On Randomization, Public Information and the Curse of Dimensionality

Charu C. Aggarwal
IBM T. J. Watson Research Center
19 Skyline Drive, Hawthorne, NY 10532
charu@us.ibm.com

Abstract

A key method for privacy preserving data mining is that of randomization. Unlike k-anonymity, this technique does not include public information in the underlying assumptions. In this paper, we will provide a first comprehensive analysis of the randomization method in the presence of public information. We will define a quantification of the randomization method which we refer to as k-randomization of the data. The inclusion of public information in the theoretical analysis of the randomization method results in a number of interesting and insightful conclusions. These conclusions expose some vulnerabilities of the randomization method. We show that the randomization method is unable to effectively achieve privacy in the high dimensional case. We theoretically quantify the degree of randomization required to guarantee privacy as a function of the underlying data dimensionality. Furthermore, we show that the randomization method is susceptible to many natural properties of real data sets such as clusters or outliers. Finally, we show that the use of public information makes the choice of perturbing distribution very critical in a number of subtle ways. Our analysis shows that the inclusion of public information in the analysis makes the goal of privacy preservation more elusive than previously thought for the randomization method.

1 Introduction

In recent years, advances in technology have lead to increased storage of data about individuals by corporations and government entities. This has increased concerns about the possibility of compromising personal information, and has spawned the research area of privacy-preserving data mining [1, 2, 5, 7, 10, 11, 12].

Two important privacy models are those of k-anonymity and randomization. In k-anonymity [11], we reduce the representational accuracy of a record, so that it cannot be linked to less than k public records containing identifier information. In randomization [1, 2], we add a perturbing distribution to the original data. Even though individual record values are distorted, it is possible to accurately reconstruct aggregate distributions and design data mining algorithms which work with these distributions. One nice characteristic of the k-anonymity model is that it is specifically designed to guarantee privacy in the presence of public information. This is not true of randomization, since the added noise is drawn from a fixed distribution. This paper is designed to introduce the analytical effects of public information into the analysis of randomization. Earlier work on randomization [8, 9] uses spectral analysis to approximately reconstruct attribute *values* without the use of public information. However, attribute value approximation is a subtly different goal from *personal* identification with the use of linkage to public databases. To our knowledge, this is the first comprehensive treatment of the randomization method in the presence of public information. We introduce the concept of k-randomization as a tool for measurement, and make the following contributions:

(1) This paper provides a first public-information sensitive methodology to analyze the randomization approach.

(2) As in the case of k-anonymity [3], the effectiveness of randomization degrades rapidly with increasing dimensionality. We quantify the required perturbation to achieve a given privacy level as a function of dimensionality.

(3) The use of public information makes the choice of perturbing distribution more critical than previously thought. We analyze two widely used perturbing distributions (gaussian and uniform) and show that gaussian perturbations have overwhelming advantages in high dimensional cases.

(4) The use of public information in the analysis exposes the susceptibility of the randomization method to many natural properties of real data sets such as clusters or outliers.

(5) The paper demonstrates that the inclusion of public information makes the randomization method vulnerable in unexpected ways. Thus, the goal of privacy preservation may be more elusive than previously thought for the randomization method.

This paper is organized as follows. In the next section,

we discuss how to quantify the risk of disclosure in the presence of public information. In section 3, we analyze the effects of dimensionality and data distribution. In section 4, we present the experimental results. Section 5 discusses the conclusions and discussions.

2 Effects of Public Information

In this section, we will introduce the concepts of likelihood fit and k-randomization which quantify the ability to re-identify the data in the presence of public information. This creates an analogous randomization framework to that of k-anonymity. We assume that the database $\mathcal{D}$ contains N records and d dimensions. The random perturbations for the different dimensions have distributions denoted by $f_{Y_1}(y) \ldots f_{Y_d}(y)$. The corresponding standard deviations of these distributions are denoted by $\sigma_1 \ldots \sigma_d$. Without loss of generality, we may assume that each of the perturbing distributions has zero mean. Let us consider a record $X = (x_1 \ldots x_d)$ to which the perturbation $Y = (y_1 \ldots y_d)$ is added. Then the perturbed data is denoted by $Z = (z_1 \ldots z_d) = (x_1 + y_1, \ldots x_d + y_d)$. Now let us consider an adversary who has access to the publicly available database $\mathcal{D}_p$. Since the perturbing distribution is publicly known, the adversary can calculate the *potential* perturbation of the record Z with respect to each record in the public database $\mathcal{D}_p$. This can be used to calculate the probability that these set of d-dimensional perturbations fit the set of distributions denoted by $f_{Y_1}(y) \ldots f_{Y_d}(y)$. The natural way of calculating the fit of a set of models to a set of observations is the *log-likelihood* fit. In the event that one of the records in the public database has an unusually high degree of fit, this allows the adversary the ability to guess whether the current record truly corresponds to any particular record in the public database.

Let us consider the public record $X = (x_1 \ldots x_d)$. We would like to calculate the likelihood that the perturbed record $Z = (z_1 \ldots z_d)$ corresponds to this publicly available record. In order to do so, the adversary can compute the *potential fit* of the perturbed record to the public database record X. Next, we define the *potential perturbation* of a given record Z to the public database record X.

Definition 2.1 *The potential perturbation $Q(Z, X)$ of a perturbed record $Z = (z_1 \ldots z_d)$ with respect to the public database record $X = (w_1 \ldots w_d)$ is denoted by $Q(Z, X) = (q_1(Z, X) \ldots q_d(Z, X)) = Z - X = (z_1 - x_1 \ldots z_d - x_d)$. The ith component of $Q(Z, X)$ is denoted by $q_i(Z, W) = z_i - x_i$.*

The above definition simply states that in order for the public database record X to correspond to the perturbed record Z, the perturbation for the ith dimension would need to be $q_i(Z, X) = z_i - x_i$. What is the likelihood that the publicly known perturbing distributions $f_{Y_i}(y)$ generate these potential perturbations over the d different dimensions? We note that the log-likelihood that the model $f_{Y_i}(y)$ fits the potential perturbation $q_i(Z, X)$ is given by $\log(f_{Y_i}(q_i(Z, X))) = \log(f_{Y_i}(z_i - x_i))$. We define the corresponding *potential fit* of the dimensions in $Q(Z, X)$ to the distributions denoted by $f_{Y_1}(y) \ldots f_{Y_d}(y)$ as the sum of the log-likelihood fits over the different dimensions.

Definition 2.2 *The potential fit $\mathcal{F}(Z, X)$ of the perturbed data Z to the record X is given by $\sum_{i=1}^{d} \log(f_{Y_i}(q_i(Z, X)))$.*

The higher the value of the log-likelihood fit, the greater the probability that the public database record X corresponds to the perturbed data Z. For a given public database $\mathcal{D}_p$, an adversary can try to match the record in $\mathcal{D}_p$ which has the highest level of fit to the perturbed record Z. We observe that the log likelihood fit is an indirect representation of the Bayes a-posteriori probability that the perturbed data record fits a particular record X:

Observation 2.1 *Consider a database $\mathcal{D}_p$ which is known to contain the true representation of the perturbed record Z with equal a-priori probability. Then the posterior probability $\mathcal{B}(Z, X, \mathcal{D}_p)$ of a particular record $X \in \mathcal{D}_p$ to correspond to Z is given by:*

$$\mathcal{B}(Z, X, \mathcal{D}_p) = \frac{e^{\mathcal{F}(Z,X)}}{\sum_{V \in \mathcal{D}_p} e^{\mathcal{F}(Z,V)}} \tag{1}$$

The above observation is easy to verify, since the perturbations over different dimensions are independent and the value of $e^{\mathcal{F}(Z,X)}$ is simply equal to the product of the corresponding probability densities. By applying the Bayes formula in conjunction with equal a-priori probability, we get the desired result. Thus, the log likelihood is an indirect representation of the Bayes probability, and the use of this particular representation is chosen for the sake of numerical and algebraic convenience.

In many cases, the log likelihood fit can provide considerable insights to an adversary in including or excluding particular database records. For example, the log likelihood fit may be a significantly better fit to one record in the public database compared to any other record. In such a case, the corresponding Bayes probability $\mathcal{B}(Z, X, \mathcal{D}_p)$ may approach 1, and the said record can be identified to a high degree of probability. Therefore, anonymity is lost. Another extreme case is one in which the perturbing distribution has a finite range (such as the uniform distribution), and the value of $f_{Y_i}(q_i(Z, X))$ to be zero. In such a case, the corresponding log likelihood fit is $-\infty$, and it is possible to exclude the record X as a fit with Z.

In general, we would like the perturbation to be sufficient, so that at least some other spurious records in the data

set have a higher fit to the correct public database record than the true record. Larger perturbations reduce the log-likelihood fit of the true record $X \in \mathcal{D}$ corresponding to Z, and increase the probability that another spurious record in $\mathcal{D}$ may have a higher log-likelihood fit than X by chance. This is desirable from the point of view of privacy preservation. When there are at least k records in $\mathcal{D}$ which have higher (or equal) log likelihood fit than X, then the record X is said to be k-randomized. In such a case, no public database can be used to distinguish X from the k other records within $\mathcal{D}$ which are a better fit to the randomized representation of X. Now, we will define the concept of k-randomization formally.

Definition 2.3 *A (randomized) record $Z \in \mathcal{D}$ with original representation X is said to be k-randomized when there are at least k records $\{X_1 \ldots X_k\} \in \mathcal{D}$ for which the following is true:*

$$\mathcal{F}(Z, X) \leq \mathcal{F}(Z, X_i) \tag{2}$$

This means that the randomized record Z cannot be used to distinguish its true representation X from the k records $X_1 \ldots X_k$ in $\mathcal{D}$. By performing k-randomization of every record in the database $\mathcal{D}$, it is possible to achieve an equivalent level of k-anonymity for the randomization approach. However, since the randomization approach does not use a trusted server and can be performed at *data collection time* (without knowledge of other records), the exact level of randomization may not be known or precisely controlled a-priori. This is different from the k-anonymity model which performs the privacy transformation in a controlled way so as to explicitly *engineer* k-anonymity. Here, our aim in defining the randomization level of a record is to use it as an *analytical tool* for judging the effectiveness of a given level of perturbation. The only a-priori control parameter is the perturbation standard deviation, and the randomization level is computed a-posteriori. Thus, the *calculated* randomization level of a point X is denoted by $kr(X)$ and is equal to the number of randomized points in the database which fit the randomized version of X at least as well as (the randomized representation of) X itself. We make the following observation about the expected value of $kr(X)$:

Observation 2.2 *Let $X = (x_1 \ldots x_d)$ be a d-dimensional point from the database $\mathcal{D}$. Let $Z = (z_1 \ldots z_d)$ represent the randomization of X. Then, the expected randomization level $E[kr(X)]$ is as follows:*

$$E[kr(X)] = \sum_{X' \in \mathcal{D}} P(\mathcal{F}(Z, X') \geq \mathcal{F}(Z, X)) \tag{3}$$

As in the case of k-anonymity, this value is at least 1 to account for the case when $X' = X$. Next, we generalize the point specific randomization level to the entire database.

Definition 2.4 *The average randomization level of the* database $\mathcal{D}$ *is defined as the average value of $kr(X)$ over all points in $\mathcal{D}$.*

Since the calculated randomization level $kr(X)$ may vary with data point X, we also define a worst-case quantification. In this context, we define the randomization level at *quantile* q.

Definition 2.5 *The randomization level of database $\mathcal{D}$ at quantile q is computed as the lowest quantile q of the randomization level array $kr(\cdot)$.*

The average and worst case behaviors provide different kinds of insights. In the next section, we will use these quantifications to analyze the effects of different kinds of data sets, dimensionality, and perturbing distributions.

3 Effects of High Dimensionality

In this section, we will analyze the effect of different perturbing distributions on the effectiveness of randomization. We will also analyze the effects of dimensionality on the effectiveness of randomization. The two most common distributions used for perturbation are the uniform and the gaussian distribution [1]. In this section, we will analyze the effects of both.

3.1 Gaussian Perturbing Distribution

The gaussian perturbation with standard deviation σ_i on the ith dimension is defined as follows:

$$f_Y(y) = \frac{1}{\sqrt{2 \cdot \pi} \sigma_i} e^{-\frac{y^2}{2 \cdot \sigma_i^2}} \tag{4}$$

Let us consider the record $X = (x_1 \ldots x_d)$ which is perturbed to the randomized record denoted by $Z = (z_1 \ldots z_d)$. Then, the log likelihood fit $\mathcal{F}(Z, X)$ is given by $\mathcal{F}(Z, X) = \sum_{i=1}^{d} \log(f_{Y_i}(q_i(Z, X))) = \sum_{i=1}^{d} \log(f_{Y_i}(z_i - x_i))$. By substituting the value of $f_{Y_i}(y)$ according to Equation 4, we get:

$$\mathcal{F}(Z, X) = -(d/2) \cdot \log(2 \cdot \pi) - \sum_{i=1}^{d} \log(\sigma_i) - \sum_{i=1}^{d} \frac{(z_i - x_i)^2}{2 \cdot \sigma_i^2} \tag{5}$$

Let us now consider another record $X' = (x'_1 \ldots x'_d) \in \mathcal{D}$ which is in the neighborhood of X. We would like to calculate the probability that the likelihood fit $\mathcal{F}(Z, X')$ is at least equal to that of $\mathcal{F}(Z, X)$. As evident from Observation 2.2, this probability $P(\mathcal{F}(Z, X') \geq \mathcal{F}(Z, X))$ plays a key role in defining the expected randomization level $E[kr(X)]$. Therefore, our future analysis will quantify the value of $P(\mathcal{F}(Z, X') \geq \mathcal{F}(Z, X))$. We will show the following result about this probability:

Lemma 3.1 *Let $X = (x_1 \ldots x_d)$ and $X' = (x'_1 \ldots x'_d)$ be the two d-dimensional points from the database $\mathcal{D}$, such that $\Delta = (\delta_1 \ldots \delta_d) = X - X'$. Let $Z = (z_1 \ldots z_d)$ represents the randomization of X and σ_i^2 be the variance of the gaussian perturbation along the ith dimension. Then, we have:*

$$P(\mathcal{F}(Z, X') \geq \mathcal{F}(Z, X)) = P\left(\sum_{i=1}^{d} \delta_i^2/(2\sigma_i^2) \leq \sum_{i=1}^{d} -\delta_i \cdot y_i/\sigma_i^2\right) \tag{6}$$

Here y_i is the random variable representing the gaussian perturbation along the ith dimension.

Proof: By substituting the values of $\mathcal{F}(Z, X)$ and $\mathcal{F}(Z, X')$ from Equation 5, and canceling the common terms, we get:

$$P(\mathcal{F}(Z, X') \geq \mathcal{F}(Z, X)) =$$
$$= P\left(\sum_{i=1}^{d} -(z_i - x'_i)^2/\sigma_i^2 \geq \sum_{i=1}^{d} -(z_i - x_i)^2/\sigma_i^2\right)$$
$$= P\left(\sum_{i=1}^{d} (z_i - x_i + \delta_i)^2/\sigma_i^2 \leq \sum_{i=1}^{d} (z_i - x_i)^2/\sigma_i^2\right)$$

The last relationship is obtained by replacing $X' = X - \Delta$, and reversing the sign of the inequality by negating both sides. Now, we note that $z_i - x_i$ is simply the value of the random perturbation y_i which is derived from a gaussian distribution. Therefore, let us replace $z_i - x_i$ by y_i for algebraic convenience. Therefore, we have:

$$P(\mathcal{F}(Z, X') \geq \mathcal{F}(Z, X)) = P\left(\sum_{i=1}^{d} (y_i + \delta_i)^2/\sigma_i^2 \leq \sum_{i=1}^{2} y_i^2/\sigma_i^2\right) \tag{7}$$
$$= P\left(\sum_{i=1}^{d} \delta_i^2/(2 \cdot \sigma_i^2) \leq -\sum_{i=1}^{d} \delta_i \cdot y_i/\sigma_i^2\right)$$

The last relationship is obtained by simple algebraic expansion of $(y_i + \delta_i)^2$ and subsequent simplification. ■

While the above lemma provides an algebraic expression for this bound, a more intuitive interpretation with respect to dimensionality and distribution needs to be constructed. In order to do so, we will make use of the well known Chebychev inequality. First, we will prove a simple lemma which we will need in a later section.

Lemma 3.2 *Let y_i be the gaussian perturbation along the ith dimension with variance σ_i^2. Let $V = -\sum_{i=1}^{d} y_i \cdot \delta_i/\sigma_i^2$. Then, we have:*

$$E[V^2] = \sum_{i=1}^{d} \delta_i^2/\sigma_i^2 \tag{8}$$

Proof: We note that $y_1 \ldots y_d$ are independent perturbations along the d dimensions. Therefore, by expanding the expression for V^2, and using independence to simplify expectation of products of random variables, we get:

$$E[V^2] = \sum_{i=1}^{d} \delta_i^2 \cdot E[y_i^2]/\sigma_i^4 + 2 \cdot \sum_{i=1}^{d} \sum_{j=i+1}^{d} \delta_i \cdot \delta_j \cdot E[y_i] \cdot E[y_j]/(\sigma_i^2 \cdot \sigma_j^2) \tag{9}$$

Since y_i is a gaussian with variance σ_i^2 about a mean of zero, we have $E[y_i] = 0$ and $E[y_i^2] = \sigma_i^2$. By substituting this in Equation 9, we get the desired result. ■

Theorem 3.1 *Let $X = (x_1 \ldots x_d)$ and $X' = (x'_1 \ldots x'_d)$ be two d-dimensional points from the database $\mathcal{D}$, such that $\Delta = (\delta_1 \ldots \delta_d) = X - X'$. Let Z represent the randomization of X and σ_i^2 be the variance of the gaussian perturbation along the ith dimension. Then, we have:*

$$P(\mathcal{F}(Z, X') \geq \mathcal{F}(Z, X)) \leq 4/(\sum_{i=1}^{d} \delta_i^2/\sigma_i^2) \tag{10}$$

Proof: As in Lemma 3.2, let us define $V = -\sum y_i \cdot \delta_i/\sigma_i^2$. From Lemma 3.1, we get:

$$P(\mathcal{F}(Z, X') \geq \mathcal{F}(Z, X)) = P\left(\sum_{i=1}^{d} \delta_i^2/(2 \cdot \sigma_i^2) \leq V\right) \tag{11}$$
$$\leq P\left(V^2 \geq (\textstyle\sum_{i=1}^{d} \delta_i^2/(2 \cdot \sigma_i^2))^2\right) \text{ (squaring both sides and recognizing that } \delta_i^2/(2 \cdot \sigma_i^2) \text{ is always positive)}$$
$$\leq E[V^2]/(\textstyle\sum_{i=1}^{d} \delta_i^2/(2 \cdot \sigma_i^2))^2 \text{ (Chebychev Inequality)}$$

By substituting the expression for $E[V^2]$ from Lemma 3.2, we get the desired result. ■

We note that the variance of the perturbing distribution along each dimension is typically chosen proportional to the corresponding variance of the original data. This is a natural choice in order to provide a similar level of perturbation over the different dimensions.

Assumption 3.1 Proportionality Assumption: *If the variance of the original data along the ith dimension is denoted by σ_i^o, then the perturbing variance σ_i is chosen such that $C_1 \cdot \sigma_i \leq \sigma_i^o \leq C_2 \cdot \sigma_i$ for some constants C_1 and C_2.*

The proportionality assumption automatically helps us reword the results of Theorem 3.1 as follows:

Theorem 3.2 *Let $X = (x_1 \ldots x_d)$ and $X' = (x'_1 \ldots x'_d)$ be two d-dimensional points from the database $\mathcal{D}$, such that $\Delta = (\delta_1 \ldots \delta_d) = X - X'$. Let $Z = (z_1 \ldots z_d)$ represents the randomization of X. Let σ_i^2 be the variance of the gaussian perturbation along the ith dimension, and $(\sigma_i^o)^2$ be the variance of the original data along dimension i. Then, under the proportionality assumption, for some constant C_3, we have:*

$$P(\mathcal{F}(Z, X') \geq \mathcal{F}(Z, X)) \leq C_3/(\sum_{i=1}^{d} \delta_i^2/(\sigma_i^o)^2) \tag{12}$$

We note that denominator of the right hand side of the relationship of Theorem 3.2 contains the term $\sum_{i=1}^{d} \delta_i^2/(\sigma_i^o)^2)$. This is simply the distance between X and X', when the original data is normalized by the variance along each dimension. Therefore, it is intuitively clear that a data point X' which is spatially close to X has a higher chance of satisfying the requirement $\mathcal{F}(Z, X') \geq \mathcal{F}(Z, X)$ which increases the randomization level of X. However, with increasing dimensionality, the concept of spatial locality becomes more problematic. According to [6], the sparsity of high dimensional data ensures that the distance to other points in the data $\sum_{i=1}^{d} \delta_i^2/(\sigma_i^o)^2)$ grows with d^* in high dimensional space, where d^* is the implicit dimensionality of the data. Therefore, even if X' is chosen to be the nearest neighbor of X, the value of $P(\mathcal{F}(Z, X') \geq \mathcal{F}(Z, X))$ tends to zero with increasing value of d. From Observation 2.2, the expected randomization level $E[kr(X)]$ is critically dependent upon this probability, and therefore, the randomization level of X also reduces with increasing dimensionality. We summarize this result as follows:

Conclusion 3.1 *The expected randomization level reduces with increasing dimensionality for a fixed level of perturbation.*

How strong is this revealing effect of high dimensionality? We note that the Chebychev inequality is extremely weak in practice. Therefore, the above results represent a fairly weak bound. In practice, it is possible to get much tighter bounds with the use of a few approximations on Lemma 3.1. We note that the right hand side of Lemma 3.1 contains $V = -\sum_{i=1}^{d} y_i \cdot \delta_i/\sigma_i^2$. Since each y_i is independent, the variance of V is equal to the sum of the individual variances. This works out to $\sigma^2(V) = \sum_{i=1}^{d} \delta_i^2/\sigma_i^2$. We further note that $E[V] = 0$. Now, we make the approximation that V is normally distributed. This may be fairly close to the truth for large values of d, since each component of V (which is $-y_i \cdot \delta_i/\sigma_i^2$) is a unit normal distribution scaled by δ_i/σ_i.

The right hand side of Lemma 3.1 can be expressed as $P(V \geq \sum_{i=1}^{d} \delta_i^2/(2 \cdot \sigma_i^2)) = 1 - \Phi((\sum_{i=1}^{d} \delta_i^2/(2 \cdot \sigma_i^2))/\sigma(V))$. Here $\Phi(\cdot)$ is the cumulative normal distribution. Since $\sigma(V) = \sqrt{\sum_{i=1}^{d} \delta_i^2/\sigma_i^2}$, we can summarize as follows:

Approximation 3.1 *Let $X = (x_1 \ldots x_d)$ and $X' = (x_1' \ldots x_d')$ be two d-dimensional points from the database $\mathcal{D}$, such that $\Delta = (\delta_1 \ldots \delta_d) = X - X'$. Let $Z = (z_1 \ldots z_d)$ represents the randomization of X. Let σ_i^2 be the variance of the gaussian perturbation along the ith dimension. Then, we have:*

$$P(\mathcal{F}(Z, X') \geq \mathcal{F}(Z, X)) = 1 - \Phi((\sqrt{\sum_{i=1}^{d} \delta_i^2/\sigma_i^2})/2) \tag{13}$$

Here $\Phi(\cdot)$ is the cumulative normal distribution. The corresponding expected randomization level of the data point X is obtained by summing $P(\mathcal{F}(Z, X') \geq \mathcal{F}(Z, X))$ over all points $X' \neq X$ in the database $\mathcal{D}$.

We note that the cumulative normal distribution $\Phi(\cdot)$ is approximately equal to 1 for an argument value greater than 3. Therefore, the expression $\sum_{i=1}^{d} \delta_i^2/\sigma_i^2$ needs to be at most 36 in order for the probability $P(\mathcal{F}(Z, X') \geq \mathcal{F}(Z, X))$ to not be (nearly) zero. Consider the case of a uniformly distributed data set in which we pick $\sigma_i = C \cdot \sigma^o$. In such a case, we can show [6] that the distance value $\sum_{i=1}^{d} \delta_i^2/\sigma_i^2$ grows as d/C^2. This means that C must grow with $\sqrt{d}$ in order for the probability $P(\mathcal{F}(Z, X') \geq \mathcal{F}(Z, X))$ to be significantly larger than zero. Since Observation 2.2 ties the probability $P(\mathcal{F}(Z, X') \geq \mathcal{F}(Z, X))$ to the expected randomization level $E(kr(X)]$, this indicates that the value of C should grow with $\sqrt{d}$ for the randomization level to be constant with increasing dimensionality. While the result of [6] is true for the case of uniform distribution of the original data, it provides the intuition that the perturbing standard deviation along each dimension should grow as the square root of the *implicit dimensionality* of the data. We summarize this result as follows:

Conclusion 3.2 *Under the proportionality assumption, the perturbing gaussian distribution along each dimension should have a standard deviation which grows with the square root of the implicit dimensionality of the underlying data in order to retain the same level of randomization.*

In practice, only a small number of data points X' (which lie in the locality of X) are likely to have dominant values for $P(\mathcal{F}(Z, X') \geq \mathcal{F}(Z, X))$ in the right hand side of Observation 2.2. The value of each of these terms depend inversely upon the normalized distance $\sum_{i=1}^{d} \delta_i^2/(\sigma_i^o)^2$ between X and X'. Thus, for data sets with the same global variance, the expected randomization level $E[kr(X)]$ is likely to be higher when non-empty localities of the data are dense and highly clustered. This provides the following result:

Conclusion 3.3 *The presence of clusters is helpful in increasing the randomization level for data sets with similar global variance.*

This is a nice property of the randomization method, since most real data sets exhibit clustered behavior. We further note that while Approximation 3.1 provides an understanding of the randomization level of each data point, it may

often be more desirable to examine the worst-case randomization behavior of the entire data set. As discussed earlier, the *local* magnitudes of the normalized distances $\sum_{i=1}^{d}(\delta_i/\sigma_i^o)^2$ have a strong inverse relationship with the expected randomization level $E[kr(X)]$. Therefore, for data sets with the same global variance, a variation in the local density distribution can affect the worst-case randomization more sharply.

Conclusion 3.4 *A data set with varying density distribution is likely to have a significantly lower worst-case randomization level than the average randomization level.*

The presence of outliers is the extreme case, since the density within the locality of an outlier is significantly lower than the average case density.

Conclusion 3.5 *The presence of outliers may reduce the worst-case randomization level without significantly affecting the average-case randomization behavior of the data.*

These results show that the randomization approach is susceptible to the presence of the density variations and outliers. The intuition for this is that unlike methods such as k-anonymity, the current methods for randomization of individual data points are applied without assumption of knowledge about the rest of the data. This is an issue which needs to be addressed in future research on randomization.

3.2 Uniform Perturbing Distribution

We assume that the perturbation along the ith dimension is uniformly distributed with range $[0, a_i]$, and the corresponding standard deviation σ_i is equal to $a_i/\sqrt{12}$. For simplicity, we assume that the range of the perturbation a_i is larger than the range of the non-perturbed data along dimension i. This is not really restrictive, since it is needed to preserve a minimum level of privacy along the ith dimension. Therefore, if $\Delta = (\delta_1 \ldots \delta_d) = X - X'$, we must have $|\delta_i| \leq a_i$.

Theorem 3.3 *Let $X = (x_1 \ldots x_d)$ and $X' = (x'_1 \ldots x'_d)$ be two d-dimensional points from the randomized database $\mathcal{D}$, such that $\Delta = (\delta_1 \ldots \delta_d) = X - X'$ and $Z = (z_1 \ldots z_d)$ represents the randomization of X. Let $[0, a_i]$ be the range of the uniform perturbation along the ith dimension. Then, we have:*

$$P(\mathcal{F}(Z, X') \geq \mathcal{F}(Z, X)) = \pi_{i=1}^{d}(1 - |\delta_i|/a_i) \quad (14)$$

Proof: Since the distribution is uniform with density $1/a_i$, the value of $\mathcal{F}(Z, X)$ is simply $d \cdot \log(1/a_i)$. Now we note that the value of $\mathcal{F}(Z, X')$ is defined as follows:

$$\mathcal{F}(Z, X') = \sum_{i=1}^{d} \log(f_Y(z_i - x'_i)) = \quad (15)$$

$$= \sum_{i=1}^{d} \log(f_Y(z_i - x_i + \delta_i)) = \sum_{i=1}^{d} \log(f_Y(y_i + \delta_i))$$

Here y_i is the uniformly distributed perturbation in the range $[0, a_i]$. We note that each of the d terms on the right hand side is either $\log(1/a_i)$ or $-\infty$ depending upon whether or not $(y_i + \delta_i)$ lies in the range $[0, a_i]$. Therefore $\mathcal{F}(Z, X')$ can *never* be larger than $\mathcal{F}(Z, X)$. The value of $\mathcal{F}(Z, X')$ can at most be equal to $\mathcal{F}(Z, X)$, if and only if for each and every dimension i, $y_i + \delta_i$ lies in the range $[0, a_i]$. Since y_i is uniformly distributed in the range $[0, a_i]$, it is easy to verify that the probability of $(y_i + \delta_i)$ lying in the range $[0, a_i]$ is $(1 - |\delta_i|/a_i)$. By using the independence of the different values of y_i, the result follows. ■

A simple corollary of the above result is as follows;

Corollary 3.1 *Let $X = (x_1 \ldots x_d)$ and $X' = (x'_1 \ldots x'_d)$ be two d-dimensional points from the randomized database $\mathcal{D}$, such that $\Delta = (\delta_1 \ldots \delta_d) = X - X'$ and $Z = (z_1 \ldots z_d)$ represents the randomization of X. Let $[0, a_i]$ be the range of the uniform perturbation along the ith dimension. Then, we have:*

$$P(\mathcal{F}(Z, X') \geq \mathcal{F}(Z, X)) \leq (1 - \sum_{i=1}^{d}(|\delta_i|/(d \cdot a_i)))^d \quad (16)$$

Proof: This corollary simply follows from Theorem 3.3 and the fact that the geometric mean of a set of non-negative values is at most equal to the arithmetic mean. ■

As in the previous case, let us examine what happens in a uniformly distributed data set, when the range a_i is chosen to be $C \cdot \sigma_i^o$ for some constant C using the proportionality assumption. In this case, the results of [6] indicate that in the high dimensional case, $\sum_{i=1}^{d} |\delta_i|/\sigma_i^o$ is expected to increase as $B \cdot d$ for some constant B. Then, we can use the result of Corollary 3.1 to derive the following;

$$P(\mathcal{F}(Z, X') \geq \mathcal{F}(Z, X)) \leq (1 - B/C)^d \quad (17)$$

Note that when the value of C is chosen to be $B \cdot d$, the value of the above expression is $(1 - 1/d)^d$. This is bounded above by $1/e$, where e is the base of the natural logarithm. By choosing C smaller than $B \cdot d$, it is possible for this probability $P(\mathcal{F}(Z, X') \geq \mathcal{F}(Z, X))$ to fall off rapidly to zero. This would result in lower randomization levels. We summarize as follows:

Conclusion 3.6 *Under the proportionality assumption, the perturbing uniform distribution along each dimension should have a range (or standard deviation) which grows* **at least** *linearly with the implicit dimensionality of the underlying data.*

Recall that the in the case of the gaussian distribution, the required standard deviation grows proportionally only with

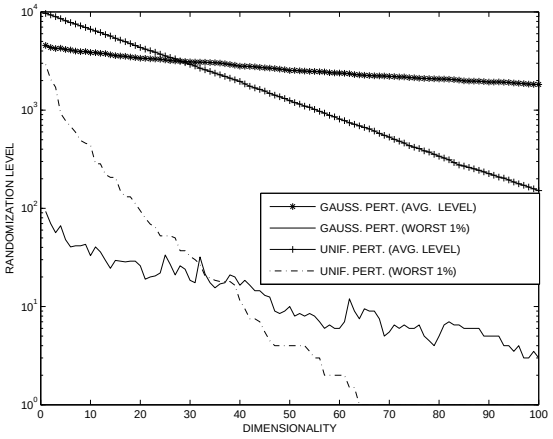

Figure 1. Randomization Level with Increasing Dimensionality, Perturbation level$= 8 \cdot \sigma^o$ ($UniDis$)

the *square-root* of the dimensionality. Therefore, a greater level of randomization (and hence information loss) may be sustained when the uniform distribution is used. We also emphasize that unlike the case of the gaussian distribution, the above intuition on the choice of the perturbing distribution is only a lower bound. This is because of the use of the inequality between the geometric and arithmetic mean. This inequality can be extremely loose in practice, when the values of $|\delta_i|/a_i$ are very different from one another over the different dimensions. Therefore, even the lower bound on the standard deviation of the perturbing distribution for the uniformly distributed case is significantly higher than the required standard deviation for the gaussian distribution. We summarize as follows:

Conclusion 3.7 *In the high dimensional case, gaussian perturbations provide higher randomization than the uniform perturbation.*

4 Experimental Analysis

We used a number of synthetic data sets for experimental analysis. In each case, we normalized the variance of each dimension to one unit by scaling. This is helpful in testing the relative effectiveness on different data sets. In each synthetic data set, a 100-dimensional base data set was generated and $N = 10000$ data points were generated. We will test the effects of varying dimensionality by picking projections of different dimensionality from the base data.

4.1 Data Sets

. The aim of generating different data sets was to expose the effects of using different kinds of data distributions on the testing process. The data sets generated were as follows:

(1) We generated a uniformly distributed set of points in the unit cube. The variance along each dimension was normalized to one unit by scaling. We denote this data set as *UniDis*.
(2) In order to test the effects of data skew, we generated a clustered data set. The centroids of $p = 5$ clusters were chosen randomly in the unit cube. Each cluster containing 2000 points, and the radius along each dimension was uniformly picked from the range $[0, 0.1]$. The clusters were generated from a gaussian distribution with the corresponding radii along each dimension. Once the data set was generated, we normalized the variance along each dimension to one unit by scaling. We refer to this data set as $EGauDis$.
(3) In order to test the effect of varying density skew, we generated a data set which was the exactly similar to $EGauDis$ in terms of cluster centroid and radii generation along each of the dimensions. The only difference was that the number of points in the ith cluster was proportional to $1/i^\theta$, where θ was the Zipf parameter. Note that the use of $\theta = 0$ creates $EGauDis$. We denote this data set by $VGauDis(\theta)$. As in the previous cases, we normalized each dimension after data set generation. The default value of θ used was 1.
(4) In order to test the effect of outliers, we again generated a data set which was the exactly similar to the data set $EGauDis$ in terms of number and position of cluster centroids, and the radii along each of the dimensions for the different clusters. The only difference was that a fraction f of the data points were picked as outliers, whereas the remaining data points were evenly distributed among the different clusters. We denote this data set by $OGauDis(f)$. Note that a choice of $f = 0$ yields the data set $EGauDis$. Unless otherwise mentioned, the value of f used was 0.1.

In combination with the use of the above data sets for the base distribution, we tested both the uniform and gaussian perturbing distributions. We assume that all dimensions have the same perturbing variance. This follows from the similarity assumption, since the variance of all base data sets were normalized to one unit along each dimension. Therefore, the standard deviation of the perturbing distribution exhibited the same proportional behavior with different data sets.

4.2 Measures

In order to test the privacy effectiveness of a given perturbation, we utilized two measures:

(1) Average Randomization Level: For each data point X, we calculated the number of data points (includ-

ing itself) which had a maximum likelihood fit which was *at least* equal to it. This is the randomization level $kr(X)$. Note that the lowest possible value of $kr(X)$ is 1. Then, the average randomization level $\mathcal{AR}(\mathcal{D})$ of the data set $\mathcal{D}$ is defined as follows:

$$\mathcal{AR}(\mathcal{D}) = \sum_{X \in \mathcal{D}} kr(X)/|\mathcal{D}| \tag{18}$$

This measure computes the average anonymity level of the records in the data.

(2) Worst Case Randomization Level: We calculated the worst q-quantile of the array $kr(X)$, and defined this as the worst case randomization $\mathcal{WR}(\mathcal{D})$. This measure calculates the maximum number of records which fit a perturbed record at least as well as the true record among the fraction q of the data which has the smallest randomization level. In many applications, the worst-case behavior may be a more important measure because the ability to discover even a small fraction of the data may be undesirable.

The average and worst case randomization provide different kinds of insights into the behavior of different data and perturbing distributions. In particular, cases in which worst case randomization level is significantly lower than the average case are interesting from the perspective of exposing the difficulty of preserving privacy in certain kinds of data sets.

4.3 Experimental Results

Our evaluation will examine the behavior of the randomization level with respect to the effect of using different data distributions, increasing dimensionality, density and outlier behavior of the data set. This particular design was chosen in order to verify the different intuitions about natural characteristics of base data sets and perturbing distributions. We first tested the effect of data dimensionality on the perturbation effectiveness. In Figure 1, we have illustrated the effect of increasing dimensionality on the $UniDis$ data set. These results were obtained by applying the technique to projections of the data of different dimensionality. The X-axis on each chart illustrates the data dimensionality, whereas the Y-axis illustrates both the average and worst case randomization levels $\mathcal{AR}(\mathcal{D})$ and $\mathcal{WR}(\mathcal{D})$ respectively for different perturbing distributions. Since the randomization level varied widely for different data sets, distributions, and dimensionalities, we made it a point to use a logarithmic scale on the Y-axis. As pointed out earlier, the base data sets were normalized so that the variance along each dimension was $\sigma^o = 1$. The corresponding perturbation variance was set of $8 \cdot \sigma^o$ in each case. Since the variance of the original data set was always the same, this set of charts helps us compare the relative behavior of different data sets and perturbing distributions with varying dimensionality.

One immediate observation from each of the (logarithmically scaled) charts in Figures 1 was that both the average and worst case randomization levels reduced rapidly with increasing dimensionality for different data sets. For example, in Figure 1, the average randomization (with uniform perturbations) for the 1-dimensional data set was 9646.1, whereas the average randomization level for the 100-dimensional case was 151.7. This means that for the 1-dimensional case, 96.46% of the original 10,000 points fit a given data point as well as the true point. On the other hand, this number reduced to only 1.51% in the 100-dimensional case. Even more interesting behavior was observed by examining the lowest 1% quantile of the data. This corresponds to $\mathcal{WR}(\mathcal{D})$. In this case, the randomization level was 2907 for the 1-dimensional case. *However, for any instantiation of the data set beyond 64 dimensions , the randomization level was only 1 for the entire lower* 1% *quantile of the data*, when uniform perturbations were used. We note that a randomization level of 1 denotes no privacy, since the data point itself contributes to a randomization level of 1. This behavior was specific to the uniform perturbing distribution, and happened in spite of a high perturbation level of $8 \cdot \sigma^o$ for each dimension.

As evident from Figure 1, the behavior of the gaussian perturbing distribution was much more robust with increasing dimensionality, even though the uniform perturbation turned out to be superior for the lower dimensional cases. For example, for the 1-dimensional case in Figure 1, the average randomization level for the gaussian perturbing distribution 4552.2 which was less than half the randomization level of the uniform distribution. However, when the dimensionality increases to 100, the average randomization level was 1824.4, which was more than an order of magnitude higher than the randomization level 151.7 for the uniform perturbations. An even more interesting case was the behavior of the worst 1%-quantile of the data. While the uniform perturbation had no privacy of the worst 1% quantile for dimensionalities beyond 64, the gaussian perturbation had a randomization level of between 5 and 10 for dimensionalities higher then 64. Thus, the results show that while the curse of dimensionality results in a reduction of privacy with increasing dimensionality, the effect was more pronounced in the uniformly distributed case. Since a better choice of perturbing distribution seems to moderate the effects of the dimensionality curse, this underlines the importance of judiciously choosing the perturbing distribution in the randomization method.

In Figure 2, in which we have illustrated the randomization level of the data set $VGauDis(\theta)$ with increasing level of skew θ. In this case, we used a perturbation level of $8 \cdot \sigma^o$, and a dimensionality of 75. A rather curious pattern emerges when we closely examine this Figure. It is clear that the average randomization levels increased with

skew, whereas the lower 1%-quantile of randomization levels worsened with increasing skew. For example, when gaussian perturbations were used, the worst-case randomization level reduced from 24 to 13 with increasing level of skew. However, the average case randomization level increased from 2353.0 to 2773.9. *The results show that the average randomization levels for the skewed data set $VGauDis(1)$ were greater than those of $EGauDis$, but randomization level of the lower 1%-quantile was lower for the skewed data set.* The explanation for this curious anomaly may be found in the fact that the randomization level of a data point depends upon its *local density*. In the case of the skewed distribution, there were always a few points with very low local density (particularly the ones belonging to the cluster with fewest points). These points had low randomization level, and therefore contributed to poor worst-case behavior. On the other hand, the majority of points belonged to high density clusters containing the most points. Therefore, the average randomization level of the skewed data set was higher.

We tested the randomization level of the data set $(OGauDis(f))$ with increasing fraction of outliers f. We note that when $f = 0$, this corresponds to the data set $EGauDis$, and when $f = 1$, then this corresponds to the data set $UniDis$. The results are illustrated in Figure 3 for a dimensionality of 75, and a perturbation level of $8 \cdot \sigma^o$. The X-axis illustrates the outlier fraction f. It is interesting to see that while the average case randomization level monotonically reduces with increasing outlier fraction, the worst-case behavior first reduces and then increases. This behavior is a little counter-intuitive and needs some explanation. The worst-case behavior is defined by the lowest 1% quantile. Therefore, only a small fraction of the data points need to be an outliers in order for the worst-case behavior to be defined by these points. This corresponds to the sharp drop off in randomization level between $f = 0$ and $f = 0.2$. An increase in outlier level beyond this point only increases the local density of the entire data space (except the clustered space) by redistributing the points from the clusters to the entire space. Since the variance along each dimension in the original data is always normalized to one unit, a redistribution from clusters to the outlier space contracts the range along each dimension in the original data (by increasing the corresponding scaling factor of the standard deviation along each dimension). In turn, this reduces the average intra-outlier distance. Since the worst case behavior is not defined by the clustered space, the worst-case behavior improves slightly with increase in outlier factor f. We note that the results for the data set $OGauDis(0.2)$ show the greatest level of difference between worst-case and average case behavior. In fact the absolute worst-case randomization level is typically even lower than the uniformly distributed data set (corresponding to $f = 1$). Furthermore,

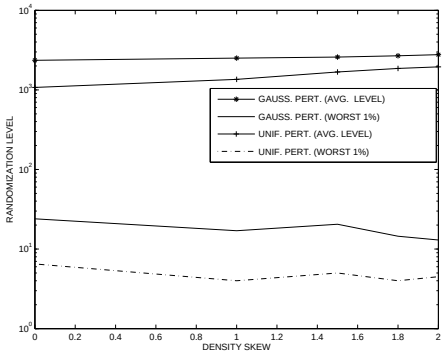

Figure 2. Randomization Level with Increasing Density Skew θ, Dimensionality$= 75$, Perturbation Level$= 8 \cdot \sigma^o$ ($VGauDis(\theta)$)

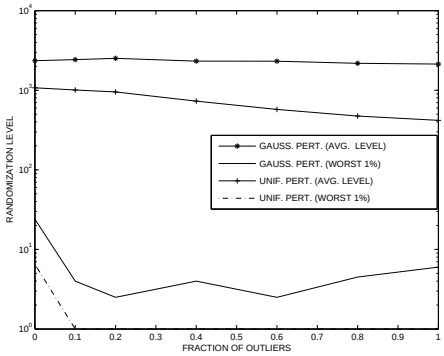

Figure 3. Randomization Level with Increasing Outlier Fraction f, Dimensionality$= 75$, Perturbation Level $= 8 \cdot \sigma^o$ ($OGauDis(f)$)

we note that this kind of data set (which has a mixture of clusters and outliers) is also the most likely in real applications. Thus, these results show that a wide variation in density distribution across the different points can have a powerful effect on the randomization level of some of the data points. This is especially the case since the randomization is done with a global perturbation level irrespective of the underlying data density. In such cases, one is caught between the two extremes of losing too much information in the dense regions (by a larger perturbation level) or that of losing privacy in the sparse regions (by choosing a smaller perturbation level).

More insight can be obtained by examining the behavior of the maximum dimensionality which retains a *fixed* randomization level for different perturbing distributions. In Figure 4, we have illustrated the maximum dimensionality of the data $UniDis$ that a given level of perturbation could support an average randomization level at (at least) 3300. In

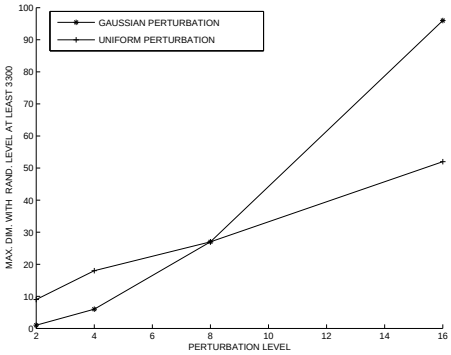

Figure 4. Maximum Dimensionality Supported by a Given Perturbation Level in $UniDis$ (Average Randomization Level at least 3300)

this case, we have illustrated the perturbation level on the X-axis, as a multiple of the standard deviation of the original data $UniDis$. The Y-axis illustrates the maximum dimensionality for which this perturbation level will support an average randomization level of at least 3300. It is clear that it is desirable to be able to support as high a dimensionality of the data as possible. It is interesting to see that for data sets of lower dimensionality, a lower level of perturbation is required with the use of uniform perturbations. However, with increasing dimensionality, the required increase in perturbation with dimensionality is much lower (sublinear) for the gaussian case. This is in agreement with our other analytical and empirical results. Thus, our results show the considerable sensitivity of the randomization technique to dimensionality as well as data and perturbing distributions. In the next section, we will discuss the implications of the inability to precisely control the privacy level with the use of the randomization technique.

5 Discussion and Future Directions

In this paper, we provide a first comprehensive treatment of the randomization approach in the presence of public information. This also provides a framework for analysis of other future members of this privacy preserving methods. We use this framework to illustrate a number of insights of the randomization method. We show the degrading effect of the dimensionality curse, and quantify the required perturbation level as a function of the dimensionality. We show that a careless choice of the perturbing distribution can degrade the privacy behavior in subtle ways because of the presence of public information. Finally, we show that many natural properties of real data sets such as clustering or outliers can significantly impact the effectiveness of the randomization approach. In summary, our results expose the high level of vulnerability of the randomization method to a variety of properties of the data sets and perturbing distributions. This shows that privacy is an extremely elusive goal for the randomization method, when public information is injected into the analysis. In future research we will propose randomization methods which are carefully designed in order to account for the effects of public information.

References

[1] Agrawal R., Srikant R. Privacy Preserving Data Mining. *Proceedings of the ACM SIGMOD Conference*, 2000.

[2] Agrawal D. Aggarwal C. C. On the Design and Quantification of Privacy Preserving Data Mining Algorithms. *ACM PODS Conference*, 2002.

[3] Aggarwal C. C. On k-anonymity and the curse of dimensionality. *VLDB Conference*, 2005.

[4] Lakshmanan L., Ng R., Ramesh G. To Do or Not To Do: The Dilemma of Disclosing Anonymized Data. *ACM SIGMOD Conference*, 2005.

[5] Evfimievski A., Gehrke J., Srikant R. Limiting Privacy Breaches in Privacy Preserving Data Mining. *ACM PODS Conference*, 2003.

[6] Hinneburg A., Aggarwal C., Keim D. What is the nearest neighbor in high dimensional space? *VLDB Conference*, 2000.

[7] Liew C. K., Choi U. J., Liew C. J. A data distortion by probability distribution. *ACM TODS*, 10(3):395-411, 1985.

[8] Huang Z., Du W., Chen B. Deriving Private Information from Randomized Data. pp. 37–48, *ACM SIGMOD Conference*, 2005.

[9] Kargupta H., Datta S., Wang Q., Sivakumar K. On the Privacy Preserving Properties of Random Data Perturbation Techniques. *ICDM Conference*, pp. 99-106, 2003.

[10] Rizvi S., Haritsa J. Maintaining Data Privacy in Association Rule Mining. *VLDB Conference*, 2002.

[11] Samarati P.: Protecting Respondents' Identities in Microdata Release. IEEE Trans. Knowl. Data Eng. 13(6): 1010-1027 (2001).

[12] Verykios V. S. et al. State-of-the-art in privacy preserving data mining. *ACM SIGMOD Record*, v.33 n.1, 2004