Forthcoming in *Management Science*

Drive More Effective Data-Based Innovations: Enhancing the Utility of Secure Databases

Databases play a central role in evidence-based innovations in business, economics, social, and health sciences. In modern business and society, there are rapidly growing demands for constructing analytically valid databases that also are secure and protect sensitive information in order to meet customer and public expectations, to minimize financial losses, and to comply with privacy regulations and laws. We propose new data perturbation and shuffling (DPS) procedures, named MORE, for this purpose. As compared with existing DPS methods, MORE can substantially increase the utility of secure databases without increasing disclosure risk. MORE is capable of preserving important nonmonotonic relationships among attributes, such as the inverted-U relationship between competition and innovation. Maintaining such relationships is often the key to determining optimal levels of policy and managerial interventions. MORE does not require data to be of particular types or have particular distributional shapes. Instead, it provides unified, flexible, and robust algorithms to mask general types of confidential variables with arbitrary distributions, thereby making it suitable for general-purpose data masking. Since MORE nests the commonly used generalized linear models as special cases, a much wider range of statistical analyses can be conducted using the secure databases with results similar to those using the original databases. Unlike existing DPS approaches which typically require a joint model for all variables, MORE requires no modeling of nonconfidential variables, and thus further increases the robustness of secure databases. Evaluation of MORE through Monte Carlo simulation studies and empirical applications demonstrates that it performs better than existing data masking methods.

Key words: Database; Digital Economy; Innovation; Nonparametric; Perturbation; Privacy; Shuffling.

"The digital age will be to the analog age what the iron age was to the stone age. "—Joel Mokyr

1. Introduction

Data Privacy Problem

Data are key to innovations in many industries and are invaluable assets for many entities, including government agencies, firms, nonprofit organizations, and academic institutions. Effective data management and analytics play a central role in gaining insights on customers, patients, and product and service providers, and in making critical policy and managerial decisions. A foundation to fulfilling the benefits of such data-based activities is the procurement, construction, analysis, sharing, and dissemination of relevant databases, which also raise data privacy concerns. Data privacy concerns arise whenever a database contains sensitive attributes that if disclosed without control to a third party, can lead to negative consequences. These privacy concerns can render these sensitive key data elements unavailable to the third party who needs them for data-based innovations. To ease the tensions between these data-based activities and privacy concerns, we propose a new methodology to provide analytically valid data that also protect privacy. Given the importance of data, our methodology can have important impacts for driving data-based innovations in many fields.

Data privacy is a very important issue in the new digital economy. For example, Blattberg et al. (2008) devote an entire chapter to customer privacy, and describe the consequences of customer privacy concerns for database marketing. Privacy has also emerged as a key concern for innovation policy (Goldfarb and Tucker 2012). Confidential data are found not only in consumer databases, but also in databases about firms or organizations (e.g., in business census or surveys). Data privacy concerns occur in almost all stages of database-related activities, from data procurement, data construction and analysis (e.g., merging data from multiple sources as noted in Mela 2011 and Qian and Xie 2013) to data dissemination that fulfills the microdata release requirements from funding agencies (e.g., NIH and NSF) or journals (Desai 2013). For respondents and the public to trust that their private information is in good hands, they expect database owners to implement adequate disclosure control.

Consistent with these expectations, privacy laws and regulations require database owners to follow certain rules to protect private information. Examples of earlier laws regarding consumer privacy are the Fair Credit Reporting Act of 1970, which established consumers' rights regarding their financial information, and the Health Insurance Portability and Accountability Act (HIPAA) of 1996 in the healthcare industry. Privacy concerns have become increasingly important with innovations in information technology and e-commerce (Kalvenes and Basu 2006). For example, public concerns about Web privacy led to the halt of data sharing between a Web ad company, DoubleClick, and a marketing database company, Abacus, in 1999 (Winer 2001). Insufficient data privacy protection can also have dire consequences. A recent example is the withdrawal of Netflix from its once very successful collaboration with academia because of a privacy breach (Mela 2011). In that case, one was able to combine the Netflix database with the Internet Movie Database and recover the identities of the customers and their rental histories. The privacy breach caught the attention of the Federal Trade Commission and prompted a class action lawsuit against Netflix. Losses in market value and consumer relationships can be large as well. As reviewed in Miller and Tucker (2011), studies found that (1) negative publicity from privacy breaches causes affected firms to lose on average 2.1 percent of their stock market values within two days of the announcement, amounting to $1.65 billion average loss in market capitalization per incident and (2) 31 percent of survey consumers say they would end their relationship with a firm as a a result of a privacy breach.

In response to the high demand for confidential data protection, various solutions for creating secure databases have been proposed. Prior studies demonstrate that an effective approach to protecting data privacy is data masking through perturbation and shuffling (Willenborg and de

Waal 2001, Muralidhar and Sarathy 2006). These data masking procedures replace the original confidential data values in a database with modified values. The resulting masked dataset is secure in the sense that the masked values, rather than the original values, of the confidential data elements are released to the third-party. In order for the secure database to be useful, these masked values should be "realistic" so that queries based on the secure database are as close as possible to those based on the original database. Formally, an ideal data masking procedure should have the following two key properties (Muralidhar and Sarathy 2003): (1) high data utility: statistical inferences using the masked confidential values in a secure database remain the same as those using the original dataset; and (2) low disclosure risk: the release of the masked values does not improve the ability of intruders to predict the original values of confidential attributes. The two requirements are often in conflict. The trend in data masking is to develop methods that have data utility as high as possible while maintaining a sufficiently low disclosure risk.

An Example: Competition and Innovation

Although much past research concerns consumer privacy, similar privacy concerns apply to firms and organizations supplying confidential data. The insights gained from data-based studies often have important implications for firms/organizations and market regulators, and are of broad interest to the economics and management community, as well as policy makers. As with consumer databases, collecting, analyzing, sharing, and disseminating data on firms or organizations also raise data privacy issues. Data masking provides an attractive approach to harnessing the power of data while minimizing the harm caused by privacy violations. When only masked datasets can be released or shared, a key question is whether their users will be able to gain the same insights as from the original databases on important questions such as "Does competition spur innovation, and if so, how?". Answers to such questions are of great interest in innovation policy. Clearly, being able to provide secure databases with maximal utility is a crucial problem in data masking and has important managerial and policy implications. As an example, we consider a dataset that provides information about the relationship between competition and innovation (C&I) using a sample of 311 firms (Aghion et al. 2005). The study measures innovation by the citation weighted patent count (*patcw*) and competition by a Competition index (*Ci*), which is 1 minus the Lerner Index. Fig 1 plots the distributions of these two attributes. The data cover seventeen two-digit SIC code industries over the period 1973-1994. Fig 2 shows an inverted-U relationship between competition and innovation: when the competition level is low, more competition increases innovation activity; but when the competition is fierce enough, it decreases innovation activity.

Privacy concerns can arise for various reasons in this context. For example, the C&I dataset contains information on patent count, research and development investment, financial cost, profits, and sales. Since the release of confidential data (e.g., patent count) can cause reidentification of firms, other sensitive information about the firms, such as financial cost, profits, sales, and R&D

investment, can be revealed to intruders (e.g., firms' competitors), which in turn can compromise the interests of respondent firms.[1]

In order for secure databases to provide accurate information while providing adequate disclosure control, effective data masking procedures should preserve distributional properties of the masked values as close as possible to those of the original sensitive values. The C&I dataset has several interesting features that allow us to demonstrate the important benefits of our proposed data masking procedures for enhancing the utility of secure databases. Due to the rich policy implications of the inverted-U relationship, this is an ideal example to demonstrate the importance of preserving nonmonotonic relationships in data masking. Moreover, the key variables in this dataset exhibit several interesting features and require data masking procedures that can account for these features properly. Fig 1 shows that the competition measure, Ci, and the innovation measure, $patcw$, are bounded and semicontinuous: Ci takes values from 0 to 1, and $patcw$ takes only nonnegative values with a large probability at the boundary value of zero. One needs to take this into account so that implausible masked values that are out of bound are not generated. There also exists skewness in these two variables, with $patcw$ being highly skewed. Our more complex analysis includes $Industry$ and $Year$, both of which are categorical discrete variables. In sum, the dataset contains various types of variables with important nonmonotonic relationships among attributes, and we consider this an ideal empirical example for comparing different data masking procedures.

These important data features are not limited to this dataset, but are frequently encountered in business applications. Other examples with complex distributional features include sensitive attributes that are continuous (home value, mortgage balance, net asset value), binomial (number of occasions using protection measures among all occasions of drug use), count data (number of times downloading pirated movies and music), and fractional and bounded (fraction of credit debts repayed, share of wallet for counterfeit products). These types of variables are very informative, and thus often exhibit complex distributional features, such as heavy tails, outliers, departure from the nominal variance in the binomial and count outcomes, boundedness, skewness, multimodality, and zero-inflation. Furthermore, the relationships among these sensitive attributes, as well as between them and other attributes, can be complex and nonmonotonic. These important data features call for general and flexible data masking procedures so that the resulting secure databases have high utility for important managerial and policy decisions.

Contribution

Prior studies have shown that data perturbation and shuffling (DPS) are an important class of data masking methods, and have a number of advantages as compared with alternative masking

[1] The firms in this dataset are public firms traded in the London Stock Exchange. Therefore, most of this sensitive information is publicly available. However, this is not the case for private businesses or organizations. Our use of this dataset is therefore mainly for illustration purposes.

methods (e.g., see reviews and comparisons provided in Muralidhar et al. 1995, and Muralidhar and Sarathy 2006). However, as discussed in Section 3, there are a range of important limitations of existing DPS methods, some of which are identified in this work. These issues limit the scopes of distributional properties, as well as the types of attributes and the relationships among attributes for which the existing DPS methods are applicable, which can affect the ability of users of secure databases to make optimal managerial and policy decisions.

Our objective in this paper is to develop a new set of nonparametric DPS procedures that *simultaneously* address all these limitations of existing procedures while retaining the nice properties of DPS methods. Our approach is a conditional distribution approach and thus satisfies the low disclosure risk requirement for an ideal data masking procedure (Muralidhar and Sarathy 2003). It provides secure databases with substantially higher data utility that have important managerial and policy implications, and therefore represents a significant advance in the area. More specifically, we develop new nonparametric DPS procedures with the following capabilities:

(1) they provide unified data masking algorithms that are not restricted by the types and distributions of confidential variables;

(2) they maintain marginal distributions of confidential variables with arbitrary distributional shapes;

(3) they maintain important and complex relationships among variables, including nonmonotonic relationships (such as the inverted-U relationship between competition and innovation);

(4) they are applicable to continuous, discrete, and semicontinuous types of confidential variables with which a much wider range of analyses can be conducted using the masked dataset with results similar to those using the original one (in particular, they nest the commonly used Generalized Linear Models for statistical data analysis as special cases);

(5) they preserve the set of the original values of confidential variables, which can lead to greater acceptance of masked data among the common users;

(6) they further increase the robustness of masked datasets because they directly model the conditional distribution of confidential variables, instead of modeling the joint distribution of all variables; and

(7) they evaluate the risk of releasing masked datasets with closed-form disclosure risk measures that are simple to calculate, regardless of the types and distributions of confidential variables.

2. Existing Data Masking Methods

A number of approaches to optimally preserve the confidentiality of sensitive information in a database have been proposed, including aggregation, coarsening, imputation, swapping, and perturbation (Willenborg and de Waal 2001). Past research suggests perturbation is a superior class of data masking methods for maximizing data utility and minimizing data disclosure risk (Sarathy et al. 2002). This approach creates perturbed values, which replace the original confidential values. Releasing the perturbed version of the datasets makes it much harder for data intruders to recover the original values of those confidential variables, thereby maintaining data privacy.

There is significant past research on data perturbation methods. Muralidhar et al. (1999) introduce the general additive data perturbation method (GADP), and demonstrate its superior performance in terms of both data utility and security over previous data perturbation methods and thus also over a range of alternative non-perturbation masking methods (Muralidhar et al. 1995). When modeling assumptions are satisfied, GADP has optimal performance. For example, it can be tuned to maintain the linear relationship exactly (Muralidhar and Sarathy 2001). A key assumption in GADP is that the attributes in a database can be modeled by a multivariate normal (MVN) distribution. When this assumption is violated, bias can be introduced into the analysis based on the perturbed dataset, which reduces the utility of secure databases.

Efforts to relax this restrictive assumption have been made recently. There are two main strategies. The first is to apply different, preferably more general and flexible, fully parametric multivariate distributions than the MVN models to perform data perturbation. Muralidhar et al. (1995) propose using a log-normal distribution to model one skewed confidential attribute. Lee et al. (2010) propose a more general data perturbation method, STDP. This method is based on a richer family of parametric multivariate skew-t (MVST) distribution that allows database managers to model skewness and heavy tails in the data and can better answer higher-level questions. Another strategy is to use nonparametric models for the joint distribution of all variables in a database. Sarathy et al. (2002) propose a MVN copula-based GADP (C-GADP) method, which relaxes the strong parametric distributional assumptions in those fully parametric data perturbation methods by using the empirical marginal distributions. As a result, for a much broader range of applications, C-GADP can preserve the marginal distributions of confidential variables. In addition, unlike GADP, C-GADP is capable of preserving monotonic nonlinear relationships among attributes.

Another approach whose performance is on par with perturbation methods is the data shuffling method (DSP) (Muralidhar and Sarathy 2006). Its idea is akin to data swapping (Dalenius and Reiss 1982), which exchanges confidential values among observations. Consequently, DSP fully preserves the marginal distributions of confidential variables. It also helps overcome the reservations about using modified confidential data (*Wall Street Journal* 2001), and can lead to greater acceptance of masked data in practice. DSP outperforms its predecessor, data swapping, by having higher data utility and lower disclosure risk.

Another very powerful class of masking approach is the multiple imputation synthetic data approach (MI, Rubin 1993, Raghunathan et al. 2003, Reiter 2005, Reiter and Raghunathan 2007).

Unlike the above DPS methods, which view the original data as the population, MI views the original data as a sample drawn from a population, and draws multiply imputed synthetic datasets from this population. In its most radical form, no unit in the released data is in the original dataset. This approach has a number of merits, including its ability to assess the inferential uncertainty introduced in the masking process. Raghunathan et al. (2003) and Reiter (2005) discuss the advantages and disadvantages of MI.

There is also active research on other types of privacy-preserving methods that do not rely on explicit statistical models. For example, Xiao and Tao (2006) develop a new data generalization framework for personalized privacy. Menon and Sarkar (2007) formulate the frequent itemset hiding problem as a mathematical integer programming problem, and develop an effective two-phase approach to solving the problem. Li and Sarkar (2011) combine recursive partitioning with bounded swapping to prevent record linkage disclosure. Bertino et al. (2005) outline the issues and approaches for privacy preserving from the perspective of computer scientists.

3. Need for More Effective Data Perturbation and Shuffling Methods

We view the research presented here as advancing data perturbation and shuffling (DPS) procedures based on statistical models. One benefit of such model-based methods is that their performance is theoretically more predictable and they are supported by statistical theory. By making the modeling assumptions explicit, the users know better when these methods perform optimally and when there is a substantial room to improve. For example, a database may contain sensitive attributes of mixed discrete and continuous types, and various types of statistical analyses can be performed on these variables. Especially important is the family of Generalized Linear Models (GLMs) (McCullagh and Nelder 1989), which include normal, binomial, Poisson, Gamma, and inverse Gaussian regression models as special cases. It is important that data perturbation methods are general and flexible to ensure that the masked dataset can generate results for this wide range of types of analyses similar to those using the original database. A trend is therefore to develop effective procedures that provide high data utility in broader applications with sufficiently low disclosure risk. Although the existing model-based DPS methods are enlightening and powerful, there exist unresolved problems which reduce the utility of the resulting secure databases. To motivate the need for new data perturbation and shuffling methods that can better address these important issues, we describe these issues below.[2]

(1) *The existing methods may not maintain the marginal distributional properties of confidential variables in general situations.* The parametric MVN or MVST distributions rely on strong distributional assumptions that may not hold for all the variables in a database. Many distributional features, such as boundedness, semicontinuity and discreteness (e.g., those features occurring

[2] To be fair to the developers of existing methods, some of these problems are reported in the literature (Muralidhar and Sarathy 2006, Lee et al. 2010) and the authors made clear that their methods are not designed to address these problems.

in the patent count and competition index data), multimodality, outliers, and heterogeneous tail behaviors, can occur in databases but cannot be accommodated by these parametric models. Consequently, the data masking procedures based on these parametric distributions cannot preserve the distributional properties of confidential variables in general cases. Similar concerns exist for copula-based methods that use parametric models for marginals. There is generally a lack of guidelines for choosing suitable marginals, and misspecification of these marginal functions can lead to similar problems (Kim et al. 2007). This is why a generally preferred approach for copula applications is to use empirical distributions for these marginals. However, such nonparametric copula modeling is at best fraught with caution for discrete data. As will be shown in a later section, substantial bias arises when applying such nonparametric copula-based masking procedures for discrete confidential variables.

(2) *The existing methods cannot preserve nonmonotonic relationships.* Nonmonotonic relationships, such as an inverted-U relationship, are of great importance for policy and management decision makers (Aghion et al. 2005, Qian 2007). Such relationships are key to determining optimal policy and managerial intervention. Despite this importance, existing DPS methods lack the ability to preserve these relationships. The correlation parameters in the MVN model are Pearson product-moment correlation coefficients, which are only suitable for measuring linear relationships. The copula-based methods (C-GADP and DSP) are more general. These methods capture the dependence among attributes by rank order correlation and are able to preserve monotonic nonlinear relationships. However, nonmonotonic relationships are not preserved. For example, when these masking methods are applied to the C&I dataset, users cannot recover the important nonmonotonic inverted-U relationship between competition and innovation (Fig 2). Consequently, suboptimal policy and managerial decisions will be made. One may consider creating subsets of the original dataset so that within each subset the relationship is monotonic. This may work well in some situations. However, this requires nonrandom subsetting of the dataset and so requires considerably more prior knowledge about relationships among attributes. Furthermore, this strategy may lead to small sample sizes in some subsets. What is needed are more flexible DPS methods that can preserve important nonmonotonic relationships and increase the utility of secure databases so that opportunities to inform optimal decision making are not missed.

(3) *The existing methods lack the ability to handle discrete and semicontinuous confidential variables.* Discrete and semicontinuous data occur frequently in the real world. Semicontinuity arises when an attribute is bounded by a lower and/or upper bound but otherwise is distributed continuously. Typically such variables have a non-zero probability occurring at the bound(s). Examples of semicontinuous variables are the competition index in our C&I example, respondents' incomes, or the amount of expenditures on a product or service in marketing surveys or consumer databases (because many respondents may have no income or no expenditure on certain products or

services). Despite the abundance of discrete and semicontinuous variables in databases, existing DPS methods are not designed for masking these types of attributes. Both GADP and STDP are based on multivariate distributions for continuous data. When applied to discrete and semicontinuous variables, these methods can lead to bias in the masked datasets. Furthermore, both MVN and MVST have the entire real line as the support, and thus can create masked datasets with meaningless masked values. One also needs to take extra care when applying C-GADP and DSP to mask discrete and semicontinuous variables because, as discussed above and as will be shown later, a nonparametric copula model can lead to biased results in data masking for these types of confidential variables. Therefore, new DPS masking methods are needed that can handle these different types of variables and construct secure databases on which a much wider range of statistical analysis, such as the widely used GLMs, yield results similar to those using the original data.

(4) *The existing methods typically require modeling nonconfidential variables, even though these variables remain unchanged before and after data masking.* Prior DPS methods (GADP, C-GADP, STDP, and DSP) require a joint model for both confidential and nonconfidential variables. Among the existing DPS methods, DSP requires minimal modeling of the nonconfidential variables. However, it still requires modeling the relationships among the nonconfidential variables, and imposes monotonic relationships among them. Intuitively, the modeling of nonconfidential variables can be avoided because they remain unchanged before and after data masking. The extra modeling of the nonconfidential variables creates two difficulties. First, it makes it harder to find a suitable joint model that can simultaneously model all the variables in a database reasonably well. As noted in Muralidhar et al. (1999), the challenge in constructing new data masking procedures for nonnormal data is the lack of multivariate distributions amenable to model manipulation and random number generation. The extra modeling of nonconfidential variables further complicates the issue. Furthermore, misspecified models for nonconfidential variables can have adverse effects on masked confidential variables. One example is noted in Lee et al. (2010), who indicate that their STDP procedure has difficulty dealing with heterogeneous tail behaviors. This is because the MVST used in STDP has only one degree of freedom parameter governing the tail behaviors of all variables. If nonconfidential variables have different tail behaviors than confidential variables, the single degree of freedom parameter has to provide a compromise between those differential tail behaviors, which can introduce bias on the tail behavior estimation of confidential variables.[3] As another example, some nonconfidential variables (e.g., the $Year$ dummies) in the C&I data are a group of binary variables for which some combinations of values cannot occur (e.g., any pair of

[3] The alternative approach that uses more general multivariate t distributions with multiple degrees of freedom parameters is unrealistic due to its unappealing parametric form and difficulty in random number generation (Lee et al. 2010).

$Year$ dummies cannot take the value of one simultaneously). A joint MVN or MVN copula model is highly questionable for modeling such nonconfidential variables.

These important limitations of existing DPS methods call for more effective data perturbation and shuffling procedures. The objective of this paper is to propose a new set of procedures, developed below, that can better address the above issues.

4. The MORE Approach to Enhancing the Utility of Secure Databases

In this section we describe our approach to data **M**asking by an **O**dds **R**atio **E**xpression (**MORE**) of the conditional distribution of confidential variables. We first provide an overview of our overall approach.

4.1. Overview

Let $\mathbf{S} = (S_1, \cdots, S_{L_S})$ be a vector of length L_S containing nonconfidential variables, $\mathbf{X} = (X_1, \cdots, X_{L_C})$ be a vector of length L_C containing confidential variables, and $\mathbf{Y} = (Y_1, \cdots, Y_{L_C})$ be a vector of the same length containing masked confidential variables. The joint distribution of the three sets of variables can be written in the following form:

$$f(\mathbf{S}, \mathbf{X}, \mathbf{Y}) = f(\mathbf{S})f(\mathbf{X}|\mathbf{S})f(\mathbf{Y}|\mathbf{S}, \mathbf{X}).$$

Our overall approach for constructing a secure database consists of the following steps:

- Condition on the nonconfidential variables in $\mathbf{S}$. That is, our approach does not require modeling $f(\mathbf{S})$ because $\mathbf{S}$ remains unchanged before and after data masking.
- Estimate an odds ratio model for $f(\mathbf{X}|\mathbf{S})$.
- Generate masked data values using $f(\mathbf{Y}|\mathbf{S}, \mathbf{X})$. In order to achieve high security of $\mathbf{Y}$, we set $f(\mathbf{Y}|\mathbf{S}, \mathbf{X}) = f(\mathbf{Y}|\mathbf{S})$. In order to maintain the distributional characteristics of confidential variables, we further set $f(\mathbf{Y}|\mathbf{S}) = f(\mathbf{X}|\mathbf{S})$.

It is important to note that the above data masking approach is a conditional distribution approach. That is, the masked values are generated from the conditional distribution $f(\mathbf{X}|\mathbf{S})$. As discussed in Muralidhar and Sarathy (2003), a conditional distribution data masking approach has good properties. It satisfies the following two requirements for ideal data masking methods. (1) Data utility requirement. Because we set $f(\mathbf{Y}|\mathbf{S}) = f(\mathbf{X}|\mathbf{S})$, the relationship between $\mathbf{Y}$ and $\mathbf{S}$ remains the same as that between $\mathbf{X}$ and $\mathbf{S}$. Because $f(\mathbf{Y}) = \int f(\mathbf{Y}|\mathbf{S})f(\mathbf{S})d\mathbf{S} = \int f(\mathbf{X}|\mathbf{S})f(\mathbf{S})d\mathbf{S} = f(\mathbf{X})$, it is readily seen that the marginal distribution of $\mathbf{Y}$ remains the same as that of $\mathbf{X}$. (2) Disclosure risk requirement. The conditional approach sets $f(\mathbf{Y}|\mathbf{S}, \mathbf{X}) = f(\mathbf{Y}|\mathbf{S})$. Therefore, given the nonconfidential variable $\mathbf{S}$, $\mathbf{X}$ and $\mathbf{Y}$ are independent of each other, meaning that knowing the masked values $\mathbf{Y}$ provides no additional information about the original values of $\mathbf{X}$, thereby effectively controlling for the disclosure risk. Of course, these requirements may not be met when the modeling assumptions in a conditional distributional data masking approach are violated by the data. As will be seen below, because our modeling approach is very general, the proposed approach provides high data utility in much broader applications than existing data masking procedures, while still effectively controlling for the disclosure risk. Furthermore, unlike prior approaches that derive $f(\mathbf{X}|\mathbf{S})$ from a joint model for $f(\mathbf{S},\mathbf{X})$, we directly model $f(\mathbf{X}|\mathbf{S})$ and bypass modeling $f(\mathbf{S})$. This further increases the robustness of masked data.

4.2. Representing a Conditional Distribution Using Odds Ratio Models

Key components for the data masking approach include the modeling and estimation of $f(\mathbf{X}|\mathbf{S})$, and the generation of masked values from $f(\mathbf{Y}|\mathbf{S})$. Our approach utilizes odds ratio models and provides a unified, flexible, and robust framework for modeling, estimation, and secure value generation for a wide variety of types of confidential variables. Therefore, the proposed methods are particularly attractive as general-purpose data masking techniques. The odds ratio model was first proposed in Chen (2004), and a Bayesian approach was first proposed in Qian and Xie (2011) that outperforms Chen (2004) for high-dimensional missing covariate problems frequently seen in business applications. These methods employ odds ratio models for regression covariates, and are designed for correcting selection bias in parametric regression model estimates due to missingness in covariates. We adopt the odds ratio models for solving data privacy problems, and study various issues specific to data privacy problems, including developing and assessing efficient data masking algorithms that can effectively address the utility issues of secure datasets as described in Section 3, developing disclosure risk measures for evaluating the risk of releasing masked datasets, and evaluating the utility of masked databases. [4]

Before moving to technical details, we first provide some intuitions behind the mathematics so that users can capture the basic idea of the approach. Consider a simple case where all variables in a dataset are independent of each other. In this situation a simple robust approach to masking a confidential variable is to generate perturbed values from its empirical distribution. Our masking approaches are akin to this idea, except that they allow for and preserve the complex relationships among variables through flexible odds ratio functions. In fact, by setting all odds ratio functions to be one, our masking approaches reduce to the simple masking approach using the empirical distributions. The odds ratio model represents the conditional distribution $f(\mathbf{X}|\mathbf{S})$ as follows:

$$f(X_1,\cdots,X_{L_C}|\mathbf{S}) = \prod_{l=1}^{L_C} f(X_l|\mathbf{S},\widetilde{\mathbf{X}}_l), \tag{1}$$

[4] Because a missing data problem does not have privacy issues, these missing data methods in Chen (2004) and Qian and Xie (2011) do not consider measuring and controlling for disclosure risk. Furthermore, they do not consider issues regarding the generation and utility of masked secure datasets.

where $\widetilde{\mathbf{X}}_l = (X_1, \cdots, X_{l-1})$ when $l > 1$, and $\widetilde{\mathbf{X}}_l$ reduces to a null set when $l = 1$. As shown above, a separate model is posited for each confidential variable. It thus offers the flexibility to model distributional characteristics that are different among attributes, such as heterogeneous tail behaviors. We then model each conditional distribution $f(X_l|\mathbf{S}, \widetilde{\mathbf{X}}_l)$ as follows:

$$f(X_l|\mathbf{S}, \widetilde{\mathbf{X}}_l) = \frac{\eta(X_l, X_{l0}; \mathbf{S}, \widetilde{\mathbf{X}}_l, \mathbf{S}_0, \widetilde{\mathbf{X}}_{l0}) f(X_l|\mathbf{S}_0, \widetilde{\mathbf{X}}_{l0})}{\int \eta(X_l, X_{l0}; \mathbf{S}, \widetilde{\mathbf{X}}_l, \mathbf{S}_0, \widetilde{\mathbf{X}}_{l0}) f(X_l|\mathbf{S}_0, \widetilde{\mathbf{X}}_{l0}) dX_l}, \tag{2}$$

where the conditional distribution is expressed as a function of two component functions that can be modeled separately. The first component function, $\eta(X_l, X_{l0}; \mathbf{S}, \widetilde{\mathbf{X}}_l, \mathbf{S}_0, \widetilde{\mathbf{X}}_{l0})$, is the odds ratio function as defined below:

$$\eta(X_l, X_{l0}; \mathbf{S}, \widetilde{\mathbf{X}}_l, \mathbf{S}_0, \widetilde{\mathbf{X}}_{l0}) = \frac{f(X_l|\mathbf{S}, \widetilde{\mathbf{X}}_l) f(X_{l0}|\mathbf{S}_0, \widetilde{\mathbf{X}}_{l0})}{f(X_{l0}|\mathbf{S}, \widetilde{\mathbf{X}}_l) f(X_l|\mathbf{S}_0, \widetilde{\mathbf{X}}_{l0})}, \tag{3}$$

and $\mathbf{S}_0, \widetilde{\mathbf{X}}_{l0}, X_{l0}$ are fixed and prespecified reference points for the odds ratio function in the sample space of $\mathbf{S}$, $\widetilde{\mathbf{X}}_l$, and X_l, respectively. Theoretically, it does not matter how these reference points are chosen. However, if a reference point is absurdly chosen to be extremely remote from the center of a distribution, it may cause numerical instability in the likelihood evaluation. Thus, for the purpose of computational stability we recommend using the mean of each variable as the reference point. Note that from the definition of the odds ratio function given in Equation (3), we have

$$f(X_l|\mathbf{S}, \widetilde{\mathbf{X}}_l) = \frac{\eta(X_l, X_{l0}; \mathbf{S}, \widetilde{\mathbf{X}}_l, \mathbf{S}_0, \widetilde{\mathbf{X}}_{l0}) f(X_l|\mathbf{S}_0, \widetilde{\mathbf{X}}_{l0})}{f(X_{l0}|\mathbf{S}_0, \widetilde{\mathbf{X}}_{l0}) / f(X_{l0}|\mathbf{S}, \widetilde{\mathbf{X}}_l)},$$

and integrating both sides of the above equation with respect to X_l we see that the denominator of the righthand side in the equation above is $\int \eta(X_l, X_{l0}; \mathbf{S}, \widetilde{\mathbf{X}}_l, \mathbf{S}_0, \widetilde{\mathbf{X}}_{l0}) f(X_l|\mathbf{S}_0, \widetilde{\mathbf{X}}_{l0}) dX_l$. Replacing the denominator with this integral immediately leads to Equation (2).

The odds ratio function, $\eta(X_l, X_{l0}; \mathbf{S}, \widetilde{\mathbf{X}}_l, \mathbf{S}_0, \widetilde{\mathbf{X}}_{l0})$, captures the dependence of X_l on $\mathbf{S}$ and $\widetilde{\mathbf{X}}_l$. It reduces to the case of independence between X_l and $(\mathbf{S}, \widetilde{\mathbf{X}}_l)$ when the odds ratio function $\eta(X_l, X_{l0}; \mathbf{S}, \widetilde{\mathbf{X}}_l, \mathbf{S}_0, \widetilde{\mathbf{X}}_{l0})$ is one for all possible values of X_l, $\widetilde{\mathbf{X}}_l$, and $\mathbf{S}$. The odds ratio function can be modeled in a generalized log-bilinear form (Chen 2004) as follows:

$$\ln \eta_\gamma(X_l, X_{l0}; \mathbf{S}, \widetilde{\mathbf{X}}_l, \mathbf{S}_0, \widetilde{\mathbf{X}}_{l0}) = \sum_{v=1}^{L_S} \sum_{m=1}^{M_v} \gamma_{lvm} (X_l - X_{l_0})(S_v - S_{v0})^m + \sum_{w=1}^{l-1} \sum_{m=1}^{M_w} \gamma_{lwm} (X_l - X_{l_0})(X_w - X_{w0})^m \tag{4}$$

As shown below, the choice of the log-bilinear odds ratio function makes clear that odds ratio models nest GLMs as special cases. The above odds ratio function includes higher-order terms up to the order of M. By increasing the value of M, this form of odds ratio function can approximate any odds ratio function arbitrarily smooth. In particular, it allows for both monotonic and non-monotonic relationships. With this added flexibility, one would need to determine optimal value(s) of M that adequately capture the relationships among attributes in a parsimonious manner. The selection of values for M can be formulated as a model selection problem. Because our approach is based on statistical models, a set of model selection methods developed in the statistical field can be applied for this purpose, such as the one based on the likelihood ratio test (Chen 2007). There is a close connection between the log-bilinear form of odds ratio function with the GLMs that are frequently used in practical database analyses. To see this, let X_l follow a GLM, and its mean, μ_l, is $g(\mu_l) = \alpha_0 + \alpha_1 S_1 + \cdots + \alpha_{L_S} S_{L_S} + \alpha_{L_S+1} X_1 + \cdots + \alpha_{L_S+l-1} X_{l-1}$, where $g(\cdot)$ is the canonical link. The corresponding odds ratio function can be derived to be

$$\ln \eta(X_l, X_{l0}; \mathbf{S}, \widetilde{\mathbf{X}}_l, \mathbf{S}_0, \widetilde{\mathbf{X}}_{l0}) = \sum_{v=1}^{L_S} \frac{\alpha_v}{a(\tau)} (X_l - X_{l0})(S_v - S_{v0}) + \sum_{v=L_S+1}^{L_S+l-1} \frac{\alpha_v}{a(\tau)} (X_l - X_{l0})(X_{v-L_S} - X_{(v-L_S)0}),$$

where $a(\tau)$ is the dispersion function in GLMs. This is a reduced form of the generalized log-bilinear odds ratio function in Equation (4), where $M = 1$. Sometimes it can be useful to consider a transformed log-bilinear odds ratio model as follows:

$$\begin{aligned} \ln \eta(X_l, X_{l0}; \mathbf{S}, \widetilde{\mathbf{X}}_l, \mathbf{S}_0, \widetilde{\mathbf{X}}_{l0}) = & \sum_{v=1}^{L_S} \gamma_{lv} (G_l(X_l) - G_l(X_{l_0}))(H_v(S_v) - H_v(S_{v0})) \\ & + \sum_{w=1}^{l-1} \gamma_{lw} (G_l(X_l) - G_l(X_{l_0}))(G_w(X_w) - G_w(X_{w0})), \end{aligned} \tag{5}$$

where $G(\cdot)$ and $H(\cdot)$ are monotone transformation functions. For example, $G_l(\cdot)$ and $H_v(\cdot)$ can be the distribution functions for X_l and S_v that are related to rank-order-based models and procedures.

We next move to modeling the second component function, $f(X_l|\mathbf{S}_0, \widetilde{\mathbf{X}}_{l0})$, the density function of X_l at the reference point $(\mathbf{S}_0, \widetilde{\mathbf{X}}_{l0})$. The odds ratio model models this function nonparametrically. Let $(x_{l1}, \cdots, x_{lK_l})$ be the unique values of X_l observed in the dataset. A nonparametric model for $f(X_l|\mathbf{S}_0, \widetilde{\mathbf{X}}_{l0})$ assigns point mass $(p_{l1}, \cdots, p_{lK_l})$ on these unique values, such that $\sum_{k=1}^{K_l} p_{lk} = 1$, and $p_{lk} > 0$ $\forall k = 1, \cdots, K_l$. To remove the model constraint, we can reparametrize the parameters in $f_\lambda(X_l|\mathbf{S}_0, \widetilde{\mathbf{X}}_{l0})$ such that $\lambda_{lk} = \ln \frac{p_{lk}}{p_{lK_l}}, \quad \forall k = 1, \cdots, K_l$. On the other hand, in GLMs the density function $f(X_l|\mathbf{S}_0, \widetilde{\mathbf{X}}_{l0})$ follows a parametric distributional model. Therefore, it is readily seen that the odds ratio model nests the commonly used parametric GLMs as special cases. It is also important to note that, because a MVN model is equivalent to a product of conditional normal models, it follows that the odds ratio model nests the MVN as a special case.

4.3. The MORE Procedures

We develop two data masking procedures using the odds ratio modeling approach. The first (second) procedure generates perturbed (shuffled) values for confidential variables. We describe these procedures below. In order to help understand these materials and the MORE procedures in general, we also provide a small example that demonstrates numerically how the MORE procedures perform data masking in the Appendix.

4.3.1. MORE-P Procedure

The first procedure creates perturbed values and uses them to replace the original values of confidential attributes. This procedure is termed as MORE-P and is described below.

- **Step 1**: Estimate the odds ratio model parameters $\boldsymbol{\theta} = (\boldsymbol{\lambda}, \boldsymbol{\gamma})$ in $f_{\boldsymbol{\theta}}(\mathbf{X} = \mathbf{x}|\mathbf{S} = \mathbf{s})$ using the observed data $(\mathbf{s}, \mathbf{x})$, where $\boldsymbol{\gamma}$ and $\boldsymbol{\lambda}$ can be considered the location and scale parameters, respectively. Because the model likelihood factorizes as in Equation (1), the estimation can proceed for each set of parameters, $\boldsymbol{\theta}_l, l = 1, \cdots, L_c$, separately by the method of maximum likelihood estimation (MLE) as follows:

$$\hat{\boldsymbol{\theta}}_l = (\hat{\boldsymbol{\lambda}}_l, \hat{\boldsymbol{\gamma}}_l) = \text{argmax}_{\boldsymbol{\lambda}_l, \boldsymbol{\gamma}_l} \sum_{i=1}^{N} \ln L_i(\lambda_l, \gamma_l | X_{il} = x_{il}; \mathbf{S}_i = \mathbf{s}_i, \widetilde{\mathbf{X}}_{il} = \widetilde{\mathbf{x}}_{il}),$$

where $i = 1, \cdots, N$ is the index for independent observations, and N denotes the sample size. The likelihood from the ith observation, $L_i(\boldsymbol{\lambda}_l, \boldsymbol{\gamma}_l | X_{il} = x_{il}; \mathbf{S}_i = \mathbf{s}_i, \widetilde{\mathbf{X}}_{il} = \widetilde{\mathbf{x}}_{il})$, is

$$L_i(\boldsymbol{\lambda}_l, \boldsymbol{\gamma}_l) \propto f_{\boldsymbol{\theta}_l}(X_{il} = x_{il} | \mathbf{S}_i = \mathbf{s}_i, \widetilde{\mathbf{X}}_{il} = \widetilde{\mathbf{x}}_{il}) = \frac{\sum_{k=1}^{K_l} 1_{\{x_{il} = x_{lk}\}} \eta_{\gamma_l}(x_{lk}, x_{l0}; \mathbf{s}_i, \widetilde{\mathbf{x}}_{il}, \mathbf{s}_0, \widetilde{\mathbf{x}}_{l0}) \exp(\lambda_{lk})}{\sum_{k=1}^{K_l} \eta_{\gamma_l}(x_{lk}, x_{l0}; \mathbf{s}_i, \widetilde{\mathbf{x}}_{il}, \mathbf{s}_0, \widetilde{\mathbf{x}}_{l0}) \exp(\lambda_{lk})} \tag{6}$$

There is generally no closed-form MLE solution so an iterative optimization algorithm is required for finding the MLEs. We find that the Quasi-Newton (QN) algorithm performs well for obtaining the MLEs of model parameters. The QN algorithm requires evaluating the derivatives of the log-likelihood function, which are derived as follows:

$$\begin{aligned}
\frac{\partial \ln L_i(\boldsymbol{\lambda}_l, \boldsymbol{\gamma}_l)}{\partial \lambda_{lk}} &= 1_{(x_{il} = x_{lk})} - \frac{\eta_{\gamma_l}(x_{lk}, x_{l0}; \mathbf{s}_i, \widetilde{\mathbf{x}}_{il}, \mathbf{s}_0, \widetilde{\mathbf{x}}_{l0}) \exp(\lambda_{lk})}{\sum_{k'=1}^{N_l} \eta_{\gamma_l}(x_{lk'}, x_{l0}; \mathbf{s}_i, \widetilde{\mathbf{x}}_{il}, \mathbf{s}_0, \widetilde{\mathbf{x}}_{l0}) \exp(\lambda_{lk'})} \\
\frac{\partial \ln L_i(\boldsymbol{\lambda}_l, \boldsymbol{\gamma}_l)}{\partial \gamma_{lvm}} &= (x_{il} - x_{l0})(s_{iv} - s_{v0})^m - \frac{\sum_{k'=1}^{N_l} \eta_{\gamma_l}(x_{lk'}, x_{l0}; \mathbf{s}_i, \widetilde{\mathbf{x}}_{il}, \mathbf{s}_0, \widetilde{\mathbf{x}}_{l0}) \exp(\lambda_{lk'})(x_{lk'} - x_{l0})(s_{iv} - s_{v0})^m}{\sum_{k'=1}^{N_l} \eta_{\gamma_l}(x_{lk'}, x_{l0}; \mathbf{s}_i, \widetilde{\mathbf{x}}_{il}, \mathbf{s}_0, \widetilde{\mathbf{x}}_{l0}) \exp(\lambda_{lk'})} \\
\frac{\partial \ln L_i(\boldsymbol{\lambda}_l, \boldsymbol{\gamma}_l)}{\partial \gamma_{lwm}} &= (x_{il} - x_{l0})(x_{iw} - x_{w0})^m - \\
&\qquad \frac{\sum_{k'=1}^{N_l} \eta_{\gamma_l}(x_{lk'}, x_{l0}; \mathbf{s}_i, \widetilde{\mathbf{x}}_{il}, \mathbf{s}_0, \widetilde{\mathbf{x}}_{l0}) \exp(\lambda_{lk'})(x_{lk'} - x_{l0})(x_{iw} - x_{w0})^m}{\sum_{k'=1}^{N_l} \eta_{\gamma_l}(x_{lk'}, x_{l0}; \mathbf{s}_i, \widetilde{\mathbf{x}}_{il}, \mathbf{s}_0, \widetilde{\mathbf{x}}_{l0}) \exp(\lambda_{lk'})}.
\end{aligned}$$

As shown above, these derivatives are given in closed form containing no integrals. Only a summation over a finite number of points is required, and can be evaluated straightforwardly. The evaluation of the model likelihood and derivatives is fed to the IMSL Fortran library routine *UMING* to perform the functional optimization.

- **Step 2**: Simulate the perturbed values of confidential variables from $f_{\boldsymbol{\theta}}(\mathbf{Y}|\mathbf{S})$. To ensure a high utility for the secure database, we set $\boldsymbol{\theta}$ to be $\hat{\boldsymbol{\theta}}$ estimated in Step 1 above. One could simulate the perturbed values for all the confidential variables altogether, which may involve evaluating the probabilities over a large number of combinatorial terms. A computationally much simplified approach is to generate perturbed values for one confidential variable at a time. For the ith observation of the lth confidential variable, Y_{il}, we generate y_{il} from the following multinomial distribution on the set of values $(x_{l1}, \cdots, x_{lK_l})$ uniquely observed in the original dataset:

$$Y_{il}|(\mathbf{S}_i = \mathbf{s}_i, \widetilde{\mathbf{Y}}_{il} = \widetilde{\mathbf{y}}_{il}) \sim multinomial([P_{il1}, \cdots, P_{ilK_l}]),$$

where $\widetilde{\mathbf{Y}}_{il} = (Y_{i1}, \cdots, Y_{i,l-1})$ if $l > 1$ and it reduces to a null set if $l = 1$ and the kth component in the multinomial probability vector $[P_{il1}, \cdots, P_{ilK_l}]$, for $k = 1, \cdots, K_l$, is given as

$$P_{ilk} = \frac{\eta_{\hat{\gamma}_l}(x_{lk}, x_{l0}; \mathbf{s}_i, \widetilde{\mathbf{y}}_{il}, \mathbf{s}_0, \widetilde{\mathbf{x}}_{l0}) \exp(\hat{\lambda}_{lk})}{\sum_{k'=1}^{K_l} \eta_{\hat{\gamma}_l}(x_{lk'}, x_{l0}; \mathbf{s}_i, \widetilde{\mathbf{y}}_{il}, \mathbf{s}_0, \widetilde{\mathbf{x}}_{l0}) \exp(\hat{\lambda}_{lk'})}. \tag{7}$$

It is important to note that the simulated values are in the set of the values observed in the dataset, and therefore can lead to greater acceptance of data perturbation methods in practical users.

- **Step 3**: Report $\mathbf{Y}$ as the perturbed values of confidential attributes in the released database.

4.3.2. MORE-S Procedure

MORE-P is only able to preserve the marginal distributions asymptotically. If one needs to preserve the marginal distributions exactly, a useful idea is data shuffling (Muralidhar and Sarathy 2006) based on which we develop a shuffling procedure, named as MORE-S. Its algorithm is described below.

- **Step 1**: Apply MORE-P to generate perturbed confidential values, named as $\mathbf{Y}^p = (Y_1^p, \cdots, Y_{l_c}^p)$.
- **Step 2**: Rank the values of the lth perturbed variable $Y_l^p, l = 1, \cdots, l_c$, and denote the vector of ranks for Y_l^p as R_l^p. Then replace the perturbed value for the ith observation, Y_{il}^P, with the original value having the same rank, $X_{(R_l^p),l}$. When ties occur in computing ranks, the values in each set of ties are assigned rank values randomly from a set of consecutive ranks such that the assigned ranks for all observations go from 1 to N (the number of observations), and there are no ties in ranks. Denote the resulting values as $Y_l, l = 1, \cdots, l_c$.
- **Step 3**: Release $\mathbf{Y} = (Y_1, \cdots, Y_{l_c})$ as the shuffled values in the secure database.

4.3.3. Security Measures Provided by MORE

In this section we develop a disclosure risk measure based on the distance/closeness between the original value and the masked value that can be useful for quantifying the security level provided by MORE. We define the expected mean perturbation distance (EMPD) for the lth confidential variable as

$$EMPD = E\left[\frac{\sum_{i=1}^N D(Y_{il}, x_{il})}{N}\right] = \frac{\sum_{i=1}^N \sum_{k=1}^{K_l} D(y_{ilk}, x_{il}) P_{ilk}}{N}, \tag{8}$$

where P_{ilk} is given in Equation (7) and D denotes a perturbation distance function. The quantity inside the bracket is the mean perturbation distance between the perturbed and original values, and the expectation is taken with respect to the distribution of perturbed values. This security measure is consistent with the concept of data perturbation that considers the original data as the population. An intuitive choice for the perturbation distance function $D(a, b)$ is the absolute difference, i.e., $|a - b|$, which is used in later sections. As seen in Equations (7) and (8), one advantage of MORE is that this disclosure risk measure can be computed simply as a by-product of masking, and has a closed-form expression regardless of the types and distributions of masked values.

The security measure is very useful for quantifying and diagnosing the expected prediction disclosure risk of releasing a secure database under a specific data masking setting. For example, a very low value of EMPD means a high prediction power, implying a high predictive disclosure risk. This is likely because there is a nonconfidential variable or a combination of several nonconfidential variables that predict well the confidential attribute. In extreme cases, the confidential attribute may collinear with the nonconfidential variables or close to be a deterministic function of the nonconfidential variables. In practice, database owners make judgments on a suitable cutoff value for disclosure risk. When the EMPD is less than this cutoff value, the secure data may not satisfy the disclosure risk requirement and the data masking setting may need to be changed, e.g., by considering highly predictive nonconfidential variables to be also sensitive attributes. The security measure can thus provide very useful information for quantifying the disclosure risk and for diagnosing what actions need to be taken to increase the security level.

4.3.4. Computational Cost and Simple Strategies to Manage It

We conclude the description of the MORE procedures with an analysis of computational cost and a discussion of recommended strategies to manage it. In the analysis, we will evaluate the impact of various elements, including the numbers of variables, of distinct values, and of estimation iteration steps, on the computational costs of MORE procedures. First, because MORE masks sensitive attributes one by one, the computational cost is the sum of that for masking each sensitive attribute. Thus, the computational cost increases additively, rather than multiplicatively, with the number of confidential variables that need to be masked. Second, when masking a generic lth confidential variable, the computational cost-determining step is the model estimation step (i.e., Step 1 of MORE-P procedure as described in Section 4.3.1) because this step requires an iterative optimization process. Note that the number of parameters involved in masking the lth sensitive attribute is $n_l = n_{\lambda_l} + n_{\gamma_l}$, where $n_{\lambda_l}(n_{\gamma_l})$ is the number of parameters in $\lambda_l(\gamma_l)$. Thus, the bottlenecks of computational time are those sensitive attributes that are continuous and have many unique values.[5] For any such bottlenecking sensitive attribute, n_{λ_l}, which equals the number

[5] It is also important to note that these bottleneck variables exclude nonconfidential variables because they are not modeled in MORE. Furthermore, confidential variables that have only a limited number of unique values (e.g., count data) in a massive dataset are also not bottlenecking variables.

of unique values of this attribute, is much larger than n_{γ_l} in big data and is the main factor affecting computational time. The well-established limited-memory Quasi-Newton method is often the choice for large-scale optimization problems. It has an acceptable linear convergence rate, and at each iteration step has a low storage and computation cost that is on the order of $O(mn_l)$ where m is typically fixed at a number between 3 and 20 (Nocedal and Wright 1999, chap. 9). More specifically, the algorithm does not need to store or compute the $n_l \times n_l$ Hessian matrix. Instead it stores only the first derivatives of size n_l from the most recent m iterations, which are used for updating parameter estimates via an inexpensive two-loop recursion scheme that requires approximately $4mn_l$ multiplications (Nocedal and Wright 1999, chap. 9).

We recommend the following strategies to manage the computational cost when using MORE to mask these bottlenecking sensitive variables in big data. One convenient way to reduce the number of the unique values and thus the number of model parameters is rounding. The analysis in Section 5.6 shows that rounding can sufficiently control computational time without affecting the utility of secure datasets. Another simple strategy to further reduce computation time, if desired, is splitting data into random subsets. Although all our computations in this paper are executed on a single processor, data masking for each random subset can be executed on different processors simultaneously using parallel computing. These recommended strategies are simple and flexible to implement and can be easily tailored to the available computational power.

5. Performance of MORE Procedures

In this section we first conduct Monte Carlo simulation studies to evaluate and compare the performance of the proposed MORE procedures with existing DPS approaches for solving data security problems, and then apply MORE to two applications. These evaluations and comparisons demonstrate the capabilities of the MORE procedures to overcome the limitations of the existing DPS methods. We first describe simulation setup.

5.1. Simulation Setup

Our setup emulates a database marketing example. We simulate data consisting of five variables for consumer data from a retail store with the following distributions. The first three variables, S_1, S_2, X_1, representing income, age, and log expenses of a consumer in the store, are simulated from a trivariate normal with a zero mean vector and a variance-covariance matrix in which the diagonal elements are 1 and the off-diagonal elements are 0.5. The fourth variable, X_2, representing the dollar amount of coupons redeemed by the consumer, is generated from the following distribution, $X_2|S_1 \sim N(\beta_{41}S_1 + \beta_{42}S_1^2, 1)$, where β_{41} and β_{42} are set to be 0 and 1, respectively. The fifth variable, X_3, representing the number of transactions made by the consumer in the store, is generated from a Poisson distribution with its rate parameter $\lambda = \exp(\beta_{51}S_1)$, where β_{51} is set to be 1. In the simulation study, the first two variables (S_1, S_2) are nonconfidential and remain unchanged after data masking. The other three variables (X_1, X_2, X_3) are confidential variables whose original values need to be masked. The above simulation setup aims to emulate the following situations under which we can study and compare the performances of different data masking procedures: (1) the classical case of a multivariate normal setting using the confidential variable X_1; (2) a confidential variable X_2 having a nonmonotonic nonlinear relationship with the other variables; and (3) a discrete and nonnormal confidential variable X_3.

For each simulated original dataset, we create datasets that mask the three confidential variables using five methods: GADP, C-GADP, DSP, MORE-P, and MORE-S.[6] Both C-GADP and DSP utilize a nonparametric MVN copula model for the joint distribution of all the variables. We then conduct a range of analyses on the masked datasets and compare the results with those based on the original dataset. In this way we can investigate the performance of different data masking procedures to maintain statistical properties of the original dataset. Using the distributional characteristics of the original data as the true value, we calculate the biases of those of the masked datasets. The bias measures the closeness of the masked datasets to the original dataset and is a sensible measure of the utility of a secure database. Note that the confidential variable X_3 is a type of count data taking only non-negative integer values. However, both GADP and C-GADP can generate masked values that are negative or non-integer. To improve the performance of these two methods for X_3, we postprocess the masked data values of X_3 from these two methods and round the masked values of X_3 to the nearest integer values. Any negative values are reassigned a value of zero. We repeat the simulation 500 times and calculate the average and standard deviation (SD) of the biases. The sample size is varied at the values of 100, 500, and 1000. We summarize and discuss the Monte Carlo simulation results below, which reveal that only the proposed MORE approaches perform well in all these situations.

5.2. Preserving Distributional Characteristics of Confidential Variables

We first study the ability of different approaches to maintain marginal distributions of confidential variables. We apply the nonparametric Komogorov-Smirnov (KS) tests to measure the overall closeness of marginal distributions of masked confidential values to those of original values. Because the simulated original dataset is considered the underlying finite population, we apply a KS test that considers the empirical distribution of the original values as the reference distribution. The KS test measures the distance between the empirical distribution of masked values and this reference distribution. The smaller value of the KS test statistic, the closeness of the distribution of the masked values to that of the original values and thus the more faithful preservation of marginal distribution of confidential variables. Therefore, the KS statistic serves as a measure of bias in the

[6] The more recent STDP approach (Lee et al. 2010) performs better for heavy tails and highly skewed data than GADP. However, as a parametric approach for continuous data, STDP shares some important limitations with GADP. Thus, for the sake of relative ease in implementation, we use GADP as the benchmark model to compare with.

overall distribution of masked values. We calculate the average and SD of the KS test statistics over all 500 simulated datasets for each data masking procedure. We also perform similar analyses for some other important but less comprehensive distributional summaries, including moments (Mean, Variance, Skewness, Kurtosis), and quantiles at 5%(Q05), 25% (Q25), 50% (Q50), 75% (Q75), and 95% (Q95). The results are summarized in Table 1. DSP and MORE-S are not included in this table because these two methods shuffle the original values among records, and as a result the marginal distributions of confidential variables are preserved exactly. We also exclude GADP from the table because it can generate biased masked data when the assumption of MVN is violated. As expected, we do find significant bias when GADP is used for masking variables X_2 and X_3. Therefore, in Table 1 we summarize the results for the two nonparametric perturbation procedures, C-GADP and MORE-P.

(1) For confidential variable X_1, which follows a multivariate normal with the two nonconfidential variables, the results in Table 1 show that both C-GADP and MORE-P perform well. The KS test statistics are small and decrease as the sample size increases, implying that as data become richer, the marginal distribution of a confidential variable can be estimated more accurately and all methods can preserve this marginal distribution arbitrarily well. We observe the same pattern for moments and quantiles. Although the results for GADP are not presented here for reasons given above, they show that both C-GADP and MORE-P perform somewhat better than the parametric GADP, even though the data are simulated from a MVN. The reason is that the *observed* original values are considered as the finite population in data masking. The two nonparametric procedures, C-GADP and MORE-P, can adapt better to the shape of the empirical distribution of the observed original values than GADP. Therefore, these two methods can better preserve the marginal distributions of the observed original confidential data.

(2) For confidential variable X_2, which has a nonlinear relationship with the nonconfidential variables, the results show that both C-GADP and MORE-P perform well in preserving the marginal distributions, although MORE-P performs noticeably better. GADP performs worse than C-GADP and MORE-P because of the restrictive linear dependence structure of the MVN model, the modeling approach adopted in GADP. Because the correlation parameters in a MVN are Pearson product-moment correlation coefficients, GADP allows only for linear relationships between attributes. Because marginals and the dependence structure are jointly estimated in the MVN model, the misspecification of the dependence structure can have adverse effects on the estimation of marginal distributions. Unlike GADP, MORE is capable of allowing for nonlinear relationships and thus preserving the marginal distributions in the presence of nonlinear relationships. It is interesting to note that the preservation of marginal distribution in C-GADP is more robust than GADP, even though the dependence structure is still of a multivariate normal nature. The reason is that C-GADP uses the nonparametric empirical distributions for marginals and therefore

forces the preservation of marginal distributions. However, as will be seen later, the nonmonotonic relationship will not be preserved in C-GADP.

(3) For confidential variable X_3, the results in Table 1 show that MORE-P performs best. In this case, we find that the KS statistics for both GADP and C-GADP remain large even when the sample size increases. Similar patterns are found for moments and quantiles. Note that the confidential variable X_3 is a count variable, which is nonnormal and discrete and takes no negative values. In this case, because the normality assumption is violated, it is not surprising that GADP performs poorly as GADP is designed for MVN data. What is less obvious is the suboptimal performance of C-GADP. There are two reasons. First, C-GADP invokes a MVN model for dependence structure. Such dependence structure can become incompatible with the nonlinear relationships that involve nonnormal data. Another important reason is that the copula model can have difficulty handling discrete data. Below we discuss more about the latter reason, an issue that is less known in the literature.

Copula modeling has been very popular recently for modeling multivariate distributions. Various parametric approaches to copula estimation have been proposed (see Qu et al. 2009 for a discussion). In practice, a nonparametric copula that uses empirical marginal distributions may be preferred because they rely less on the correct specification of marginal distributions. Copula modeling proves to be very successful for modeling multivariate continuous data. In fact, all previous applications of copula-based data masking procedures (C-GADP and DSP) have focused on continuous variables. On the other hand, the literature on copula-based data masking has been scarce for discrete data, despite the abundance of discrete variables in business fields. One difficulty is that many key properties in the copula theory for continuous data do not hold for discrete data. Genest and Neslehova (2007) review various facts about copula modeling for discrete data, and provide both theoretical derivations and empirical examples demonstrating how discreteness in a probability distribution invalidates various familiar properties that are fundamental to copula theory in the continuous case. The cause of difficulties in discrete data cases lies in the fact that the inverse of a discrete distribution function has plateaus. Therefore, for the discrete data, although the existence of a copula representation is guaranteed, a copula representation compatible with the data is not unique. This nonuniqueness creates an identifiability issue. Consequently, copula inference (and particularly nonparametric rank-based inference) from discrete data is fraught with identification difficulties. Although the estimation through a parametric copula remains possible in some situations, it is unclear under what conditions identifiability is achieved. Furthermore, such a fully parametric copula is computationally more intensive (Qu et al. 2009), and does not have the same level of robustness as its nonparametric counterpart.

Because of the identifiability issue, our simulation results show significant bias in the distributions of masked values. The moments and quantiles of the distributions are also substantially different from those of the original values. These biases do not reduce with larger sample size. In contrast, our approach does not have this problem because no inverse mapping from a discrete distribution function is required.

5.3. Preserving Nonmonotonic Relationships

We next study the performance of different data masking procedures in their ability to preserve various types of relationships among attributes. For the confidential variable X_1, we compute its Pearson correlation coefficients with S_1 (i.e., ρ_{31}) and with S_2 (i.e., ρ_{32}). For the confidential variable X_2, we fit a quadratic regression model that regresses X_2 on S_1 and S_1^2, and obtain the estimates for the corresponding regression parameters β_{41} and β_{42}. For the confidential variable X_3, we fit a Poisson regression model with a log link that regresses X_3 on S_1, and obtain the estimate for the corresponding regression parameter β_{51}. These analyses are performed on both the original simulated datasets and those masked datasets for each data masking procedure. We then compute the average and SD of the biases over all 500 simulated datasets.

Because the confidential variable X_1 and the nonconfidential variables S_1 and S_2 are jointly distributed as a MVN, the Pearson correlation coefficients ρ_{31} and ρ_{32} fully capture the linear relationships between X_1 and S_1 and S_2, respectively. In this linear relationship case, we expect and do find that all data masking procedures perform well. Although a linear relationship is a simple and parsimonious way to describe relationships among variables, it is by no means a universally applicable one. As discussed in an earlier section, nonmonotonic relationships are critical for making optimal policy and management decisions. Such nonlinear relationships can occur for the confidential variable X_2, with which S_1 has a U-shaped relationship. In this case, GADP, C-GADP, and DSP all lead to significant bias in the estimates of regression parameters (β_{41} and β_{42}) because these methods are not designed to maintain the nonmonotonic relationship. Instead of identifying a U-shaped relationship, these methods assert there is no relationship between X_2 and S_1. This is not surprising because the MVN model, the basis of these masking procedures, cannot model nonmonotonic relationships. In contrast, the MORE procedures allow for general nonmonotonic relationships, and thus can preserve the potentially complex but important relationships of variables in the masked datasets. A nonlinear relationship could also arise in a nonlinear regression model, as for the case of confidential variable X_3. The methods GADP, C-GADP, and DSP all have sizable biases for the parameter β_{51} that remains when the sample size increases. These biases arise for two reasons. First, as discussed above, these methods cannot handle the discrete confidential data. Second, they have a restrictive dependence structure that cannot adequately capture the nonlinear relationship in a Poisson regression model. In contrast, the results in Table 2 show that MORE-P and MORE-S perform well in preserving these different relationships. Because GADP, C-GADP, and DSP are not designed for preserving these more complex relationships, we omit results for them in the table.

5.4. Disclosure Risk

While as shown above MORE can substantially enhance the utility of secure databases, a remaining important point is to ensure that these utility improvements do not compromise the security of resulting databases. As explained in Section 4.1, MORE ensures that the masked values in Y are independent of X given S, thereby causing no increase in disclosure risk, given that the released data preserve statistical properties of the original data. As described in Section 4.3, MORE has two equivalent ways to generate the masked values Y. The first is to generate the perturbed values for all the confidential variables together, and the other is to generate perturbed values one at a time using a sequence of conditional distributions. Both ways satisfy the conditional independence requirement and are just different approaches to generating values from the same distribution. [7]

To validate the above theoretical justification, we perform simulation studies to evaluate the value disclosure risk of MORE that uses the conditional simulation approach. We simulate datasets of sample size $n(=100, 500, 1000)$ from a bivariate normal distribution with a specified correlation coefficient $(\rho = 0.0, 0.2, 0.4, 0.6, 0.8, 0.95)$. Both variables were to be masked. We then compute the proportion of variability in a confidential variable explained by the corresponding masked variable. In this case, since S is null, the confidential and masked variables should be independent of each other if the conditional independence assumption is satisfied, and therefore we expect that the masked variable should have no power to explain the confidential variable. Table 3 reports the average proportion of explained variability for each combination of correlation coefficient and sample size for DSP, MORE-P, and MORE-S. It is evident that all three methods provide essentially the same level of excellent protection from the value disclosure risk. In all cases, the proportion of variability explained in the confidential variable using the masked variable is essentially zero.

Table 1 also compares the disclosure risk level between two nonparametric perturbation methods, C-GADP and MORE-P. Both approaches are conditional, and have similar EMPD for X_1.[8] On the other hand, for X_2 and X_3 C-GADP has a larger EMPD. This is expected because C-GADP does not preserve statistical properties of these two confidential variables and thus can have a lower disclosure risk. As explained in Section 4.3, the EMPD can be very useful for quantifying the disclosure risk and informing data owners when actions are needed to increase the security level, e.g., by masking nonconfidential variables that are highly predictive of the confidential variables. With suitable adjustment of the security level, MORE can provide maximal utility for legitimate users and adequate disclosure control for intruders, thereby achieving optimal user-intruder information equilibrium (Muralidhar and Sarathy 2003).

[7] Scheuer and Stoller (1962) show that to generate a vector of random variables, one approach is to generate values for a sequence of conditional distributions one at a time in the same way as used in the conditional simulation approach of MORE. We choose the conditional simulation approach purely because of its computational simplicity, although it will generate secure values from the same distribution as the joint simulation approach.

[8] There is no close-form expression of EMPD for C-GADP. We therefore generate a large number (e.g., 100) of perturbed datasets using C-GADP and use the sample mean of the mean perturbation distance as an estimate of EMPD.

5.5. Application of MORE to the Competition and Innovation Data

We next apply the MORE procedures to the competition and innovation dataset. As shown below, the proposed data masking procedures can properly account for important features in this dataset. The methods are flexible to preserve the nonmonotonic relationship between innovation and competition in the original dataset. Because masked values are in the set of the observed values, the bound requirement is automatically satisfied, which ensures no out-of-bound values are generated. Because of the nonparametric distributional modeling, features such as skewness are also preserved.

For our illustrative purpose, in our data masking analysis we will consider the masking of the confidential variable $X = ptcw$ and the nonconfidential variables $S = (Year, Industry, Ci)$. We first consider a simplified analysis that includes only Ci in S. Figure 3 presents the PP-plots that compare the distributions of masked values of *patcw* from GADP, C-GADP, and MORE-P with the distribution of the original values of *patcw*. Both GADP and C-GADP can generate negative perturbed values of *patcw* that are not meaningful. We thus reassign any negative values to a value of zero. The shuffling methods, DSP and MORE-S, are not reported because they preserve the marginal distribution exactly. A PP-plot compares the cumulative distribution functions from two datasets and provides a graphical evaluation of whether two datasets agree closely. When the masked data resemble the original data, the points in the plot should be close to the straight diagonal dotted line. Figure 3 shows a large discrepancy between the distribution of the masked values using GADP and that of the original values. C-GADP performs much better in preserving the marginal distribution of *patcw* but there is still a noticeable discrepancy for small values of *patcw*. MORE-P performs best because the curve formed by the points is almost indistinguishable from the diagonal line, indicating the excellent preservation of the marginal distribution.

We also perform KS tests to formally evaluate the ability of these methods to preserve the marginal distribution of *patcw*. Table 4 summarizes the results, which show the KS test rejects the null hypothesis that the perturbed values generated using GADP come from the distribution of the original values with a p-value of $< .0001$. There are also substantial biases in the moments and quantiles compared with those calculated from the original values. As a comparison, C-GADP performs much better with a substantially smaller KS statistic and a borderline significant p-value (0.0418). The moments and quantiles of the C-GADP are also much closer to those of the original values as compared with GADP, indicating that the nonparametric marginal method is performing better. MORE-P performs best with much smaller KS statistics and a highly nonsignificant p-value of 0.9989, indicating the masked values preserve the marginal distribution very well.

As a test to verify whether the masked values can preserve the important inverted-U relationship, we run a Poisson regression model on the original values and the masked values of *patcw* for each data masking procedure.[9] The Poisson regression model regresses *patcw* on Ci and Ci^2. This

[9] Note that *patcw* is a citation weighted patent count and therefore can take noninteger positive values. The Poisson regression routines in many software packages (e.g., R and Stata) are extended to handle noninteger response values, even though the classical Poisson regression requires integer values.

mimics a scenario to assess the ability of different data masking procedures to allow a user of a secure database to discover the inverted-U relationship between competition and innovation using the masked values of *patcw*. The regression results using the original data and masked data are reported in Table 5. The prior data masking procedures, GADP, C-GADP, and DSP, are all based on a joint MVN-type model and are not designed for maintaining a nonmonotonic inverted-U relationship. Therefore, using the masked values based on these methods will not preserve the inverted-U relationship. In fact, the analyses from these methods assert no relationship between competition and innovation, as seen from the nonsignificance of regression parameters for Ci and Ci^2. In contrast, MORE-P produces the regression parameters that are closest to those based on the original values and that can preserve the inverted-U relationship. Figure 2 compares the fitted regression curves using each data masking procedure with that using the original data. The figure shows that only MORE-P and MORE-S are able to maintain the inverted-U relationship existing in the original dataset.

We next consider a refined analysis that includes $(Year, Industry, Ci)$ in S. Note that $Year$ and $Industry$ are categorical variables, which are typically incorporated into analyses as a set of dummy variables. This would create additional 21 (for $Year$) + 16 (for $Industry$) dummy variables. This substantially enlarges the modeling work for a joint data masking approach that models these variables, even though these variables remain unchanged after data masking. On the other hand, MORE procedures condition on these variables. Table 6 reports the analysis results. As compared with the simpler analysis that excludes $Year$ and $Industry$, the performance of all data masking procedures improves, although the order of different data masking procedures on their performance remains the same. The KS statistics reduce in size and the corresponding p-values increase, indicating better preservation of the distribution of the original values. The preservation of moments and quantiles also improves for all methods. The results based on Poisson regressions on the original and masked values are summarized in Table 7. In addition to Ci and Ci^2, the Poisson regressions control for year and industry dummies. The parameter estimates in the column of “Original” are the same as those reported in Aghion et al. (2005) and show a strong inverted-U pattern between competition and innovation. As in the simplified analysis, prior methods (GADP, C-GADP, and DSP) cannot preserve the nonmonotonic inverted-U relationship and essentially conclude no relationship between these two variables. In contrast, MORE-P and MORE-S maintain this important relationship. Consistent with our simulation results, because MORE-P preserves statistical distributional properties better, it has a somewhat smaller EMPD than GADP and C-GADP for the simplified analyses, as shown in Table 4. The difference in EMPD among different methods is reduced for the refined analyses, as reported in Table 6, likely because of a much improved fit for all methods in the refined analyses.

We further compare MORE with the multiple imputation synthetic data (MI) approach to data masking, as described in Raghunathan et al. (2003) and Reiter (2005). In order to maintain complex relationships among attributes, we use their parametric MI procedure in which data owners need to specify different types of parametric imputation models for masking mixed types of confidential variables, and then employ different algorithms tailored for generating synthetic data for different types of variables. This is in contrast with MORE, which provides a single nonparametric masking model and algorithm applicable for mixed types of confidential attributes, which is important for general purpose automatic data masking. When generating synthetic datasets, we set the imputation models to be the same as the analysis models, i.e., to be the Poisson regression models for *patcw* conditional on S as specified for fitting data in Tables 5 and 7. This mimics a scenario in which data owners know the types of analyses to be performed on masked datasets. Notice that our odds ratio perturbation models nest the Poisson imputation models as special cases. As shown next, this benefit has important consequences in preserving a much wider range of statistical properties of sensitive attributes.

One hundred synthetic datasets are drawn from the posterior predictive distribution of the sensitive attribute *patcw* using flat priors on the Poisson imputation model parameters. The results from these synthetic datasets are combined using the rules developed by Raghunathan et al. (2003). Because the MI uses the same imputation models as the analysis models, we expect the synthetic datasets to preserve the inverted-U relationship between competition and innovation. This is confirmed by the results under columns “MI” in Tables 5 and 7, which show comparable performance of MORE-P/S and MI, and both are capable of preserving this relationship. The parameter estimates from MI are somewhat closer to the estimates from fitting the original data than MORE. This is because MI takes the average of the 100 estimates from synthetic datasets and thus eliminates the variability of the estimates, which can be appreciable unless sample size is very large. The standard errors from MI are somewhat larger than from MORE, reflecting this added variability. On the other hand, Tables 4 and 6 report the means of the marginal distribution characteristics over 100 synthetic datasets, showing that these synthetic datasets do not preserve the marginal distributions of the confidential variable *patcw*. In both tables, the average KS test statistics and P-values under the columns “MI” show high statistical significance, suggesting a large discrepancy between overall distributions of synthetic datasets and that of the original data. Many summary measures, except mean, show substantial differences from the original ones, which can affect inferences based on these measures. Overall, these synthetic datasets do not preserve the marginal distribution of *patcw* as well as MORE does. This analysis demonstrates the value of the robustness of MORE in preserving a much wider range of statistical properties of sensitive attributes.

The primary advantage of MI over MORE-P/S methods proposed here is its ability to incorporate the added variability that results from the data masking process into statistical inference of secure datasets. A disadvantage of MI is the increased analytical complexity and the need to store and process multiple synthetic datasets (Raghunathan et al. 2003). For instance, simple masking algorithms in DPS may need to be replaced with computationally more expensive algorithms in MI. Avoiding negative standard error estimates in MI would require generating a large number of synthetic datasets or using more a complicated formula involving numerical integrations. Storing and processing a large number of synthetic datasets for masking massive datasets may not always be desirable for data users. Thus, the choice of which class of methods to use in practice will depend on a number of factors, and it is helpful to offer users both types of methods. For situations in which the variability added from data masking and its impact on statistical inference are relatively small (e.g., with a large sample size), one may favor MORE-P/S for its relative analytical simplicity and low cost, robustness, and ability to nest GLMs as special cases. In situations in which the variability introduced in data masking must be properly incorporated, it is desirable to overcome this limitation of MORE-P/S by developing an MI version of MORE. Because of substantial conceptual and implementation differences between the two classes of methods, we leave this development for future research.

5.6. An Additional Application: An Organizational Data Warehouse Example

Our second application is on an organization data warehouse example described in Muralidhar and Sarathy (2006). This dataset contains six variables with three nonconfidential ones (Gender, Marital Status, and Age) and three confidential (Home Value, Mortgage Balance, and Total Net Asset Value). A large dataset with 50,000 observations was generated using the procedure described in Muralidhar and Sarathy (2006). The authors use the dataset to illustrate the applicability of DSP to large datasets and the capability of DSP to maintain the monotonic relationships among variables in this dataset. Unlike the prior application, the relationships between variables have nonlinear but monotonic relationships. One approach is to use the polynomial odds ratio functions to approximate these nonlinear relationships. Another approach is to model the relationships using the transformed log-bilinear form as specified in Equation (5). We adopt this simpler second approach here and use the distribution functions as the transformation functions. After using MORE-P to generate the perturbed values on the transformed variables, we perform data shuffling to obtain the masked dataset. In this dataset, each of the three confidential variables has a large number ($> 10,000$) of unique values. To speed up the computation, we create five random subsets with equal sample sizes (i.e., 10,000 observations per random subset). For each subset, we round the transformed variables to the third decimal place to further reduce the number of unique values. Table 8 reports the analysis result as was done in Muralidhar and Sarathy (2006). As we see, both DSP and MORE-S perform similarly well and both can maintain the rank-order correlations accurately. We also consider the cases where only Home Value (MORE-S1), Mortgage (MORE-S2), or Total Net Asset Value (MORE-S3) is confidential and needs to be masked. As we see from Table 8, MORE performs well in all these cases where different variables are selected for masking. Computationally, DSP requires no iterative optimization of model parameters, but MORE does. Therefore, unlike the C&I dataset, when the requirements for using DSP are satisfied (such as in this data warehouse example), DSP offers a computationally simpler approach, is as effective as MORE-P or MORE-S in maintaining security and utility, and needs not be replaced.

To investigate the effectiveness of the rounding strategy as recommended in Section 4.3.4, we conduct an alternative analysis that rounds the transformed variables to the first decimal place. As seen in Table 8, the different levels of rounding produce very similar results.[10] For this large dataset, on a desktop computer with a 2.7GHZ Intel Xeon Processor and 4GB memory, rounding to the first decimal place reduces computational time from about 3 minutes to only 22 seconds, which approaches the computation time of DSP (12 seconds). This analysis demonstrates that the simple rounding strategy can very effectively reduce computational time with negligible effects on the utility of secure datasets.

6. Discussion

Modern evidence-based research and practice in many areas, such as database marketing, strategy, business census, and public health, depend heavily on the construction, sharing, and dissemination of effective databases. There are rapidly increasing demands for secure database construction methods that can protect sensitive information while maintaining the high utility of databases shared with third parties. In this article we have developed new data perturbation and shuffling procedures for building secure databases more effectively. As compared with existing ones, our procedures substantially enhance the utility of secure databases while maintaining low disclosure risk. The proposed methods have a number of features particularly attractive for general-purpose data masking. The modeling, estimation, and generation of masked sensitive values in MORE do not rely on the distributional properties of any variables in the dataset, and are applicable to mixed continuous, semicontinuous, and discrete confidential variables. An important benefit of MORE is that the data masking models nest GLMs, a large family of statistical models widely used in business, economics, social science, and health database analyses, as special cases. Therefore, a much wider range of statistical analysis can be conducted on the resulting secure databases with results similar to those using the original ones. Furthermore, MORE is capable of preserving nonmonotonic relationships among attributes, addressing an important issue unresolved from prior research in model-based data perturbation and shuffling methods. MORE directly models the conditional distribution of those confidential data elements, and therefore can further increase the robustness of secure databases.

[10] Because the results for MORE-S1, MORE-S2,and MORE-S3 with different levels of rounding are also very similar, we omit results when rounding to the first decimal place in Table 8.

These benefits of the proposed methods have rich managerial and policy implications. As shown in our empirical application regarding an inverted-U relationship between competition and innovation, the flexibility of the odds ratio function allows one to reproduce this nonmonotonic relationship. Because of the robust nonparametric distribution modeling ability, MORE is capable of preserving the distributional characteristics of confidential variables and can automatically account for important data features, such as bounded support, discreteness, outliers, skewness, and heavy tails. As a result of these benefits, MORE preserves the inverted-U relationship between competition and innovation while the existing DPS methods create masked datasets that lead to the conclusion of no relationship between competition and innovation, a difference with huge managerial and policy implications. As indicated in Aghion et al. (2005), such an inverted-U relationship provides crucial insights into the impact of competition and closeness in technology space on innovation. For market regulators, such effects would be critical for antitrust and competition policy applications and potential reforms. For the manager of a firm who plans to enter a market, the inverted-U relationship can be very important to predict the responses of incumbent firms in the market and to formulate entry strategies. As demonstrated in our application, secure databases using different methods can lead to different estimated competition-innovation relationships and substantively different or even opposite predictions of incumbent firms' response behavior after entry. In the application, MORE substantially enhances the utility of secure databases and provides opportunities to make optimal policy and managerial decisions for the users of these databases.

Appendix. Illustration of the MORE Methods in a Small Dataset

We illustrate the steps in applying MORE procedures for data masking using a small simulated dataset that has one nonconfidential variable S and one confidential variable X. For ease of illustration, a random sample of 10 observations is generated from a bivariate normal distribution with mean zero and a variance-covariance matrix in which the diagonal elements are one and off-diagonal elements are 0.5. The purpose is to use this small dataset to demonstrate numerically the steps involved in applying MORE procedures.

- **A.1**: Estimate the odds ratio model parameters $\theta = (\lambda, \gamma)$ in $f_\theta(X|S)$. The reference point (X_0, S_0) in the odds ratio model is chosen to be the mean of (X, S). In this dataset, the unique values of X are

$$\left[-1.306\ -0.415\ -0.196\ 0.055\ 0.160\ 0.314\ 0.457\ 0.778\ 1.226\ 1.617\right]$$

In our model estimation, we have the estimates of parameters in $f_\lambda(X|S_0)$ assigned to this set of unique values as

$$\hat{\lambda} = \left[-0.077\ 0.165\ 0.199\ 0.224\ 0.230\ 0.233\ 0.231\ 0.205\ 0.120\ 0.000\right],$$

and the estimate of the parameter γ in the log odds ratio function: $\ln \eta(X, X_0, S, S_0) = \gamma(X - X_0)(S - S_0)$ is $\hat{\gamma} = 0.954$.

- **A.2**: Evaluate the security level of the current data masking setting. The EMPD is computed to be 0.821, which is larger than 50% of the standard deviation of the sensitive variable X. The EMPD value is useful for database owners to determine whether the current data masking setting has sufficient control of disclosure risk by comparing this EMPD value with a prechosen cutoff value. Our example assumes that this EMPD value is larger than the cutoff value, indicating the current data masking setting has sufficient control for disclosure risk. If EMPD is less than this cutoff value indicating insufficient disclosure control, the data masking setting needs to be revised. For example, database owners may consider masking the nonconfidential attribute S, too, which will increase the EMPD value and further reduce the disclosure risk.
- **A.3**: Generate the perturbed values of the confidential variable X. For the first observation, [S, X] = [-0.787, -1.306], we generate the masked value of X from a multinomial distribution on the set of the values shown in A.1, and the probability vector in the multinomial distribution is computed as

$$P_1 = \left[0.201\ 0.141\ 0.126\ 0.109\ 0.103\ 0.093\ 0.084\ 0.066\ 0.045\ 0.031\right].$$

The random number generated from the above multinomial distribution represents the perturbation value for the sensitive attribute X of the first observation, which is -0.196, and is used to replace the original value of -1.306. The random number generation is then repeated for every observation, leading to the final perturbed dataset.

- **A.4**: For MORE-S, we perform the data shuffling step. For the first observation, we replace its perturbed X value with the value of X in the original dataset that has the same rank as that of this perturbed value in the perturbed dataset. This shuffling step replaces the perturbed X value (-0.196) for the first observation with the shuffled X value of 0.055. The shuffling operation is repeated for every observation, leading to the final shuffled dataset.

References

Aghion, P., Bloom, N., Blundell, R., Griffith, R., and Howitt, P. (2005) "Competition and Innovation: An Inverted U Relationship," *Quarterly Journal of Economics,* **120**, 701–728.

Bertino, E., Byun, J., and Li, N. (2005) *Privacy-Preserving Database Systems* in Foundations of Security Analysis and Design III, Lecture Notes in Computer Science, Springer-Verlag, Berlin.

Blattberg, R.C., Kim, B. and Neslin, S.A. (2008) *Database Marketing: Analyzing and Managing Customers,* Spring, New York.

Chen, H. Y. (2004) "Nonparametric and Semiparametric Models for Missing Covariates in Parametric Regression," *Journal of the American Statistical Association,* **99**, 1176–1189.

Chen, H. Y. (2007) "A Semiparametric Odds Ratio Model for Measuring Association," *Biometrics,* **63**, 413–421.

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

Desai, P. (2013) "Marketing Science Replication and Disclosure Policy," *Marketing Science,* **32**, 1-3.

Genest, C., and Neslehova, J. (2007) "A Primer on Copulas for Count Data," *Astin Bulletin,* **37(2)**, 475–515.

Goldfarb, A., and Tucker, C. (2012) "Privacy and Innovation," *Innovation Policy and the Economy,* **Vol. 11**, Josh Lerner and Scott Stern (Eds), NBER.

Kalvenes, J., and Basu, A. (2006), "Design of Robust Business-to-Business Electronic Marketplaces with Guaranteed Privacy," *Management Science*, **52**:1721-1736.

Kim, G., Silvapulle, M.J., and Silvapulle, P. (2007), "Comparison of Semiparametric and Parametric Methods for Estimating Copulas," *Computational Statistics & Data Analysis*, **51**:2836-2850.

Lee, S., Genton, M.G., and Arellano-Valle, R.B. (2010), "Perturbation of Numerical Confidential Data Via Skew-t Distributions," *Management Science*, **56**:318-333.

Li, X., and Sarkar, S. (2011), "Protecting Privacy Against Record Linkage Disclosure: A Bounded Swapping Approach for Numeric Data," *Information Systems Research*, **22**: 774-789.

McCullagh, P., and Nelder, J.A. (1989) *Generalized Linear Models,* 2nd Ed., Chapman & Hall/CRC, BR.

Mela, C. (2011) "Data Selection and Procurement," *Marketing Science,* **30**: 965-976.

Menon, S., and Sarkar, S. (2007), "Minimizing Information Loss and Preserving Privacy," *Management Science*, **56**:318-333.

Miller, A., and Tucker, C. (2011), "Encryption and the Loss of Patient Data," *Journal of Policy Analysis and Management*, **30**:534-556.

Muralidhar, K., Batra, D., and Kirs, P.J. (1995) "Accessibility, Security, and Accuracy in Statistical Database: The Case for the Multiplicative Fixed Data Perturbation Approach," *Management Science,* **41(9)**, 1549-1564.

Muralidhar, K., Parsa, R., Sarathy, R. (1999) "A General Additive Data Perturbation Method for Database Security," *Management Science,* **45(10)**, 1399-1415.

Muralidhar, K., Sarathy, R., and Parsa, R. (2001) "An Improved Security Requirement for Data Perturbation with Implications for E-Commerce," *Decision Sciences,* **32**, 683-698.

Muralidhar, K., and Sarathy, R. (2003) "A Theoretical Basis for Perturbation Methods," *Statistics and Computing,* **13**, 329-335.

Muralidhar, K., and Sarathy, R. (2006) "Data Shuffling – A New Masking Approach for Numerical Data," *Management Science,* **52(5)**, 658-670.

Nocedal , J., and Wright, S.J. (1999) "Numerical Optimization," Springer Series in Operation Research, NY.

Qian, Y (2007), "Do National Patent Laws Stimulate Domestic Innovation in a Global Patenting Environment? A Cross-Country Analysis of Pharmaceutical Patent Protection, 1978-2002," *Review of Economics and Statistics,* **89(3)**, 436–453.

Qian, Y and Xie, H (2011), "No Customer Left Behind: A Distribution-Free Bayesian Approach to Account for Missing Xs in Marketing Models," *Marketing Science,* **30(4)**, 717–736.

Qian, Y and Xie, H (2013), "Which Brand Purchasers Are Lost to Counterfeits?," *Marketing Science,* Forthcoming.

Qu, L., Qian, Y. and Xie, H. (2009). "Copula Density Estimation by Total Variation Penalized Likelihood," *Communications in Statistics: Simulation and Computation*, **38**:9, 1891-1908.

Raghunathan, T. E., Reiter, J. P., and Rubin, D. B. (2003) "Multiple imputation for statistical disclosure limitation, " *Journal of Official Statistics,* **19**, 1–16.

Reiter, J. P. (2005) "Releasing multiply-imputed, synthetic public use microdata: An illustration and empirical study, " *Journal of the Royal Statistical Society, Series A,* **168**, 185–205.

Reiter, J. P., and Raghunathan, T. E. (2007) "The multiple adaptations of multiple imputation," *Journal of the American Statistical Association,* **102**,1462–1471.

Rubin, D.B. (1993) "Discussion: Statistical Disclosure Limitation," *Journal of Official Statistics,* **9**, 461–468.

Sarathy, R., Muralidhar, K., Parsa, R. (2002) "Perturbing Nonnormal Confidential Attributes: The Copula Approach." *Management Science,* **48(12)**, 1613-1627.

Scheuer, E.M. and Stoller, D.S. (1962) "On the generation of normal random vectors." *Technometrics*, **4**, 278-281.

Wall Street Journal. (2001) "Bureau Blurs Data to Keep Names Confidential," February, B1–B2.

Willenborg, L., and de Waal, T. (2001) *Elements of Statistical Disclosure Control,* Springer, New York.

Winer, R.S. (2001) "A Framework for Customer Relationship Management," *California Management Review,* **43(4)**, 89–105.

Xiao, X., and Tao, Y. (2006) "Personalized Privacy Preservation." *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 229-240.

Table 1 Simulation Study Results on the Performance of Perturbation Methods to Preserve the Marginal Distributions of the Confidential Variables.

		X_1		X_2		X_3	
Criteria	N	C-GADP	MORE-P	C-GADP	MORE-P	C-GADP	MORE-P
KS	100	0.0807 (0.0229)	0.0734 (0.0222)	0.0845 (0.0236)	0.0702 (0.0236)	0.2293 (0.0405)	0.0475 (0.0205)
	500	0.0374 (0.0104)	0.0349 (0.0106)	0.0395 (0.0119)	0.0332 (0.0096)	0.2241 (0.0175)	0.0215 (0.0090)
	1000	0.0269 (0.0075)	0.0251 (0.0072)	0.0281 (0.0083)	0.0237 (0.0074)	0.2240 (0.0128)	0.0160 (0.0065)
Mean	100	0.0044 (0.0765)	-0.0031 (0.0755)	-0.0140 (0.1561)	0.0002 (0.0966)	0.3569 (0.1725)	-0.0008 (0.1215)
	500	0.0033 (0.0366)	0.0001 (0.0386)	0.0014 (0.0736)	0.0034 (0.0484)	0.3892 (0.0700)	0.0091 (0.0577)
	1000	0.0050 (0.0247)	0.0002 (0.0269)	0.0053 (0.0535)	0.0087 (0.0341)	0.4036 (0.0490)	0.0121 (0.0433)
Var	100	-0.0765 (0.1160)	-0.0043 (0.1362)	-0.2742 (0.7066)	0.0160 (0.2915)	-2.8628 (4.2329)	0.0564 (1.3124)
	500	-0.0217 (0.0571)	-0.0003 (0.0590)	-0.0691 (0.3414)	0.0208 (0.1679)	-2.8060(2.2193)	0.1472 (0.7497)
	1000	-0.0109 (0.0403)	0.0043 (0.0425)	-0.0414 (0.2525)	0.0683 (0.1360)	-2.6266 (1.2602)	0.2528 (0.7049)
Skewness	100	-0.0053 (0.2258)	0.0095 (0.2074)	-0.1218 (0.4899)	0.0030 (0.2239)	-1.1302 (1.1871)	-0.0435 (0.3991)
	500	0.0048 (0.1058)	-0.0136 (0.1071)	-0.0577 (0.3453)	0.0010 (0.1505)	-1.6904 (1.7420)	-0.0386 (0.3646)
	1000	-0.0010 (0.0709)	-0.0048 (0.0722)	-0.0209 (0.2755)	0.0148 (0.1128)	-1.9032 (1.8016)	0.0516 (0.4020)
Kurtosis	100	-0.0141 (0.4156)	-0.0002 (0.4202)	-0.5334 (2.7035)	0.0031 (0.9473)	-6.7014 (10.8616)	-0.4007 (3.4543)
	500	-0.0190 (0.2041)	-0.0003 (0.2029)	-0.4323 (2.9978)	-0.0291 (0.9852)	-17.6801 (33.3058)	-0.7367 (5.4429)
	1000	-0.0190 (0.1499)	-0.0079 (0.1501)	-0.1797 (2.3016)	-0.0235 (0.7853)	-23.5202 (43.1230)	0.0634 (7.3234)
Q05	100	0.0633 (0.2084)	0.0024 (0.2216)	0.0792 (0.2510)	-0.0055 (0.2537)	0.0019 (0.0425)	0.0000 (0.0000)
	500	0.0192 (0.0931)	-0.0088 (0.1018)	0.0222 (0.1068)	0.0056 (0.1074)	0.0000 (0.0000)	0.0000 (0.0000)
	1000	0.0119 (0.0631)	-0.0080 (0.0659)	0.0196 (0.0738)	0.0022 (0.0753)	0.0000 (0.0000)	0.0000 (0.0000)
Q25	100	0.0256 (0.1217)	0.0009 (0.1251)	0.0292 (0.1645)	-0.0055 (0.1660)	0.9925 (0.0842)	0.0095 (0.1063)
	500	0.0077 (0.0561)	0.0014 (0.0577)	0.0098 (0.0732)	0.0000 (0.0717)	1.0000 (0.0000)	0.0000 (0.0000)
	1000	0.0085 (0.0387)	-0.0033 (0.0381)	0.0118 (0.0548)	-0.0018 (0.0510)	1.0000 (0.0000)	0.0000 (0.0000)
Q50	100	0.0086 (0.1134)	-0.0078 (0.1040)	0.0057 (0.1578)	-0.0003 (0.1465)	0.6250 (0.4778)	-0.0170 (0.2175)
	500	0.0036 (0.0517)	0.0015 (0.0545)	0.0067 (0.0788)	0.0015 (0.0639)	0.8180 (0.3784)	0.0000 (0.0000)
	1000	0.0035 (0.0369)	0.0008 (0.0382)	0.0072 (0.0533)	-0.0004 (0.0475)	0.9200 (0.2678)	0.0000 (0.0000)
Q75	100	-0.0229 (0.1207	-0.0081 (0.1269)	-0.0299 (0.2300)	-0.0056 (0.1710)	0.5325 (0.5144)	0.0235 (0.3989)
	500	-0.0011 (0.0580)	-0.0010 (0.0612)	-0.0032 (0.1150)	0.0017 (0.0836)	0.8910 (0.3071)	0.0395 (0.2354)
	1000	0.0034 (0.0384)	0.0017 (0.0425)	0.0037 (0.0759)	0.0094 (0.0579)	0.9810 (0.1339)	0.0160 (0.1195)
Q95	100	-0.0778 (0.2157)	0.0004 (0.2079)	-0.1500 (0.7898)	0.0278 (0.4471)	-0.5749 (0.9834)	0.0505 (1.0712)
	500	-0.0133 (0.963)	-0.0052 (0.0881)	-0.0398 (0.3470)	0.0284 (0.2232)	-0.4443 (0.5724)	0.0701 (0.5464)
	1000	-0.0031 (0.0662)	0.0017 (0.0651)	-0.0196 (0.2267)	0.0608 (0.1566)	-0.4380 (0.5086)	0.0323 (0.4550)
EMPD	100	0.8726(0.0884)	0.8801 (0.0883)	1.7414 (0.2264)	1.0762 (0.1099)	1.4071 (0.2133)	1.0833 (0.1645)
	500	0.9125 (0.0415)	0.9160 (0.0383)	1.7953 (0.1030)	1.1336 (0.0505)	1.4298 (0.0970)	1.1670 (0.0756)
	1000	0.9147 (0.0274)	0.9191 (0.0272)	1.7999 (0.0687)	1.1529 (0.0384)	1.4327 (0.0657)	1.1839 (0.0512)

Table 2 Simulation Study Results on the Performance of MORE to Preserve Relationships Among Attributes.

Criteria	N	MORE-P	MORE-S
$\rho_{31} = 0.5$	100	0.0040 (0.0654)	0.0002 (0.0660)
	500	-0.0017 (0.0318)	-0.0027 (0.0318)
	1000	-0.0064 (0.0220)	-0.0067 (0.0220)
$\rho_{32} = 0.5$	100	-0.0015 (0.0681)	-0.0057 (0.0697)
	500	-0.0016 (0.0317)	-0.0024 (0.0318)
	1000	-0.0062 (0.0220)	-0.0066 (0.0220)
$\beta_{41} = 0$	100	0.0014 (0.1031)	0.0046 (0.1024)
	500	-0.0016 (0.0518)	-0.0027 (0.0537)
	1000	-0.0003 (0.0384)	-0.0006 (0.0385)
$\beta_{42} = 1$	100	-0.0094 (0.0767)	-0.0098 (0.0476)
	500	-0.0253 (0.0443)	-0.0294 (0.0305)
	1000	-0.0368 (0.0329)	-0.0489 (0.0256)
$\beta_{51} = 1$	100	0.0001 (0.0872)	-0.0207 (0.0523)
	500	-0.0038 (0.0381)	-0.0146 (0.0253)
	1000	-0.0090 (0.0285)	-0.0200 (0.0233)

Table 3 Value Disclosure Risk.

N	ρ	DSP		MORE-P		MORE-S	
		Proportion of Variability Explained					
		$X_1\|Y_1$	$X_2\|Y_2$	$X_1\|Y_1$	$X_2\|Y_2$	$X_1\|Y_1$	$X_2\|Y_2$
100	0.0	0.000116	0.000111	0.000091	0.000100	0.000092	0.000102
	0.2	0.000101	0.000113	0.000109	0.000110	0.000107	0.000111
	0.4	0.000097	0.000093	0.000105	0.000113	0.000106	0.000113
	0.6	0.000109	0.000103	0.000096	0.000095	0.000096	0.000094
	0.8	0.000099	0.000094	0.000111	0.000097	0.000113	0.000096
	0.95	0.000088	0.000083	0.000094	0.000103	0.000093	0.000103
500	0.0	0.000004	0.000003	0.000004	0.000003	0.000004	0.000003
	0.2	0.000004	0.000004	0.000003	0.000004	0.000004	0.000004
	0.4	0.000004	0.000004	0.000004	0.000005	0.000004	0.000005
	0.6	0.000004	0.000004	0.000004	0.000004	0.000004	0.000004
	0.8	0.000004	0.000004	0.000004	0.000004	0.000004	0.000004
	0.95	0.000006	0.000006	0.000004	0.000004	0.000004	0.000005
1000	0.0	0.000001	0.000001	0.000001	0.000001	0.000001	0.000001
	0.2	0.000001	0.000001	0.000001	0.000001	0.000001	0.000001
	0.4	0.000001	0.000001	0.000001	0.000001	0.000001	0.000001
	0.6	0.000001	0.000001	0.000001	0.000001	0.000001	0.000001
	0.8	0.000001	0.000002	0.000001	0.000001	0.000001	0.000001
	0.95	0.000001	0.000001	0.000001	0.000001	0.000001	0.000001

Table 4 Marginal Distributions of *ptcw* with Original and Perturbed Values When $S = (Ci)$.

Criteria	Original	GADP	C-GADP	MORE-P	MI
KS		0.1836 ($<$ .0001)	0.1045 (0.0418)	0.0282 (0.9989)	0.4375 (0.0000)
Mean	6.6554	7.3322	6.7049	6.9243	6.7513
Var	71.1421	48.0904	61.3625	80.2012	7.0243
Skewness	1.6354	0.7864	1.5200	1.7185	0.3809
Kurtosis	5.2310	2.9131	4.6862	5.5888	3.1173
Q05	0.0000	0.0000	0.0000	0.0000	2.7450
Q25	0.4453	0.1282	1.0789	0.4551	4.9425
Q50	3.3498	6.0219	3.8915	3.2436	6.5350
Q75	9.3848	11.9146	9.5400	9.6585	8.2450
Q95	25.0600	20.4321	23.5460	26.8393	11.2510
EMPD	—	8.34	8.05	6.45	6.70

Table 5 Estimates of Relationship Between Innovation and Competition When $S = (Ci)$.

Criteria	Original	GADP	C-GADP	DSP	MORE-P	MORE-S	MI
Competition (Ci)	165.12**	-64.36	33.29	41.06	199.70**	175.29**	169.53**
	(54.77)	(47.77)	(51.44)	(55.09)	(52.11)	(52.65)	(70.08)
Competition^2 (Ci^2)	-88.55**	34.51	-18.28	-19.36	-109.53**	-96.43**	-90.94**
	(29.08)	(25.36)	(27.32)	(29.16)	(27.73)	(28.02)	(37.09)

Table 6 **Marginal Distributions of $ptcw$ with Original and Masked Values When $S = (Industry, Year, Ci)$.**

Criteria	Original	GADP	C-GADP	MORE-P	MI
KS		0.1186 (0.0137)	0.1017 (0.0514)	0.0254 (0.9998)	0.1151 (0.0292)
Mean	6.6554	7.1618	6.5441	6.4748	6.7370
Var	71.1421	57.3740	62.9778	65.7977	65.0568
Skewness	1.6354	1.2019	1.5614	1.5507	1.5874
Kurtosis	5.2310	3.6744	4.6564	4.7391	5.7508
Q05	0.0000	0.0000	0.0000	0.0000	0.0000
Q25	0.4453	1.0212	0.6715	0.3764	0.9850
Q50	3.3498	4.8538	3.5306	3.4114	4.0650
Q75	9.3848	10.6395	9.3014	9.3850	10.4000
Q95	25.0600	22.4616	24.9842	24.4802	23.7090
EMPD	—	3.5459	3.2766	3.2263	3.1820

Table 7 **Estimates of Relationship Between Innovation and Competition When $S = (Industry, Year, Ci)$.**

Criteria	Original	GADP	C-GADP	DSP	MORE-P	MORE-S	MI
Competition (Ci)	387.46***	-67.42	6.54	-125.95*	411.50***	430.31***	395.05***
	(67.74)	(59.31)	(62.52)	(59.59)	(71.26)	(70.30)	(74.44)
Competition^2 (Ci^2)	-204.55***	38.25	-2.09	66.96*	-215.13***	-225.47***	-194.72***
	(36.17)	(31.71)	(33.44)	(31.89)	(38.01)	(37.49)	(39.41)

Table 8 **Rank Order Correlation**

Correlation	Original	DSP	MORE-S^a	MORE-S^b	MORE-S1^a	MORE-S2^a	MORE-S3^a
Home Value and							
Gender	-0.00373	-0.00701	-0.00402	-0.00313	-0.00657		
Marital Status	-0.00187	-0.00060	-0.00621	-0.00753	0.00307		
Age	0.57146	0.60600	0.57029	0.56050	0.56977		
Mortgage Balance	0.58229	0.59072	0.58574	0.58327	0.57843	0.57820	
Total Net Asset Value	0.68129	0.68126	0.68866	0.68745	0.67873		0.68076
Mortgage balance and							
Gender	-0.00409	-0.00491	-0.00574	-0.000895		-0.00758	
Marital Status	-0.00093	0.00001	-0.00469	-0.00743		-0.00243	
Age	0.28334	0.29953	0.28344	0.29004		0.27983	
Total Net Asset Value	0.78156	0.78341	0.77916	0.78088		0.77841	0.78046
Total Net Asset Value and							
Gender	-0.00510	-0.00696	-0.00737	-0.00987			-0.00658
Marital Status	0.07426	0.07639	0.07858	0.06200			0.07410
Age	0.37367	0.38865	0.37754	0.37890			0.37487

Note: a and b denote rounding to the third and first decimal places, respectively.

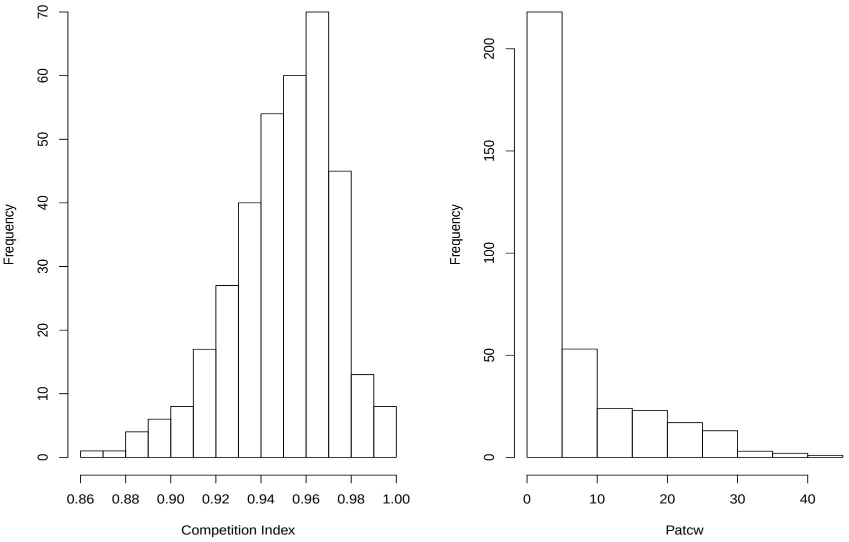

Figure 1 Histograms of Ci and $Patcw$ in the Original Dataset.

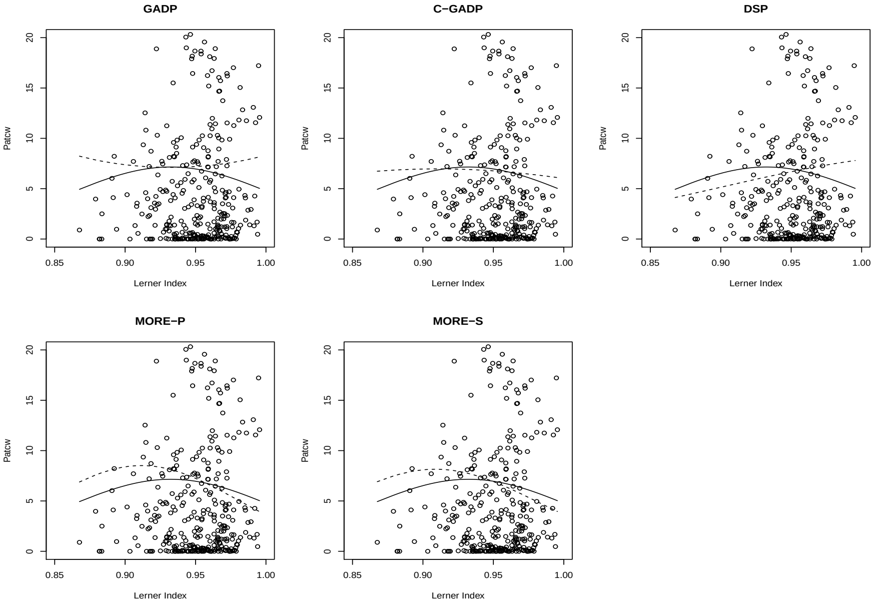

Figure 2 Comparison of the Relationships Between Competition and Innovation Estimated with the Original Dataset and Masked Datasets. The solid curve represents the inverted-U relationship estimated using the original dataset. The dotted curve represents the relationship estimated using the masked dataset generated by the data masking method given on the top of the plot. The points in the plots represent the original data values and only observations with $patcw < 20$ are plotted.

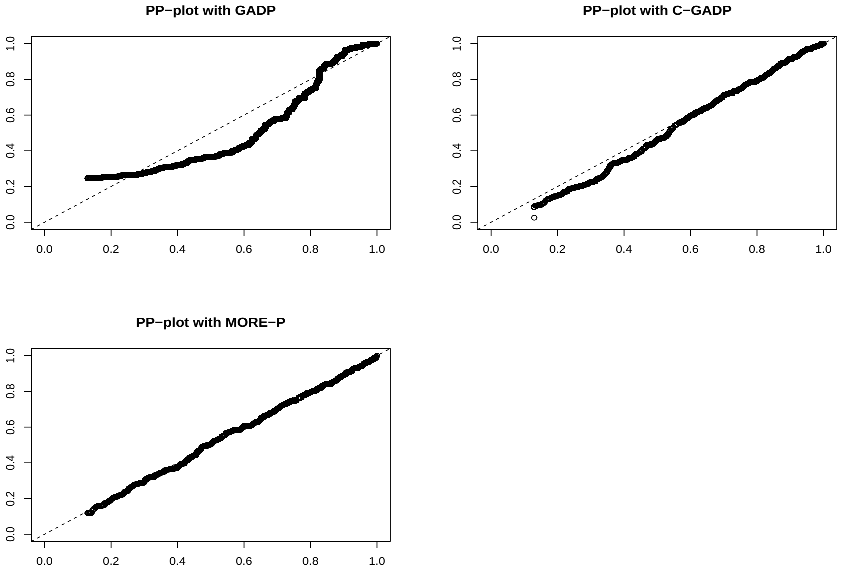

Figure 3 PP Plots for the Masked Values of $patcw$ for GADP, C-GADP, and MORE-P.