Towards Value Disclosure Analysis in Modeling General Databases

Xintao Wu, Songtao Guo
University of North Carolina at Charlotte
(xwu,sguo)@uncc.edu

Yingjiu Li
Singapore Management University
yjli@smu.edu.sg

ABSTRACT
The issue of confidentiality and privacy in general databases has become increasingly prominent in recent years. A key element in preserving privacy and confidentiality of sensitive data is the ability to evaluate the extent of all potential disclosure for such data. This is one major challenge for all existing perturbation or transformation based approaches as they conduct disclosure analysis on the perturbed or transformed data, which is too large, considering many organizational databases typically contain a huge amount of data with a large number of categorical and numerical attributes. Instead of conducting disclosure analysis on perturbed or transformed data, our approach is to build an approximate statistical model first and analyze various potential disclosure in terms of parameters of the model built. As the model learned is the only means to generate data for release, all confidential information which snoopers can derive is contained in those parameters.

Keywords
Privacy, Disclosure Analysis, General Location Model

1. INTRODUCTION
The issue of confidentiality and privacy in general databases has become increasingly prominent in recent years. Disclosures that can occur as a result of inferences by snoopers include two classes; identity disclosure and value disclosure. Identity disclosure relates to the disclosure of the identity of an individual in the database while value disclosure relates to the disclosure of the value of a certain confidential attribute of that individual. To prevent disclosures, various randomization based approaches (e.g., [1, 2, 8, 9]) have been investigated.

A key element in preserving privacy and confidentiality of sensitive data is the ability to evaluate the extent of all potential disclosure for such data. In other words, we need to be able to answer to what extent confidential information in a perturbed or transformed database can be compromised by attackers or snoopers. This is a major challenge for current randomization based approaches. The authors, in [7], argued that randomization schemes might not be secure and developed a random matrix-based spectral filtering technique to retrieve original data from the perturbed data. Recently the authors, in [4], showed an upper bound for reconstruction error of this spectral filtering technique. Hence attackers or snoopers may even exploit this bound to assess their estimates.

The context in this paper is the evaluation of privacy and confidentiality residing in general databases which contain both categorical attributes and numerical attributes. In [10], the authors proposed a general framework for modeling general databases using the General Location model. One advantage of the general location model is it can be used to conduct both identity disclosure and value disclosure respectively since it integrates both categorical attributes and numerical attributes in one model.

The general location model is defined in terms of the marginal distribution of categorical attributes and the conditional distribution of numerical attributes given each cell determined by categorical attributes. The former is described by a multinomial distribution on the cell count when we summarize the categorical part as a multi-dimensional contingency table. The numerical attributes of tuples in each cell are assumed to follow a multivariate normal distribution with its parameters μ, Σ, where μ is a vector of means and Σ is a covariance matrix. It is no wonder that those parameters (e.g., μ, Σ) may be used by attackers or snoopers to derive some confidential information. For example, from one distribution such as "the wages of customers from zip=28223 and race = Asian follow a normal distribution with mean 70k and standard variance 10k," snoopers can safely derive a 95% coverage interval as [50.4k, 89.6k]. This derived coverage interval may violate customers‘ privacy requirement. This paper continues this line of work and focuses on value disclosure which can occur as a result of inferences by attackers or snoopers from the multivariate normal distributions. Furthermore, we consider various factors in general databases and conduct disclosure analysis for the following scenarios.

- All numerical attributes contained in databases are sensitive attributes. Various correlations exist among those attributes. We call this as basic disclosure scenario.
- Databases contain other non-confidential numerical attributes apart from those confidential ones. Here we
assume non-confidential attributes are non-perturbed as they may be retrieved accurately by snoopers from other public sources. One problem arises as the snoopers may exploit the relationship between non-confidential attributes and confidential attributes to predict individual values of confidential attributes. We call this conditional disclosure scenario.

• Databases contain many linear combinations among both confidential and non-confidential numerical attributes. The combinations here can be either known or hidden. Many organizational databases typically contain numerous attributes that could lead themselves to potentially thousands of linear combinations. In this case, the level of security provided for linear combinations of confidential attributes could be very low even if the level of security provided for a single confidential attribute is adequate. We call this linear combinatorial scenario.

Furthermore, we extend the concept of a uni-variate confidence interval to a multi-variate confidence region to measure privacy and confidentiality for multiple confidential attributes simultaneously. The advantage of confidence region is that it also takes into consideration various correlations among those attributes.

2. PRELIMINARIES

In this section we present some known results about density contour of multi-variate normal distribution from statistics.

A word on the notation and terminology used in this paper: bold-face lower-case variables, e.g., \( z \), represent vectors; calligraphic upper-case alphabets, e.g., \( Z \), refer to sets of attributes; bold-face upper-case variables, e.g., \( \Sigma \), refer to matrices; \( \mathbf{a} \) refer to transpose of vector \( a \); \( \Sigma \) refers to a modified matrix \( \Sigma_{ij} \) refers to the inverse of the matrix \( \Sigma \); \( | \Sigma | \) refers to the determinant of covariance matrix \( \Sigma \); and \( \lambda_i, \mathbf{e}_i \) represents the matrix’s \( i \)-th eigenvalue and eigenvector respectively.

Result 1 (Constant probability density contour). ([6], page 134) Let \( Z \) be distributed as \( N_p(\mu, \Sigma) \) with \( | \Sigma | > 0 \). Then, the \( N_p(\mu, \Sigma) \) distribution assigned probability \( 1 - \alpha \) to the solid ellipsoid \( \{ z : (z - \mu)' \Sigma^{-1} (z - \mu) \leq \chi^2_\alpha (\alpha) \} \), where \( \chi^2_\alpha (\alpha) \) denotes the upper \((100\alpha\text{-th})\) percentile of the \( \chi^2_\alpha \) distribution with \( p \) degrees of freedom. The ellipsoid is centered at \( \mu \) and have axes \( \pm \sqrt{\lambda_i} \mathbf{e}_i \), where \( c^2 = \chi^2_\alpha (\alpha) \) and \( \Sigma \mathbf{e}_i = \lambda_i \mathbf{e}_i \), \( i = 1, \ldots, p \).

The multi-variate normal density is constant on surfaces where the squared distance \( (z - \mu)' \Sigma^{-1} (z - \mu) \) is constant \( c^2 \). The chi-square distribution determines the variability of the sample variance. Probabilities are represented by volumes under the surface over regions defined by intervals of the \( z_i \) values. The axes of each ellipsoid of constant density are in the direction of the eigenvectors of \( \Sigma^{-1} \) and their lengths are proportional to the square roots of the eigenvalues \( \lambda_i \) of \( \Sigma \).

Result 2. (Volume of Ellipsoid). ([3]) The volume of an ellipsoid \( \{ z : (z - \mu)'A^{-1}(z - \mu) \leq 1 \} \) determined by one positive definite \( p \times p \) matrix \( A \) is given by \( \text{vol}(E) = \eta | A^{1/2} | \), where \( \eta \) is the volume of the unit ball in \( \mathbb{R}^p \).

Result 3 shows the general result concerning the projection of an ellipsoid onto a line in a \( p \)-dimensional space.

Result 3 (Projection of Ellipsoid). ([6], page 203) For a given vector \( \ell \neq 0 \), and \( z \) belonging to the ellipsoid \( \{ z : z'A^{-1}z \leq c^2 \} \) determined by a positive definite \( p \times p \) matrix \( A \), the projection (shadow) of \( \{ z A^{-1} z \leq c^2 \} \) on \( \ell \) is

\[
\frac{c\sqrt{\ell'\ell}}{\sqrt{\ell' \Sigma x \ell}} \text{ which extends from 0 along } \ell \text{ with length } c\sqrt{\ell' \ell}.
\]

When \( \ell \) is a unit vector, the shadow extends \( c\sqrt{\ell' \ell} \) units, so \( | z \ell | \leq c\sqrt{\ell' \ell} \).

3. VALUE DISCLOSURE ANALYSIS

Value disclosure represents the situation where snoopers are able to estimate or infer the value of a certain confidential numerical attribute of an entity or a group of entities with a level of accuracy greater than a pre-specified level. In our scenario, all numerical attribute values are modeled by multi-variate normal distributions. Here multi-variate normal distribution itself is not considered confidential information, only the parameters \( \mu, \Sigma \), which are used for data generation, provides adequate security for confidential numerical attributes for an entity or a group of entities. The second issue is how to modify \( \mu, \Sigma \) when they violate privacy and confidentiality requirements.

3.1 Basic Disclosure Scenario

In [2], privacy is measured in terms of confidence intervals for each single numerical attribute. Given confidence \( c\% \), for each randomized value \( z \), an interval \( [z - w_1, z + w_2] \) is defined such that for all nonrandomized values \( x \),

\[
P[z - w_1 \leq x \leq z + w_2 | z = x + y, y \sim F_y] \geq c\%
\]

The shortest width \( w = w_1 + w_2 \) for a confidence interval is used as the amount of privacy at \( c\% \) confidence level. In the \( p \)-dimensional space, an \( c\% \) confidence region will be an ellipsoidal region given by its probability density contour. This region consists of values of \( x \) (i.e., a vector over all numerical attributes) that may be accepted at the \( 1 - c\% \) level of significance.

From Result 1, we know the ellipsoid \( \{ z : (z - \mu)' \Sigma^{-1} (z - \mu) \leq \chi^2_\alpha (\alpha) \} \), which is yielded by the paths of \( z \) values, contains a fixed percentage, \((1 - \alpha)100\%\) of customers. Although snoopers may use various techniques to estimate and predict the confidential values of individual customers, however, all confidential information which snoopers can learn is the bound of ellipsoid in our scenario.

Assume \( \hat{E} \) is the ellipsoid from the original data \( z \) at one given confidence level \( 1 - \alpha \) and \( \hat{E} \) is the ellipsoid from the modified distribution (or perturbed data). Equation 1 defines the measure of disclosure of \( z \) when \( \hat{z} \) is given.

\[
D(z, \hat{z}) = \frac{| \text{vol}(E \cap \hat{E}) |}{| \text{vol}(E \cup \hat{E}) |}
\] (1)

Here compromise is said to occur when \( D(z, \hat{z}) \) is greater than \( \tau \), specified by the database owner. The greater the \( D(z, \hat{z}) \), the closer the estimates are to the true distribution, and the higher the chance of disclosure. In other words, if the ellipsoid learned by snoopers is close enough to that
We replace \(\{\) from spectral decomposition

**Proof.** From Result 2, we know the volume of an ellipsoid \(\{x: (z - \mu)^\top \Sigma^{-1} (z - \mu) \leq \chi^2_p(\alpha)\}\) is given by

\[
\text{vol}(E) = \eta(\sqrt{\lambda^p_1}) |\Sigma^{1/2}| \\
or
\text{vol}(E) = \eta(\sqrt{\lambda^p_1}) \prod_{i=1}^{p} \sqrt{\lambda_i}
\]

where \(\eta\) is the volume of the unit ball in \(\mathbb{R}^p\), and \(\lambda_i\) is the \(i\)-th eigenvalue of matrix \(\Sigma\).

Proof. From Result 2, we know the volume of an ellipsoid \(\{x: (z - \mu)^\top A^{-1} (z - \mu) \leq 1\}\) is given by \(\text{vol}(E) = \eta|A^{1/2}|\).

We replace \(A\) with \(\Sigma\) as shown in Proposition 1.

Figure 1: A constant density contour for a bi-variate normal distribution

**Proposition 1 (Volume of Density Contour).** The volume of an ellipsoid \(\{x: (z - \mu)^\top \Sigma^{-1} (z - \mu) \leq \chi^2_p(\alpha)\}\) is given by

\[
\text{vol}(E) = \eta(\sqrt{\lambda^p_1}) |\Sigma^{1/2}| \\
or
\text{vol}(E) = \eta(\sqrt{\lambda^p_1}) \prod_{i=1}^{p} \sqrt{\lambda_i}
\]

where \(\eta\) is the volume of the unit ball in \(\mathbb{R}^p\), and \(\lambda_i\) is the \(i\)-th eigenvalue of matrix \(\Sigma\).

Proof. From Result 2, we know the volume of an ellipsoid \(\{x: (z - \mu)^\top A^{-1} (z - \mu) \leq 1\}\) is given by \(\text{vol}(E) = \eta|A^{1/2}|\).

We replace \(A\) with \(\Sigma\), then we get

\[
\text{vol}(E) = \eta(\sqrt{\lambda^p_1}) |\Sigma^{1/2}| \\
or
\text{vol}(E) = \eta(\sqrt{\lambda^p_1}) \prod_{i=1}^{p} \sqrt{\lambda_i}
\]

**Proposition 2 (Simultaneous Confidence Intervals).** Let \(Z\) be distributed as \(N_p(\mu, \Sigma)\) with \(\Sigma > 0\). The projection of this ellipsoid \(\{z: (z - \mu)^\top \Sigma^{-1} (z - \mu) \leq \chi^2_p(\alpha)\}\) on axis \(z_i = (0, \cdots, 1, \cdots, 0)\) (only the \(i\)-th element is 1, all other elements are 0) has bound:

\[
[\mu_i - \sqrt{\chi^2_p(\alpha)}\sigma_{ii}, \mu_i + \sqrt{\chi^2_p(\alpha)}\sigma_{ii}]
\]

**Proof.** From Result 3, we know the projection of an ellipsoid \(\{x: x^\top A^{-1} x \leq c^2\}\) on a given unit vector \(\ell\) has length \(\text{len} = c\sqrt{\ell^\top \Sigma \ell}\). We replace \(A\) with

\[
\Sigma = \begin{pmatrix}
\sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\
\sigma_{12} & \sigma_{22} & \cdots & \sigma_{2p} \\
\vdots & \vdots & \ddots & \vdots \\
\sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp}
\end{pmatrix}
\]

replace \(\ell\) as \(z_i = (0, \cdots, 1, \cdots, 0)^\top\), and replace \(c\) as \(\sqrt{\chi^2_p(\alpha)}\).

Considering the center of this ellipsoid, we have the bound as \(\mu_i - \sqrt{\chi^2_p(\alpha)}\sigma_{ii}, \mu_i + \sqrt{\chi^2_p(\alpha)}\sigma_{ii}\).

It is easy to see from Proposition 2 that the confidence interval for each attribute (by projecting on each axis) is only dependent on \(\mu_i, \sigma_{ii}\), while it is independent with covariance values \(\sigma_{ij}\), where \(i \neq j\). Figure 2 illustrates how the shape of ellipse changes when we vary its covariance matrix \(\Sigma\) using one bi-variate normal distribution example. For example, by varying \(\sigma_{12}\) while fixing \(\sigma_{11}, \sigma_{22}\) as shown in Figure 2(a), the axes of ellipse rotate and the ratio between these two axes also changes. However, the projection of ellipse on axis \(z_1, z_2\) does not change as it only depends on \(\sigma_{11}\) and \(\sigma_{22}\).
of customers which are modeled by one multi-variate normal distribution. To check whether a given distribution of \( z \) may incur value disclosure for one attribute \( z \), we can simply compare the disclosure measure \( d(z, z) \) with \( \tau \), specified by the database owner. If disclosure occurs, we need to modify parameters \( \mu, \Sigma \). As we know from Proposition 2, the mean vector \( \mu \) determines the center of ellipsoid or the center of projection interval while the covariance matrix \( \Sigma \) determines the size of ellipsoid or the length of projection interval. As the change of \( \mu \) will significantly affect the data distribution (it will affect the accuracy of analysis or mining subsequently), in the remainder of this paper we focus only on how to change variance matrix \( \Sigma \) to satisfy user’s security requirement. From the bound \( [\mu - \sqrt{\chi^2(\alpha)\sigma_{ii}}, \mu + \sqrt{\chi^2(\alpha)\sigma_{ii}}] \), we can easily derive \( \sigma_{ii} \) to satisfy privacy requirements on the confidential attribute \( z \).

Please note that both the ellipsoid \( \mathcal{E} \) and the confidence interval \([z^l, z^u]\) from discussions above are specified for a group of customers which are modeled by one multi-variate normal distribution with the same parameters. Hence both \( \mathcal{E} \) and \([z^l, z^u]\) are privacy specifications at the aggregate level. In practice, each customer \( j \) may specify his own privacy interval \([z^l(j), z^u(j)]\) which contains his confidential value \( z(j) \). In this scenario, the database owner is required to prevent snoopers to derive or estimate the confidential value falling into its privacy interval. We use \([\hat{z}_i, \hat{z}_u]\) to denote the numerical attribute \( z \)’s confidence interval learned by snoopers through projecting the ellipsoid on its axis. If the derived confidence interval \([\hat{z}_i, \hat{z}_u]\) by snoopers is close to the customer \( j \)’s privacy interval \([z^l(j), z^u(j)]\), we say individual value disclosure occurs for customer \( j \). Currently we are working on how to evaluate this individual value disclosure in database modeling.

### 3.2 Conditional Scenario

Consider a database with \( k \) numerical, confidential attributes \( X \) and \( l \) non-confidential attributes \( S \) where \( p = k + l \). Security is measured by the degree to which a snooper can determine the values of confidential attributes in a specific record through the use of relationships between the non-confidential and confidential attributes. One question we ask here is how much information is contained in non-confidential attributes and how it affects the variability of confidential numerical attributes.

**Proposition 3 (Conditional normal distribution).**

\[(\{6\}, \text{page 131})\] Let \( Z = \begin{pmatrix} X \\ S \end{pmatrix} \) be distributed as \( N_p(\mu, \Sigma) \) with \( \mu = \left( \begin{array}{c} \mu_X \\ \mu_S \end{array} \right) \), \( \Sigma = \left( \begin{array}{cc} \Sigma_{XX} & \Sigma_{XS} \\ \Sigma_{SX} & \Sigma_{SS} \end{array} \right) \) and \( |\Sigma_{SS}| > 0 \). Then the conditional distribution of \( X \) given \( S \) is normal with mean \( \hat{\mu}_X + \Sigma_{XS}\Sigma_{SS}^{-1}(\Sigma_{SS}^{-1}\Sigma_{SX} - \Sigma_{SS}^{-1}\Sigma_{SS} \Sigma_{SS}^{-1}\Sigma_{SX}) \) and covariance \( \Sigma_{XX} - \Sigma_{XS}\Sigma_{SS}^{-1}\Sigma_{SX} \).

Proposition 3 shows the conditional distribution of \( X \) given \( S \) is also a multi-variate normal distribution. Furthermore, the conditional covariance \( \Sigma_{XX} - \Sigma_{XS}\Sigma_{SS}^{-1}\Sigma_{SX} \) does not depend upon the values of the conditioning variables. Hence we can simply apply results from Proposition 1 and 2 by replacing \( \Sigma \) with the new conditional covariance \( \Sigma_{XX} - \Sigma_{XS}\Sigma_{SS}^{-1}\Sigma_{SX} \) to conduct conditional value disclosure analysis.

In general, given a confidential variable \( x \) whose variance is \( \sigma_x^2 \), the largest eigenvalue (\( \lambda \)) of \( \Sigma_{XX}^{-1}\Sigma_{XS}\Sigma_{SS}^{-1}\Sigma_{SX} \) gives the proportion of the variance (fluctuation) of variable \( x \) that is predictable from non-confidential attributes \( S \). The eigenvalue is a measure of how well the non-confidential attributes can predict the confidential attribute \( x \). e.g. \( \lambda = 0.85^2 \) means that 85% of the total variation in \( x \) can be explained by the linear relationship between \( S \) and \( x \). The other 15% of the total variation in \( x \) remains unexplained. Hence, a rough estimate of the smallest standard error can be determined as \( \sqrt{(1 - \lambda)}\sigma_x\). Based on this estimate of standard error, a rough 95% confidence interval for \( x \) is as:

\[ \hat{\mu}_x \pm 1.96\sqrt{(1 - \lambda)}\sigma_x \]

### 3.3 Combinatorial Scenario

Many organizational databases typically contain numerous attributes that could lead themselves to potentially thousands of linear combinations. In this case, the threat of combinatorial disclosure can be magnified further. Previous research[9] has shown that even if the level of security provided for a single confidential attribute is adequate, the level of security provided for linear combinations of confidential attributes could be very low. For example, the prediction of the linear combination Total Income = Wages + Interests + Dividends is likely to have a high level of accuracy than that of each individual attribute. Thus, there is great necessity to evaluate a very large number of linear combinations.
The approach we apply here is based on Canonical Correlation Analysis (CCA) [6] which was previously used in [9]. In this paper, we focus on how to modify $\Sigma_{XX}$ or $\Sigma_{XS}$ to achieve combinatorial privacy requirements when a threshold $\lambda$ is specified by the database owner. CCA can be used to measure the maximum proportion of the variance that can be explained in any linear combination of confidential attributes $X$, using a linear combination of known non-confidential attributes $S$. The main task of CCA is to summarize the associations between the $X$ and $S$ sets in terms of a few carefully chosen covariances rather than the pq covariances in $\Sigma_{XS}$. We denote the respective linear combinations by $u = a'x$ and $v = b's$. The correlation between $u$ and $v$ is given by

$$\text{Corr}(u, v) = \frac{a' \Sigma_{xs} b}{(|a' \Sigma_{xx} a|(|b' \Sigma_{ss} b|)^{1/2})}$$

Out of the infinite number of linear combinations, we find that set of linear combinations which maximizes the correlation $\text{Corr}(u, v)$. The canonical variate pair $u_i = e_i' \Sigma_{xx}^{-1/2} x$ and $v_i = e_i' \Sigma_{ss}^{-1/2} s$ maximizes $\text{Corr}(u_i, v_i) = \sqrt{\lambda_i}$, where $i = 1, \cdots, l$. Here $\lambda_1 \geq \cdots \geq \lambda_l$ are the eigenvalues of $\Sigma_{ss}^{-1/2} \Sigma_{xs} \Sigma_{xx}^{-1/2} \Sigma_{ss}^{-1/2}$, and $e_1, \cdots, e_l$ are the associated normalized eigenvectors as shown in Equation 3.

$$\Sigma_{ss}^{-1/2} \Sigma_{xs} \Sigma_{xx}^{-1/2} \Sigma_{ss}^{-1/2} = A = \lambda_1 e_1 e_1' + \cdots + \lambda_l e_l e_l'$$

(3)

The largest eigenvalue $\lambda_1$ is the squared canonical correlation coefficient, which represents the most general measure of inferential value disclosure for any combination. In other words, $1 - \lambda_1$ represents the worst-case security. When some $\lambda_i$ is greater than the threshold $\lambda$, specified by database owners, it means some combinatorial disclosure exists for one potential combination of confidential attributes. In this case, we need to change parameters, $\Sigma_{xx}$ or $\Sigma_{ss}$, so that all new eigenvalues should be less than or equal to the threshold $\lambda$.

Our approach here is that we set those eigenvalues $\lambda_i$ as $\bar{\lambda}$ (hence no combinatorial disclosure exists) and keep the other eigenvalues $(\lambda_i < \bar{\lambda})$ and all eigenvectors unchanged. We get a new matrix $A$ after applying the inverse of spectral decomposition as shown in Equation 3. The derived matrix $A$ is guaranteed to satisfy users’ security requirement for all the possible combinations. Furthermore, as we keep all eigenvectors and those other eigenvalues $(\lambda_i \leq \bar{\lambda})$ unchanged in our approach, the density contour of modified distribution will be closest to that of the original one.

From Equation 3, we know $A$, which will satisfy users’ security requirements, is determined by $\Sigma_{xx}$ and $\Sigma_{ss}$. So we can adjust either $\Sigma_{xx}$ or $\Sigma_{ss}$ to achieve $A$. Note that $\Sigma_{ss}$ should be kept unchanged as we assume data of non-confidential attributes are non-perturbed. To adjust $\Sigma_{xx}$, we simply set $\Sigma_{xx}$ as $\Sigma_{xx} - \bar{\lambda} \Sigma_{xx}^{-1/2} \Sigma_{ss}^{-1/2} \Sigma_{xs}$.

However, there is no direct method to adjust $\Sigma_{ss}$. From $\Sigma_{ss}^{-1/2} \Sigma_{ss}^{-1/2} \Sigma_{ss}^{-1/2} \Sigma_{ss}^{-1/2} = A$, we expand the left side of equation and get an $l \times l$ matrix. Each element of this matrix is a quadratic function $f_{ij}(x_{11}, \cdots, x_{ik})$ which equals to the corresponding $\tilde{a}_{ij}$. Then we get $l \times l$ sub-quadratic equations with $l \times k$ variables. The problem becomes the following optimization problem:

**Problem 1.** Minimize $F(x_{11}, \cdots, x_{ik}) = \sum_{i=1}^{l} \sum_{j=1}^{k} (f_{ij}(x_{11}, \cdots, x_{ik}) - \tilde{a}_{ij})^2$ subject to $x_{ij} \geq 0$.

4. CONCLUSIONS AND FUTURE WORK

In this paper we have investigated various scenarios which may exist in general databases and presented how to satisfy users’ privacy requirements by adjusting parameters of the model learned. This paper presents a start on methods and metrics for evaluating privacy disclosure at the aggregate level in terms of statistical model built from the underlying data. We are working on how to conduct individual value disclosure analysis when different privacy intervals are specified by different individuals. We are also working on evaluating how the information loss due to statistical modeling affects the utility of generated data and formalizing methods and metrics when different statistical models are applied.

5. ACKNOWLEDGMENTS

This work was supported in part by U.S. National Science Foundation CCR-0310974.

6. REFERENCES