On $k$-anonymity and the curse of dimensionality
Introduction

- An important method for privacy preserving data mining is that of *anonymization*.

- In anonymization, a record is released only if it is indistinguishable from a pre-defined number of other entities in the data.

- We examine the anonymization problem from the perspective of inference attacks over all possible combinations of attributes.
Public Information

- In $k$-anonymity, the premise is that public information can be combined with the attribute values of anonymized records in order to identify the identities of records.

- Such attributes which are matched with public records are referred to as *quasi-identifiers*.

- 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.
Example

- Consider the following 2-dimensional records on \((\text{Age, Salary}) = (26, 94000)\) and \((29, 97000)\).

- Then, if age is generalized to the range 25-30, and if salary is generalized to the range 90000-100000, then the two records cannot be distinguished from one another.

- In \(k\)-anonymity, we would like to provide the guarantee that each record cannot be distinguished from at least \((k - 1)\) other records.

- In such a case, even public information cannot be used to make inferences.
The $k$-anonymity method

- The method of $k$-anonymity typically uses the techniques of generalization and suppression.

- Individual attribute values and records can be suppressed.

- Attributes can be partially generalized to a range (retains more information than complete suppression).

- The generalization and suppression process is performed so as to create at least $k$ indistinguishable records.
The condensation method

• An alternative to generalization and suppression methods is the condensation technique.

• In the condensation method, 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.

• There are some advantages in the use of pseudo-data, since it does not require any modification of the underlying data representation as in a generalization approach.
High Dimensional Case

- Typical anonymization approaches assume that only a small number of fields which are available from public data are used as quasi-identifiers.

- These methods typically use generalizations on domain-specific hierarchies of these small number of fields.

- In many practical applications, large numbers of attributes may be known to particular groups of individuals.

- Larger number of attributes make the problem more challenging for the privacy preservation process.
Challenges

• The problem of finding optimal $k$-anonymization is NP-hard.

• This computational problem is however secondary, if the data cannot be anonymized effectively.

• We show that in high dimensionality, it becomes more difficult to perform the generalizations on partial ranges in a meaningful way.
Anonymization and Locality

• All anonymization techniques depend upon some notion of spatial locality in order to perform the privacy preservation.

• Generalization based locality is defined in terms of ranges of attributes.

• Locality is also defined in the form of a distance function in condensation approaches.

• Therefore, the behavior of the anonymization approach will depend upon the behavior of the distance function with increasing dimensionality.
Locality Behavior in High Dimensionality

- It has been argued 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 (Beyer et al).

- In such a case, the concept of spatial locality becomes ill-defined.

- Privacy preservation by anonymization becomes impractical in very high dimensional cases, since it leads to an unacceptable level of information loss.
## Notations and Definitions

<table>
<thead>
<tr>
<th>Notation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>$d$</td>
<td>Dimensionality of the data space</td>
</tr>
<tr>
<td>$N$</td>
<td>Number of data points</td>
</tr>
<tr>
<td>$\mathcal{F}$</td>
<td>1-dimensional data distribution in $(0,1)$</td>
</tr>
<tr>
<td>$X_d$</td>
<td>Data point from $\mathcal{F}^d$ with each coord. drawn from $\mathcal{F}$</td>
</tr>
<tr>
<td>$dist_d^k(x, y)$</td>
<td>Distance between $(x^1, \ldots x^d)$ and $(y^1, \ldots y^d)$ using $L_k$ metric</td>
</tr>
<tr>
<td>$| \cdot |_k$</td>
<td>Distance of a vector to the origin $(0, \ldots, 0)$ using the function $dist_d^k(\cdot, \cdot)$</td>
</tr>
<tr>
<td>$E[X], \operatorname{var}[X]$</td>
<td>Expected value and variance of a random variable $X$</td>
</tr>
<tr>
<td>$Y_d \to_p c$</td>
<td>A sequence of vectors $Y_1, \ldots, Y_d$ converges in probability to a constant vector $c$ if: $\forall \epsilon &gt; 0 \lim_{d \to \infty} P[dist(Y_d, c) \leq \epsilon] = 1$</td>
</tr>
</tbody>
</table>
Range based generalization

• In range based generalization, we generalize the attribute values to a range such that at least $k$ records can be found in the generalized grid cell.

• In the high dimensional case, most grid cells are empty.

• But what about the non-empty grid cells?

• How is the data distributed among the non-empty grid cells?
Illustration

(a)

(b)
Attribute Generalization

- 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.
Assumption

- For a data point $X_d$ to maintain $k$-anonymity, its 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.

- It is interesting to compute the conditional probability of $k$-anonymity in a randomly chosen grid cell, given that it is non-empty.

- Provides intuition into the probability of $k$-anonymity in a multi-dimensional partitioning.
Result (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 $\binom{N}{q} \cdot f^q \cdot (1 - f^d) \cdot (N-q)$.

- Simple binomial distribution with parameter $f^d$. 
Result (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:

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

(1)

- $P(B_k|B_1) = P(B_k \cap B_1)/P(B_1) = P(B_k)/P(B_1)$

- **Observation**: $P(B_k|B_1) \leq P(B_2|B_1)$

- **Observation**: $P(B_2|B_1) = \frac{1-N \cdot f^d \cdot (1-f^d)^{(N-1)} - (1-f^d)^N}{1-(1-f^d)^N}$
Result

• Substitute \( x = f^d \) and use L’Hopital’s rule

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

• Expression tends to zero as \( d \to \infty \)

• 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:

\[
\lim_{d \to \infty} P(B_k|B_1) = 0 \quad (2)
\]
Probability of 2-anonymity with increasing dimensionality (f=0.5)
The Condensation Approach

- Previous analysis is for range generalization.

- Methods such as condensation use multi-group cluster formation of the records.

- In the following, we will find a lower bound on the information loss for achieving 2-anonymity using any kind of optimized group formation.
Information Loss

- 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 from database $\mathcal{D}$.

- 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.

- We define the relative condensation loss $\mathcal{L}(S)$ for that group of $k$ entities as follows:

$$
\mathcal{L}(S) = \frac{M(S)}{M(\mathcal{D})}
$$

(3)
Observations

- 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.

- In the following analysis, we will show how the value of $\mathcal{L}(S)$ is affected by the dimensionality $d$. 
Assumptions

- We 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 $O_d$. We establish some convergence results.

- We will also generalize the results to the case of $N = n$ data points.
Lemma

• 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: $\lim_{d \to \infty} E [r_d - q_d] = C$, where $C$ is some constant.

• Multiply numerator and denominator by $r_d + q_d$ and proceed.
• Let $A_d = (P_1 \ldots P_d)$ 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 $O_d$.

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

• Analyze the convergence behavior of the numerator and denominator separately in conjunction with Slutsky's results.
Generalization to $N$ points

- Let $F^d$ 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: $C''' \leq \lim_{d \to \infty} E [r_d - q_d] \leq (n - 1) \cdot C'''$, where $C'''$ is some constant.

- Direct extension of previous result.
Lemma

- 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: $\lim_{d \to \infty} E \left[ \frac{r_d - q_d}{r_d} \right] = 0$, where $C'''$ is some constant.

- This result can be proved by showing that $r_d \to p \sqrt{d}$.

- Note that the distance of each point to the origin in $d$-dimensional space increases at this rate.
Information Loss for High Dimensional Case

- We note that the information loss \( M(S)/M(D) \) for 2-anonymity can be expressed as \( 1 - E\left[ \frac{r_d - q_d}{r_d} \right] \).

- This expression converges to 1 in the limiting case as \( d \to \infty \).

- We are approximating \( M(D) \) 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.
Result

- Bounds for 2-anonymity are lower bounds on the general case of $k$-anonymity.

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

$$\lim_{d \to \infty} E[M(S)/M(D)] = 1$$  \hspace{1cm} (4)
Experimental Results

- 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.
Market Basket Data Sets

- We also tested the anonymization behavior with a number of market basket data sets.

- These data sets were generated using the data generator, except that the dimensionality was reduced to only 100 items.

- 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 is 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.
Experimental Results
Experimental Results

Fraction of data points preserving 2-anonymity vs dimensionality

Minimum info. loss for preserving 2-anonymity vs dimensionality
Conclusions and Summary

- Analysis of $k$-anonymity in high dimensionality.

- Earlier work has shown that $k$-anonymity is computationally difficult (NP-hard).

- This work shows that in high dimensionality, even the usefulness of $k$-anonymity methods becomes questionable.