Published in final edited form as:
Manage Sci. 2013 April 1; 59(4): . doi:10.1287/mnsc.1120.1584.

Class Restricted Clustering and Micro-Perturbation for Data Privacy

Xiao-Bai Li and
Department of Operations and Information Systems, University of Massachusetts Lowell, Lowell, Massachusetts 01854

Sumit Sarkar
School of Management, University of Texas at Dallas, Richardson, Texas 75080

Xiao-Bai Li: xiaobai_li@uml.edu; Sumit Sarkar: sumit@utdallas.edu

Abstract

The extensive use of information technologies by organizations to collect and share personal data has raised strong privacy concerns. To respond to the public's demand for data privacy, a class of clustering-based data masking techniques is increasingly being used for privacy-preserving data sharing and analytics. Traditional clustering-based approaches for masking numeric attributes, while addressing re-identification risks, typically do not consider the disclosure risk of categorical confidential attributes. We propose a new approach to deal with this problem. The proposed method clusters data such that the data points within a group are similar in the non-confidential attribute values whereas the confidential attribute values within a group are *well distributed*. To accomplish this, the clustering method, which is based on a minimum spanning tree (MST) technique, uses two risk-utility tradeoff measures in the growing and pruning stages of the MST technique respectively. As part of our approach we also propose a novel cluster-level micro-perturbation method for masking data that overcomes a common problem of traditional clustering-based methods for data masking, which is their inability to preserve important statistical properties such as the variance of attributes and the covariance across attributes. We show that the mean vector and the covariance matrix of the masked data generated using the micro-perturbation method are unbiased estimates of the original mean vector and covariance matrix. An experimental study on several real-world datasets demonstrates the effectiveness of the proposed approach.

Keywords

Privacy; confidentiality; clustering; minimum spanning tree; microaggregation; data perturbation; information theory

1. Introduction

As data-sharing and data-mining technologies are being increasingly used in areas such as healthcare research, crime analysis, 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 Center for Medicare and Medicaid Services, a federal agency, provides healthcare researchers with individual Medicare and Medicaid claims data (http://www.cms.hhs.gov/). This rapid transition towards electronic medical records (EMR) and data sharing has raised pressing concerns about privacy. Indeed, there is evidence that EMR has caused medical identity disclosure 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 fierce public protests, including several law suits and legal complaints against AOL, and AOL removed the data from the website within days (Zeller 2006). In another incident, Netflix had recently awarded $1 million to a research team led by two AT&T employees for winning a contest to improve the predictive accuracy of the company's movie recommendation system by over 10%. The contest, which lasted for three years, was considered by many to be a great research and business success. However, Netflix had to cancel plans for a sequel when it was discovered that the de-identified data released for the contest, which included movie recommendations and choices made by customers, could in fact be used to re-identify the customers (Lohr 2010). 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 private 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 data 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, Duncan and Mukherjee 2000, Gopal et al. 2002, Garfinkel et al. 2002, Kadane et al. 2006). A related approach is *cell suppression* in publishing tabular data (Cox 1980, Fischetti and Salazar 2001). Both query restriction and cell suppression focus on how to release summary statistics (e.g., count, sum and average) without disclosing individuals' confidential data. When the data released is intended for more sophisticated data 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, Lee et al. 2010, Oganian 2010); *data swapping*, which involves exchange of attribute values between different records (Dalenius and Reiss 1982, 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 and Sweeney 1998, Sweeney 2002, Garfinkel et al. 2007). All these methods attempt to preserve the utility of the masked data, as measured by various data quality metrics.

There are two types of privacy disclosure widely accepted in the literature (Duncan and Lambert 1989). They are *identity disclosure* (or *re-identification*), which occurs when a data intruder is able to match a record in a dataset to an individual; and *value disclosure* (also referred to as *confidential class disclosure* in this paper), which occurs when an intruder is able to predict the confidential value(s) of an individual record. Related to these two types of disclosures, 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 non-confidential attributes can often be used to identify individuals by matching data from different sources, resulting in identity disclosure. Such attributes are collectively called a *quasi-identifier* (QI) in the literature. For example, Sweeney (2002) found 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 data privacy research, it is typically assumed that the explicit identifiers have already been removed from the data. Data masking is applied to QI attributes to limit re-identification and to confidential attributes to limit value disclosure.

A class of grouping-based anonymization approaches has gained considerable popularity in the data privacy research community. The basic idea behind these approaches is to partition a dataset into groups of similar records and then anonymize the QI attribute values at the group level so that the records within a group are indistinguishable. There are two popular approaches along this line: microaggregation and *k*-anonymity. They have both had significant impacts on privacy related policies and practices. The microaggregation approach has long been used by the U.S. Internal Revenue Service and European Union's Eurostat for data publishing and sharing purposes (Defays 1997). The findings of *k*-anonymity studies have been used to justify some of the current privacy rules in the Health Insurance Portability and Accountability Act (DHHS 2000). This research examines a problem associated with the microaggregation approach. Typically, microaggregation replaces the values of the QI attributes with the group averages, while keeping the confidential attribute values unchanged.[1] This approach, however, is problematic in terms of confidential value disclosure. We focus on the problem when the confidential attributes are categorical.

To illustrate, consider an example dataset containing nine patient records plotted in Figure 1. For simplicity, we consider a single categorical confidential attribute, called the *class* attribute. There are two numeric QI attributes, Age and Weight, and the class attribute, Test Result, with two values: positive and negative. In Figure 1, a circle represents 'positive' and a square represents 'negative'. Suppose the minimum group size is three. Microaggregation will cluster the data such that the data points are closer to each other (in Age and Weight) within a group than to the points outside the group. This will result in three groups as shown by the three loops. 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.

Despite available methodologies to protect privacy, the disclosure of confidential information is quite widespread for the types of data we study. To verify this observation, we applied a traditional microaggregation method to a medical research dataset (Diabetes) of 768 patient records and a salary survey dataset (Offer) of 443 records (details of these datasets are provided in Section 6). We found that when the data were masked using

[1]Microaggregation can also be used for masking confidential attributes if they are numeric. However, this paper addresses the disclosure problem when the confidential attributes are categorical. So, we assume in this paper that data masking is applied to the QI attributes.

microaggregation, 21.2% of the records in the medical data were assigned to groups that had a homogeneous value in the confidential attribute. For the salary data, this number is 19.9%. Both results indicate very high disclosure risk, and have the potential to cause serious privacy breaches for the organizations that intend to share these data for legitimate research and analysis.

Microaggregation is widely used in practice. However, this confidential class disclosure problem has not been investigated in the microaggregation literature. Although a related problem has been studied in the *k*-anonymity related literature, the problem we investigate calls for a new approach suitable to the microaggregation framework where the QI attributes are anonymized with numeric aggregation or perturbation, rather than categorical generalization or suppression as done in *k*-anonymity. On another front, microaggregation approaches have been criticized as lacking formal justification in preserving statistical properties of the data (Winkler 2007). In particular, the use of the group-mean-substitution method results in bias in the variance and covariance of the data, which adversely affects the ability to conduct meaningful analysis using the data. This research provides an effective approach to address this important data quality issue as well.

To tackle the confidential value-disclosure problem for microaggregation, we develop a novel clustering approach that considers simultaneously the following two objectives when clustering the data: (1) the data points within a group are more homogeneous with respect to the QI attribute values (to ensure data quality); and (2) the confidential class values within a group are *well distributed*, i.e., the distribution of confidential class values within each group is close to the distribution of the class values for the entire data (to limit value-disclosure risk). A clustering algorithm generally does not consider the second objective, not to mention the combination of the two incongruent objectives. As such, the proposed clustering approach is new to the microaggregation literature. The approach is based on a minimum spanning tree (MST) technique, where the first objective is represented using a traditional Euclidean distance measure while the second objective is represented using an entropy-based distance measure. The two measures are integrated to form two separate composite tradeoff measures, one that is used in the MST growing stage, and the other used in the pruning stage.

To overcome the data quality problem with the traditional mean-substitution method, we propose a novel cluster-level micro-perturbation method in masking data (of course, perturbation also helps protect against re-identification risk). The proposed micro-perturbation method is innovative and elegant in that it uses the data from the entire dataset to estimate the group-level covariance matrix and thus avoids a singular matrix problem that would very often make the estimation infeasible if only within-group data is used. We show that the mean vector and the covariance matrix of the masked data generated using the micro-perturbation method are unbiased estimates of the original mean vector and covariance matrix. This provides a strong theoretical justification for the micro -perturbation method.

This study has important management and policy implications. Information privacy has become an imperative issue for managers and policy makers because it has a significant impact on an organization’s ability to leverage its data assets in order to create value for itself and its partners. The proposed approach overcomes the limitations inherent in current practice and offers enhanced protection for privacy. For example, for the medical and salary datasets mentioned earlier, the proposed approach can mask the data with no group having a homogeneous class value. This can enable data managers and privacy officers of organizations to share such data with business partners with minimal risk of disclosure of sensitive information (i.e., value-disclosure). Furthermore, as we show in the paper, our

approach leads to lower identity-disclosure risk and higher data quality than current microaggregation approaches. This will further reduce risks of privacy violations and allow organizations to safely share and publish high-quality data.

In the next section, we provide an in-depth review of related work. In Section 3, we discuss disclosure risks when applying a clustering-based technique for masking data. In Section 4, we develop an algorithm, based on a minimum spanning tree technique, for clustering data with a confidential class. Section 5 elaborates on the micro-perturbation method and its theoretical justification. Section 6 describes a set of experiments conducted on real-world datasets that demonstrates the effectiveness of our approach. Section 7 discusses the managerial and policy implications of this work. We conclude the paper and provide directions for future research in Section 8.

2. Related Work

There has been a large amount of research in the areas of *k*-anonymity and microaggregation in recent years. We provide details of the important works in these areas that have some bearing on our research.

The *k*-anonymity approach (Samarati and Sweeney 1998) is a grouping-based anonymization technique designed primarily for anonymizing categorical data. 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, *k*-anonymity focuses on re-identification risk only and does not consider confidential value disclosure. It generalizes different but similar QI attribute values into the same value within a group. 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 value-disclosure risk if their confidential values are the same.

To address this issue, Machanavajjhala et al. (2006) propose a privacy principle called *l*-diversity, which requires, in addition to *k*-anonymity, that the confidential attribute should include at least *l* well-diversified values in the *k*-anonymized data. An entropy measure is proposed to represent the diversity of the confidential values. The notion of *l*-diversity, however, does not consider the overall distribution of the confidential attribute. So, when the overall distribution is unbalanced, the *l*-diversity requirement may be difficult to satisfy. Furthermore, since the overall distribution is usually public information, the confidential value-disclosure risk can be high when the distribution of the *l*-diversified data deviates significantly from the overall distribution. To overcome this problem, Li et al. (2007) propose another privacy principle called *t*-closeness, which requires that, for each subset with the same QI attribute values, the distance between the distribution of the confidential attribute in the subset and the overall distribution cannot be larger than a threshold value *t*. Both *l*-diversity and *t*-closeness principles consider only the confidential value disclosure. The re-identification risk is still handled by the *k*-anonymity approach. As such, the *l*-diversity and *t*-closeness principles are typically implemented on a *k*-anonymity algorithm as additional constraints to the *k*-anonymity requirement. These multiple constraints can be hard to satisfy; they cause large group sizes, which is undesirable in terms of information loss (Machanavajjhala et al. 2006, Li et al. 2007).

There are several follow-up studies along the lines of *k*-anonymity, *l*-diversity and *t*-closeness. Ghinita et al. (2007) propose a fast algorithm to achieve *k*-anonymity along with *l*-diversity, but the algorithm is not designed to implement the *t*-closeness principle.

Rebollo-Monedero et al. (2010) propose a framework that masks data to minimize an information-theoretic measure similar to *t*-closeness. However, their approach is not based on grouping of data and thus does not address the confidential class disclosure caused by data grouping, which is the problem investigated in this study.

The *k*-anonymity, *l*-diversity and *t*-closeness approaches focus primarily on categorical data. When an attribute is originally captured in numeric form, these approaches convert the numeric values into intervals and then treat the intervals as categorical values. The conversion causes information loss and arithmetic operations are no longer applicable to the converted interval values. The information loss problem is more significant when considering that the hierarchies of the categories (intervals) for generalizing numeric values are typically specified a priori. Due to these concerns, the conversion practice is not desirable for numeric QI attributes and has in fact been criticized as "completely unsuitable for continuous attributes" (Domingo-Ferrer and Torra 2005, p.195). Microaggregation is preferred to *k*-anonymity for such data.

Microaggregation masks data by first clustering the data into groups of similar records and then replacing the QI attribute values with a group-level aggregated value such as the group average (Domingo-Ferrer and Mateo-Sanz 2002, Laszlo and Mukherjee 2005). The basic idea behind microaggregation in terms of (re-identification) disclosure protection is similar to that of *k*-anonymity with the distinction that the former applies primarily to numeric data while the latter to categorical data. Because the data is masked within each cluster, data utility for the entire dataset is expected to be reasonably preserved. Domingo-Ferrer and Torra (2005) propose a microaggregation framework that can deal with both numeric and categorical data (by representing categorical values in 0–1 format). Li and Sarkar (2006a) propose a clustering-based method for masking numeric data using an efficient kd-tree technique. Other computational issues with microaggregation are addressed in Domingo-Ferrer et al. (2008).

To preserve data utility, all these clustering-based approaches attempt to cluster the data such that the data points within a group are more similar than those between groups. As a result, confidential attribute values in a group are also similar, even when they are not used in clustering the data. This leads to higher disclosure risk for the confidential data, as illustrated in Figure 1. This problem is similar to that examined by *t*-closeness studies except that in *t*-closeness the masked QI attributes are categorical. In practice, there are usually many numeric QI attributes (either continuous or discrete), such as age, weight, height, education level, work experience in years, etc. As discussed earlier, the *t*-closeness approach is not appropriate in this situation since the underlying *k*-anonymity algorithm requires conversion of numeric values into categories. Furthermore, this difference in data type affects the identity-disclosure risk. When numeric QI attributes are generalized into categories and released, a privacy intruder can correctly identify which group a target individual belongs to if the intruder knew the actual QI values of the individual. However, this is not true with microaggregation when the numeric QI attributes are replaced by group averages, because a data point is not necessarily assigned to the group whose center (group average) is the closest to the point; this is because of the group size constraint imposed in microaggregation. Thus, due to the information loss and identity-disclosure risk concerns, an effective approach for the problem should be based on a microaggregation approach that maintains the QI attributes in their numeric form. Existing *t*-closeness methods are not satisfactory in this respect.

A criticism leveled against microaggregation approaches is that, due to their nonparametric nature, there is a lack of analytical justification for the statistical properties of the masked data (Winkler 2007). Most noticeably, the use of group averages to replace individual values

in microaggregation results in a reduction in variance and distortion in relationships between attributes in the masked data. The only study we found that formally addresses this issue is the one by Domingo-Ferrer and Gonzalez-Nicolas (2010). They propose a method, called *R-microhybrid*, that replaces the original data with synthetic data generated based on the mean vector and covariance matrix of the data in each group. *R*-microhybrid preserves exactly the mean vector and covariance matrix in each group. Consequently, the mean vector and covariance matrix of the entire dataset are also preserved. A method proposed by Burridge (2003) is used for synthetic data generation at the group level; other data generation methods that preserve the mean and covariance matrix can also be used. *R*-microhybrid implicitly assumes that all QI attributes are *continuous*. When the QI attributes include binary type (e.g., gender) or discrete numeric type with limited values (e.g., education level), the within-group covariance matrix can easily become singular due to identical values of these attributes within the group. In this case, *R*-microhybrid cannot compute the covariance matrix required to generate the synthetic data. Even if all the QI attributes are truly continuous, *R*-microhybrid will still require the number of records in a group to be larger than the number of the QI attributes, in order to compute the covariance matrix.

It should be noted that *R*-microhybrid does not consider the confidential value disclosure problem examined in this research. Another issue with *R*-microhybrid pertains to its idea of preserving exactly the mean vector and covariance matrix at the group level. Given that the mean vector and covariance matrix for the entire dataset are generally considered to be public knowledge, it is certainly desirable to preserve such overall statistics in the masked dataset. However, preserving these statistics *exactly in each group* can increase disclosure risk. Burridge (2003) points out that his data obfuscation method will not work well when the attributes are highly linearly correlated, because in this situation the synthetic data points generated will be very close to the original points. This scenario is more likely to occur for the group-level data since non-linear and non-monotonic relationships are removed or significantly reduced at the group level.

In this study, we examine ways to address the confidential value disclosure and statistical property preservation problems discussed above. The bi-objective clustering approach we develop is clearly different from traditional clustering approaches that only try to cluster data into groups of similar data points without considering the confidential value distribution. To preserve the statistical properties of the data, we derive the amount of bias in variance-covariance statistics caused by traditional microaggregation methods and use the derived statistics for micro-perturbation. Unlike *R*-microhybrid, our method does not require any condition on group size (with respect to dimensionality).

3. Confidential Class Disclosure Risk in Microaggregation

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

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

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. 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:

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

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

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

is a constant (where $\bar{\mathbf{X}}$ is the overall mean vector). It is then more convenient for microaggregation 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). Note that *SSE* decreases as the number of records in each group decreases. So, minimizing *SSE* effectively forces the group size to be as close to *m* as possible. In microaggregation (as well as in *k*-anonymity), the re-identification risk is measured by the group size. We will also use the group size for the same purpose.

To deal with the confidential class disclosure problem described in Section 1, we propose an approach that 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 the information loss as measured by *SSE/SST*, subject to the group size constraint. We call this a "class restricted" clustering approach. Figure 2 shows how the same data points in Figure 1 will be clustered by our proposed approach. With this grouping, the class distribution for each group is identical to the overall class distribution (i.e., 1/3 'positive' and 2/3 'negative').

Based on the *t*-closeness principle, the value-disclosure risk of a record in a group should be viewed as low when the class distribution of the group is close to the overall class distribution, while the risk is high when the class distribution of the group is further away from the overall class distribution. To measure the value-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; $F_k > 0$, $\forall k$), 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:

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

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 value-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 not easy to normalize the *KLD* values because the maximum *KLD* value for a given dataset depends on how the data are grouped. Since our problem involves a tradeoff between a traditional distance measure (e.g., Euclidean distance) used for clustering data and the group divergence measure, it is desirable that this divergence measure is normalized and also meets the basic properties 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:

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

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

Because *JSD* is a convex 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 its square root satisfies the triangle inequality condition. *JSD* values range between zero and one. Therefore, it can be regarded as a normalized distance measure. Below, we calculate the *JSD* value for each group in Figures 1 and 2 as follows:

Let subscripts 1 and 2 represent the circle and square classes, respectively. Then, the overall class distribution is $F_1 = 3/9 = 0.333$, and $F_2 = 6/9 = 0.667$. For the all-circle group in Figure 1, $f_{\text{all-circle, 1}} = 3/3 = 1$, and $f_{\text{all-circle, 2}} = 0$. So,

$$M_{\text{all}-\text{circle},1}=(1+0.333)/2=0.667, \text{and } M_{\text{all}-\text{circle},2}=(0+0.667)/2=0.333.$$
$$KLD_{\text{all}-\text{circle}}(f,M)=(1)\log(1/0.667)+(0)\log(0/0.333)=0.585.$$
$$KLD_{\text{all}-\text{circle}}(F,M)=(0.333)\log(0.333/0.667)+(0.667)\log(0.667/0.333)=0.333.$$

Therefore,

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

Similarly, for each of the two all-square groups in Figure 1,

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

Each group in Figure 2 has one circle and two squares. Therefore,

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

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

The mixed group in Figure 2 has the lowest value-disclosure risk because its class distribution is the same as the overall class distribution. The all-circle (positive) group in Figure 1 has the highest risk because its class distribution differs the most from the overall distribution. The *JSD* measure is used along with the Euclidean distance in our proposed class restricted clustering method.

4. 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 as 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 a result, the *k*-means clustering approach is not appropriate for microaggregation. A hierarchical clustering approach, however, can be adopted for microaggregation. A representative of such an approach is the one proposed by Laszlo and Mukherjee (2005), which is based on partitioning a minimum spanning tree (MST). 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 (record) 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) for the nine vertices.

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 contain 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 in masked data as measured by the Euclidean distance, which is essentially the same as for traditional microaggregation. In this study, we use normalized Euclidean distances, which scales the numeric QI attribute values to the range [0,1]. For categorical QI attributes, the difference between two attribute values is defined as zero if they are the same, and one otherwise, which is a standard practice in clustering (Domingo-Ferrer and Torra 2005). The distance is then normalized by dividing it by the number of attributes. The second objective is to minimize the group 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 eventually 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-distributed 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 3, 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 sum of the Euclidean distance and the *JSD* distance.

To compute the *JSD* distance, we need to specify a group of related data points. 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). If there are multiple vertices that have the same "degree of separation" from u, the Euclidean distance is used to break such a tie. In Figure 3, the 2-nearest neighbors of $e(2, 5)$ are the two end-points of the edge, vertices 2 and 5. The 3-nearest neighbors of $e(2, 5)$ can be either $\{2, 5, 1\}$ or $\{2, 5, 3\}$. The latter is selected since $e(2, 3)$ is shorter than $e(2, 1)$ in terms of the Euclidean distance. The 4-nearest neighbors of $e(2, 5)$ are vertices $\{2, 5, 1, 3\}$. 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 is partitioned to form the subtrees (clusters).

For each candidate edge, our algorithm identifies its nearest neighbors 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 corresponding to each candidate edge: the Euclidean distance and the *JSD* distance. Both distances are normalized to have values in the 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 normalized Euclidean distance of e. Let $B(e)$ represent the group of nearest neighbors of e in the partial MST (with a prespecified group size b). The *composite distance* (*CD*) of e is defined as

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

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 distributed, 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. We discuss the effect of the parameter α in Section 6.

The *CD* measure can be used instead of the Euclidean distance to build a class-restricted MST. The measure, however, 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 the group p is defined as

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

For example, in Figure 1, assuming the parent group is the entire dataset, 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 in Figure 2.

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.,

$$WJSD_p(f,F) \geq JSD_p(f,F), \forall p. \quad (8)$$

The proofs of this lemma and all other lemmas and theorems are provided in the Appendix. With the 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:

$$r_e=\frac{WJSD_{p_e}(\cdot)-JSD_{p_e}(\cdot)}{L_e}. \quad (9)$$

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 L_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 4. 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 practically still of order $O(N^2)$, since b is typically several orders of magnitude smaller than N. The edge-cutting operation in Step 2 involves identifying 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.

We should point out that there is a key difference between our MST-based approach and that of the *t*-closeness-based approaches in terms of the problem and model formulation. In our approach, both the composite distance (6) and the divergence/length ratio (9) are measures of *tradeoff* between value-disclosure risk and information loss, which are set as objectives to be optimized in the two phases of the clustering process respectively. In approaches based on *t*-closeness, the objective is only to minimize information loss; value-disclosure risk is considered as a constraint by specifying the *t* parameter, in addition to the minimum group size constraint. These approaches rely only on increasing the group size to satisfy both constraints. As a result, *t*-closeness-based approaches often cause large group sizes that are undesirable (Li et al. 2007). In our MST-based approach, disclosure control is achieved not only by increasing group size, but also by forming groups with a good mix of the class values. Therefore, it should be able to attain a better tradeoff between value-disclosure risk and information loss.

5. Micro-Perturbation Approach

In most clustering-based data masking methods, identity-disclosure control is achieved by replacing the values of the QI attributes 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). This mean-substitution method, however, results in a reduction in variance and distortion in relationships between attributes in the masked data. To overcome this problem, we propose a group-level micro-perturbation method that preserves the mean vector and covariance matrix of the data masked. The covariance matrix, which includes variance and covariance components, is an important measure of variation in each attribute and of the 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 in general larger than that of the original data. The perturbation methods are usually applied to the entire dataset instead of to a partitioned set (an exception is the method proposed by Domingo-Ferrer and Gonzalez-Nicolas (2010) that we discuss next). Since the mean-substitution method causes reduction in variance and distortion in covariance, we introduce a novel approach that adds noise after aggregation to each partitioned group to offset the reduction in variance and the distortion in covariance. 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 represents the distortion in

variance-covariance statistics caused by the group-average substitution. We provide in Theorems 1 and 2 below some important analytical results with regard to the parameters of the noise to be added at a group level.

Theorem 1

Let **X** be the original data to be perturbed and **Y** be the perturbed version of **X**. If perturbed data is generated for each group g using a multivariate distribution with the group mean vector $\mathbf{x}^g$, then the sample mean vector on the entire perturbed data, $\overline{\mathbf{Y}}$, is an unbiased estimator of the original mean vector $\boldsymbol{\mu}$; i.e.,

$$E(\overline{\boldsymbol{Y}})=\boldsymbol{\mu}. \quad (10)$$

Next, we discuss the sample covariance parameters for the perturbed data. Domingo-Ferrer and Gonzalez-Nicolas (2010) propose a data masking method that preserves the mean vector and covariance matrix in each group. We have earlier discussed a key limitation of their method; that is, when the attributes are of binary type or discrete numeric type with limited values, the within-group covariance matrix can easily become singular due to identical values of these attributes within a group. There exist a few estimation methods to deal with the singular covariance matrix problem (e.g., Ledoit and Wolf 2003), but none of them are unbiased for the original covariance matrix. We propose an unbiased estimator based on the unique nature of the clustering-based approach. This unbiased estimator can be used for binary and numeric attributes (continuous or discrete), regardless of what condition exists between the group size and the number of attributes.

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 $s(\overline{x}_j^g, \overline{x}_h^g)$ be the (*j*,*h*) element of the sample covariance matrix when the original data values in group *g* (*g* = 1,…,*G*) are replaced by the respective group average $\mathbf{x}^g$. In other words, $s(\overline{x}_j^g, \overline{x}_h^g)$ is an element of the covariance matrix calculated from the *entire dataset* that has been masked by a traditional microaggregation method (note that it is not the "within-group" covariance, which of course becomes zero in this situation). Theorem 2 below provides group-level covariance parameters for micro-perturbation that lead to perturbed data whose sample covariance matrix is an unbiased estimator of the original covariance matrix.

Theorem 2

If perturbed data is generated for each group independently using a multivariate distribution with mean vector $\mathbf{x}^g$ and covariance matrix $\mathbf{S}^\Delta$ whose (*j*,*h*) element is

$$s_{jh}^{\Delta}=\frac{N-1}{N-G}\left[s(X_j,X_h)-s(\overline{x}_j^g,\overline{x}_h^g)\right], \quad (11)$$

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

$$E(\mathbf{S}_y)=\sum. \quad (12)$$

Corollary 1

If perturbed data is generated for each group independently using a distribution with the parameters specified in Theorem 2, then the sample variances on the perturbed data are unbiased estimates of the original variances.

The corollary 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 $\mathbf{x}^g$ and covariance matrix $\mathbf{S}^\Delta$. The complete algorithm, which is called CAMP-CREST (for Clustering And Micro-Perturbation with Class REstricted Spanning Tree), is given in Figure 5. Note that the confidential class attribute is not subject to masking. To mask J quasi-identifier attributes, Step II of the algorithm takes $O(NJ)$ time, while Step I takes $O(N^2)$ time as explained earlier. Since J is much smaller than N, the time complexity for the entire CAMP-CREST algorithm is of order $O(N^2)$.

The use of the multivariate normal distribution in generating perturbed data may 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. In micro-perturbation (as well as in microaggregation), the number of clusters is typically much larger than the number of data points in a cluster. Therefore, the distribution of the entire 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 represented by its center, then the joint distribution of the clusters remains the same before and after micro-perturbation, because perturbation is performed within each cluster and, by Theorem 1, the center of each cluster remains statistically unchanged. The choice of a distribution for micro-perturbation affects data distribution only within each cluster, and has little impact on the distribution across clusters (i.e., the “macro” structure of the distribution). 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.

Micro-perturbation uses the same covariance matrix $\mathbf{S}^\Delta$ to perturb the data for all groups.[2] One may argue that the covariance matrix used for each group should be as close to the original covariance matrix for the group as possible. However, as we have explained earlier, released data generated this way may be very close to the original data. Our method avoids this problem because $\mathbf{S}^\Delta$ is derived independent of “local” covariance information. Using the same covariance matrix for all groups indeed distorts the covariance structure within each individual group. But this is necessary in order to enhance protection against re-identification. The objective of a data masking approach with respect to data quality is to preserve the statistical properties of the entire dataset, rather than those of individual groups. In our approach, this objective is achieved by capturing the “macro” structure of the distribution with a clustering mechanism and by using perturbation parameters that are designed to preserve data quality for the entire dataset. The task at the group level focuses instead on reducing re-identification risk.

6. Experiments

We conducted a set of experiments on three real-world datasets to evaluate the proposed approach. We describe the data and performance measures first, and then present and analyze the experimental results.

[2] $\mathbf{S}^\Delta$ is the same for a given group size m, but it varies with different m values.

6.1 Experimental Setup

We first describe the datasets used in our experiments. 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, called Offer, 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 mainly of binary or discrete numeric type and thus can be easily handled by the algorithms used in the experiment. The confidential class attribute is salary offered; its numeric values were grouped into two classes with approximately balanced class distribution.

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, 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 by the data provider, making the data appropriate for numeric operations.

The problem we study is new to the literature, and there are no existing approaches that can be compared directly. The problem addressed by the *t*-closeness approaches appears to be the closest to the confidential class disclosure problem we study. The *t*-closeness principle (Li et al. 2007) is not tied to a specific *k*-anonymity or microaggregation algorithm. We implemented, as our *benchmark*, a *t*-closeness approach on a microaggregation algorithm. This algorithm first constructs an MST from the full dataset. It then iteratively cuts the longest removable edge in the MST to form subgroups, where an edge is said to be removable if the two subgroups formed by cutting this edge satisfy both the minimum group size and *t*-closeness conditions. The benchmark algorithm is different from that of Laszlo and Mukherjee (2005) in that it considers not only the *k*-anonymity/microaggregation requirements but also the *t*-closeness principle in forming clusters. Thus, the benchmark combines stat e-of-the-art procedures for preventing re-identification and value disclosure, and is appropriate for comparison with the proposed method. The class attribute is not used in constructing the MST in the benchmark. The mean-substitution method is used to mask the data.

Because Step 1 of the CREST algorithm (Figure 4) 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 good 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 class attributes are not masked.

To assess re-identification risk, we use a measure, called *record linkage*, proposed by Pagliuca and Seri (1999). This measure 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 unmasked record. The record linkage measure is defined as the percentage of records that are either "linked" or "second closely linked". In our experiment, the record linkage value is calculated on the QI attributes. So, a smaller value for this measure indicates lesser re-identification risk. We use this record linkage measure since it has been used in many microaggregation and related studies (e.g., Domingo-Ferrer and Torra 2001, and Li and Sarkar 2006, among others).

To assess whether the class distribution in each group is well distributed, which relates to the confidential value disclosure risk, we use a measure based on the classical chi-square statistic, defined as:

$$X^2 \frac{1}{G} \sum_{g=1}^{G} \sum_{k=1}^{C} (n_{gk} - \frac{n_g}{N} N_k)^2 / (\frac{n_g}{N} N_k), \quad (13)$$

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 (in terms of class count) and the ideal class distribution for the group, averaged over all groups. Clearly, the smaller the X^2 value, the smaller the value-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 methods being compared.

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):

$$ABIM = \frac{1}{J} \sum_{j=1}^{J} \left| \frac{\overline{Y}_j - \overline{X}_j}{\overline{X}_j} \right|, \quad (14)$$

$$ABISD = \frac{1}{J} \sum_{j=1}^{J} \left| \frac{s(Y_j) - s(X_j)}{s(X_j)} \right|, \quad (15)$$

$$ABICO = \frac{1}{J(J-1)/2} \sum_{j=1}^{J} \sum_{h=j+1}^{J} \left| \frac{r(Y_j, Y_h) - r(X_j, X_h)}{r(X_j, X_h)} \right|, \quad (16)$$

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 (the seeds are needed in the micro-perturbation part of the CAMP-CREST algorithm). 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 types of disclosure risk and the information loss measures (i.e., the two disclosure risk measures and three information loss measures mentioned above). Without an appropriate setup, it is possible that one method outperforms another method on some measures while underperforming on the other measures, leading to difficulty in making unambiguous comparisons. To facilitate the comparisons across multiple criteria, we used the record linkage measure as the primary control factor in the experiments. In general, the record linkage value increases as the group size decreases. Therefore, 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. We started with a very small group size, calculated the record linkage value on the masked data, and then increased the group size to get the record linkage at a target level (the reported experiments consider the record linkage rate at two levels, approximately 5% and 1%). A smaller group size generally results in smaller values for the *ABIM*, *ABISD* and *ABICO* measures. For the benchmark, given a specified minimum group size, a smaller t value typically results in larger group sizes and thus a smaller record linkage value. Therefore, we also adjusted the t value (along with the group size) to get the record linkage at the target levels. For the same record linkage level, the X^2 values with the benchmark are typically larger than those with the proposed methods. In this way, we have ensured that the proposed methods performed no worse than the benchmark in terms of both disclosure risk measures. It is then easy to evaluate the performances of the three methods on the information loss measures.

6.2 Performance Evaluation

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. For example, the number of linked records with the benchmark for Offer is 20 when the record linkage rate is 4.97%. It is clear that the proposed methods can achieve much better data quality with smaller disclosure risks than the benchmark. Note that the X^2 values with Partial CAMP-CREST are larger than those with CAMP-CREST, other things approximately equal. As mentioned earlier, Partial CAMP-CREST replaces Step 1 (growing the MST using the composite measure) of the CREST algorithm with a standard MST procedure; i.e., it sets $\alpha = 1$ in Equation (6). The results for the complete CAMP-CREST were obtained using the default parameter value $\alpha = 0.5$. Therefore, the impact of this parameter can be observed by comparing the results of the two methods.

The X^2 values associated with CAMP-CREST are substantially smaller than those with the benchmark, which indicates lower confidential value disclosure risk. The differences between the benchmark and CAMP-CREST can be explained by the difference in the problem and model formulation. As mentioned earlier, for the benchmark, value-disclosure is limited only by increasing the group size to satisfy both the group size and t-closeness constraints. In CAMP-CREST, value-disclosure is limited not only by increasing the group size, but also by forming groups with a good mix of the classes in the MST growing stage. As a result, CAMP-CREST has attained lower value-disclosure risks with about the same or smaller re-identification risks.

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

perturbed data. Such deviations 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 the mean-substitution-based benchmark method. This indicates that the joint distributions of the dataset are much better preserved by the proposed methods than by the mean-substitution method. Note that there are only small differences between CAMP-CREST and Partial CAMP-CREST in all three information loss measures. This is because they both implement the same micro-perturbation approach for all the QI attributes, on which the three measures are computed.

When the record linkage rate is reduced from about 5% to 1%, the *ABISD* and *ABICO* values associated with the benchmark increase considerably. This pattern does not occur for CAMP-CREST and Partial CAMP-CREST; instead, the values of all three information loss measures remain relatively stable as the record linkage rate decreases. Thus, the proposed methods can attain lower re-identification risk without much loss in data quality, which is an important advantage.

In terms of computing time, the benchmark and Partial CAMP-CREST are about the same. CAMP-CREST runs slower than the other two algorithms although all three algorithms run in $O(N^2)$ time. As discussed earlier, this is due to the search for the nearest neighbors in Step 1 of the CREST algorithm (Figure 4). 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.

7. Managerial and Policy Implications

The proposed approach has significant management and policy implications. To further demonstrate the implications on re-identification and value-disclosure risks, we used a traditional microaggregation method to mask the Diabetes and Offer datasets with a minimum group size of 5. In the masked Diabetes dataset, 83 out of 768 records (10.8%) were correctly linked to the original records, and perhaps more strikingly, 163 records (21.2%) were assigned to groups that had a homogeneous value in the class attribute (Test Result). For the Offer data, 71 out of 443 records (16.0%) were correctly linked and 88 records (19.9%) were assigned to groups having a homogeneous value in the class attribute (Salary Offered). With the proposed technique, we could easily generate masked data for both datasets such that none of the groups had homogeneous class values and there were fewer than 1% linked records. This was achieved with the same or lesser information loss.

The unbiasedness property of our micro-perturbation approach has an important implication in the tradeoff between re-identification risk and data quality, as explained below using the Diabetes data. We generated masked data with record linkage rate that decreases from 10% to 1% with a 1% decrement each time, by gradually increasing the group size. We then recorded the corresponding *ABICO* values using the benchmark and micro-perturbation, respectively. The results are shown in Figure 5. It can be clearly observed that while the *ABICO* value with the benchmark increases considerably as record linkage decreases, the *ABICO* value with micro-perturbation only increases a little. So, with the proposed methods the user can afford to mask the data to a great extent to keep re-identification risk low without sacrificing much on data quality. Thus, the proposed method has a significant advantage over the benchmark in terms of tradeoff between re-identification risk and data quality.

In terms of practical implementation, the proposed technique is easy to use even by non-technical users because the only essential input parameter is the minimum group size. In microaggregation and *k*-anonymity, it is common to use a group size of between 3 and 10

records (Domingo-Ferrer and Mateo-Sanz 2002, Laszlo and Mukherjee 2005). In general, the smaller the group size, the closer the masked record to the original record, and the higher the re-identification risk. A user can start with a small group size (e.g., three records), calculate the record linkage value on the masked data, and then increase the group size to get the re-identification risk at a planned level. The α parameter is a weighting measure that is easy to understand (Partial CAMP-CREST algorithm does not need the parameter at all). The *t* parameter in *t*-closeness, however, is not intuitive; its upper bound is not obvious, and it often results in group sizes that are several multiples of the specified minimum size. Therefore, data privacy policies and procedures can be easily set by managers and policy makers with the proposed technique. In addition, we also provide alternative choices for the user to either perturb the data or produce *k*-anonymized data, and also to speedup runtime for large datasets.

As data-sharing is being increasingly considered in practice, there is a rising public sentiment that individual privacy is being severely eroded. The proposed approach will reduce the risks of re-identification of individuals from de-identified data, while also protecting the values of their confidential attributes. This should alleviate individuals' concerns about loss of privacy and confidentiality and increase their willingness to participate in research that uses personal data. The proposed approach will also reduce organizations' concerns about potential privacy violations (e.g., the Netflix case) and enable organizations to safely share and publish high-quality data for legitimate research and analysis. Eventually, this will help promote adoption of privacy-enhancing technology in privacy-sensitive programs such as electronic medical records.

When the distributions of the confidential class values are highly unbalanced, it is possible that for a certain class there is only one record in a group. In this case, if an intruder has insider information and is able to identify the cluster where the record belongs, the individuals in such a group maybe subject to a high value-disclosure risk. Managers and policy makers should be aware of this risk – if necessary, they should use a relatively large group size to ensure a safe class count in every group when the class distribution is highly unbalanced.

We should point out that the proposed methodology is designed primarily for applications with numeric QI attributes. Similar to microaggregation, our methodology requires that categorical QI attributes be coded in 0–1 format. Therefore, if the QI attributes are largely categorical, it may be more effective to use a *k*-anonymity approach coupled with the *t*-closeness method.

8. Conclusions and Extensions

Microaggregation and *k*-anonymity have been widely used in practice for data sharing and publishing. *K*-anonymity based approaches such as *t*-closeness are not suitable for masking numeric attributes. Existing microaggregation approaches have overlooked the confidential class disclosure problem and lacked justification in preserving statistical properties of the data. Our proposed method 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 regardless of the actual distribution of the data. Our empirical study has demonstrated that the method can lead to significantly improved performance over a viable alternative approach.

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 QI 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 QI 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 QI 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 (e.g., Domingo-Ferrer and Mateo-Sanz 2002, Li and Sarkar 2006a). We plan to examine these alternative approaches in future.

Acknowledgments

The authors are grateful to the department editor, associate editor, and three anonymous reviewers for their insightful comments and suggestions that have improved the paper considerably. Xiao-Bai Li's research was supported in part by Grant Number R01LM010942 from the National Library of Medicine (NLM) and the National Institutes of Health (NIH). The content is solely the responsibility of the authors and does not necessarily represent the official views of NLM or NIH.

References

Adam NR, Wortmann JC. Security-control methods for statistical databases: A comparative study. ACM Computing Surveys. 1989; 21(4):515–556.

Aggarwal, CC.; Yu, PS., editors. Privacy-Preserving Data Mining: Models and Algorithms. Springer; New York: 2008.

Agrawal, R.; Srikant, R. Proceedings of 2000 ACM SIGMOD International Conference on Management of Data. ACM Press; New York: 2000. Privacy-preserving data mining; p. 439-450.

Asuncion, A.; Newman, DJ. UCI Machine Learning Repository. 2007. Retrieved August 2008, from http://www.ics.uci.edu

Burridge J. Information preserving statistical obfuscation. Statistics and Computing. 2003; 13:321–327.

Chowdhury DS, Duncan GT, Krishnan R, Roehrig SF, Mukherjee S. Disclosure detection in multivariate categorical databases: Auditing con dentiality protection through two new matrix operators. Management Science. 1999; 45(12):1710–1723.

Cox LH. Suppression methodology and statistical disclosure control. Journal of the American Statistical Association. 1980; 75(370):377–385.

Deb P, Trivedi PK. Demand for medical care by the elderly: A finite mixture approach. Journal of Applied Econometrics. 1997; 12(3):313–336.

Defays D. Protecting micro-data by micro-aggregation: The experience in Eurostat. Questiio. 1997; 21:221–231.

Dalenius T, Reiss SP. Data swapping: A technique for disclosure control. Journal of Statistical Planning and Inference. 1982; 6(1):73–85.

Department of Health and Human Services (DHHS). Standards for privacy of individually identifiable health information. Federal Register. 2000 Dec 28.65(250)

Dixon, P. Medical identity theft: The information crime that can kill you. The World Privacy Forum. 2006. Retrieved April 2009, from http://www.worldprivacyforum.org/pdf/wpf_medicalidtheft2006.pdf

Domingo-Ferrer J, Mateo-Sanz JM. Practical data-oriented microaggregation for statistical disclosure control. IEEE Transactions on Knowledge and Data Engineering. 2002; 14(1):189–201.

Domingo-Ferrer, J.; Torra, V. A quantitative comparison of disclosure control methods for microdata. In: Doyle, P.; Lane, J.; Theeuwes, J.; Zayatz, L., editors. Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies. North-Holland; Amsterdam, Netherlands: 2001. p. 111-134.

Domingo-Ferrer J, Torra V. Ordinal, continuous and heterogeneous *k*-anonymity through microaggregation. Data Mining and Knowledge Discovery. 2005; 11(2):195–212.

Domingo-Ferrer J, Sebé F, Solanas A. A polynomial-time approximation to optimal multivariate microaggregation. Computers and Mathematics with Applications. 2008; 55(4):714–732.

Domingo-Ferrer J, Gonzalez-Nicolas U. Hybrid microdata using microaggregation. Information Sciences. 2010; 180(15):2834–2844.

Duncan GT, Lambert D. The risk of disclosure for microdata. Journal of Business and Economic Statistics. 1989; 7(2):201–217.

Duncan GT, Mukherjee S. Optimal disclosure limitation strategy in statistical databases: Deterring tracker attacks through additive noise. Journal of American Statistical Association. 2000; 95(451): 720–729.

Fischetti M, Salazar JJ. Solving the cell suppression problem on tabular data with linear constraints. Management Science. 2001; 47(7):1008–1027.

Galletta, D. MIS faculty salary survey results. 2004. Retrieved March 2004, from http://www.pitt.edu/~galletta/salsurv.html

Garfinkel R, Gopal R, Goes P. Privacy protection of binary con dential data against deterministic, stochastic, and insider threat. Management Science. 2002; 48(6):749–764.

Garfinkel R, Gopal R, Thompson S. Releasing individually identifiable microdata with privacy protection against stochastic threat: An application to health information. Information Systems Research. 2007; 18(1):23–41.

Ghinita, G.; Karras, P.; Kalnis, P.; Mamoulis, N. Proceedings of the 33rd International Conference on Very Large Data Bases. ACM Press; New York: 2007. Fast data anonymization with low information loss; p. 758-769.

Gopal R, Garfinkel R, Goes P. Confidentiality via camouflage: The CVC approach to disclosure limitation when answering queries to databases. Operations Research. 2002; 50(3):501–516.

Kadane JB, Krishnan R, Shmueli G. A data disclosure policy for count data based on the COM-Poisson distribution. Management Science. 2006; 52(10):1610–1617.

Kaelber DC, Jha AK, Johnston D, Middleton B, Bates DW. A research agenda for personal health records (PHRs). Journal of American Medical Informatics Association. 2008; 15(6):729–736.

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

Laszlo M, Mukherjee S. Minimum spanning tree partitioning algorithm for microaggregation. IEEE Transactions on Knowledge and Data Engineering. 2005; 17(7):902–911.

Ledoit O, Wolf M. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. Journal of Empirical Finance. 2003; 10(5):603–621.

Lee S, Genton MG, Arellano-Valle RB. Perturbation of numerical confidential data via skew-*t* distributions. Management Science. 2010; 56(2):318–333.

Li, N.; Li, T.; Venkatasubramanian, S. Proceedings of the 23rd IEEE International Conference on Data Engineering. IEEE Computer Science Society; Washington, DC: 2007. *t*-Closeness: Privacy beyond *k*-anonymity and *l*-diversity; p. 106-115.

Li XB, Sarkar S. A tree-based data perturbation approach for privacy-preserving data mining. IEEE Transactions on Knowledge and Data Engineering. 2006a; 18(9):1278–1283.

Li XB, Sarkar S. Privacy protection in data mining: A perturbation approach for categorical data. Information Systems Research. 2006b; 17(3):254–270.

Liew CK, Choi UJ, Liew CJ. A data distortion by probability distribution. ACM Transactions on Database Systems. 1985; 10(3):395–411.

Lin J. Divergence measures based on the Shannon entropy. IEEE Transactions on Information Theory. 1991; 37(1):145–151.

Lohr S. Netflix cancels contest after concerns are raised about privacy. New York Times. 2010 Mar 13.:B3.

Machanavajjhala, A.; Gehrke, J.; Kifer, D.; Venkitasubramaniam, M. Proceedings of 22nd IEEE International Conference on Data Engineering. IEEE Computer Science Society; Washington, DC: 2006. *l*-diversity: Privacy beyond *k*-anonymity; p. 24-35.

Oganian, A. Multiplicative noise protocols. In: Domingo-Ferrer, J.; Magkos, E., editors. Privacy in Statistical Databases 2010, Lecture Notes in Computer Science 6344. Springer; Berlin: 2010. p. 107-117.

Pagliuca D, Seri G. Some results of individual ranking method on the system of enterprise accounts annual survey. Esprit SDC Project, Deliverable MI-3/D2. 1999

Prim RC. Shortest connection networks and some generalizations. Bell System Technical Journal. 1957; 36:1389–1401.

Rebollo-Monedero D, Forne J, Domingo-Ferrer J. From *t*-closeness-like privacy to postrandomization via information theory. IEEE Transactions on Knowledge and Data Engineering. 2010; 22(11): 1623–1636.

Samarati, P.; Sweeney, L. Proceedings of the IEEE Symposium on Research in Security and Privacy. IEEE Computer Science Society; Oakland, CA: 1998. Protecting privacy when disclosing information: *k*-anonymity and its enforcement through generalization and suppression.

Sweeney L. k-Anonymity: A model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems. 2002; 10(5):557–570.

Teltzrow, M.; Kobsa, A. Designing Personalized User Experiences in eCommerce. Kluwer Academic Publishers; Dordrecht, Netherlands: 2004. Impacts of user privacy preferences on personalized systems: A comparative study; p. 315-332.

US General Accounting Office. Data Mining: Federal Efforts Cover a Wide Range of Uses. 2004. Retrieved May 2006, from http://www.gao.gov/new.items/d04548.pdf

Winkler, WE. 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). 2007. Retrieved March 2008, from http://www.census.gov/srd/papers/pdf/rrs2007-21.pdf

Zeller T. AOL executive quits after posting of search data. International Herald Tribune. 2006 Aug 23.:13.

Appendix. Proofs of Lemmas and Theorems

Proof of Lemma 1

If $q(x)$ is a convex function of x, then by definition, for any convex combination of $q(x_1)$ and $q(x_2)$,

$$\lambda_1 q(x_1)+\lambda_2 q(x_2) \geq q(\lambda_1 x_1+\lambda_2 x_2).$$

This property can be generalized, by induction, to the s-value case; i.e., with $\sum_{g=1}^{s}\lambda_g=1$,

$$\sum_{g=1}^{s}\lambda_g q(x_g) \geq q(\sum_{g=1}^{s}\lambda_g x_g). \quad (17)$$

Let f_{gk} be the relative frequency in subgroup g (g = 1,…,s) that are of class k, and $f_{pk}=\sum_{g=1}^{s}(n_g/n)f_{gk}$ be the corresponding frequency in parent p. Denote M_{gk} and M_{pk} similarly in the context of *JSD*. Note that F is fixed for a given dataset. It follows from Definition 2 that $KLD_g(f,M)$ and $KLD_g(F,M)$ are convex functions of f. This fact is used below by replacing $q(x_g)$ in (17) with an appropriate *KLD* function. Continuing with the notation above, we have,

$$\begin{aligned}
WJSD_p(f,F) &= \sum_{g=1}^{s}\frac{n_g}{n}JSD_g(f,F)\\
&\overset{\text{by(5)}}{=} \frac{1}{2}\sum_{g=1}^{s}\frac{n_g}{n}[KLD_g(f,M)+KLD_g(F,M)]\\
&\overset{\text{by(4)}}{=} \frac{1}{2}\sum_{g=1}^{s}\frac{n_g}{n}\left[\sum_{k=1}^{C}f_{gk}\log\frac{f_{gk}}{M_{gk}}+\sum_{k=1}^{C}F_k\log\frac{F_k}{M_{gk}}\right]\\
&=\frac{1}{2}\sum_{k=1}^{C}\left[\sum_{g=1}^{s}(\frac{n_g}{n})(f_{gk}\log\frac{f_{gk}}{M_{gk}})+\sum_{g=1}^{s}(\frac{n_g}{n})(F_k\log\frac{F_k}{M_{gk}})\right]\\
&\overset{\text{by(17)}}{\geq} \frac{1}{2}\sum_{k=1}^{C}\left[\left(\sum_{g=1}^{s}(\frac{n_g}{n})(f_{gk})\right)\log\frac{\sum_{g=1}^{s}(n_g/n)f_{gk}}{\sum_{g=1}^{s}(n_g/n)M_{gk}}+F_k\log\frac{F_k}{\sum_{g=1}^{s}(n_g/n)M_{gk}}\right]\\
&=\frac{1}{2}(\sum_{k=1}^{C}f_{pk}\log\frac{f_{pk}}{M_{pk}}+\sum_{k=1}^{C}F_k\log\frac{F_k}{M_{pk}})\\
&=\frac{1}{2}[KLD_p(f,M)+KLD_p(F,M)]=JSD_p(f,F).
\end{aligned}$$

Proof of Theorem 1

Let J be the number of attributes to be perturbed, 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 an attribute to be perturbed, 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 $x_{ij}^g(i=1,\ldots,n_g)$ 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. For j = 1,…, J,

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

So,

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

Proof of Theorem 2

Consider the perturbed data, for each j, h in {1,…, J},

$$y_{ij}^g-\overline{Y}_j=(\overline{y}_j^g-\overline{Y}_j)+(y_{ij}^g-\overline{y}_j^g), \text{and}$$
$$y_{ih}^g-\overline{Y}_h=(\overline{y}_h^g-\overline{Y}_h)+(y_{ih}^g-\overline{y}_h^g).$$

Multiplying the left and right hand sides of the above two equations respectively, we have:

$$(y_{ij}^g-\overline{Y}_j)(y_{ih}^g-\overline{Y}_h)=$$
$$(\overline{y}_j^g-\overline{Y}_j)(\overline{y}_h^g-\overline{Y}_h)+(\overline{y}_j^g-\overline{Y}_j)(y_{ih}^g-\overline{y}_h^g)+(y_{ij}^g-\overline{y}_j^g)(\overline{y}_h^g+\overline{Y}_h)+(y_{ij}^g-\overline{y}_j^g)(y_{ih}^g-\overline{y}_h^g). \quad (18)$$

Summing over all the records (first within a group and then over all groups), and noting that the summations for the middle two terms in the right-hand side of (18) equal zero, we get:

$$\sum_{g=1}^{G}\sum_{i=1}^{n_g}(y_{ij}^g-\overline{Y}_j)(y_{ih}^g-\overline{Y}_h)=\sum_{g=1}^{G}\sum_{i=1}^{n_g}(\overline{y}_j^g-\overline{Y}_j)(\overline{y}_h^g-\overline{Y}_h)+\sum_{g=1}^{G}\sum_{i=1}^{n_g}(y_{ij}^g-\overline{y}_j^g)(y_{ih}^g-\overline{y}_h^g). \quad (19)$$

Dividing both sides of (19) by $N-1$, we have:

$$\frac{1}{N-1}\sum_{i=1}^{N}(y_{ij}-\overline{Y}_j)(y_{ih}-\overline{Y}_h)=\frac{1}{N-1}\sum_{g=1}^{G}n_g(\overline{y}_j^g-\overline{Y}_j)(\overline{y}_h^g-\overline{Y}_h)+\frac{1}{N-1}\sum_{g=1}^{G}\sum_{i=1}^{n_g}(y_{ij}^g-\overline{y}_j^g)(y_{ih}^g-\overline{y}_h^g),$$

which can be written as

$$s(Y_j,Y_h)=s(\overline{y}_j^g,\overline{y}_h^g)+\frac{1}{N-1}\sum_{g=1}^{G}(n_g-1)s(y_j^g,y_h^g), \quad (20)$$

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 (20), we have,

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

or

$$E[s(Y_j,Y_h)]=E[s(\overline{y}_j^g,\overline{y}_h^g)]+\frac{N-G}{N-1}E[s(y_j^g,y_h^g)], \quad (21)$$

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 (11) and substituting the result in (21) that

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

Since perturbed data is generated using $\mathbf{x}^{\mathcal{G}}$,

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

Therefore,

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

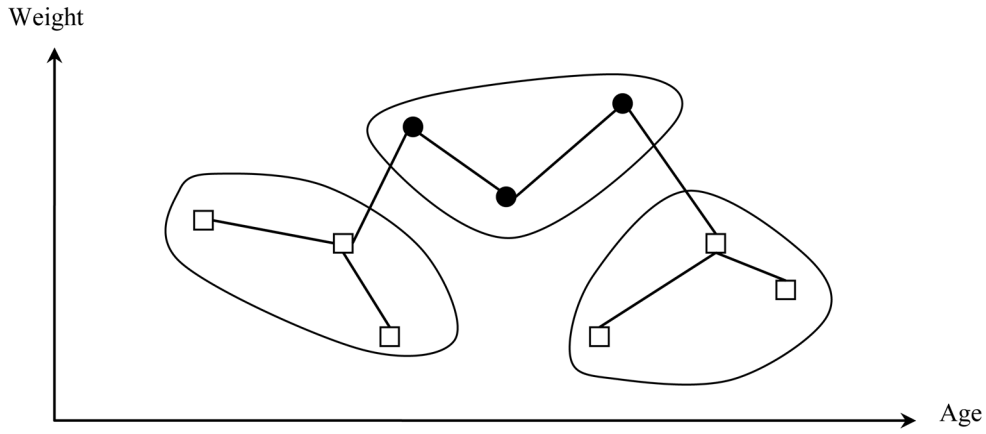

Figure 1.
An Illustrative Example for Clustering

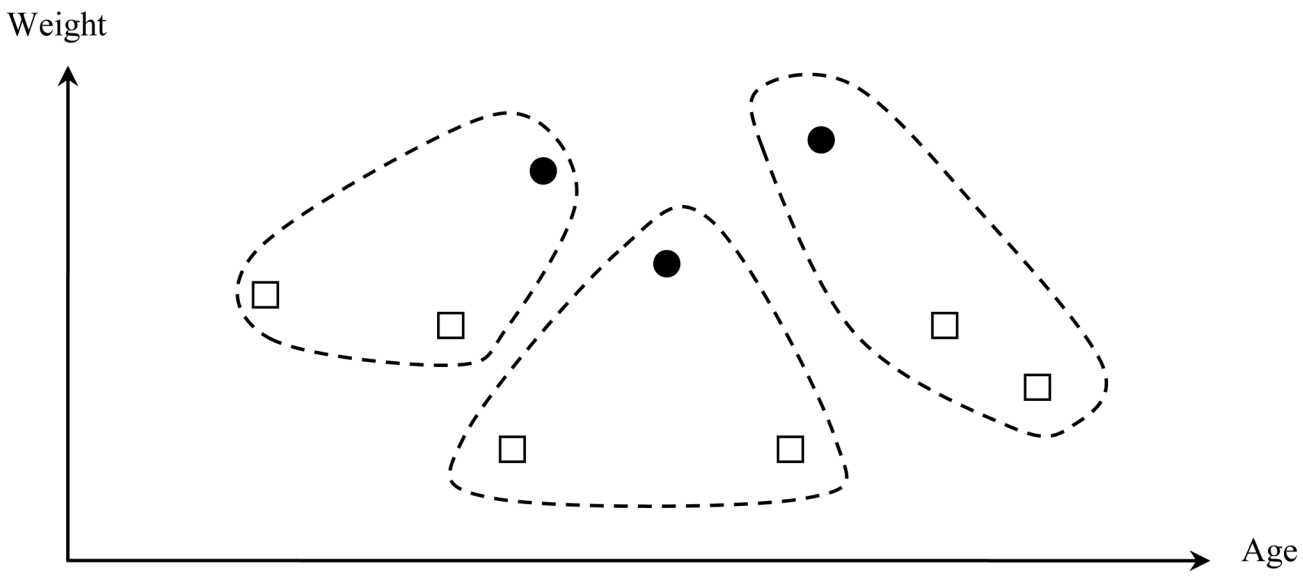

Figure 2.
Clustering of Data in Figure 1 by Proposed Approach

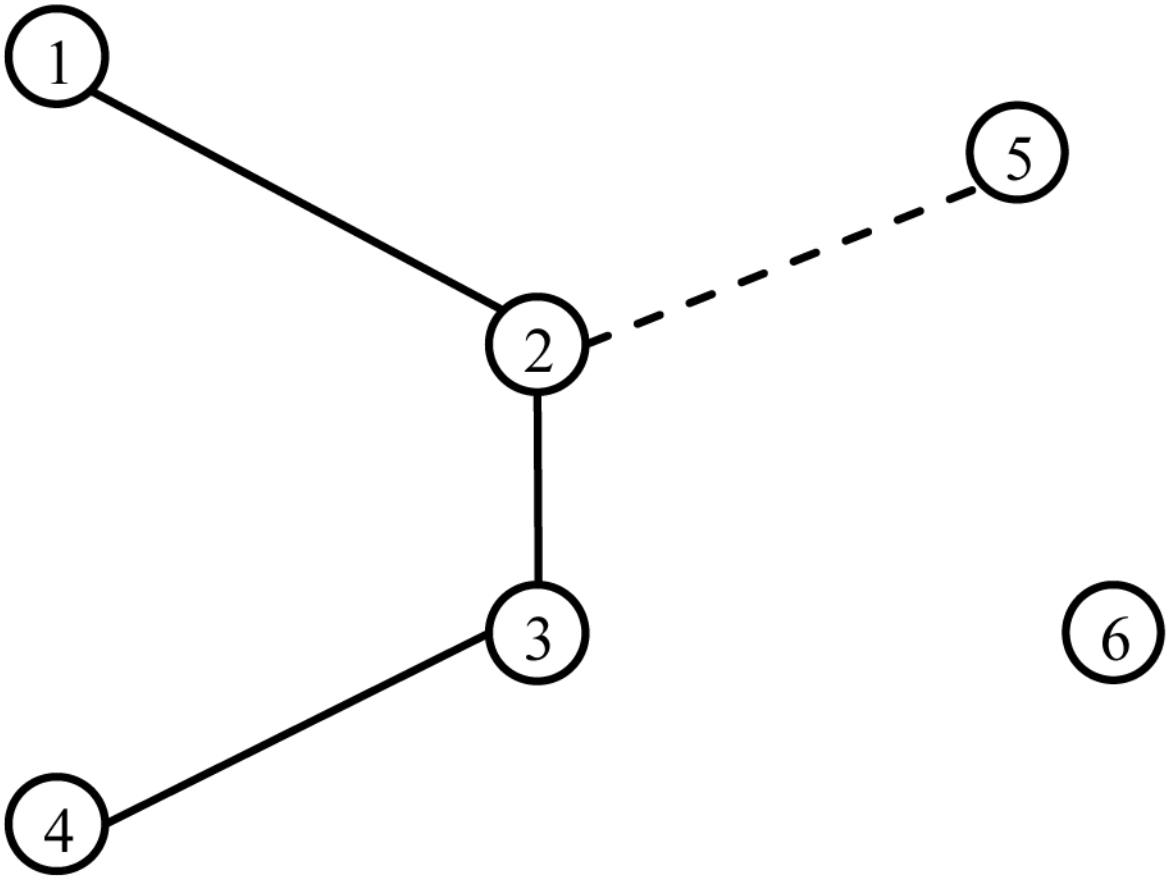

Figure 3.
An Illustrative Example for Nearest Neighbors

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 4.
CREST Algorithm

I. Run CREST algorithm described in Figure 4.

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

Figure 5.
CAMP-CREST Algorithm

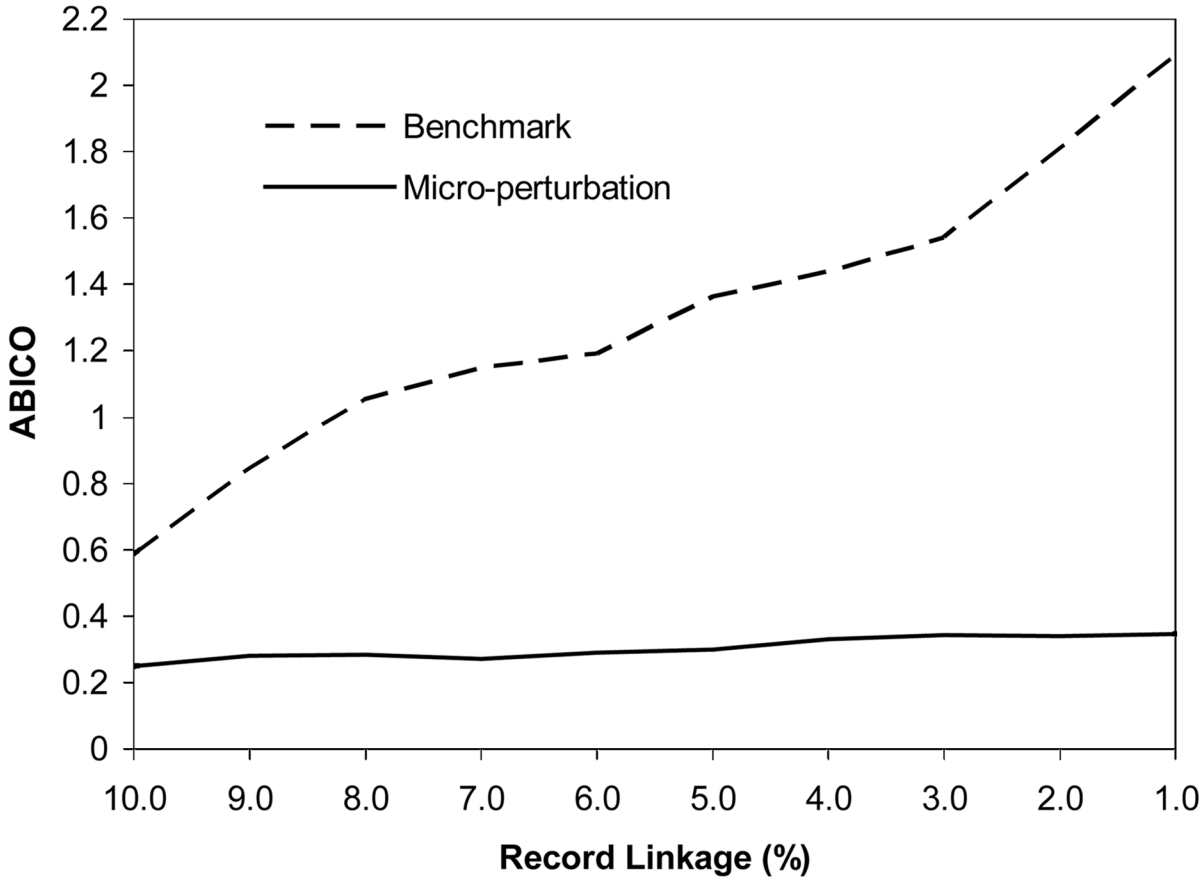

Figure 5.
Tradeoff between Re-identification Risk and Data Quality on Diabetes Data

Table 1

Results of Primary Experiments

Data	Method	Time (seconds)	Linkage (%)	X^2	*ABIM* (%)	*ABISD* (%)	*ABICO* (%)
Offer (403 records)	Benchmark	0.6	4.97	4.34	0	47.75	438.16
	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
Diabetes (768 records)	Benchmark	1.3	4.30	2.27	0	36.10	118.42
	Partial CAMP-CREST	1.3	4.24	2.04	1.54	2.94	24.98
	CAMP-CREST	4.1	4.17	1.37	1.46	3.68	28.62
Medicare (4406 records)	Benchmark	33.4	4.29	2.65	0	37.59	392.54
	Partial CAMP-CREST	34.9	4.21	2.64	2.70	2.54	127.51
	CAMP-CREST	126.2	4.09	1.89	2.11	2.96	118.51
Offer (403 records)	Benchmark	0.6	1.13	7.00	0	69.39	1343.74
	Partial CAMP-CREST	0.6	0.95	6.60	3.37	2.52	108.31
	CAMP-CREST	1.5	0.86	5.90	3.09	2.70	106.14
Diabetes (768 records)	Benchmark	1.3	1.04	3.59	0	55.23	171.87
	Partial CAMP-CREST	1.3	0.86	3.35	1.96	2.18	29.94
	CAMP-CREST	4.5	0.86	2.56	1.66	1.86	32.36
Medicare (4406 records)	Benchmark	33.1	0.98	3.39	0	50.93	692.33
	Partial CAMP-CREST	34.2	0.97	3.29	2.53	1.15	141.82
	CAMP-CREST	132.6	0.96	2.31	2.79	1.18	136.40

Manage Sci. Author manuscript; available in PMC 2013 December 02.