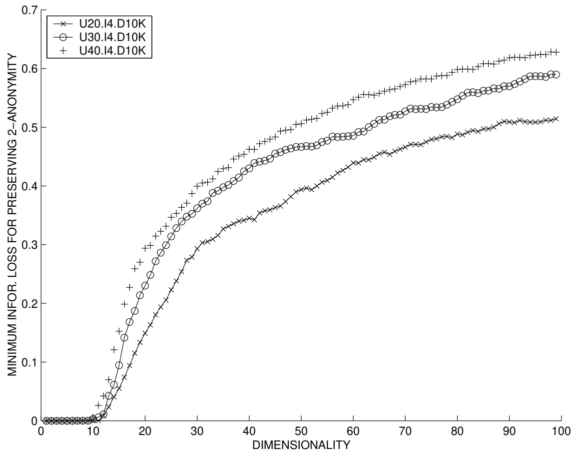


Figure 6: Minimum Information Loss for 2-Anonymity (Market Basket Data)

amount of skew in the data.

We tested the two measures on the bounds for the privacy preservation process using projections of different dimensionality from the generated data set. Since the original data set was 50-dimensional, projections up to 50 dimensions could be generated. In Figure 3, we have illustrated the behavior of a generalization approach in which each attribute is divided into only two ranges. The number of dimensions on the X-axis represents those which are partially specified using these two ranges, whereas all other dimensions are *fully suppressed*. On the Y-axis, we have illustrated the percentage of data points which maintain 2-anonymity using this generalization. We note that all other data points (which violate the 2-anonymity condition) would need to be suppressed. A high percentage of suppression is never acceptable from a data mining point of view [14]. It is interesting to see that while a greater number of clusters (and corresponding skew) in the underlying data helps the anonymization, the percentage of data points which continue to preserve privacy falls off rapidly with increasing data dimensionality. When the data sets contained more than 45 dimensions, *almost all the data points* violated the 2-anonymity condition. Another interesting characteristic of the results of Figure 3 is that for the case of 1 cluster, the shape of the corresponding curve resembles that of Figure 1. The main difference is that in this case, the rate of privacy preservation falls off much more rapidly. This is because the results in Figure 1 only represent upper bounds on the true probability of privacy preservation.

In Figure 4, we have illustrated the minimum information loss for data sets of different dimensionalities. This corresponds to the loss $\mathcal{L}(\cdot)$ as defined earlier in this paper. It is easy to see from Figure 4 that the level of information loss increases rapidly with increasing dimensionality. As in the previous case, the data sets with a smaller number of clusters were more difficult cases. Therefore, the information loss is higher in these cases as well. This is because the presence of greater number of closely clustered regions in the data helps in creating masked groups of anonymized data with lower information loss. However, the overall trends show that even the clustered behavior of the data cannot compensate for the sparsity effects in high dimensionality. This means that either a large portion of the attributes have to be *completely masked* in such cases, or the effectiveness of the anonymization process needs to be compromised. On the other hand, the complete suppression of a large number of attributes reduces the effectiveness of data mining algorithms on the anonymized data.

We also tested the anonymization behavior with a number of market basket data sets. These data sets were generated using the data generator discussed in [5], except that the dimensionality was reduced to only 100 items. This was done in order to moderate the sparsity of the data. This is because most of the interesting variations in privacy behavior are observed within this range. For a larger number of items, the data is too sparse to exhibit any kind of anonymization based privacy. In order to anonymize the data, each customer who bought an item was masked by also including other random customers as buyers of that item. Thus, this experiment to useful to illustrate the effect of our technique on categorical data sets. As a result, for each item, the masked data showed that 50% of the customers had bought it, and the other 50% had not bought it. Using this approach, we checked the probability of a customer preserving 2-anonymity, when an increasing number of items were open to inference attacks. The results are illustrated[5] in Figure 5. It is clear that in each case, the fraction of customers preserving 2-anonymity dropped off rapidly when even 15 to 20 items were open to inference attacks. In the case of the market basket data sets the privacy reduction is much more dramatic than in the Gaussian clustered data set. It is also interesting to note that while the perturbation approach [4] has also been applied to the market basket problem, it has not been tested for robustness in the presence of inference attacks which use a combination of attributes. This is an interesting issue which we will investigate in future work.

In Figure 6, we have illustrated the minimum information loss $\mathcal{L}(\cdot)$ for any condensation strategy preserving 2-anonymity. The results show that an information loss of $50 - 60\%$ is achieved rapidly for even cases where a small set of 50 to 60 items are open to inference attacks. As in the previous figure, the privacy reduction is much more dramatic for the case of the market basket data set. This is because while the market basket data set contains correlations between some subsets of items, an individual transaction may contain many such independent subsets. This opens the data to inference attacks. Thus, the results show that the anonymity model is open to inference attacks when a large number of concepts exist in the data. This also corresponds to a high *implicit dimensionality* of the underlying data. The behavior of privacy preserving data mining algorithms with increasing dimensionality is similar to that of other data mining algorithms, which fail to perform effectively in the high dimensional case because of data sparsity. This sparsity also makes the data more susceptible to inference attacks. Thus, this paper illustrates that the curse of high dimensionality is also relevant to the problem of privacy preserving data mining.

[5]The notations for the data sets in Figures 5 and 6 are the same as those in [5], except that we have replaced T with U in order to denote the fact that the data set contains 100 items instead of 1000 items. Thus $U20.I4.D10K$ represents a transaction with 20 items, potential basket size 4, and 10K records.

4 Conclusions and Summary

This paper discusses the effects of the curse of high dimensionality on privacy preserving data mining algorithms. Since k-anonymity models attempt to retain partial information about different dimensions simultaneously they are more open to inference attacks. This paper shows that in many high-dimensional cases, the level of information loss required in order to preserve even 2-anonymity may not be acceptable from a data mining point of view. This is because the specifics of the inter-attribute behavior have a very powerful revealing effect in the high dimensional case. We also conjecture that in such cases, it may be more effective to use perturbation techniques [4] which do not preserve such inter-attribute information but work with aggregate distributions on individual dimensions. Another possibility is to use selective information hiding in conjunction with conceptual reconstruction techniques [3]. Our future work will analyze the effectiveness of different kinds of privacy models in the high dimensional case.

References

[1] C. C. Aggarwal, 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, A. Hinneburg, D. A. Keim. On the Surprising Behavior of Distance Metrics in High Dimensional Space. *Proceedings of the ICDT Conference*, pp. 420–434, 2001.

[3] C. C. Aggarwal, S. Parthasarathy. Mining Massively Incomplete Data Sets through Conceptual Reconstruction. *Proceedings of the ACM KDD Conference*, pp. 227–232, 2001.

[4] R. Agrawal, R. Srikant. Privacy Preserving Data Mining. *Proceedings of the ACM SIGMOD Conference*, pp. 439–450, 2000.

[5] R. Agrawal, R. Srikant. Fast Algorithms for Mining Association rules in Large Databases. *Proceedings of the VLDB Conference*, pp. 487–499, 1994.

[6] D. Agrawal, 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. J. Bayardo, R. Agrawal. Data Privacy through Optimal k-Anonymization. *Proceedings of the ICDE Conference*, pp. 217–228, 2005.

[8] K. Beyer, J. Goldstein, U. Shaft, R. Ramakrishnan. When is Nearest Neighbors Meaningful? *Proceedings of the ICDT Conference*, pp. 217–235, 1999.

[9] C. Clifton, D. Marks. Security and Privacy Implications of Data Mining. *Proceedings of the ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, pp. 15–19, May 1996.

[10] J. Vaidya, C. Clifton. Privacy Preserving Association Rule Mining in Vertically Partitioned Data. *Proceedings of the ACM KDD Conference*, pp. 639–644, 2002.

[11] L. F. Cranor (Ed.) *Special Issue on Internet Privacy.* Communications of the ACM, 42(2), February 1999. Springer-Verlag, Lecture Notes in Computer Science 1676, 1999.

[12] A. Evfimievski, R. Srikant, R. Agrawal, J. Gehrke. Privacy Preserving Mining Of Association Rules. *Proceedings of the ACM KDD Conference*, pp. 217–228, 2002.

[13] A. Meyerson, R. Williams. On the Complexity of optimal k-anonymity. *Proceedings of the ACM PODS Conference*, pp. 223–228, 2004.

[14] P. Samarati, 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.