Association for Information Systems
AIS Electronic Library (AISeL)

ICIS 2010 Proceedings

International Conference on Information Systems (ICIS)

2010

DATA CLUSTERING AND MICRO-PERTURBATION FOR PRIVACY-PRESERVING DATA SHARING AND ANALYSIS

Xiao-Bai Li
University of Massachusetts Lowell, xiaobai_li@uml.edu

Sumit Sarkar
University of Texas at Dallas, sumit@utdallas.edu

DATA CLUSTERING AND MICRO-PERTURBATION FOR PRIVACY-PRESERVING DATA SHARING AND ANALYSIS

Completed Research Paper

Xiao-Bai Li
College of Management
University of Massachusetts Lowell
Lowell, MA 01854, U.S.A.
xiaobai_li@uml.edu

Sumit Sarkar
School of Management
University of Texas at Dallas
Richardson, TX 75080, U.S.A.
sumit@utdallas.edu

Abstract

Clustering-based data masking approaches are widely used for privacy-preserving data sharing and data mining. Existing approaches, however, cannot cope with the situation where confidential attributes are categorical. For numeric data, these approaches are also unable to preserve important statistical properties such as variance and covariance of the data. We propose a new approach that handles these problems effectively. The proposed approach adopts a minimum spanning tree technique for clustering data and a micro-perturbation method for masking data. Our approach is novel in that it (i) incorporates an entropy-based measure, which represents the disclosure risk of the categorical confidential attribute, into the traditional distance measure used for clustering in an innovative way; and (ii) introduces the notion of cluster-level micro-perturbation (as opposed to conventional micro-aggregation) for masking data, to preserve the statistical properties of the data. We provide both analytical and empirical justification for the proposed methodology.

Keywords: Data privacy, data mining, minimum spanning tree, microaggregation, data perturbation

Introduction

As data-sharing and data-mining technologies are being increasingly used in areas such as medical and healthcare research, counter-terrorism, credit and loan evaluation, and customer relationship management, there are growing concerns about their threats to individual privacy. A study by the US General Accounting Office (2004) reported that 61% of the data mining projects run by federal agencies used personal information, and 67% of the data mining projects from the private sectors involved personal information. In the healthcare industry, there has been a rapid growth of computerization of healthcare records, and over 70 million Americans have some portion of their medical records in electronic format (Kaelber 2008). The American Recovery and Reinvestment Act of 2009 includes $20 billion funding for facilitating electronic medical records (EMR). This rapid transition towards EMR and data sharing has raised pressing concerns about privacy. Indeed, there is evidence that EMR has caused medical identity theft to increase considerably (Dixon 2006).

Mishandling of privacy issues can seriously hurt an organization's credibility and reputation. In a widely-publicized incident, AOL released on its website in August 2006 a file containing 20 million search queries for over 650,000 users. According to AOL, the intention was to provide data for research into online browsing behavior. The identities of the users were not included in the data; however, it was soon found that many users in the file could be easily re-identified. This caused such fierce public protests, including several law suits and legal complaints against AOL, that AOL removed the data from the website within days (Zeller 2006). Concerns about privacy have also caused data quality and integrity to deteriorate. According to Teltzrow and Kobsa (2004), 82% of online users have refused to give personal information and 34% have lied when asked about their personal habits and preferences.

Various approaches have been proposed to address the public's concerns about privacy (Adam and Wortmann 1989; Aggarwal and Yu 2008). A conventional approach is *query restriction*, which focuses on designing statistical databases and forming restrictions for accessing confidential data (Chowdhury et al. 1999; Gopal et al. 2002; Garfinkel et al. 2002). When the data released is for data mining and statistical analysis, a dataset containing individual records is usually required. In this situation, query restriction methods are not applicable and the common practice is to mask the data before it is released. Data masking methods broadly include *noise-based perturbation*, which adds noise to the sensitive data to disguise their true values (Agrawal and Srikant 2000; Liew et al. 1985; Traub et al. 1984); *data swapping*, which involves exchange of attribute values between different records (Dalenius and Reiss 1982; Reiss 1984; Li and Sarkar 2006b); and *generalization* and *suppression*, which generalizes the original values to a higher level category or removes the values if generalization is inappropriate (Samarati 2001; Sweeney 2002; Cox 1980). All these methods attempt to preserve the utility of the masked data, as measured by various data quality metrics.

From a privacy viewpoint, the attributes of data on individuals can be classified into three categories: (i) *explicit identifiers*, which can be used to directly identify an individual, including name, social security number, phone number, and credit card number; (ii) *confidential attributes*, which contain private information that an individual typically does not want revealed, such as salary, medical test results, and sexual orientation; and (iii) *non-confidential attributes*, which are normally not considered as confidential by individuals, such as age, gender, race, education, and occupation. However, the values of some of these attributes can often be used to identify individuals by matching data from different sources, resulting in identity disclosure. Such attributes are called *quasi-identifier* (QI) in the literature. For example, Sweeney (2002) found out that 87% of the population in the United States can be uniquely identified with three attributes – gender, date of birth, and 5-digit zip code – which are accessible from voter registration records available to the public. In privacy-preserving data mining research, it is typically assumed that the explicit identifiers have already been removed from the data. Data masking is applied to QI or confidential attributes. We assume the same setting in this study.

A method for data privacy protection, called *k-anonymity* (Samarati 2001; Sweeney 2002), has recently gained considerable popularity. The basic idea behind *k*-anonymity is to mask the values of the QI attributes such that the values of these attributes for any individual matches those of at least $k - 1$ other individuals in the same dataset. In this way, the identity of an individual is expected to be better protected. However, it is still likely for an intruder to discover the confidential information of individuals in the *k*-anonymized data (Machanavajjhala 2006). The problem is that *k*-anonymity protects identity disclosure by generalizing different but similar QI attribute values into the same value. The new values produced by the generalization operation are still correct with respect to the generalized categories. Since confidential attribute values remain unchanged in *k*-anonymity, individuals in a group, who have the same generalized QI values, are subject to high disclosure risk if their confidential values in the group are the

same. Further, *k*-anonymity focuses primarily on categorical data. When an attribute is originally captured in numeric form, the technique converts its values into intervals and then treats the intervals as categorical values.

A popular approach for masking numeric data is to use a clustering-based technique. A typical representative of this approach is *microaggregation*, which masks data by first clustering the data into groups of similar records and then replacing the QI or confidential attribute values with a group-level aggregated value. The basic idea of microaggregation in terms of disclosure protection is similar to that of *k*-anonymity (with the distinction that the former applies to numeric data while the latter to categorical data). Univariate microaggregation (Defays and Nanopoulos 1993) involves sorting records by each attribute to be masked, clustering adjacent records into groups of small sizes, and replacing the individual values in each group with the group average. Univariate microaggregation does not consider the relationships between attributes; so the masked data might not be appropriate for data mining. Multivariate microaggregation groups data using a clustering technique that is based on a multi-dimensional distance measure (Domingo-Ferrer and Mateo-Sanz 2002; Laszlo and Mukherjee 2005). As a result, the relationships between attributes are expected to be better preserved. Li and Sarkar (2006a) propose a clustering-based method for masking numeric data using a kd-tree technique, which is more efficient than the multivariate microaggregation method. A limitation of these clustering-based approaches is that they apply primarily to numeric data. When categorical confidential data are present, a clustering-based approach can increase the disclosure risk of the confidential data. This problem is similar to the confidential value disclosure problem in *k*-anonymity mentioned above. We illustrate this situation in detail in the next section. Due to its nonparametric nature, another criticism leveled against clustering-based approach is that it lacks analytical justification for preserving statistical properties of the data (Winkler 2007).

In this study, we examine the problem of using microaggregation when categorical confidential data are present. We propose a new method that effectively protects or limits identity and confidentiality disclosure in this situation. The proposed approach adopts a minimum spanning tree (MST) technique for clustering data and takes into account the disclosure risk of the categorical confidential attribute when forming data groups. To reduce identity disclosure risk, our approach uses a micro-perturbation method that perturbs the numeric non-confidential data at a cluster level. Together, the clustering and micro-perturbation methods ensure that the statistical properties of the original data are well preserved. The main contributions of this research are summarized as follows:

- *The originality of the problem*. This research is the first one to investigate the privacy problem that arises when a clustering-based approach is used to mask numeric data with categorical confidential data. We demonstrate that the disclosure risk of the confidential data can increase in such a situation.
- *The novelty of the approach*. Our approach is novel in that (i) it incorporates an entropy-based measure, which represents the disclosure risk of the categorical confidential attribute, into the traditional distance measure in an innovative way for clustering data; and (ii) it introduces the notion of cluster-level micro-perturbation (as opposed to conventional micro-aggregation) for masking data, which is new to the literature.
- *Theoretical justification of the methodology*. Many of the existing data masking methods are parametric, depending on assumptions about the properties of the data, such as normality or monotonicity. Current nonparametric approaches lack theoretical results to justify their validity. We show that the mean vector and covariance matrix of the masked data generated using the proposed micro-perturbation method are unbiased estimates of the true mean vector and covariance matrix. These results are true not only for the clustering method proposed in this research, but also for any other clustering-based technique if the micro-perturbation method is used.

The rest of the paper is organized as follows. In the next section, we describe the confidentiality disclosure problem that can occur when applying a clustering-based technique for data masking, and propose entropy-based measures to represent the disclosure risk. In the follow-up section, we develop the MST-based algorithm for clustering data with confidential class restriction. The micro-perturbation method and its theoretical justification are elaborated subsequently. We then describe a set of experiments conducted on real-world datasets. The final section concludes the paper and provides directions for future research.

Class Restricted Microaggregation Problem

Microaggregation involves partitioning a dataset of *N* records into groups such that each group contains at least *m* records. That is,

$$n_g \geq m, \forall g; \text{ and } \sum_{g=1}^{G} n_g = N, \tag{1}$$

where G is the number of groups and n_g is the number of records in group g. The purpose of partitioning data into groups is to use the group-level aggregated data in place of individual values for data release. As such, microaggregation attempts to minimize information loss due to the aggregation, subject to the group size constraint. Let $\mathbf{x}_{gi}$ $(i = 1,...,n_g; g = 1,...,G)$ be the ith record in group g and $\bar{\mathbf{x}}_g$ be the mean vector for group g. The information loss can be measured using the within-group sum of squared errors:

$$SSE = \sum_{g=1}^{G} \sum_{i=1}^{n_g} (\mathbf{x}_{gi} - \bar{\mathbf{x}}_g)'(\mathbf{x}_{gi} - \bar{\mathbf{x}}_g). \tag{2}$$

For a given dataset, the total sum of squared errors,

$$SST = \sum_{i=1}^{N} (\mathbf{x}_i - \bar{\mathbf{X}})'(\mathbf{x}_i - \bar{\mathbf{X}}), \tag{3}$$

is a constant (where $\bar{\mathbf{X}}$ is the overall mean vector). It is then more convenient to use *SSE/SST*, which is a value between 0 and 1, for measuring information loss. Therefore, microaggregation problems can be viewed as minimizing *SSE/SST* subject to the group size constraint in (1).

To prevent re-identification disclosure, microaggregation typically masks the values of the QI attributes (similar to *k*-anonymity), while keeping the confidential attribute values unchanged. When the confidential attributes are categorical, however, this approach is problematic. For simplicity, we consider only a single categorical confidential attribute, which we call the *class* attribute.

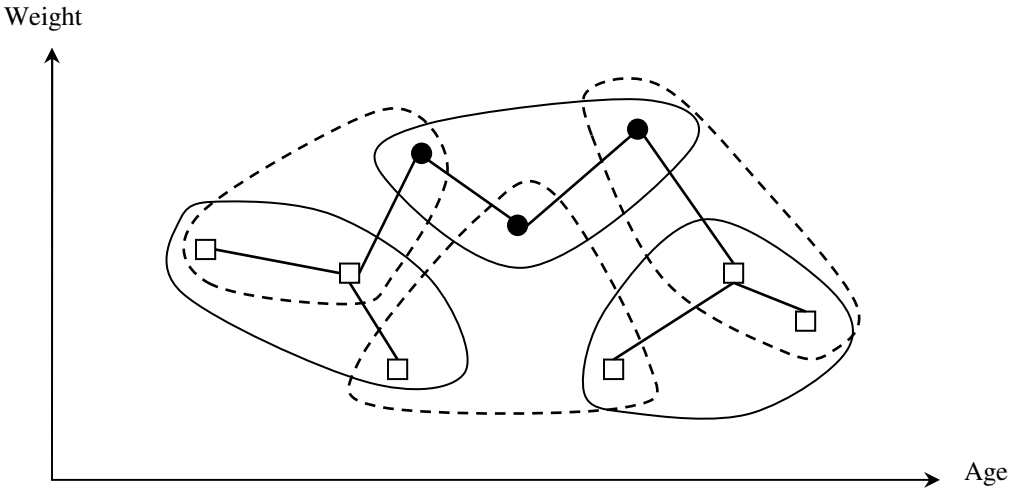

Figure 1. An Illustrative Example for Clustering

To illustrate the problem, consider an example dataset containing nine patient records. There are two numeric QI attributes, Age and Weight, and one class attribute, Test Result, with two values: positive and negative. The nine data points are plotted in Figure 1, where a circle represents 'positive' and a square represents 'negative'. Suppose the minimum group size is three. By minimizing *SSE/SST* (calculated based on Age and Weight only), microaggregation (as well as *k*-anonymity) forms three groups as shown by the loops formed by solid lines, since the data points are closer to each other within each group than to the points outside the group. The problem with this grouping is that the confidential class value in each group becomes homogeneous. It is easy to infer the test result of a patient when it is released along with the corresponding group-average age and weight values. For example, consider a privacy intruder who knows the age and weight of the patient represented by the circle having the largest

value for Weight (the upper right circle in Figure 1). The intruder can then identify that this patient belongs to the group that tested positive, because the centroid (which is released) of this group is the closest, among all groups, to that data point, in terms of distance calculated based on age and weight.

The method we present attempts to cluster the data such that the frequency distribution of the class values within each group is as close to the overall distribution as possible. At the same time, we still want to minimize information loss as measured by *SSE/SST*, subject to the group size constraint. A result of three groups with this approach is shown by the loops enclosed in dashed lines in Figure 1, where the class distribution for each group is the same as the overall class distribution (i.e., 1/3 'positive' and 2/3 'negative').

When the confidential class attribute is present, the disclosure risk of a record in a group should be viewed as high when the records in the group have purer class values, while the risk is low when the class distribution of the group is close to the overall class distribution (since knowing this distribution does not help much in determining the class value of a record in such a group). To measure the disclosure risk for records in a group with this desired property, we first consider a measure, based on the well-known Kullback-Leibler divergence (*KLD*) (also known as relative entropy, Kullback 1959), as defined below.

Definition 1. Let C be the number of classes of the confidential attribute. Let F_k and f_{gk} $(k = 1,...,C;\ g = 1,...,G)$, where $\sum_{k=1}^{C} F_k = 1$ and $\sum_{k=1}^{C} f_{gk} = 1$, be the frequency distributions of the class values in the full dataset and in a group g, respectively. The *group KL-divergence* of g is defined as:

$$KLD_g(f, F) = \sum_{k=1}^{C} f_{gk} \log \frac{f_{gk}}{F_k}. \tag{4}$$

KLD is a convex function of f (F is fixed for a given dataset), which attains its minimum value of zero if and only if $f_{gk} = F_k, \forall k$ (Kullback 1959). This property satisfies a requirement for the risk measure described above – the disclosure risk should be at the minimum when the frequency distribution of the class values in a group is the same as the overall distribution. However, *KLD* is not a true distance metric because it does not satisfy the properties of symmetry and triangle inequality (Lin 1991). It is also difficult to normalize the *KLD* values because there is not a clear way to define the maximum *KLD* value. Since our problem involves a tradeoff between a distance measure (e.g., Euclidean distance) used for clustering data in a traditional way and the group divergence measure, it is desirable that this divergence measure also meets the basic property of a distance metric. Given this consideration, we propose using the Jensen–Shannon divergence (*JSD*) measure, first introduced by Lin (1991), as follows.

Definition 2. The *group JS-divergence* for group g is defined as:

$$JSD_g(f, F) = \frac{1}{2}[KLD_g(f, M) + KLD_g(F, M)], \tag{5}$$

where M is the average of f and F, i.e., $M_{gk} = (f_{gk} + F_k)/2,\ k = 1,...,C;\ g = 1,...,G$.

Because *JSD* is a nonnegative linear combination of two *KLD* measures, it is also convex and has the same attractive property as that of *KLD* when it reaches the minimum value of zero. In addition, *JSD* is symmetric and satisfies the triangle inequality condition. Furthermore, *JSD* has an upper-bound of one. Therefore, it can be regarded as a normalized distance measure. To illustrate, we calculate the *JSD* value for each group in Figure 1 as follows:

Let subscripts 1 and 2 represent the circle and square classes, respectively. Then, $F_1 = 3/9 = 0.333$, and $F_2 = 6/9 = 0.667$. For the all-circle group,

$f_{\text{all-circle},1} = 3/3 = 1$, and $f_{\text{all-circle},2} = 0$.

So,

$M_{\text{all-circle},1} = (1 + 0.333)/2 = 0.667$, and $M_{\text{all-circle},2} = (0 + 0.667)/2 = 0.333$.

$KLD_{\text{all-circle}}(f, M) = (1)\log(1/0.667) + (0)\log(0/0.333) = 0.585$.

$KLD_{\text{all-circle}}(F, M) = (0.333)\log(0.333/0.667) + (0.667)\log(0.667/0.333) = 0.333$.

Therefore,

$$JSD_{\text{all-circle}}(f, F) = \frac{1}{2}[KLD_{\text{all-circle}}(f, M) + KLD_{\text{all-circle}}(F, M)] = 0.459.$$

Similarly, for each of the two all-square groups,

$$JSD_{\text{all-square}}(f, F) = 0.191.$$

For each of the groups with one circle and two squares (i.e., enclosed by dashed lines),

$$JSD_{\text{mixed}}(f, F) = 0,$$

since $M_{\text{mixed},k} = f_{\text{mixed},k} = F_k \ (k = 1, 2)$.

The mixed group has the lowest disclosure risk because its class distribution is the same as the overall class distribution. The all-circle (positive) group has the highest disclosure risk because its class distribution differs the most from the overall distribution. The *JSD* measure will be used along with the Euclidean distance in our proposed class restricted clustering method.

Class Restricted Minimum Spanning Tree for Clustering

Domingo-Ferrer and Mateo-Sanz (2002) have shown that the globally optimal microaggregation problem, as characterized by equations (1), (2) and (3), is computationally prohibitive. Several clustering-based approaches have been developed to solve the problem efficiently. The objective functions of traditional clustering problems are essentially the same as that of microaggregation. The constraints for clustering problems, however, are somewhat different from those of microaggregation (which is shown in equation 1). For instance, the well-known *k*-means clustering approach has an equality constraint on the number of groups, while microaggregation has a lower bound constraint on the number of records in each group. As such, *k*-means clustering is not appropriate for microaggregation. Among a few efficient microaggregation algorithms, we are interested in the one proposed by Laszlo and Mukherjee (2005), which is based on partitioning a minimum spanning tree (MST). The use of MST for data clustering was initially proposed by Zahn (1971). Given a graph of N vertices, a spanning tree contains a group of $N - 1$ edges that connect all vertices of the graph. An MST is a spanning tree with minimum total edge length. When the MST is used for data clustering, each vertex represents a data point and the length of an edge is the distance between the two related data points. Figure 1 shows an MST with a group of eight edges (line segments).

In the context of microaggregation, an edge in an MST is said to be *removable* if all of the subtrees (subgroups) formed by cutting this edge contains no fewer than the specified minimum number of vertices (which is m in equation 1). The algorithm by Laszlo and Mukherjee (2005) first constructs an MST from the full dataset. It then iteratively cuts the longest removable edge in the MST to form clusters for microaggregation. The method, however, does not address the confidential class issue.

In our class-restricted microaggregation problem, there are two objectives. The first is to minimize information loss as measured by *SSE/SST*, which is the same as the traditional microaggregation. *SSE/SST* is essentially a function of the Euclidean distances between different records. In this study, we use normalized Euclidean distances. That is, for numeric data, attribute values are normalized to the range [0, 1]; for categorical data, the difference between two attribute values is defined as zero if they are the same, and one otherwise (as is the standard practice in clustering). The second objective is to minimize the overall class divergence after clustering. This aspect can be captured by the *JSD* measure after the groups are formed. During the construction of an MST, however, it is not known how the groups will be structured. Since groups are formed by iteratively cutting the edges in the MST, it is desirable for neighboring vertices in the MST to have well-represented class values. This idea is implemented in our proposed MST-based algorithm, described next.

We use Prim's algorithm (Prim 1957) for building an MST, which expands the tree by adding a vertex with the smallest edge length to a vertex already in the partially completed MST. In the process of constructing an MST, let $e(u, v)$ be a *candidate edge* that connects a vertex u already in the partial MST and a vertex v not in the partial MST. In Figure 2, for example, vertices 1 through 4 are already in the partial MST (connected by solid lines) while vertices 5 and 6 are not. So, $u \in \{1, 2, 3, 4\}$ and $v \in \{5, 6\}$. Let u be vertex 2 and v be vertex 5. Then $e(2, 5)$ is a candidate edge (out of many possible candidate edges). Prim's algorithm uses a data structure called priority queues

to efficiently identify all candidate edges for a partial MST. The candidate edge with the smallest length is then selected and added to the partial MST. Our algorithm follows the same idea in identifying and selecting candidate edges except that the edge length is defined as a weighted measure of the Euclidean distance and the *JSD* distance. To compute the *JSD* distance for a candidate edge, we need to specify a group of related data points.

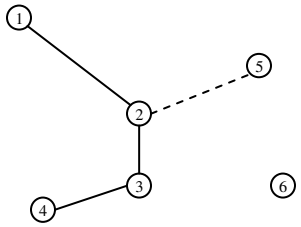

Figure 2. An Illustrative Example for Nearest Neighbors

Let $b \geq 2$ be a prespecified number (a natural choice for b would be the group size m). We define the *b-nearest neighbors* of a candidate edge $e(u, v)$ to be a set of vertices that include u, v, and the first $b - 2$ vertices that are encountered by a breadth-first search within the partial MST starting from vertex u (at the beginning, the partial MST may contain fewer than $b - 2$ vertices, in which case all vertices in the partial MST will belong to the nearest neighbors). In Figure 2, the 2-nearest neighbors of $e(2, 5)$ are vertices 2 and 5. The 3-nearest neighbors of $e(2, 5)$ are either {2, 5, 1} or {2, 5, 3}, depending on which of vertices 1 and 3 is encountered first by the search (the Euclidean distance can also be use to break such a tie, but it incurs a slightly higher computational cost). The 4-nearest neighbors of $e(2, 5)$ are vertices {2, 5, 1, 3}. Note that even if the Euclidean distance between vertices 2 and 1 is longer than that between vertices 2 and 4, the breadth-first search will select vertex 1 instead of 4, because vertex 1 is closer to vertex 2 in terms of "degree of separation." We define the nearest neighbors based on the breath-first search because records in such a neighborhood are likely to be grouped together eventually when the MST are partitioned to form the subtrees (clusters).

For each candidate edge, our algorithm identifies its nearest neighboring vertices in the partial MST and calculates the *JSD* value based on the class distribution of the corresponding (nearest neighbor) records. Consequently, there are two "distance" measures for each candidate edge: the Euclidean distance and the *JSD* distance. Both distances are normalized to have values in range [0, 1]. The class restricted MST is built using a composite measure representing the tradeoff between these two aspects.

Definition 3. Given a partial MST and a candidate edge e, let L_e be the Euclidean distance of e. Let $B(e)$ represent the group of nearest neighbors of e (with a prespecified group size b). The *composite distance* (*CD*) of e is defined as

$$CD_e = \alpha L_e + (1 - \alpha) JSD_{B(e)}, \tag{6}$$

where $\alpha \in [0,1]$ is a weight parameter and $JSD_{B(e)}$ follows from Definition 2.

The weight parameter α represents the tradeoff between the Euclidean and *JSD* distances. In general, the larger the α value is, the more similar a *CD*-based MST is to a traditional MST, which implies more emphasis on grouping similar data records together and less emphasis on the divergence of the classes in a group. If the class distribution in $B(e)$ is already well represented, then $JSD_{B(e)}$ will be small and the *CD* value will depend largely on the Euclidean distance. In our MST-based algorithm, the default α value is set to 0.5 to assign an equal weight to the Euclidean and *JSD* distances.

The *CD* measure can be used instead of the Euclidean distance to build a class-restricted MST. The measure is computed dynamically during the process of MST construction and it depends on the partial MSTs and candidate edges. It becomes undefined once the MST is complete. Hence, it cannot be used for cutting the edges of the MST to form the data clusters. While it is possible to simply use the Euclidean distance for the purpose of removing edges,

for our problem it will be more appropriate to use a distance measure that also considers the class distribution. We describe such a measure next.

When a set of data is partitioned, the ensuing subsets typically become more homogeneous in class values. To measure this difference in homogeneity before and after partitioning, we define the weighted *JSD* of a parent group as follows.

Definition 4. Let p be a parent group containing s subgroups, labeled as 1, …, s. Let n and n_g $(g = 1,...,s)$ be the number of records in p and in each subgroup, respectively. The *weighted JSD* of p is defined as

$$WJSD_p(f,F) = \sum_{g=1}^{s} \frac{n_g}{n} JSD_g(f,F). \tag{7}$$

For example in Figure 1, assuming the parent is the entire dataset, with the subgroups enclosed by the solid-lined loops, the weighted *JSD* is

$$WJSD_p(\cdot) = (3/9)(0.585) + (3/9)(0.191) + (3/9)(0.191) = 0.323,$$

whereas it is zero with the subgroups enclosed by the dash-enclosed loops.

The weighted *JSD* has the following property with respect to the group *JSD*.

Lemma 1. *The weighted JSD of the subgroups is always larger than or equal to the parent group JSD; i.e.,*

$$WJSD_p(f,F) \geq JSD_p(f,F), \forall p. \tag{8}$$

Sketch of Proof. If $q(x)$ is a convex function of a random quantity x, then it follows by generalizing the notion of convex combination that, with $\sum_{g=1}^{s} \lambda_g = 1$,

$$\sum_{g=1}^{s} \lambda_g q(x_g) \geq q(\sum_{g=1}^{s} \lambda_g x_g). \tag{9}$$

As mentioned earlier, *KLD* is a convex function. So, it follows by replacing $q(x_g)$ with $KLD_g(\cdot)$ and using (4), (5) and (7) that the left-hand side of (8) can be represented as the left-hand side of (9). Similarly, the right-hand side of (8) can be represented as the right-hand side of (9). □

With weighted *JSD* and its property described in Lemma 1, we can define a measure that takes both the Euclidean distance and class distribution into account for partitioning the MST.

Definition 5. Let e be an edge in the MST, L_e be the Euclidean distance length of e, and p_e be the parent group before cutting e. The *divergence/length ratio* is defined as:

$$r_e = \frac{WJSD_{p_e}(\cdot) - JSD_{p_e}(\cdot)}{L_e}. \tag{10}$$

The numerator, $WJSD_{p_e}(\cdot) - JSD_{p_e}(\cdot)$, is the increase in class divergence due to cutting e, resulting in two subgroups. So, r_e will be small when the increase in divergence is small and/or the Euclidean distance of e is large. Therefore, given an MST, the edge with minimum r_e value should be cut first to obtain two subtrees (representing two subgroups of data). This process continues for each of the subtrees until no edge is removable (an edge is removable if all of the ensuing subtrees contains no fewer than the specified minimum number of data points). This will result in clustered data that satisfy the group size constraint in (1).

Our algorithm, called CREST (for Class REstricted Spanning Tree), is described in Figure 3. In terms of computational complexity, Prim's algorithm for constructing MST is of order $O(N^2)$, where N is the number of records in the dataset. Finding the b nearest neighbors for computing the composite distance takes $O(b)$ time. So, Step 1 takes $O(bN^2)$ time, which is still of order $O(N^2)$, since b is much smaller than N. The edge-cutting

operation in Step 2 involves getting the edge with minimum r_e and, if the edge is cut, updating the counts of the related records and classes. This step takes $O(N)$ time in the worst case scenario. So, the worst-case time complexity for the whole edge-cutting phase (Steps 2 and 3) is of order $O(N^2)$, which is also the case for the entire algorithm.

1. Construct an MST using Prim's algorithm, where the composite distance defined in (6) is used as the distance measure.
2. Identify the edge in the MST having the minimum r_e value. Cut it if it is removable; otherwise, do not consider this edge in later iterations.
3. Repeat Step 2 until no removable edge is available.

Figure 3. CREST Algorithm

Micro-Perturbation Approach

In most clustering-based data masking methods, disclosure control is achieved by replacing the values of a QI or confidential attribute in a group with the group-average value of the corresponding attribute (Domingo-Ferrer and Mateo-Sanz 2002; Laszlo and Mukherjee 2005; Li and Sarkar 2006a). Taking the average, however, results in a reduction in variance and potentially serious distortion in relationships between attributes in the masked data. In addition, the subset average tends to be close to some original values in the same subset and thus may not provide sufficient protection against disclosure attacks. To overcome these problems, we propose a noise-based micro-perturbation method that preserves the mean vector and covariance matrix of the data while providing better disclosure protection than average-based microaggregation methods. The covariance matrix, which includes variance and covariance components, is an important measure of variation in each attribute and of relationships between different attributes.

Noise-based perturbation methods add noise to the original data to disguise their true values. One limitation of this approach is that the perturbation mechanisms typically depend on some assumptions about the properties of the data, such as normality or monotonicity (Liew et al. 1985; Agrawal and Srikant 2000). This can cause data utility to deteriorate when the assumptions are violated. Another limitation is that the variance of the perturbed data is always larger than that of the original data. The perturbation methods are usually applied to the entire dataset (Traub et al. 1984; Agrawal and Srikant 2000), instead of to a partitioned set. Since average-based aggregation reduces variance, we introduce a novel approach that adds noise after aggregation to each partitioned group to offset the reduction in variance. This idea can be implemented by directly replacing the data for each group using a statistical distribution with the mean equal to the group-average and an appropriate amount of noise that essentially preserves the variances and covariances of the original data. We provide some important analytical results with regard to the parameters of noise to be added at a group level in Theorems 1 and 2 below.

For convenience, we assume that all non-class attributes are to be perturbed. Let J be the number of non-class attributes, G be the number of total groups after clustering, and n_g be the number of records in group g. Let X_j $(j = 1,...,J)$ be a non-class attribute, x_{ij} $(i = 1,...,N)$ be the value of X_j in the ith record, and $\overline{X}_j$ be the overall mean of the X_j values. Let x_{ij}^g $(i = 1,...,n_g)$ be the value of X_j in the ith record in group g, and $\overline{x}_j^g$ be the mean of the X_j values in group g. For the perturbed data, we replace X with Y in the notation.

Theorem 1. *For each group g, if perturbed data is generated using a multivariate distribution with mean vector* $\overline{\mathbf{x}}^g = [\overline{x}_1^g, ..., \overline{x}_J^g]'$ *at the group level, then the sample mean vector on the entire perturbed data,* $\overline{\mathbf{Y}}$*, is an unbiased estimator of the true mean vector* $\boldsymbol{\mu}$*; i.e.,*

$$E(\overline{\mathbf{Y}}) = \boldsymbol{\mu}. \tag{11}$$

Proof. For $j = 1, \ldots, J$,

$$\bar{Y}_j = \frac{1}{N}\sum_{i=1}^{N} y_{ij} = \frac{1}{N}\sum_{g=1}^{G}(\sum_{i=1}^{n_g} y_{ij}^g) = \frac{1}{N}\sum_{g=1}^{G}(n_g \bar{y}_j^g) .$$

So,

$$E(\bar{Y}_j) = E\left[\frac{1}{N}\sum_{g=1}^{G}(n_g \bar{y}_j^g)\right] = E\left[\frac{1}{N}\sum_{g=1}^{G}(n_g \bar{x}_j^g)\right] = E(\bar{X}_j) = \mu_j . \quad \square$$

Next, we discuss the covariance parameters for the perturbed data. After all groups are formed, a seemingly straightforward approach is to compute sample covariance matrix for each group and then use it in the multivariate distribution for generating perturbed data at the group level. This approach, however, is good only if the group size *m* is sufficiently larger than the number of attributes *J*. But this is often not true for a clustering-based method. When $m < J$, in particular, the sample covariance matrix becomes singular and unstable, and thus not appropriate for generating simulated data. There exist a few estimation methods to deal with the singular covariance matrix problem (e.g., Ledoit and Wolf 2003), but none of them is unbiased for the true covariance matrix. We propose an unbiased estimator based on the unique nature of the clustering-based approach.

Let $\mathbf{S}_x$ be the sample covariance matrix with its (*j*, *h*) element $s_{jh} = s(X_j, X_h)$ being the covariance between X_j and X_h. Let $s(\bar{x}_j^g, \bar{x}_h^g)$ be the (*j*, *h*) element of the sample covariance matrix when the original data values are replaced by group averages. Theorem 2 below provides group-level covariance parameters for micro-perturbation, which results in perturbed data whose sample covariance matrix is an unbiased estimator of the true covariance matrix.

Theorem 2. *If perturbed data is generated for each group independently using a multivariate distribution with mean vector* $\bar{\mathbf{x}}^g = [\bar{x}_1^g, ..., \bar{x}_J^g]'$ *and covariance matrix* $\mathbf{S}^g$ *whose* (*j, h*) *element is*

$$s_{jh}^g = \frac{N-1}{N-G}\left[s(X_j, X_h) - s(\bar{x}_j^g, \bar{x}_h^g)\right], \tag{12}$$

then the sample covariance matrix based on the entire perturbed data, $\mathbf{S}_y$*, is an unbiased estimator of the true covariance matrix* $\mathbf{\Sigma}$*; i.e.,*

$$E(\mathbf{S}_y) = \mathbf{\Sigma} . \tag{13}$$

Sketch of Proof. Similar to the decomposition of sum of squares in analysis of variance (ANOVA), the total sum of cross-products for the perturbed data can be decomposed as

$$\sum_{g=1}^{G}\sum_{i=1}^{n_g}(y_{ij}^g - \bar{Y}_j)(y_{ih}^g - \bar{Y}_h) = \sum_{g=1}^{G}\sum_{i=1}^{n_g}(\bar{y}_j^g - \bar{Y}_j)(\bar{y}_h^g - \bar{Y}_h) + \sum_{g=1}^{G}\sum_{i=1}^{n_g}(y_{ij}^g - \bar{y}_j^g)(y_{ih}^g - \bar{y}_h^g). \tag{14}$$

Dividing both sides of (14) by $N - 1$, we have:

$$s(Y_j, Y_h) = s(\bar{y}_j^g, \bar{y}_h^g) + \frac{1}{N-1}\sum_{g=1}^{G}(n_g - 1)s(y_j^g, y_h^g) , \tag{15}$$

where the second term on the right is the within-group covariance that is ignored when replacing the original values with group-averages. Taking expectations on both sides of (15), we have,

$$E[s(Y_j, Y_h)] = E[s(\bar{y}_j^g, \bar{y}_h^g)] + \frac{N-G}{N-1}E[s(y_j^g, y_h^g)] , \tag{16}$$

where $E[s(y_j^g, y_h^g)]$ represents the expected covariance within group *g*, which is the covariance used to generate the perturbed data for group *g*. It follows by taking expectations on (12) and substituting the result in (16) that

$$E[s(Y_j, Y_h)] = E[s(\bar{y}_j^g, \bar{y}_h^g)] + \left\{E[s(X_j, X_h)] - E[s(\bar{x}_j^g, \bar{x}_h^g)]\right\}.$$

Since perturbed data is generated using $\bar{\mathbf{x}}^g$,

$$E[s(\bar{y}_j^g, \bar{y}_h^g)] = E[s(\bar{x}_j^g, \bar{x}_h^g)].$$

Therefore,

$$E[s(Y_j, Y_h)] = E[s(X_j, X_h)] = \Sigma_{jh}, \ \forall j, h. \qquad \square$$

Corollary 1: *If perturbed data is generated using the distribution parameters specified in Theorem 2, then the sample variances on the perturbed data are unbiased estimates of the true variances.*

The proof follows immediately from the fact that the variances are the diagonal elements of the covariance matrix. The micro-perturbation procedure can be implemented by generating perturbed data for each group using a distribution (e.g., multivariate normal) with mean vector $\bar{\mathbf{x}}^g$ and covariance matrix $\mathbf{S}^g$. The complete algorithm, which is called CAMP-CREST (for Clustering And Micro-Perturbation with Class REstricted Spanning Tree), is given in Figure 4. Step II of the algorithm takes $O(NJ)$ time, while Step I takes, as explained earlier, $O(N^2)$ time. Since *J* is much smaller than *N*, the time complexity for the entire CAMP-CREST algorithm is of order $O(N^2)$.

I. Run CREST algorithm described in Figure 3.

II. For each group formed from Step I, replace the quasi-identifier attribute values with perturbed values generated using a multivariate normal distribution $N(\bar{\mathbf{x}}^g, \mathbf{S}^g)$.

Figure 4. CAMP-CREST Algorithm

The use of the multivariate normal distribution in generating perturbed data seems to suggest that the perturbed data follows a multivariate normal distribution, which would be problematic if the original data is not normally distributed. This is not the case, however. The distribution of the perturbed data is dictated by the distribution of the clusters, not by the distribution used for micro-perturbation. If we view each cluster geometrically as a packed data object in the relevant space, then the joint distribution of these data packs remains the same before and after micro-perturbation, because perturbation is performed within each pack and, by Theorem 1, the center of each pack remains statistically unchanged. The choice of a distribution form for micro-perturbation affects data distribution only within each pack, and has no impact on the distribution of the packs. As a result, the joint distribution of the entire dataset is reasonably preserved. This is achieved without assuming any knowledge about statistical distributions of the original data. If the data indeed follows a multivariate normal distribution, then based on Theorems 1 and 2, the original joint distribution is completely preserved, because the statistical properties of a multivariate normal distribution can be fully captured by its mean vector and covariance matrix.

Experiments

We conducted experiments on three real-world datasets to evaluate the proposed method. The Association for Information Systems conducts annual surveys of MIS faculty salary offers (Galletta 2004). We selected the offer data from 1999 to 2002 (attributes are consistent for these four years and somewhat different for the other years). The dataset consists of 443 records of faculty members who received offers during the period. There are 11 attributes, including salary offered, position, course load, number of years teaching, region, year indicator, etc. They are of numeric, ordinal or binary type and thus can be easily handled by the algorithms used in the experiment. The confidential class attribute is salary offered, which was originally numeric. To see the impact of multiple classes, we created two versions of datasets based on this data. In the first set, the salary values are grouped into two classes with approximately balanced class distributions (called Offer2). The second set has three approximately balanced classes (Offer3).

The second dataset, Diabetes, contains 768 records of female patients, with 9 attributes, including age, number of times pregnant, and several numeric medical measures (Asuncion and Newman 2007). The confidential class attribute is test result, which has two classes: positive (34.9%) and negative (65.1%).

The third dataset, Medicare, contains 4,406 records of individuals who are covered by the Medicare insurance program (Deb and Trivedi 1997). It has 22 attributes, including age, gender, race, education, marital status, family income, employment status, number of visits to a physician office, number of visits to an emergency room, number of hospital stays, and additional health insurance coverage information, etc. The confidential class attribute is individuals' chronic conditions, which has three classes: zero chronic disease (23.3%), one disease (34.0%), and multiple diseases (42.7%). All of the non-class categorical attribute values had been preprocessed with 0-1 coding, making the data available for numeric operations.

We compare CAMP-CREST to the standard MST-based microaggregation method by Laszlo and Mukherjee (2005).[1] The class attribute is not used in constructing the MST. Because Step 1 of the CREST algorithm (Figure 3) takes relatively longer time (due to the search for the nearest neighbors) than the remaining part of the CAMP-CREST algorithm, we also tested an alternative algorithm that replaces this step with a standard MST procedure. We call this variant a "Partial CAMP-CREST" algorithm, which relies on Step 2 of the CREST algorithm to obtain better group-level class distributions during the MST partitioning phase. For simplicity, we assume all non-confidential attributes are QI attributes and thus subject to masking. The confidential attributes are not masked.

We use two measures to assess disclosure risk. The first measure assesses if the class distribution in each group is well represented, which relates to the confidential value disclosure risk. The measure is based on the classical chi-square statistic, defined as:

$$X^2 = \frac{1}{G}\sum_{g=1}^{G}\sum_{k=1}^{C}(n_{gk} - \frac{n_g}{N}N_k)^2 \Big/ (\frac{n_g}{N}N_k), \tag{17}$$

where N_k and n_{gk} are the number of records with the kth class in the full dataset and in group g, respectively. The X^2 statistic measures the closeness between the class distribution of a group and the ideal class distribution for the group, averaged over all groups. Clearly, the smaller the X^2 value, the smaller the disclosure risk for the individual class values in a group, as the class distribution in the group is closer to the overall class distribution. This measure is related to the clustering part of the proposed method.

The second measure is related to the micro-perturbation part of the proposed method, and it concerns re-identification risk. This measure, called *record linkage*, was proposed by Pagliuca and Seri (1999). It uses the Euclidean distance between a record shown on an original data file and that shown on the corresponding masked file. A record in the masked file is said to be "linked" if the record closest to it in the original file is indeed the corresponding unmasked record. A record in the masked file is "second closely linked" if the second closest record in the original file is the corresponding one. The record linkage measure is defined as the percentage of records that are either "linked" or "second closely linked". A smaller value for this measure indicates lesser re-identification risk.

Data quality is measured by information loss due to data masking. Univariate information loss is measured using two metrics, average absolute bias in mean (ABIM) and average absolute bias in standard deviation (ABISD). Multivariate information loss is measured using average absolute bias in correlation (ABICO). These measures are defined below, based on Adam and Wortmann (1989), and Domingo-Ferrer and Torra (2001):

$$\text{ABIM} = \frac{1}{J}\sum_{j=1}^{J}\left|\frac{\overline{Y}_j - \overline{X}_j}{\overline{X}_j}\right|, \tag{18}$$

$$\text{ABISD} = \frac{1}{J}\sum_{j=1}^{J}\left|\frac{s(Y_j) - s(X_j)}{s(X_j)}\right|, \tag{19}$$

$$\text{ABICO} = \frac{1}{J(J-1)/2}\sum_{j=1}^{J}\sum_{h=j+1}^{J}\left|\frac{r(Y_j, Y_h) - r(X_j, X_h)}{r(X_j, X_h)}\right|, \tag{20}$$

[1] We have not selected the microaggregation method proposed by Domingo-Ferrer and Mateo-Sanz (2002), since Laszlo and Mukherjee (2005) have shown that the performance of this method is similar to the MST-based microaggregation.

where $s(X_j)$ and $s(Y_j)$ are standard deviations calculated on the original and masked data respectively, and $r(X_j, X_h)$ and $r(Y_j, Y_h)$ are the correlations between attributes *j* and *h* calculated on the original and masked data respectively (the number of such correlations is $J(J-1)/2$). The absolute values are taken in the above definitions to prevent positive and negative biases over different attributes from canceling out. A small ABIM, ABISD or ABICO value indicates that the means, standard deviations, or correlations for the masked data are on average close to those for the original data. Clearly, the smaller the ABIM, ABISD and ABICO values, the smaller the information loss in mean, standard deviation and correlation.

The results of CAMP-CREST and Partial CAMP-CREST vary somewhat with different random number seeds. Therefore, these two algorithms were run five times for each dataset, with a new random number being generated each time. The average results are reported. Comparisons of different masking methods should be made in terms of both disclosure risk and information loss measures. As described above, there are two disclosure risk measures and three information loss measures. To facilitate the comparisons across multiple criteria, we used record linkage, which measures re-identification risk, as the control factor in the experiments. For each dataset, we adjusted the group sizes for the different methods to produce masked data such that the record linkage values for all three methods were about the same (to remain on the conservative side we ensured that CAMP-CREST had the smallest values). Effort was also made to ensure that no record linkage value from any method is larger than 5% in order to provide reasonable disclosure protection. The performances of the three methods are then examined on the X^2 and information loss measures.

Table 1. Results of Experiments

Data	Method	Time (seconds)	Linkage (%)	X^2	ABIM (%)	ABISD (%)	ABICO (%)
Offer2	MST	0.6	4.97	5.07	0	47.84	440.85
(403 records)	Partial CAMP-CREST	0.6	4.83	4.14	2.55	3.44	112.16
	CAMP-CREST	1.4	4.74	3.59	2.28	2.87	113.71
Offer3	MST	0.6	4.97	7.70	0	47.84	440.85
(403 records)	Partial CAMP-CREST	0.6	4.97	5.56	2.87	2.74	104.97
	CAMP-CREST	2.2	4.70	4.88	2.84	3.35	95.20
Diabetes	MST	1.3	2.73	3.08	0	36.44	156.49
(768 records)	Partial CAMP-CREST	1.3	2.68	2.39	1.22	2.63	30.83
	CAMP-CREST	4.4	2.42	1.77	1.20	2.81	28.42
Medicare	MST	31.8	3.22	3.10	0	39.88	444.68
(4406 records)	Partial CAMP-CREST	33.5	3.19	2.83	2.24	2.08	111.48
	CAMP-CREST	127.3	2.72	1.95	2.01	2.27	113.77

The results of the experiments are shown in Table 1. The total number of records in each dataset is shown below the dataset name. The record linkage rate is obtained by dividing the number of linked records by the total number of records in the dataset (note that the data receiver does not know which linked records are indeed correctly matched) . For example, the number of linked records with the MST method for Offer2 is 20. It is clear that, for all the datasets, the X^2 values associated with CAMP-CREST are substantially smaller than those with the other two methods, which indicates, as mentioned earlier, lower confidential value disclosure risk. In most cases, the X^2 values with Partial CAMP-CREST are also considerably smaller than those with standard MST. It is also observed from the results of Offer2 and Offer3 that the difference in the X^2 values becomes larger when the number of classes increases. This is not surprising because the class distribution is likely to be distorted more in this situation if the distribution is not considered in the clustering procedure.

In terms of the information loss measures, CAMP-CREST and Partial CAMP-CREST show small deviations in mean (ABIM) on all datasets, which is due to the randomness from simulated data. The approximately 2% deviations in mean are generally acceptable in practice (Liew et al. 1985; Aggarwal and Yu 2008). For the ABISD

and ABICO measures, both CAMP-CREST and Partial CAMP-CREST significantly outperform standard MST. This indicates that the joint distributions of the dataset are better preserved by the proposed methods than standard MST. Note that there are only small differences between these two methods in all three information loss measures. This is because they both implement the same micro-perturbation approach for the non-class attribute data, on which the three measures are computed.

In terms of computing time, MST and Partial CAMP-CREST are about the same. CAMP-CREST runs several times slower than these two algorithms. As discussed earlier, this is due to the search for the nearest neighbors in Step 1 of the CREST algorithm (Figure 3). Therefore, for large amounts of data, if runtime is a concern, the Partial CAMP-CREST can be a good alternative for the complete CAMP-CREST.

Conclusions and Extensions

We have presented a confidential-class-restricted clustering and micro-perturbation method for privacy-preserving data sharing and data analysis. Our proposed method prevents or limits potential privacy disclosure risks that can occur when a traditional clustering-based technique such as microaggregation is used. We have shown analytically that the proposed method preserves some important statistical properties of the data regardless of the actual distributions of the data. Our empirical study has demonstrated that the method can lead to significantly improved performance over existing approaches. The proposed approach has important management and policy implications. As data-sharing and data-mining techniques are being increasingly used in areas such as healthcare and medical research, crime analysis, database marketing, and customer relationship management, there is a rising public sentiment that individual privacy is being severely eroded. Our proposed approach addresses this imperative issue and provides a solution to resolve the conflict between data sharing and privacy protection.

We have assumed in this study that there is only one confidential class attribute in the data. The proposed approach can be extended to handle multiple confidential class attributes. We suggest two approaches. The first is to consider all confidential class attributes together as one compound class attribute. Suppose, for instance, there is another confidential class attribute representing test result for another disease in the example in Figure 1, which also has two values: positive and negative. A compound attribute can be created, which would have four categories, formed by different combinations of test results for the two diseases. The transformed dataset would have two non-class attributes (Age and Weight) and one (compound) class attribute. The proposed method can then be applied to this transformed dataset. The second approach is to run CAMP-CREST multiple times, each time dealing with one class attribute without considering the remaining class attributes. The micro-perturbed non-class attribute values for each record from multiple runs will be different while the values of all class attributes are unchanged. In the final release version, an aggregated value (e.g., average) over the results of multiple runs can be used for the respective non-class attribute value for each record.

In this study, we have considered applying an information divergence measure to the MST-based clustering method to deal with the confidential class restriction problem. The same idea could also be applied to other clustering methods used for microaggregation. We plan to examine such alternative approaches in future.

References

Adam, N.R., and Wortmann, J.C. 1989. "Security-Control Methods for Statistical Databases: A Comparative Study," *ACM Computing Surveys* (21:4), pp. 515-556.

Aggarwal, C.C., and Yu, P.S. (eds.) 2008. *Privacy-Preserving Data Mining: Models and Algorithms*, New York: Springer.

Agrawal, R., and Srikant, R. 2000. "Privacy-Preserving Data Mining," in *Proceedings of 2000 ACM SIGMOD International Conference on Management of Data*, Dallas, TX, pp. 439-450.

Asuncion, A., and Newman, D.J. 2007. *UCI Machine Learning Repository*. Retrieved from http://www.ics.uci.edu.

Chowdhury, D.S., Duncan, G.T., Krishnan, R., Roehrig, S.F., and Mukherjee, S. 1999. "Disclosure Detection in Multivariate Categorical Databases: Auditing Confidentiality Protection Through Two New Matrix Operators," *Management Science* (45:12), pp. 1710-1723.

Cox, L.H. 1980. "Suppression Methodology and Statistical Disclosure Control," *Journal of the American Statistical Association* (75:370), pp. 377-385.

Deb, P., and Trivedi, P.K. 1997. "Demand for Medical Care by the Elderly: A Finite Mixture Approach," *Journal of Applied Econometrics* (12:3), pp. 313-336.

Dalenius, T, and Reiss, S.P. 1982. "Data Swapping: A Technique for Disclosure Control," *Journal of Statistical Planning and Inference* (6:1), pp. 73-85.

Defays, D., and Nanopoulos. P. 1993. "Panels of Enterprises and Confidentiality: The Small Aggregates Method," in *Proceedings of Statistics Canada Symposium 92 on Design and Analysis of Longitudinal Surveys*, Ottawa, Canada, pp. 195-204.

Dixon P. 2006. "Medical Identity Theft: The Information Crime That Can Kill You," *The World Privacy Forum*. Retrieved from http://www.worldprivacyforum.org/pdf/wpf_medicalidtheft2006.pdf.

Domingo-Ferrer, J., and Mateo-Sanz. J.M. 2002. "Practical Data-Oriented Microaggregation for Statistical Disclosure Control," *IEEE Transactions on Knowledge and Data Engineering* (14:1), pp. 189-201.

Domingo-Ferrer, J., and Torra, V. 2001. "A Quantitative Comparison of Disclosure Control Methods for Microdata," in *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, P. Doyle, J. Lane, J. Theeuwes, and L. Zayatz (eds.), Amsterdam, Netherlands: North-Holland, pp. 111-134.

Galletta, D. 2004. "MIS Faculty Salary Survey Results," Retrieved from http://www.pitt.edu/~galletta/salsurv.html.

Garfinkel, R., Gopal, R., and Goes, P. 2002. "Privacy Protection of Binary Confidential Data Against Deterministic, Stochastic, and Insider Threat," *Management Science* (48:6), pp. 749-764.

Gopal, R., Garfinkel, R., and Goes, P. 2002. "Confidentiality Via Camouflage: The CVC Approach to Disclosure Limitation When Answering Queries to Databases," *Operations Research* (50:3), pp. 501-516.

Kaelber, D.C., Jha, A.K., Johnston, D., Middleton, B., and Bates, D.W. 2008. "A Research Agenda for Personal Health Records (PHRs)," *Journal of American Medical Informatics Association* (15:6), pp. 729-736.

Kullback, S. 1959. *Information Theory and Statistics*, New York: John Wiley & Sons.

Laszlo, M., and Mukherjee, S. 2005. "Minimum Spanning Tree Partitioning Algorithm for Microaggregation," *IEEE Transactions on Knowledge and Data Engineering* (17:7), pp. 902-911.

Ledoit, O., and Wolf, M. 2003. "Improved Estimation of the Covariance Matrix of Stock Returns with an Application to Portfolio Selection," *Journal of Empirical Finance* (10:5), pp. 603-621.

Li, X.-B., and Sarkar, S. 2006a. "A Tree-Based Data Perturbation Approach for Privacy-Preserving Data Mining," *IEEE Transactions on Knowledge and Data Engineering* (18:9), pp. 1278-1283.

Li, X.-B., and Sarkar, S. 2006b. "Privacy Protection in Data Mining: A Perturbation Approach for Categorical Data," *Information Systems Research* (17:3), pp. 254-270.

Liew, C.K., Choi, U.J., and Liew, C.J. 1985. "A Data Distortion by Probability Distribution," *ACM Transactions on Database Systems* (10:3), pp. 395-411.

Lin, J. 1991. "Divergence Measures Based on the Shannon Entropy," *IEEE Transactions on Information Theory* (37:1), pp. 145-151.

Machanavajjhala, A., Gehrke, J., Kifer, D., and Venkitasubramaniam, M. 2006. "*l*-Diversity: Privacy Beyond *k*-Anonymity," in *Proceedings of 22nd IEEE International Conference on Data Engineering*, Atlanta, GA, pp. 24-35.

Pagliuca, D., and Seri, G. 1999. "Some Results of Individual Ranking Method on the System of Enterprise Accounts Annual Survey, Esprit SDC Project, Deliverable MI-3/D2.

Prim R.C. 1957. "Shortest Connection Networks and Some Generalizations," *Bell System Technical Journal* 36, pp. 1389-1401.

Reiss, S.P. 1984. "Practical Data-Swapping: The First Steps," *ACM Transactions on Database Systems* (9:1), pp. 20-37.

Samarati, P. 2001. "Protecting Respondents' Identities in Microdata Release," *IEEE Transactions on Knowledge and Data Engineering* (13:6), pp. 1010-1027.

Sweeney, L. 2002. "*k*-Anonymity: A Model for Protecting Privacy," *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems* (10:5), pp. 557-570.

Teltzrow, M., and Kobsa, A. 2004. "Impacts of User Privacy Preferences on Personalized Systems: A Comparative Study," in *Designing Personalized User Experiences in eCommerce*, Dordrecht, Netherlands: Kluwer Academic Publishers, pp. 315-332.

Traub, J.F., Yemini, Y., and Wozniakowski, H. 1984. "The Statistical Security of a Statistical Database," *ACM Transactions on Database Systems* (9:4), pp. 672-679.

US General Accounting Office. 2004. *Data Mining: Federal Efforts Cover a Wide Range of Uses*. Retrieved from http://www.gao.gov/new.items/d04548.pdf.

Winkler, W.E. 2007. "Examples of Easy-to-Implement, Widely Used Methods of Masking for Which Analytic Properties Are Not Justified," Census Bureau Research Report Series (Statistics #2007-21). Retrieved from http://www.census.gov/srd/papers/pdf/rrs2007-21.pdf.

Zahn, C.T. 1971. "Graph-Theoretical Methods for Detecting and Describing Gestalt Clusters," *IEEE Transactions on Computers* (C-20:4), pp. 68-86.

Zeller T. 2006. "AOL Executive Quits After Posting of Search Data," *International Herald Tribune*, Aug. 23, p. 13.