HHS Public Access
Author manuscript
MIS Q. Author manuscript; available in PMC 2016 January 06.

Published in final edited form as:
MIS Q. 2014 September ; 38(3): 679–698.

Digression and Value Concatenation to Enable Privacy-Preserving Regression

Xiao-Bai Li and
Department of Operations and Information Systems, Manning School of Business, University of Massachusetts Lowell, Lowell, MA 01854 U.S.A. {xiaobai_li@uml.edu}

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

Abstract

Regression techniques can be used not only for legitimate data analysis, but also to infer private information about individuals. In this paper, we demonstrate that regression trees, a popular data-analysis and data-mining technique, can be used to effectively reveal individuals' sensitive data. This problem, which we call a "regression attack," has not been addressed in the data privacy literature, and existing privacy-preserving techniques are not appropriate in coping with this problem. We propose a new approach to counter regression attacks. To protect against privacy disclosure, our approach introduces a novel measure, called *digression*, which assesses the sensitive value disclosure risk in the process of building a regression tree model. Specifically, we develop an algorithm that uses the measure for pruning the tree to limit disclosure of sensitive data. We also propose a dynamic value-concatenation method for anonymizing data, which better preserves data utility than a user-defined generalization scheme commonly used in existing approaches. Our approach can be used for anonymizing both numeric and categorical data. An experimental study is conducted using real-world financial, economic and healthcare data. The results of the experiments demonstrate that the proposed approach is very effective in protecting data privacy while preserving data quality for research and analysis.

Keywords

Privacy; data analytics; data mining; regression; regression trees; anonymization

INTRODUCTION

Predictive analytics techniques using personal data have been deployed by organizations in a variety of domains, including marketing research, financial analysis, human behavior study, and healthcare research. While these techniques are generally used by organizations to better understand and serve their customers, there are growing concerns about invasions to privacy from the use of these techniques.

In a widely-circulated article, Charles Duhigg, a *New York Times* reporter, wrote how Target Corporation used predictive analytics to conduct targeted marketing (Duhigg 2012). Perhaps the most intriguing story in the article is how Target identified pregnant customers. The

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

analytics team started with the shopping records of the Target's baby gift registry, where the customers had voluntarily disclosed their due dates. The team discovered a set of products, such as unscented lotion and soap, and calcium and zinc supplements, that pregnant women bought in large quantities during different periods of their pregnancy. These items enabled Target to calculate a "pregnancy prediction" score and estimate the due date for other customers with similar purchase behaviors, and to send coupons timed to specific stages of the pregnancy. The model worked very well, perhaps too well in that Target seemed to know things that even close family members of a targeted woman did not know. In one instance, a father whose teenage daughter was receiving coupons for baby products walked into a Target store and complained: "She's still in high school, and you're sending her coupons for baby clothes and cribs? Are you trying to encourage her to get pregnant?" It turns out the man later apologized to the store manager that he was not aware of his daughter's pregnancy. This story took the media by storm, with more than one million views on the Internet within days (KDnuggets 2012). The reactions from the public were mostly negative. As a privacy expert put it: "This is the exciting possibility of Big Data, but for privacy, it is a recipe for disaster" (Ohm 2012).

In this case, Target was using its in-house data for analysis. When the personal data is shared with a third party, privacy concerns become even more serious. However, sharing and selling of personal data are common today. As an example, the Center for Medicare and Medicaid Services, a federal agency, sells individual Medicare and Medicaid claims data to third parties for analysis (http://www.resdac.org/). The center's operations follow the guidelines of the Health Insurance Portability and Accountability Act (HIPAA). However, studies have shown that the HIPAA rules may be insufficient in protecting patient privacy (Sweeney 2002). In fact, secondary use of private data has long been a cause for serious concern, and studies have found the majority of the public react negatively to their use (Culnan 1993; Angst and Agarwal 2009).

This research concerns regression, which is one of the most widely used predictive techniques in business environments. More specifically, our research investigates a privacy disclosure problem involving the use of a popular regression technique called *regression trees*. Introduced by Breiman et al. (1984), regression trees build prediction models based on recursive partitioning of data. In contrast to the classic linear regression model, regression trees are nonparametric in nature and thus very effective in dealing with nonlinear and non-monotonic relationships in data. They can easily handle both numeric and categorical predictor variables. The regression tree models can be converted into a set of rules that are easy to understand and interpret. These desirable features have led to their wide use in predictive data mining and analysis. An excellent example is the regression tree diagram published by *The New York Times* during the 2008 democratic primary election (Cox 2008). The regression tree used a set of demographic, geographic, economic and political variables to predict the number of votes (counties) that Barack Obama and Hillary Clinton would win. The tree diagram not only showed prediction outcomes, but also clearly described the decision rule leading to each outcome.

Regression trees, however, can also be used as a tool to effectively reveal private information about individuals. We call this use of regression trees for "mining" personal

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

information a *regression attack*. To understand such situations, we first distinguish the attributes of data on individuals from a privacy perspective. Typically, the attributes can be classified into three categories (Machanavajjhala et al. 2006; Li et al. 2007; LeFevre et al. 2008): (1) *explicit identifiers*, which can be used to directly identify an individual, including name, social security number, and phone number; (2) *sensitive attributes*, which contain private information that an individual typically does not want revealed, such as income, medical test results, and sexual orientation; and (3) *non-sensitive attributes*, such as age, gender, education, and occupation; the values of such attributes can often be obtained from public sources. Some non-sensitive attributes can be used to identify individuals by matching data from different sources, resulting in identity disclosure. Such attributes are collectively called a *quasi-identifier* (QI) in the literature. For example, Sweeney (2002) found that 87% of the population in the United States can be uniquely identified with three attributes – gender, date of birth, and 5-digit zip code – which are accessible from voter registration records available to the public. In data privacy research and practice, the explicit identifiers are typically removed from the data (a process referred to as *de-identification*). Data anonymization is applied to QI attributes to further prevent or limit the identity disclosure. With identity information properly protected, the sensitive attributes are typically released with their original values. For instance, this scheme of handling the three types of attributes is adopted by HIPAA. We follow the same scheme in this study.

In analyzing privacy disclosure risk, the literature recognizes two types of disclosure: *identity disclosure* (or *re-identification*) and *value disclosure* (Duncan and Lambert 1989; Lambert 1993). Re-identification occurs when a data intruder is able to match a record in a de-identified dataset to an actual individual. The finding that 87% of the US population can be uniquely identified by gender, date of birth and zip code is an example of re-identification. Value disclosure occurs when an intruder is able to predict the sensitive value(s) of an individual record, with or *without* knowing the identity of the individual. For example, suppose all new faculty members in a unionized college receive the same starting salary and the college releases the average salary of new faculty. Then the release discloses the salary of each new faculty member, even though the individuals are not identified. Thus, a technique that protects against identity disclosure does not necessarily prevent value disclosure.

The Target example discussed earlier is not an identity-disclosure problem because Target already had the explicit identifiers of the customers (due to their using Target credit cards or purchasing Target items online, etc.); so there was no re-identification issue involved. Instead, the problem is about value disclosure, i.e., predicting the status and timing of the pregnancy based on the information independent of the identities. Suppose a retail store does not have an in-house analytics team and hires a third-party consulting firm to perform a similar study. Even if the customer data provided by the store are anonymized, the consulting firm can still build a predictive model for value disclosure. Consequently, the group of customers whose demographic and purchase profiles fit the prediction model well will be subject to high disclosure risk. Once a customer is determined to belong to this group, the sensitive information of the customer is very likely to be compromised even when

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

it is not possible to identify the record in the group that represents this customer (i.e., re-identification is not feasible).

While regression analysis and regression trees are widely used in data mining and business analytics, the regression attack problem has not been addressed in the data privacy literature. As we elaborate later, existing privacy-preserving data-analysis and data-mining techniques are not appropriate for dealing with this problem; some of them could even make a regression attack easier. To fill this research gap, we propose a regression-tree-based approach that can be used by organizations to protect individuals' private information against regression attacks while preserving the utility of the released data for legitimate data analysis. We introduce a novel measure, called digression, to assess the value-disclosure risk in constructing regression trees for data partitioning – specifically, an algorithm is developed that uses the measure for pruning the tree to limit disclosure of sensitive data. The approach can be used with one or more numeric sensitive attributes, and can handle both numeric and categorical QI attributes.

To anonymize QI attribute values, a common practice is to generalize the QI values using a pre-defined generalization hierarchy (Samarati and Sweeney 1998; Aggarwal and Yu 2008). Using a pre-defined hierarchy for generalization, while effective in preventing re-identification, is quite inflexible and can lead to undesirable information loss. We propose a dynamic value-concatenation method, which merges categorical values based on the hierarchical structure of the regression tree itself. This method has the potential to significantly improve the utility of the anonymized data. Both the disclosure risk (digression) measure and associated tree-pruning algorithm, and the value-concatenation method, are new to the literature. The proposed approach, which we call MART (Multivariate Anonymization with Regression Trees), is computationally very efficient and is much faster than traditional *k*-anonymity algorithms. It is therefore well-suited for applications with large datasets.

The rest of the paper is organized as follows. In the next section, we provide a small example to illustrate the regression attack problem. We then discuss prior and current research related to the problem. Following that, we develop the regression-tree-based data partitioning technique and the dynamic value-concatenation method. We then describe a set of experiments conducted on real-world data to demonstrate the effectiveness of our approach. The final section elaborates the implications of this research and provides directions for future research.

AN ILLUSTRATIVE EXAMPLE

A regression attack can be accomplished by building a regression tree using the sensitive attributes as the response variables and the QI attributes as predictors. The regression tree can then be used to systematically reveal or infer the individuals' sensitive information based on the values of the QI attributes. This can be done even after the dataset is anonymized using well-known anonymization techniques.

Consider an example dataset containing information on 14 individuals, as shown in Table 1. There are two numeric QI attributes (Age and Years of Education (YearsEdu)), one

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

categorical QI attribute (Occupation, with four categories), and two numeric sensitive attributes (Income and Asset). Given this dataset, a privacy intruder can build a regression tree based on the methods of Breiman et al. (1984) and De'ath (2002), using the two sensitive attributes as the responses. The resulting tree is shown in Figure 1, where a leaf node (rectangle) represents a partitioned subset (the records included in the subset are listed and the ranges of the Income and Asset values for the node are shown below the node). A split criterion is specified along with the edge representing the split. With this tree, it is easy for the intruder to infer an individual's sensitive information from Table 1 even though the identity information is not included. For example, if the intruder knew that an individual with less than 15 years of education and age more than 43 years is included in the dataset, then the intruder will find that this record is located in node 4, which includes record numbers 3, 4 and 5. The ranges of the Income and Asset values for the records in this node are very narrow. Moreover, the intruder may further split this node into child nodes to get more specific Income and Asset values if the intruder has more specific information about the age, occupation or years of education for the target individuals.

Regression trees are not the only means for an intruder to snoop for sensitive information. For instance, the intruder can issue an ad-hoc query to directly search for any targets if the intruder knew a few attribute values of the targets. However, because regression trees partition the data based on the relationships between the QI and sensitive attributes, regression attacks can compromise the data privacy more systematically and intrusively in several aspects. With regression trees, sensitive values can often be revealed using only a small number of the QI attributes, whereas an ad-hoc query often requires more attributes, depending on the sequence in which the attributes are considered. For instance, the regression tree in Figure 1 uses two QI attributes (YearsEdu and Age) to predict the sensitive values of the first two records, but an ad-hoc query may involve all three QI attributes if the search starts with the Occupation attribute. Furthermore, a regression attack can simultaneously identify a large number of target individuals. Finally, a regression tree shows which targets' sensitive values can be determined more easily and which QI attributes are the critical attributes for disclosure. So, even if the intruder did not have enough information for positive disclosure, regression trees can help the intruder identify potential targets, or gather additional data on important QI attributes. For example, the regression tree in Figure 1 indicates that YearsEdu may be the most important attribute for finding the sensitive attribute values. In short, a regression attack can find structural information for privacy disclosure that an ad-hoc query cannot; it is a systematic, efficient and proactive technique for revealing private information.

Some of the existing anonymization approaches are vulnerable to regression attacks. We describe the problem here with a well-known technique called *k*-anonymity (Samarati and Sweeney 1998). A *k*-anonymity approach aims at anonymizing the values of the QI attributes such that the values of these attributes for any individual matches those of at least $k - 1$ other individuals in the same dataset. With *k*-anonymity, a dataset is partitioned into groups with at least *k* records in each group; the QI attribute values are then anonymized using the same generalized value within a group to make the records in the group indistinguishable. For a numeric attribute, *k*-anonymity replaces the original values in a

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

group with the group range. For a categorical attribute, it generalizes the values based on a user-defined hierarchy. Figure 2 shows a generalization hierarchy for the Occupation attribute in the data in Table 1 (the value of the root node can be any expression that covers all Occupation values; we use the symbol * following the convention in the *k*-anonymity literature).

When the data is intended for regression analysis, a regression tree technique appears to be a natural choice for dividing the data into groups for *k*-anonymity. LeFevre et al. (2008) propose a method called *Regression Mondrian* based on this idea. Table 2 shows the anonymized data using Regression Mondrian on the example data. When $k = 2$, the dataset is partitioned into six groups (separated by both dash-lines and solid-lines); when $k = 4$, it is partitioned into three groups (separated by solid-lines only). It can be observed from Table 2 that for many of the 2-anonymized groups, the sensitive Income and Asset values are very close within the groups. As a result, the intruder can still obtain the sensitive information fairly accurately for the individuals in these groups even though the intruder may not be able to positively identify the individuals (which is indicative of value disclosure). For example, if the intruder knew that an individual has less than 15 years of education and is older than 43 years, the intruder can still find the same sensitive information from the anonymized data as from the original data using a regression attack (this is because the attack based on the anonymized data makes the same partition of the data as in Table 2). Similar situations also occur for some of the 4-anonymized groups (e.g., the narrow Income range for the group with records numbered 10, 13, 8 and 14). Further, because a regression tree technique attempts to partition the data such that the values of a response attribute are close to each other within a group (to increase prediction accuracy), the use of regression trees for *k*-anonymity could actually make a regression attack easier.

As mentioned earlier, a limitation of *k*-anonymity relates to its use of user-defined generalization hierarchies for categorical attributes. In this example, if the Occupation attribute in a group contains ‘unskilled’ and any other value, the value will have to be replaced by the general symbol *, based on the pre-defined hierarchy in Figure 2. For instance, the original Occupation values for records #6 and #7 are ‘technical’ and ‘unskilled’, respectively. When $k = 2$, they are grouped together (Node 7 in Figure 1). The generalized value for ‘technical’ and ‘unskilled’ is the symbol * based on the hierarchy in Figure 2. This causes the utility of the released data to deteriorate significantly.

In general, the tradeoff between data utility and anonymity is associated with all privacy-preserving data-sharing techniques. Our approach addresses this issue directly and, as shown subsequently, leads to a better tradeoff than existing techniques.

RELATED WORK

Information privacy has been studied extensively from different perspectives in multiple disciplines (Smith et al. 2011). This work focuses on the analysis and design aspects of privacy-preserving technology (Aggarwal and Yu 2008; Garfinkel et al. 2007; Sweeney 2002). Figure 3 provides a conceptual view of the technology in the process of collecting, processing and utilizing data, with privacy-related activities highlighted. A central idea

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

behind all approaches along this line of technology is to process and alter the data such that, while the identifying and sensitive information for the individuals in the data are well protected, the utility of the data is reasonably preserved in the data released for research and analysis.

A significant development in the literature on data privacy is the *k*-anonymity framework, proposed by Samarati and Sweeney (1998). As described earlier, the *k*-anonymity approach uses generalization and suppression methods to alter the values of QI attributes such that the values of these attributes for any individual matches those of at least $k - 1$ other individuals. In this way, the re-identification risk for an individual is reduced. *K*-anonymity is a general-purpose technique for privacy-preserving data publishing. Its original framework is not designed to preserve the relationships between the sensitive attributes and the QI attributes. From a data utility perspective, therefore, it may not be effective when the anonymized data is used for predictive data mining and analysis.

Privacy issues have been studied extensively in the predictive data mining and data analysis area (e.g., Agrawal and Srikant 2000; Aggarwal and Yu 2008). A number of studies develop privacy-preserving data-mining approaches under the *k*-anonymity framework. For instance, a top-down refinement method for classification problems is proposed in Fung et al. (2007). A set of *k*-anonymity-based algorithms for various data-mining tasks, developed by Friedman et al. (2008), covers classification, clustering and association rule mining (but not regression).

Besides *k*-anonymity, Li and Sarkar (2009) investigate the problem of using classification trees for privacy disclosure and propose a method to protect against such a “classification attack.” The sensitive data considered in that study is categorical and the related approach is applicable to classification analysis only. This study, however, considers sensitive numeric data and the approach we propose is intended for regression application. In another paper, Li and Sarkar (2011) propose a multivariate partitioning method for anonymizing data, which can be used for regression analysis. That work focuses on protecting privacy against record linkage, an identity-disclosure problem. It does not consider the value-disclosure risk under regression attacks. Furthermore, the method proposed in Li and Sarkar (2011) assumes that all data attributes are of numeric type. Our proposed approach, however, allows non-sensitive attributes to be of type numeric, categorical or both. As such, it is conceptually more general, and more widely applicable for real-world scenarios.

For regression applications, LeFevre et al. (2008) and Fu et al. (2010) propose *k*-anonymity based approaches using regression trees (along with approaches using classification trees for classification applications). The method proposed by LeFevre et al. (2008) first builds a regression tree with the minimum leaf size *k*, and then applies generalization and suppression schemes to satisfy the *k*-anonymity requirement. The objective of the study by Fu et al. (2010) is to preserve the regression tree model while anonymizing data. Their study focuses on the conditions that result in the same tree structure when the original or anonymized data are used, and the computational procedure to satisfy these conditions. Neither of these two studies, however, has considered sensitive value disclosure that is vulnerable to a regression attack.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

The *k*-anonymity approach focuses on re-identification risk only and does not consider value-disclosure risk. It generalizes different but similar QI attribute values into the same value within a group. The new values produced by the generalization operation are still correct with respect to the generalized categories. The sensitive attribute values (which can be numeric or categorical) remain unchanged in *k*-anonymity. However, these values become more similar within a group. As a result, individuals in a group, who have the same generalized QI values, are subject to high risk of value disclosure.

To address this issue, a privacy principle called *l*-diversity has been proposed (Machanavajjhala et al. 2006). The *l*-diversity principle requires that a sensitive attribute should include at least *l* well-represented values in the *k*-anonymized data. For example, a typical instantiation of *l*-diversity requires that, for each group, at most 1/*l* of the records have the most frequent sensitive value. The notion of *l*-diversity, however, does not consider the overall distribution of the sensitive attribute. So, when the overall distribution is unbalanced, the *l*-diversity requirement may be difficult to satisfy. Furthermore, since the overall distribution is usually public information, the sensitive value disclosure risk can be high when the distribution of the *l*-diversified data deviates significantly from the overall distribution. To overcome these problems, another privacy principle called *t*-closeness has been proposed (Li et al. 2007). This principle requires that, for each group, the distance between the distributions of the sensitive attribute in the group and the overall distribution cannot be larger than a threshold value *t*.

The *l*-diversity and *t*-closeness approaches, however, focus on situations where sensitive attributes are categorical. The *l*-diversity measure is not appropriate for evaluating the disclosure risk of numeric values. For example, every record in the example dataset in Table 1 has a distinct Income or Asset value, so the anonymized data in Table 2 would satisfy any *l*-diversity requirement. However, it is clear that the sensitive value disclosure risks for most records are high even though the *l*-diversity requirement is satisfied. The *t*-closeness measure, although also designed mainly for categorical attributes, can deal with a single numeric attribute. However, it is not appropriate for multiple correlated numeric attributes because the measure is defined for each attribute independently. Furthermore, the *t*-closeness measure concerns value-disclosure risk only; it does not explicitly consider the prediction error issue. As a result, the anonymized data might not be suitable for regression analysis. In short, because regression attacks involve multiple numeric sensitive attributes and are tied to prediction tasks, the *l*-diversity and *t*-closeness approaches are not appropriate to counter regression attacks.

There has been considerable research in the area of statistical databases (SDB) on inference disclosure control (Denning and Schlörer 1983; Adam and Wortmann 1989). Inference disclosure is similar to regression attacks in that they both attempt to reveal sensitive values without requiring identity disclosure. However, an SDB is designed to provide summary statistics, not individual records, to the user. An SDB user cannot retrieve a complete dataset and is typically limited to a few types of queries to obtain aggregate statistics. Therefore, inference control in SDB focuses on query restriction and output perturbation to prevent or limit inference disclosure by queries. This study considers situations where a dataset containing individual records is released to a third party for regression and other predictive

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

analyses. Clearly, inference control methods such as query restriction and output restriction are not applicable to our problem.

Several studies in the area of privacy-preserving association rule mining refer to the use of association rules to infer sensitive information as an "inference" or "inference attack" problem (Verykios et al. 2004; Oliveira and Zaiane 2006; Menon and Sarkar 2007; Atzori et al. 2008). In a broad sense, such an inference attack resembles a regression attack because they both attempt to find sensitive relationships across attribute-values without requiring identity disclosure and in both cases the disclosure of sensitive information is not necessarily deterministic. However, those studies typically assume that some association rules discovered from the data are confidential to the organization that owns the data and need to be protected when the data is shared. The problem thus is about confidentiality of organizational knowledge rather than individual privacy. Further, association rule mining requires all attributes to be categorical; thus, techniques developed to deal with such inference problems are not applicable to regression problems, which concern predictions of numeric values.

Similar problems have also been discussed in some privacy studies on social network analysis and graph mining (Zheleva and Getoor 2009; Cormode et al. 2010; Heatherly et al. 2013). The problems studied in these works examine inferences that can be made without identity disclosure, e.g., the attributes of individuals or the existence of links across entities. As is the case for association rule mining, all confidential attributes/values considered are categorical; none of these studies involves prediction of numeric attributes with regression, and nor do they consider regression attacks.

In summary, the data privacy literature has not addressed the regression attack problem. Given the widespread use of regression techniques, it is important to develop an approach to counter such an attack.

MART: MULTIVARIATE ANONYMIZATION WITH REGRESSION TREES

The notion of regression trees was introduced by Breiman et al. (1984). Similar to classification trees (also known as decision trees), regression trees adopt a divide-and-conquer strategy to build prediction models. We call a regression tree with a single response (dependent) variable a *univariate regression tree* and one with multiple response variables a *multivariate regression tree*. Given the problem this study focuses on, it is natural to set the sensitive attributes as response variables and use the QI and other non-sensitive attributes as regression predictors.

Δ-Digression: A Value-Disclosure Risk Measure

A commonly used splitting criterion for growing regression trees is the sum of squared errors (*SSE*). Consider the single response attribute case. Let n_t be the number of records in node t. Let $y_i(t)$ ($i = 1, \ldots, n_t$) be the value of the response attribute in the ith record in node t, and $\overline{y}(t)$ be the mean of the response attribute values in node t. The univariate *SSE* at node t is defined as

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$e(t)=\sum_{i=1}^{n_t}[y_i(t)-\overline{y}(t)]^2. \quad (1)$$

When a node is split, the combined *SSE* for the child nodes is always smaller than the *SSE* for the parent node. Suppose node *t* is split into *m* child nodes, $t_1, \ldots, t_m$. The reduction in *SSE*, which is $e(t) - [e(t_1) + \cdots + e(t_m)]$, serves as a criterion to select the splitting attribute and splitting value. The algorithm searches over all possible trial-splits for each non-response attribute, and the trial-split that maximizes the reduction in *SSE* is selected to split the data. The process continues until a stopping criterion (e.g., the minimum leaf size) is met. This produces a complete regression tree.

There are limited studies of multivariate regression trees in the literature. The splitting criteria proposed in these studies are some multivariate versions of the *SSE*. We use a measure, based on De'ath (2002) and LeFevre et al. (2008), that directly extends the univariate *SSE* to the multivariate case. For a problem with *r* response attributes, let $\mathbf{y}_i(t) = [y_{i1}(t), \ldots, y_{ir}(t)]'$ be the values of the response attributes in the *i*th record in node *t*, and $\overline{\mathbf{y}}(t)$ be the mean vector of the response attributes in node *t*. All response values are normalized to the range [0, 1] to remove the impact of the varying scales in different response attributes. The multivariate *SSE* at node *t* is defined as

$$e(t)=\sum_{i=1}^{n_t}[\mathbf{y}_i(t)-\overline{\mathbf{y}}(t)]'[\mathbf{y}_i(t)-\overline{\mathbf{y}}(t)]. \quad (2)$$

With this measure, a multivariate regression tree can be built in a manner similar to a univariate regression tree. Multivariate regression trees attempt to minimize prediction errors for the multiple responses. This explains why each subset partitioned by the multivariate regression tree in Figure 1 contains data points that are close to each other in the Income and Asset values.

An important stage in constructing a regression tree is pruning. For a traditional regression tree, the purpose of pruning is to avoid the over-fitting problem. Therefore, the usual pruning method in regression trees aims at minimizing the prediction error. We consider, in our problem, both prediction error and disclosure risk while selecting nodes for pruning. Clearly, the sensitive value disclosure risk of a record at a node is high when the variation in the sensitive attribute values of the records at the node is low. Based on the *t*-closeness principle (Li et al. 2007), the risk is low when the conditional distributions (conditioned on the non-sensitive attributes) of the sensitive attributes at the node are close to the overall distributions of the sensitive attributes. The *t*-closeness principle assumes that the overall distributions are public information. In other words, when anonymized data is released, it is expected that the overall parameters, such as the means and covariances of the response attributes for the entire dataset, is the same as or close to the original parameter values. Indeed, in many cases, such original parameters are released with the data.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

To measure the disclosure risk in terms of the closeness between a conditional distribution and the overall distribution, we propose a measure, based on the *scatter matrix* of the response attributes. The scatter matrix, which is the covariance matrix multiplied by the sample size, includes sum of squared errors (or variance) and cross-product (or covariance) components. It is an important measure of variation in each attribute and of relationships between different attributes (we choose to use scatter matrix instead of the covariance matrix merely for convenience, because regression trees use *SSE* instead of variance for measuring errors and the risk-utility tradeoff measure we propose involves comparing the risk measure with *SSE*). A significant difference between the scatter matrix on the data at a node and the overall scatter matrix can reveal useful information about the data at the node. The measure below evaluates this "digression" of the scatter structure from the overall scatter matrix.

Definition 1—Let **S** be the scatter matrix of the response attributes on the entire dataset and S_{jk} be the (j,k) element of **S**. Let $\mathbf{S}(t)$ be the scatter matrix calculated on the subset data at node t and $s_{jk}(t)$ be its (j,k) element. Let $\mathbf{D}(t)$ be a scatter difference matrix with its (j,k) element being $d_{jk}(t) = S_{jk} - s_{jk}(t)$. The *node digression* in scatter, denoted as $\Delta(t)$, is defined as the determinant of $\mathbf{D}(t)$, i.e.,

$$\Delta(t)=|\mathbf{D}(t)|. \quad (3)$$

The determinant of a scatter matrix is a single number that captures the characteristics of both variance and covariance information in a scatter matrix (Johnson and Wichern 2002, p. 125). The node digression measures the amount of deviation between the variance-covariance structure on the subset at the node and that on the entire dataset (when there is only one attribute, the node digression simply measures the variance aspect of the deviation). A small digression indicates a small deviation from the overall distribution, which implies a low disclosure risk and thus is desirable. If there are no perfect correlations between response attributes (which is almost always the case in real-world data), then the node digression has the following property:

Lemma 1: *The node digression is always a positive number; i.e.,*

$$\Delta(t)=|\mathbf{D}(t)|>0, \forall t. \quad (4)$$

The proofs of this lemma and all other mathematical properties are provided in the Appendix. Since the node digression is meant to measure the deviation of the covariance matrix on the subset at the node from that on the entire dataset, it is not meaningful for the measure to be negative (in a sense similar to the notion of variance or standard deviation). Lemma 1 justifies the node digression measure from this aspect. The result also enables us to define a digression measure for a group of nodes, and compare it with data quality measures.

When a node is split, the response values in its child nodes typically become closer to each other. Therefore, the parent node digression should be smaller than the digression of a child node. To formally describe this property, we first define some terms.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Definition 2—A *branch* B_t is a subsection of a tree that starts at an internal node, *t*, and includes all of its leaf or non-leaf descendant nodes.

In Figure 1, branch B_5 consists of nodes 5 (the root of B_5), 6, 7, 8, 9, 10, and 11.

Definition 3—Let B_t be a branch having *m* leaves ($\ell = 1, \ldots, m$). The *branch digression* of B_t is defined as the sum of its leaf node digressions, i.e.,

$$\Delta(B_t)=\sum_{\ell=1}^{m}\Delta(\ell). \quad (5)$$

We will use the term Δ-*digression* to generally refer to both the node digression and branch digression. The branch digression has the following property with respect to the node digression.

Lemma 2: *The node digression for a leaf* ℓ *is always greater than that for its parent node t. Hence, the branch digression for* B_t *is always greater than the node digression for t; that is,*

$$\Delta(\ell)>\Delta(t), \forall \ell, t \Rightarrow \Delta(B_t)>\Delta(t), \forall t. \quad (6)$$

Lemma 2 states that a split of a node always increases digression. In other words, Δ-digression increases monotonically in the depth of the node (with respect to its ancestor nodes). So, pruning a branch into a leaf always reduces digression.

Next, we define the error for a node *t* and a branch B_t. In fact, the *node error* $e(t)$ is simply the *SSE* of node *t* as defined in Equations (1) and (2).

Definition 4—The *branch error* $e(B_t)$ is defined as the sum of its leaf node errors:

$$e(B_t)=\sum_{\ell=1}^{m}e(\ell). \quad (7)$$

It is well known that a split always reduces errors, i.e., $e(B_t) < e(t)$ (Breiman et al. 1984). To assess the tradeoff between disclosure risk and regression error due to a split, we propose the following measure:

Definition 5—The *error-digression* measure for an internal node *t* is defined as:

$$q_t=\frac{e(t)-e(B_t)}{\Delta(B_t)-\Delta(t)}. \quad (8)$$

We describe next how this criterion is used in the proposed pruning algorithm.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Error-Digression Pruning

During the pruning process, we want the increase in error to be as small as possible to preserve prediction accuracy; at the same time, we want the decrease in digression as large as possible (which implies that the scatter matrix at the leaf node after pruning is as close to the overall scatter matrix as possible) to reduce disclosure risk. So, to achieve the best tradeoff between error and digression, the branch having the smallest q_t value should be pruned first.

The proposed pruning algorithm is recursive in nature. At each iteration, it calculates the value of q_t for each branch in the current tree. The branch that has the smallest value of q_t is pruned. The process continues until some pre-specified stopping criterion is satisfied. An obvious choice of a stopping criterion is the minimum number of records in a leaf. As mentioned earlier, however, this parameter, like the *k* parameter in *k*-anonymity, only measures re-identification risk. To measure the probability of sensitive value disclosure risk, we propose using a measure for testing the equality of two covariance matrices, based on the likelihood ratio test statistic (Morrison 1990, p.292), as shown below:

$$L_t = n_t \left(\log \left| \tilde{\mathbf{\Sigma}} \right| - \log \left| \tilde{\mathbf{\Sigma}}_t \right| + \text{trace}(\tilde{\mathbf{\Sigma}}_t \tilde{\mathbf{\Sigma}}^{-1}) - r \right), \quad (9)$$

where $\tilde{\mathbf{\Sigma}}$ and $\tilde{\mathbf{\Sigma}}_t$ are the sample covariance matrices for the entire dataset and the subset at node *t*, respectively, and *r* is the number of response attributes. The L_t statistic follows a chi-squared distribution with $r(r + 1)/2$ degrees of freedom. Therefore, the disclosure risk of the records in node *t* can be evaluated based on the *p*-value associated with L_t. We also use an adjusted L_t for small node size (Morrison 1990, p.292).

The proposed *error-digression pruning* (EDP) algorithm is provided in Figure 4. This algorithm, like usual decision tree algorithms, runs very fast. The time complexity is of $O(N \log N)$ for tree growing and $O(|T|^2)$ for tree pruning, where *N* is the number of records in the dataset and $|T|$ is the number of internal nodes in the unpruned tree *T*. In summary, the MART algorithm has the same time complexity as that of a typical regression tree algorithm, which is much faster than a traditional *k*-anonymity algorithm (Samarati and Sweeney 1998; Sweeney 2002).

We explain the EDP procedure using the example shown in Figure 1 and Table 1. We provide details of the computations for node 9. The node and branch errors are:

$$e(9) = 0.0537 \text{ and } e(B_9) = 0.0155.$$

The node and branch digressions are:

$$\Delta(9) = 0.0494 \text{ and } \Delta(B_9) = 0.1017.$$

The error-digression ratio and the *p*-value of the likelihood ratio test statistic (denoted p_9) are:

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$q_9=0.7297 \text{ and } p_9=0.0089.$$

Note that the response attribute values are normalized when calculating these measures. For the other internal nodes, we have

$$q_2=2.0392, q_6=2.9492, q_5=12.5078; \text{ and}$$

$$p_2=0.0365, p_6=0.1065, p_5=0.1310.$$

Suppose $k = 2$ and $\alpha = 0.05$. Then, node 9 will be pruned first, followed by node 2. This will result in a pruned tree that includes nodes 1, 2, 5, 6, 7, 8 and 9, with leaf nodes 2, 7, 8 and 9. So, given the minimum node size value *k*, the results of the EDP procedure are often different from those of *k*-anonymity. For instance, with *k*-anonymity there are 6 groups when $k = 2$, while the EDP procedure partitions the data into 4 groups (leaves) as described above. This, however, does not imply that the proposed method will always produce groups of larger size than a *k*-anonymity approach. The user can set a small *k* parameter along with a reasonable α value.

Categorical Value Concatenation

After the data are partitioned into subsets, the QI attribute values are altered to protect against re-identification. For numeric QI attributes, traditional *k*-anonymity approaches replace the original QI values in a subset with the range values of the attributes in the subset (Samarati and Sweeney 1998; Sweeney 2002). LeFevre et al. (2008) also suggest alternative values such as mean and median for replacement. In this study, we focus on anonymizing categorical QI attributes. Numeric QI attribute values can be anonymized using one of the existing replacement methods.

For categorical QI attributes, traditional *k*-anonymity approaches use generalization and suppression methods for anonymization. Typically, a user-defined generalization hierarchy is required. The use of pre-defined hierarchies may be ineffective in preserving data utility. For example, with the pre-defined hierarchy shown in Figure 2, many categorical values in the anonymized data are essentially suppressed (Table 2). To overcome this problem, we propose a dynamic value-concatenation method that merges categorical values based on the hierarchical structure of regression trees.

We adopt the binary split method used in Breiman et al. (1984) for splitting a categorical attribute. Many decision tree algorithms use a multi-way split method for categorical attributes, which routes each category into a branch. This method is not effective for our purpose. For the illustrative example, suppose such a multi-way split is made on the Occupation attribute at a node having all four Occupation values. The node will be divided into four branches, one for each Occupation value. A generalization based on this hierarchy will force the suppression of all values. Binary splits, on the other hand, allow more flexibility for generalization. For example, the node may be divided into two branches, one

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

with the Occupation value ‘unskilled’, and the other with the rest of the three Occupation values. Consequently, a generalization not involving suppression may be performed for the records in the second branch (even based on the pre-specified hierarchy in Figure 2).

For an attribute with c categories, there are $2^{c-1} - 1$ possible binary partitions of these categories (e.g., there are 7 different ways to partition the four Occupation attribute values in our example into two groups). When c is large, it is computationally prohibitive to find the best partition. However, for regression trees, Breiman et al. (1984) show that there is an efficient way to order the categories in a certain sequence so that there are only $c - 1$ (instead of $2^{c-1} - 1$) possible partitions. This method is used in our splitting algorithm.

The value-concatenation method is very easy to implement. It simply concatenates all categorical values that appear at a leaf of the pruned tree and then treats the concatenated value as one category. If there is a single category in the leaf, then no concatenation is needed. The results of using the value-concatenation method for the data in Table 1 are shown in Table 3. It is clear that data quality is better preserved with this method than with the pre-defined generalization hierarchy (see Figure 2 and Table 2). The semantics of the concatenated values are also clear. For example when $k = 4$, the occupation for the four records in the last group are ‘managerial’ or ‘professional’. It is not necessary to provide a generalized term for the category.

It is important to note that the value-concatenation method does not cause higher re-identification risk than the user-defined hierarchy even though it can provide more detailed information in the released data. Based on the k-anonymity principle, the reidentification risk is the same for the anonymized data in Table 3 as for that in Table 2. Given the parameter k (as a constraint), the objective of a k-anonymity-based approach is to minimize information loss caused by generalization and suppression. So, for the same k, an anonymized dataset with more detailed information in the QI attributes (e.g., Table 3) has better data quality than that with less detailed information (e.g., Table 2).

EXPERIMENTAL STUDY

An experimental study was conducted on several real-world financial, economic and healthcare datasets (these applications are well-documented as having some privacy implications). The proposed approach is compared with a current state-of-the-art technique. Performances are evaluated in terms of re-identification and value-disclosure risks under regression attacks, as well as data quality for performing regression analysis using two regression methodologies, linear regression and regression trees.

Because the experimental evaluation is conducted in the context of regression analysis, we select the response attributes such that they are most likely to be the output variables for prediction. Furthermore, the response attributes in each dataset are considered as the sensitive attributes since this is how regression attacks would be conducted. This is appropriate for assessing the tradeoff between anonymizing QI information and preserving data quality for regression analysis. Suppose a non-sensitive attribute is set as a response. If the relationship between a sensitive attribute that is not a response and this non-sensitive response is insignificant and negligible, then it is likely that the sensitive attribute will not

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

appear (or will appear very infrequently) as a splitting attribute for the tree. Consequently, it will be difficult to evaluate the impact of anonymization on the relationships between this sensitive attribute and the other non-sensitive attributes. We describe the data below.

Offer

The Association for Information Systems conducts annual surveys of MIS faculty salary offers (Galletta 2004). We selected the offer data from 1999 to 2002 (attributes are consistent for these four years and somewhat different for the other years). This dataset consists of 509 applicants who received offers during the period. There are 13 attributes, with three of them numeric and 10 categorical. The attributes are: salary offered, position, course load, number of years teaching, education, public or private, campus type, campus region, school's highest degree, accreditation, respondent accepted offer or not, respondent revealed identity or not, and the year of survey. Salary offered and course load were considered as the response (and sensitive) attributes.

Alcohol

This dataset was taken from Kenkel and Terza (2001), who study factors affecting individuals' drinking behaviors. It includes data on 2,467 male individuals, each with 17 attributes (3 numeric and 14 categorical): age, race, education, marital status, region, employment type, income, drinking frequency, having health insurance or not, insurance type, insurance source, having activity limit or not, having diabetes or not, having heart condition or not, having stroke history or not, visiting same doctor or not, and doctor's advice on drinking. The attribute drinking frequency was the response attribute in the original study (Kenkel and Terza 2001). We have added the attribute income as the second response (and sensitive) attribute.

Credit

This is a credit evaluation dataset (Bache and Lichman 2013). It has 1,000 records of customers, with 20 attributes (7 numeric and 13 categorical), used by a bank to evaluate credit applications. Some attributes are demographic or economic in nature, and include age, gender, marital status, length of employment, occupation type, housing status, housing liability, length of residence, other personal property status, foreign worker or not, having a phone number or not. Other attributes are banking and credit related, and include checking account status, savings account status, credit history, credit purpose, number of existing credits at this bank, other debtors, credit duration, installment, and credit amount. The attributes credit duration, installment and credit amount were considered as the response (and sensitive) attributes.

Experiment Design and Performance Measures

We compare our proposed MART method with the Regression Mondrian (RM) method proposed by LeFevre et al. (2008), which is, to our knowledge, the only existing data anonymization method that considers multi-response regression. As discussed earlier, there are two key differences between MART and RM: (1) MART considers sensitive value disclosure while RM does not; and (2) for categorical QI attributes, MART uses dynamic

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

value-concatenation while RM uses generalization that requires a user-defined hierarchy. We defined a generalization hierarchy for each categorical attributes in a dataset, based on the ideas provided by LeFevre et al. (2008). For simplicity, we assume all non-sensitive attributes are QI attributes and thus are subject to anonymization. For numeric QI attributes, we replace the original values by the group averages for both MART and RM methods. The values of sensitive attributes are not changed, following the *k*-anonymity protocol.

In the *k*-anonymity studies, re-identification risk is measured by minimum group size *k*, which often serves as a control measure to facilitate comparisons on the other risk and utility measures. We followed this common practice and used three typical group size values for RM and MART: *k* = 10, 20, and 30. The performances of the two techniques are then evaluated on a sensitive value disclosure risk and a data utility measure, which are described next.

To assess the sensitive value disclosure risk, we use a measure called relative squared distance (*RSD*), based on Liew et al. (1985). The *RSD* for a sensitive attribute Y_j is defined as:

$$RSD_j=\frac{1}{M}\left[\sum_{t=1}^{M}\left(\sum\nolimits_{i=1}^{n_t}[y_{ij}^t-\overline{y}_j^t]^2/\sum\nolimits_{i=1}^{n_t}[y_{ij}^t-\overline{Y}_j]^2\right)\right], \quad (10)$$

where M is the total number of groups (leaves), n_t is the number of records in group t, y_{ij}^t is the value of Y_j in the ith record in group t, $\overline{y}_j^t$ is the mean of the Y_j values in group t, and $\overline{Y}_j$ is the overall mean of the Y_j values (all values are normalized). The rationale for this measure is that once an intruder has used a regression attack and identified a target group t, the intruder will most likely use the group average $\overline{y}_j^t$ to estimate y_{ij}^t. So the numerator evaluates the closeness of the disclosure. The denominator represents the closeness when $\overline{Y}_j$ is used, which can be assumed as public information. Clearly, a larger *RSD* value implies a smaller disclosure risk (i.e., more difficult for the intruder to determine the sensitive values after identifying the group). For multiple attributes, the *RSD* measure is calculated as the average of the individual RSD_j.

Data utility is measured by the mean absolute percentage error (*MAPE*), defined for a response attribute Y_j as

$$MAPE_j=\frac{1}{H}\sum_{i=1}^{H}\left|\frac{y_{ij}-\hat{y}_{ij}}{y_{ij}}\right|, \quad (11)$$

where H is the number of records in the test set (we describe how to separate test data from training data next), y_{ij} is the value of the jth response attribute for the ith record in the test set, and $\hat{y}_{ij}$ is the estimate of y_{ij} based on the regression model built on the anonymized training data. For multiple responses, the *MAPE* value is calculated as the average of the individual $MAPE_j$. As *MAPE* measures the relative distance between the predictions of the model built from the anonymized data and the values in the test data, a smaller *MAPE* value is desirable.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Two regression methods, linear regression and regression trees, were used in the experiment for testing the performance in data quality. We built regression models using the anonymized data and then evaluated the utility of the regression models based on prediction accuracy. More specifically, we designed a 10-fold cross-validation experimental methodology, which is similar to the experimental scheme used by LeFevre et al. (2008), described below:

1. Divide the entire dataset into 10 equal-sized blocks using random sampling. The experiment will run 10 times, each using one of the blocks in turn as a test set and the remaining data as a training set.
2. For each run, reserve a block and call it the *original test set*; call the remaining data the *original training set*. Apply an anonymization technique (i.e., MART or RM) to the original training set to obtain an *anonymized training set*. During this process, a tree structure for partitioning data is created.
3. Build a regression model (i.e., a linear regression equation or a regression tree) using the anonymized training set.
4. Partition the original test set using the tree structure created in Step 2. Anonymize the partitioned test data to obtain an *anonymized test set*.
5. Test the regression model built in Step 3 using the anonymized test set and compute prediction accuracy or error accordingly.
6. Repeat Steps 2 through 5 for each of the 10 blocks. Report the average results over the 10 cross-validation runs.

Experimental Results

The results of the experiments are shown in Table 4. As mentioned above, we report the average *MAPE* results over the 10 cross-validation runs. For comparison, we also report the average *MAPE* results using the original data. It is observed that, for the same group size, the *RSD* values with MART are larger than those with RM in all datasets. This indicates that given the same re-identification risk, MART produces anonymized data with lower value-disclosure risk for the sensitive attributes than RM does. This can be explained by the use of the Δ-digression measure in MART for reducing the value-disclosure risk.

With respect to data utility, the *MAPE* value resulting from MART is smaller than that from RM in each scenario, using either linear regression or regression trees, which indicates that overall MART outperforms RM for regression analysis. One reason for this is that MART uses dynamic value-concatenation method to generalize categorical QI attribute values, which is better in preserving data quality than the pre-defined generalization hierarchies. The differences in the *MAPE* results between MART and RM are relatively small in some cases but fairly large in others. To examine if the differences are statistically significant, we performed a paired *t* test (Mitchell 1997) for each scenario, using significance levels of $\alpha = 0.05$ and $\alpha = 0.1$. The results are shown in Table 4. Overall, the differences are statistically significant in about half the comparisons at $\alpha = 0.05$ and in more than half at $\alpha = 0.1$.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Both RM and MART algorithms ran very fast and completed the procedures within one or two seconds. They are much faster than the traditional *k*-anonymity algorithms (Samarati and Sweeney 1998; Sweeney 2002). The runtimes for the two algorithms were almost the same, which is expected because they use similar regression tree algorithms.

Experiment on a Large Dataset

We have provided a computational complexity analysis indicating that the proposed MART algorithm is suitable for large-data applications. The datasets used in the primary experiment above are small or moderate in size. To test the performance of MART in a large-data setting, we conducted an additional experiment using a census dataset (Bache and Lichman 2013), which contains 95,130 individual records with 42 attributes (8 numeric and 34 categorical).

We use a classical *k*-anonymity algorithm developed by Sweeney (2002) as the *baseline* algorithm for comparison. We also included the RM algorithm for completeness (RM uses regression trees as well and is as efficient as MART). While the computational times for MART and RM do not depend much on the number of QI attributes, most of the traditional *k*-anonymity algorithms, including the baseline approach, have exponential time complexity in the number of QI attributes. Therefore, it is practically very difficult to run the baseline with many QI attributes. Consequently, from the 42 attributes, we selected age, gender, race, education, occupation and marital status in the Census data to be the QI attributes (these are also frequently considered as QI attributes in other *k*-anonymity studies). The wage attribute was considered as the response (and sensitive) attribute. Because it is very time consuming to run the baseline algorithm, we only tested for group size $k = 30$. Also, we performed a 2-fold cross validation procedure (instead of 10-fold cross validation). Given the large size of this data, we believe the results would not differ much if a different group size and number of folds were used.

The results of the experiment on the Census data are shown in Table 5. It is very clear that MART and RM run much faster than the baseline and are well-suited for large data applications. MART is slightly slower than RM because of the extra computation related to the digression values. MART and RM also outperform the baseline in terms of data quality for both linear regression and regression trees. This is because the baseline algorithm is not designed to preserve the relationships between the predictors (QI attributes) and the responses (sensitive attributes) for regression analysis. Furthermore, the *RSD* value with MART is considerably larger than that with RM, indicating that MART produces anonymized data with lower value-disclosure risk for the sensitive attributes than RM does. The *RSD* value with MART is also slightly better than that of the baseline. This experiment used only a single response variable (wage). Therefore, the results also demonstrate that the proposed approach is effective for a regression problem with one response variable.

Discussion

While MART outperforms RM in all the experiments, the performance of these approaches vary considerably in terms of data utility (*MAPE*) across the different datasets when

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

compared to those on the original data. For the Offer data, the *MAPE* results produced by MART and RM are very close to those based on the original data. For the Alcohol data, the results are a little further apart. For the Credit data, however, the error results based on the anonymized data by MART and RM are considerably larger than those on the original data. This suggests that it is relatively hard to preserve data utility for the Credit data when it is anonymized. A likely explanation is that the relationships between the responses (sensitive attributes) and the predictors (QI attributes) are very sensitive to changes in the QI values in the Credit data. There is another factor, however, that impacts the ability to preserve data utility. In the reported experiments, we have assumed that all of the non-sensitive attributes in the data were the QI attributes and thus were subject to anonymization. This assumption is reasonable for the purpose of experimental evaluation because it avoids potential bias due to the selection of the QI attributes, and is used in a consistent manner for the different approaches. In practical situations, it is usually unnecessary to anonymize all non-sensitive attributes. To further investigate this scenario, we randomly selected half the non-sensitive attributes as the QI attributes and then anonymized them using MART. The resulting *MAPE* values dropped to about 0.42 (from around 0.46 ~ 0.47 when all the non-sensitive attributes are anonymized), which is much closer to the *MAPE* values on the original data (about 0.37 ~ 0.38). This suggests that the utility of anonymized data depends on the strength of the relationships between the sensitive and non-sensitive attributes and on the number of non-sensitive attributes being anonymized. Therefore, when using MART, the user can begin by anonymizing all non-sensitive attributes. If this causes considerable deterioration in data utility, then the user can restrict the QI attributes to achieve acceptable levels of data utility.

The size of the dataset also impacts the ability to preserve data utility. For example, for the large dataset (Census data) the *MAPE* results from MART and RM are very close to those based on the original data. Given a *k* value, a dataset with larger size allows more groups. Subsequently, there will be more variation in the QI attribute values across the groups, and the generalization of the QI values within each group will have a relatively small impact on the characteristics of the entire dataset. As a result, it should be easier to preserve the data utility for larger datasets when they are anonymized.

When there is an outlier (i.e., an extreme value) in a predictor/QI attribute, traditional regression trees may create a leaf node containing only the outlier. This situation will usually not occur in MART because the final group size will be greater than one. However, if the group size *k* is considered only at the pruning stage, it is possible that the outlier will be merged with other nodes at a very high level in the tree, potentially resulting in a group size much larger than *k*. This can increase the prediction error considerably. One way to address this problem is for the user to first inspect whether there exist outliers in the data. If outliers are detected, a larger stopping size, say *m* ($1 < m \leq k$), should be specified to grow the regression tree. This will force MART to select the splits that ensure at least *m* records in each child node, avoiding any node that contains only an outlier.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

CONCLUSIONS AND IMPLICATIONS FOR FUTURE RESEARCH AND PRACTICE

Regression techniques have been widely used not only as a tool for business analytics in private and public domains, but also as a research method in management and social science studies, which often involve using personal data. Therefore, the regression attack problem we investigate is vitally important. This kind of an attack has not been examined in prior research, and extant approaches to preserve privacy are not designed to address this issue. To fill this research gap, we have presented a novel approach for protecting against sensitive value disclosure from such an attack. We have also proposed a dynamic value-concatenation method to limit re-identification risk.

We have shown analytically that the proposed Δ-digression measure has some important properties that help to evaluate value-disclosure risk when multiple numeric sensitive attributes are targeted. In addition, the proposed value-concatenation method better preserves data utility than user-defined generalization schemes used in existing approaches. Our experimental study demonstrates that the proposed approach is very effective in protecting data privacy and preserving data quality. Our approach can be applied to applications that have numeric and/or categorical data types. That enhances the breadth of applicability of our approach which has been a limitation for several related approaches that attempt to restrict disclosure of private information.

Future research could investigate alternative methods to anonymize the partitioned data. Particularly, the proposed value-concatenation method can be extended to include frequency information into the concatenated categories. For example in Table 3, when $k = 4$, the Occupation values for the five records in the first group can be coded as 'unskilled4+technical1' (based on the original count in Table 1). When the data is anonymized with this "weighted-value-concatenation" method, the frequency distributions of the categorical attributes can be completely preserved (it is easy to code a program that decomposes the concatenated values). This method would work well for data released for simple publishing purposes such as summary statistics reporting. However, it can be difficult to use for more advanced analysis such as regression, because there will be significantly more concatenated categories than the original ones. Therefore, developing a weighted value-concatenation method for predictive data mining and analytics is an interesting challenge deserving further study.

This work has important implications for future research beyond a strict regression setting. It will be useful to develop an integrated framework to deal with "predictive data mining attacks" that considers both classification attacks (Li and Sarkar 2009) and regression attacks. In this framework, re-identification risks can be assessed independent of the type of the sensitive attributes (i.e., numeric or categorical). The assessment for value-disclosure risks will depend on the type of the sensitive attributes: the digression measure proposed in this work can be used for sensitive numeric attributes while the entropy-based divergence measure proposed by Li and Sarkar (2009) can be used for sensitive categorical attributes. The most challenging situation is when the sensitive attributes include correlated numeric and categorical data. This problem clearly warrants more extensive research.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Our work has significant practical implications. Using the approach proposed in this work, organizations can assess the disclosure risks of the data to be released and take actions to reduce the risks. The first step is to identify the sensitive attributes in the data. In general, the sensitive attributes contain private information that an individual typically does not want revealed. Given a dataset, the sensitive attributes are those that cannot be found from public or external sources and they typically constitute the centerpiece of the information in the data (e.g., salary in a salary survey). After identifying the sensitive attributes, the proposed MART algorithm can be applied to the data to identify which non-sensitive attributes are the important QI attributes that can be used to re-identify individuals. MART also provides a measure to assess the value-disclosure risk for the individuals in a group and the value of the measure increases as the group size decreases. Therefore, the risks of both identity disclosure and value disclosure can be controlled by adjusting the group size. As a final step, the proposed value-concatenation method, which provides better data utility than the traditional generalization method, can be applied to anonymize the grouped data for release.

ACKNOWLEDGEMENTS

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

APPENDIX

Proof of Lemma 1

Let *M* be the total number of subsets partitioned by the tree, and n_t ($t = 1, \ldots, M$) be the number of records in subset *t*. Consider any two responses Y_j and Y_k. Let $y_{ij}^t (i=1, \ldots, n_t)$ be the value of Y_j in the *i*th record in subset *t*, $\overline{y}_j^t$ be the mean of the Y_j values in subset *t*, and $\overline{Y}_j$ be the overall mean of the Y_j values. Notation for Y_k is denoted similarly. Consider

$$y_{ij}^t - \overline{Y}_j = (\overline{y}_j^t - \overline{Y}_j) + (y_{ij}^t - \overline{y}_j^t), \text{ and}$$

$$y_{ik}^t - \overline{Y}_k = (\overline{y}_k^t - \overline{Y}_k) + (y_{ik}^t - \overline{y}_k^t).$$

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

$$(y_{ij}^t - \overline{Y}_j)(y_{ik}^t - \overline{Y}_k) = (\overline{y}_j^t - \overline{Y}_j)(\overline{y}_k^t - \overline{Y}_k) + (\overline{y}_j^t - \overline{Y}_j)(y_{ik}^t - \overline{y}_k^t) + (y_{ij}^t - \overline{y}_j^t)(\overline{y}_{ik}^t - \overline{Y}_k) + (y_{ij}^t - \overline{y}_j^t)(y_{ik}^t - \overline{y}_k^t). \quad \text{(A1)}$$

Summing over all the records (first within a subset and then over all subsets), and noting that the summations for the middle two terms in the right-hand side of (A1) equal zero, we get:

$$\sum_{t=1}^{M} \sum_{i=1}^{n_t} (y_{ij}^t - \overline{Y}_j)(y_{ik}^t - \overline{Y}_k) = \sum_{t=1}^{M} \sum_{i=1}^{n_t} (\overline{y}_j^t - \overline{Y}_j)(\overline{y}_k^t - \overline{Y}_k) + \sum_{t=1}^{M} \sum_{i=1}^{n_t} (y_{ij}^t - \overline{y}_j^t)(y_{ik}^t - \overline{y}_k^t). \quad \text{(A2)}$$

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

The term on the left is the scatter S_{jk} defined in Definition 1. The first term on the right is the between-subset scatter while the second term on the right is the sum of within-subset scatters, which can be written as $\sum_{t=1}^{M}s_{jk}(t)$ (following notation in Definition 1). Let t' be the node under consideration. Then,

$$d_{jk}(t')=S_{jk}-s_{jk}(t')=\sum_{t=1}^{M}\sum_{i=1}^{n_t}(\overline{y}_j^t-\overline{Y}_j)(\overline{y}_k^t-\overline{Y}_k)+\sum_{t\neq t'}s_{jk}(t). \quad \text{(A3)}$$

If the response values $y_{ij}^{t'}$ and $y_{ik}^{t'}(i=1,\ldots,n_{t'})$ are replaced by $\overline{y}_j^{t'}$ and $\overline{y}_k^{t'}$ respectively, then,

$$s_{jk}(t')=\sum_{i=1}^{n_{t'}}(y_{ij}^{t'}-\overline{y}_j^{t'})(y_{ik}^{t'}-\overline{y}_k^{t'})=0. \quad \text{(A4)}$$

In this case, the sum of within-subset scatters can still be written as $\sum_{t=1}^{M}s_{jk}(t)$, and $d_{jk}(t')$ in (A3) can be expressed in a form analogous to (A2). In other words, $\mathbf{D}(t')$ is the scatter matrix when the response values in node t' are replaced by the subset averages. Since the determinant of a scatter matrix is always positive, this completes the proof.

Proof of Lemma 2

Let B_t be a branch rooted at t with m leaves. Let n_ℓ ($\ell = 1, \ldots, m$) be the number of records in leaf ℓ. Let $y_{ij}^{\ell}(i=1,\ldots,n_\ell)$ be the value of Y_j in the ith record in leaf ℓ, $\overline{y}_j^{\ell}$ be the mean of the Y_j values in leaf ℓ, and $\overline{y}_j$ be the mean of the Y_j values in B_t's root node t. Denote these quantities similarly for another attribute Y_k. Following the same algebraic manipulation in the proof of Lemma 1, we have

$$\sum_{\ell=1}^{m}\sum_{i=1}^{n_\ell}(y_{ij}^{\ell}-\overline{y}_j)(y_{ik}^{\ell}-\overline{y}_k)=\sum_{\ell=1}^{m}\sum_{i=1}^{n_\ell}(\overline{y}_j^{\ell}-\overline{y}_j)(\overline{y}_k^{\ell}-\overline{y}_k)+\sum_{\ell=1}^{m}\sum_{i=1}^{n_\ell}(y_{ij}^{\ell}-\overline{y}_j^{\ell})(y_{ik}^{\ell}-\overline{y}_k^{\ell}). \quad \text{(A5)}$$

The term on the left is $s_{jk}(t)$ while the second term on the right can be written as $\sum_{\ell=1}^{m}s_{jk}(\ell)$. Denote the first term on the right (the between-leaf scatter) as b_{jk}. Then, (A5) can be written as

$$s_{jk}(t)=b_{jk}+\sum_{\ell}s_{jk}(\ell). \quad \text{(A6)}$$

Now, consider any leaf ℓ'. Equation (A6) can be written as

$$s_{jk}(t)=b_{jk}+\sum_{\ell\neq\ell'}s_{jk}(\ell)+s_{jk}(\ell'). \quad \text{(A7)}$$

Rearranging (A7) and adding S_{jk} to both sides, we have

Author Manuscript
Author Manuscript
Author Manuscript
Author Manuscript

$$S_{jk}-s_{jk}(\ell')=S_{jk}-s_{jk}(t)+b_{jk}+\sum_{\ell\neq\ell'}s_{jk}(\ell). \quad \text{(A8)}$$

That is

$$d_{jk}(\ell')=d_{jk}(t)+b_{jk}+\sum_{\ell\neq\ell'}s_{jk}(\ell). \quad \text{(A9)}$$

Let $\mathbf{D}(\ell')$, $\mathbf{D}(t)$ and $\mathbf{b}$ be the matrices with their (j, k) element being $d_{jk}(\ell')$, $d_{jk}(t)$ and $b_{jk}+\sum_{\ell\neq\ell'}s_{jk}(\ell)$, respectively. Then,

$$\mathbf{D}(\ell')=\mathbf{D}(t)+\mathbf{b}, \quad \text{(A10)}$$

It follows from the Minkowski determinant theorem (Marcus and Minc 1992) that

$$|\mathbf{D}(\ell')|=|\mathbf{D}(t)+\mathbf{b}| \geq |\mathbf{D}(t)|+|\mathbf{b}| \quad \text{(A11)}$$

Based on the same argument as in the proof of Lemma 1, **b** is a form of scatter matrix and thus $|\mathbf{b}| > 0$. Therefore,

$$|\mathbf{D}(\ell')|>|\mathbf{D}(t)|.$$

REFERENCES

Adam NR, Wortmann JC. Security-Control Methods for Statistical Databases: A Comparative Study. ACM Computing Surveys. 1989; 21(4):515–556.

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

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

Angst CM, Agarwal R. Adoption of Electronic Health Records in the Presence of Privacy Concerns: The Elaboration Likelihood Model and Individual Persuasion. MIS Quarterly. 2009; 33(2):339–370.

Atzori M, Bonchi F, Giannotti F, Pedreschi D. Anonymity Preserving Pattern Discovery. The VLDB Journal. 2008; 17(4):703–727.

Bache, K.; Lichman, M. UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science; 2013. http://archive.ics.uci.edu/ml.

Breiman, L.; Friedman, JH.; Olshen, RA.; Stone, CJ. Classification and Regression Trees. Belmont, CA: Wadsworth; 1984.

Cormode G, Srivastava D, Yu T, Zhang Q. Anonymizing Bipartite Graph Data Using Safe Groupings. The VLDB Journal. 2010; 19(1):115–139.

Cox A. Decision Tree: The Obama-Clinton Divide. The New York Times. 2008 Apr 16. http://www.nytimes.com/imagepages/2008/04/16/us/20080416_OBAMA_GRAPHIC.html.

Culnan M. ‘How Did They Get My Name?’: An Exploratory Investigation of Consumer Attitudes toward Secondary Information Use. MIS Quarterly. 1993; 17(3):341–363.

De’ath G. Multivariate Regression Trees: A New Technique for Modeling Species–Environmental Relationships. Ecology. 2002; 83(4):1105–1117.

Denning DE, Schlörer J. Inference Control for Statistical Databases. Computer. 1983; 16(7):69–82.

Duhigg C. How Companies Learn Your Secrets. The New York Times Magazine. 2012 Feb 16.:10.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

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

Friedman A, Schuster A, Wolff R. Providing *k*-Anonymity in Data Mining. International Journal on Very Large Data Bases. 2008; 17(4):789–804.

Fu Y, Chen Z, Koru G, Gangopadhyay A. A Privacy Protection Technique for Publishing Data Mining Models and Research Data. ACM Transactions on Management Information Systems. 2010; 1(1): 7, 1–7, 20. Article 7.

Fung BCM, Wang K, Yu PS. Anonymizing Classification Data for Privacy Preservation. IEEE Transactions on Knowledge and Data Engineering. 2007; 19(5):711–725.

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

Garfinkel R, Gopal R, Thompson S. Releasing Individually Identifiable Microdata with Privacy Protection against Stochastic Threat: An Application to Health Information. Information Systems Research. 2007; 18(1):23–41.

Heatherly R, Kantarcioglu M, Thuraisingham B. Preventing Private Information Inference Attacks on Social Networks. IEEE Transactions on Knowledge and Data Engineering. 2013; 25(8):1849–1862.

Johnson, RA.; Wichern, DW. Applied Multivariate Statistical Analysis. Upper Saddle River, NJ: Prentice Hall; 2002.

Kenkel DS, Terza JV. The Effect of Physician Advice on Alcohol Consumption: Count Regression with an Endogenous Treatment Effect. Journal of Applied Econometrics. 2001; 16(2):165–184.

KDnuggets. New Poll: Was Target Wrong in Using Analytics to Find Pregnant Women? 2012 http://www.kdnuggets.com/2012/02/index.html.

Lambert D. Measures of Disclosure Risk and Harm. Journal of Official Statistics. 1993; 9(2):313–331.

LeFevre K, DeWitt DJ, Ramakrishnan R. Workload-Aware Anonymization Techniques for Large-Scale Datasets. ACM Transactions on Database Systems. 2008; 33(3):17, 1–17, 47. Article 17.

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

Li X-B, Sarkar S. Against Classification Attacks: A Decision Tree Pruning Approach to Privacy Protection in Data Mining. Operations Research. 2009; 57(6):1496–1509.

Li X-B, Sarkar S. Protecting Privacy Against Record Linkage Disclosure: A Bounded Swapping Approach for Numeric Data. Information Systems Research. 2011; 22(4):774–789.

Liew CK, Choi UJ, Liew CJ. A Data Distortion by Probability Distribution. ACM Transactions on Database Systems. 1985; 10(3):395–411.

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

Marcus, M.; Minc, H. A Survey of Matrix Theory and Matrix Inequalities. New York: Dover; 1992.

Menon S, Sarkar S. Minimizing Information Loss and Preserving Privacy. Management Science. 2007; 53(1):102–116.

Mitchell, TM. Machine Learning. New York: McGraw-Hill; 1997.

Morrison, DF. Multivariate Statistical Methods. New York: McGraw-Hill; 1990.

Ohm P. Don't Build a Database of Ruin. Harvard Business Review. 2012 blog network, August 23. http://blogs.hbr.org/cs/2012/08/dont_build_a_database_of_ruin.html.

Oliveira SRM, Zaiane OR. A Unified Framework for Protecting Sensitive Association Rules in Business Collaboration. International Journal of Business Intelligence and Data Mining. 2006; 1(3):247–287.

Samarati, P.; Sweeney, L. Proceedings of the IEEE Symposium on Research in Security and Privacy. Washington, DC: IEEE Computer Society; 1998. Protecting Privacy When Disclosing Information: *k*-Anonymity and Its Enforcement through Generalization and Suppression; p. 19

Smith HJ, Dinev T, Xu H. Information Privacy Research: An Interdisciplinary Review. MIS Quarterly. 2011; 35(4):989–1015.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

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

Verykios VS, Elmagarmid AK, Bertino E, Saygin Y, Dasseni E. Association Rule Hiding. IEEE Transactions on Knowledge and Data Engineering. 2004; 16(4):434–447.

Zheleva, E.; Getoor, L. Proceedings of the 18th International Conference on World Wide Web. New York: ACM Press; 2009. To Join or Not to Join: The Illusion of Privacy in Social Networks with Mixed Public and Private User Profiles; p. 531-540.

Biographies

Xiao-Bai Li is a professor in the Department of Operations and Information Systems at the University of Massachusetts Lowell. He received his Ph.D. from the University of South Carolina in 1999. His research focuses on data mining, information privacy, and information economics. He has received funding for his research from National Institutes of Health (NIH) and National Science Foundation (NSF). His work has appeared in *Information Systems Research, Management Science, Operations Research, IEEE Transactions on Knowledge and Data Engineering, IEEE Transactions on Systems, Man, and Cybernetics, IEEE Transactions on Automatic Control, Communications of the ACM, Decision Support Systems, INFORMS Journal on Computing*, among others.

Sumit Sarkar is the Charles and Nancy Davidson Chair and Professor of Information Systems in the Naveen Jindal School of Management at the University of Texas at Dallas. He received his PhD from the Simon School of Business at the University of Rochester. His research interests are in personalization and recommendation technologies, sponsored search, data privacy, information quality, data integration, and software release strategies. His research has appeared in *Management Science, Information Systems Research*, *ACM Transactions on Database Systems*, *Operations Research*, *IEEE Transactions on Knowledge and Data Engineering*, and *The INFORMS Journal on Computing,* among others. He has served as the conference co-chair for the *Workshop on Information Technology and Systems* (WITS) in 1999, the program co-chair for the *International Conference on Information Systems* (ICIS) in 2001, and the conference co-chair for the IEEE *International Conference on Services Computing* (IEEE SCC) in 2009. He has been a visiting faculty member at the National University of Singapore and the Indian School of Business, and a visiting scientist at IBM Research Laboratories.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

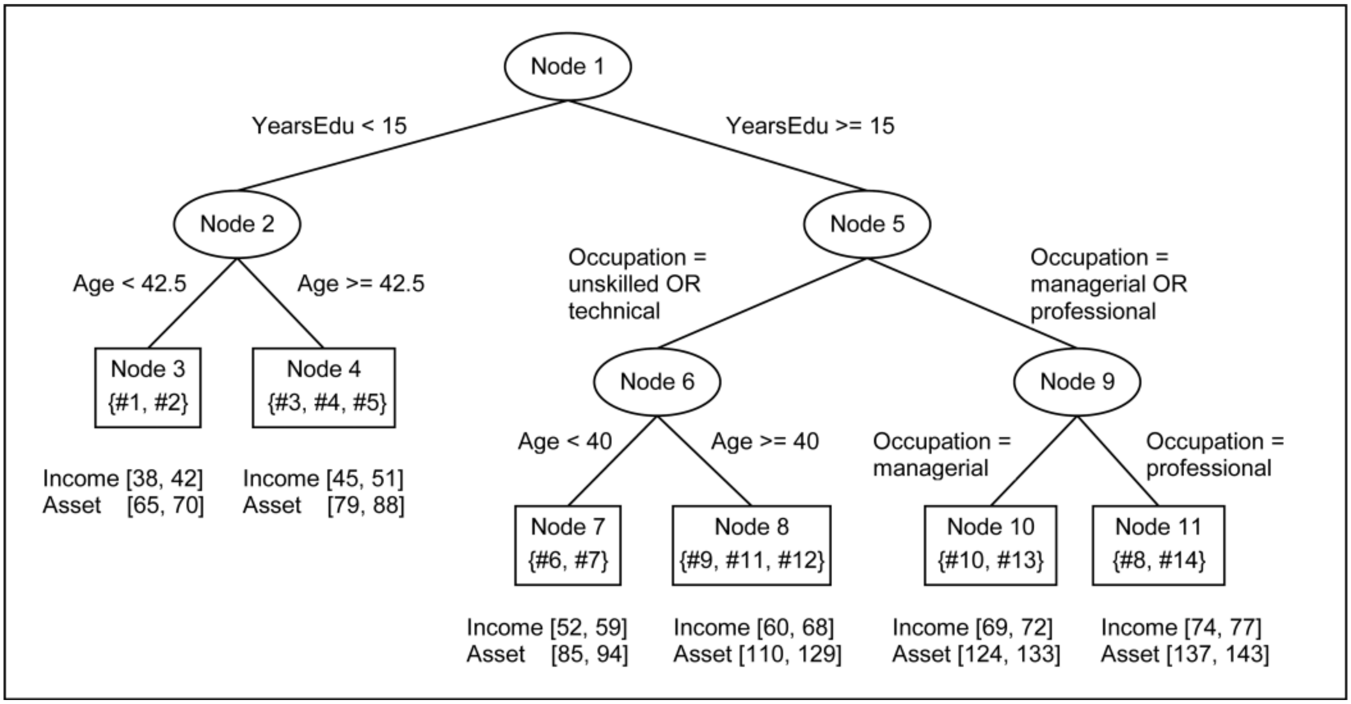

Node 1
YearsEdu < 15
YearsEdu >= 15
Node 2
Node 5
Age < 42.5
Age >= 42.5
Occupation = unskilled OR technical
Occupation = managerial OR professional
Node 3 {#1, #2}
Node 4 {#3, #4, #5}
Income [38, 42] Asset [65, 70]
Income [45, 51] Asset [79, 88]
Node 6
Node 9
Age < 40
Age >= 40
Occupation = managerial
Occupation = professional
Node 7 {#6, #7}
Node 8 {#9, #11, #12}
Node 10 {#10, #13}
Node 11 {#8, #14}
Income [52, 59] Asset [85, 94]
Income [60, 68] Asset [110, 129]
Income [69, 72] Asset [124, 133]
Income [74, 77] Asset [137, 143]

Figure 1.

A Regression Tree Built on Data in Table 1

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

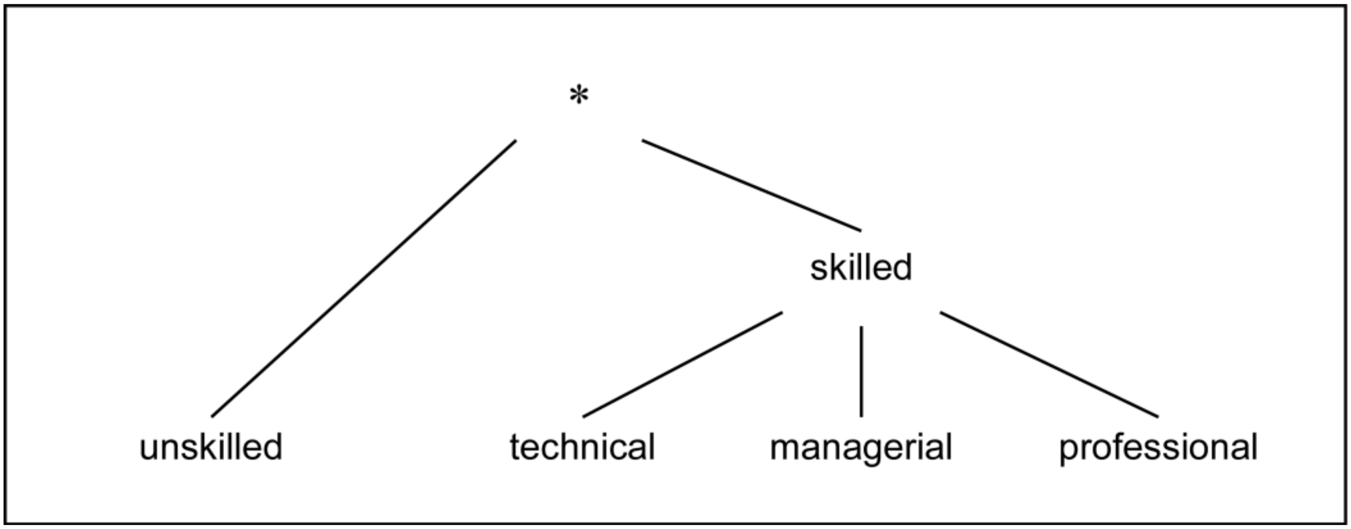


Figure 2.
Generalization Hierarchy for Occupation Attribute

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

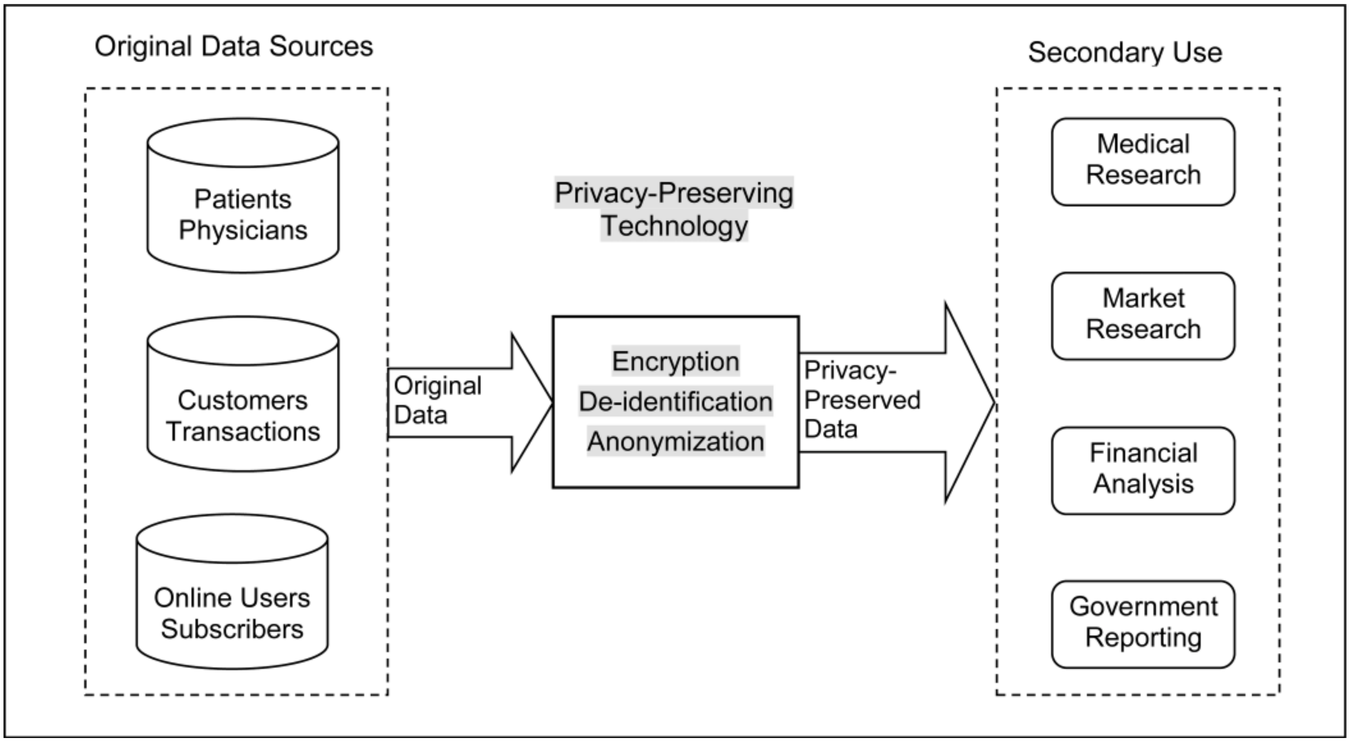
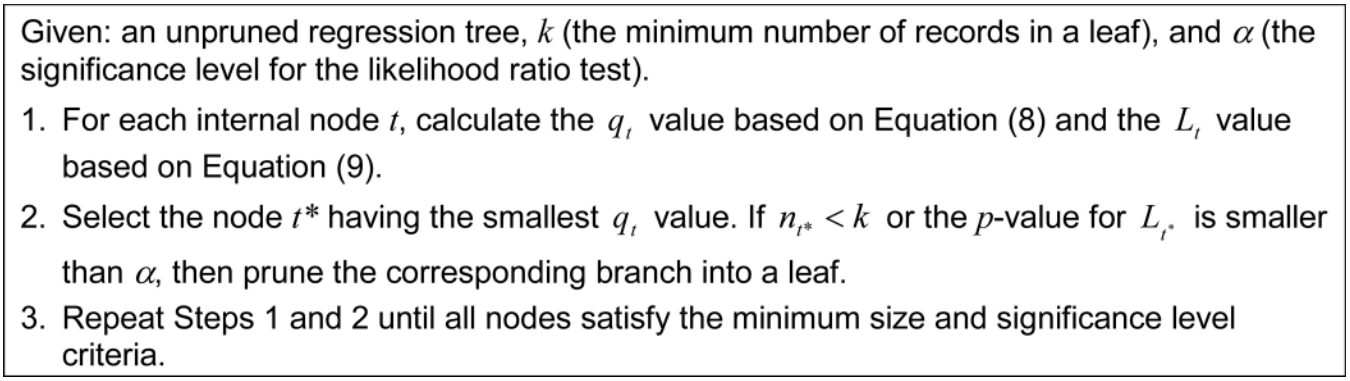


Figure 3.

A Conceptual View of Privacy-Preserving Technology

Author Manuscript Author Manuscript Author Manuscript Author Manuscript



Given: an unpruned regression tree, k (the minimum number of records in a leaf), and α (the significance level for the likelihood ratio test).

1. For each internal node t, calculate the q_t value based on Equation (8) and the L_t value based on Equation (9).
2. Select the node t^* having the smallest q_t value. If $n_{t*} < k$ or the p-value for L_{t^*} is smaller than α, then prune the corresponding branch into a leaf.
3. Repeat Steps 1 and 2 until all nodes satisfy the minimum size and significance level criteria.

Figure 4.
The Error-Digression Pruning (EDP) Algorithm

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 1

Illustrative Example: Original Data

No.	Age	YearsEdu	Occupation	Income ($000)	Asset ($000)
1	27	12	unskilled	38	65
2	39	14	unskilled	42	70
3	46	14	unskilled	45	79
4	59	12	technical	50	84
5	64	13	unskilled	51	88
6	33	16	technical	59	94
7	35	16	unskilled	52	85
8	42	18	professional	74	137
9	45	18	technical	66	116
10	30	18	managerial	69	124
11	48	16	technical	68	129
12	62	16	unskilled	60	110
13	56	17	managerial	72	133
14	51	20	professional	77	143

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 2

Illustrative Example: *k*-Anonymized Data

	***k* = 2**			***k* = 4**				
No.	**Age**	**YearsEdu**	**Occupation**	**Age**	**YearsEdu**	**Occupation**	**Income ($000)**	**Asset ($000)**
1	[27–39]	[12–14]	unskilled	[27–64]	[12–14]	*	38	65
2	[27–39]	[12–14]	unskilled	[27–64]	[12–14]	*	42	70
3	[46–64]	[12–14]	*	[27–64]	[12–14]	*	45	79
4	[46–64]	[12–14]	*	[27–64]	[12–14]	*	50	84
5	[46–64]	[12–14]	*	[27–64]	[12–14]	*	51	88
6	[33–35]	16	*	[33–62]	[16–18]	*	59	94
7	[33–35]	16	*	[33–62]	[16–18]	*	52	85
9	[45–62]	[16–18]	*	[33–62]	[16–18]	*	66	116
11	[45–62]	[16–18]	*	[33–62]	[16–18]	*	68	129
12	[45–62]	[16–18]	*	[33–62]	[16–18]	*	60	110
10	[30–56]	[17–18]	managerial	[30–56]	[17–20]	skilled	69	124
13	[30–56]	[17–18]	managerial	[30–56]	[17–20]	skilled	72	133
8	[42–51]	[18–20]	professional	[30–56]	[17–20]	skilled	74	137
14	[42–51]	[18–20]	professional	[30–56]	[17–20]	skilled	77	143

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 3

An Illustrative Example: Anonymized Data Using Value-Concatenation

	k = 2			*k* = 4				
No.	**Age**	**YearsEdu**	**Occupation**	**Age**	**YearsEdu**	**Occupation**	**Income ($000)**	**Asset ($000)**
1	[27–39]	[12–14]	unskilled	[27–64]	[12–14]	unskilled+technical	38	65
2	[27–39]	[12–14]	unskilled	[27–64]	[12–14]	unskilled+technical	42	70
3	[46–64]	[12–14]	unskilled+technical	[27–64]	[12–14]	unskilled+technical	45	79
4	[46–64]	[12–14]	unskilled+technical	[27–64]	[12–14]	unskilled+technical	50	84
5	[46–64]	[12–14]	unskilled+technical	[27–64]	[12–14]	unskilled+technical	51	88
6	[33–35]	16	unskilled+technical	[33–62]	[16–18]	unskilled+technical	59	94
7	[33–35]	16	unskilled+technical	[33–62]	[16–18]	unskilled+technical	52	85
9	[45–62]	[16–18]	unskilled+technical	[33–62]	[16–18]	unskilled+technical	66	116
11	[45–62]	[16–18]	unskilled+technical	[33–62]	[16–18]	unskilled+technical	68	129
12	[45–62]	[16–18]	unskilled+technical	[33–62]	[16–18]	unskilled+technical	60	110
10	[30–56]	[17–18]	managerial	[30–56]	[17–20]	managerial+professional	69	124
13	[30–56]	[17–18]	managerial	[30–56]	[17–20]	managerial+professional	72	133
8	[42–51]	[18–20]	professional	[30–56]	[17–20]	managerial+professional	74	137
14	[42–51]	[18–20]	professional	[30–56]	[17–20]	managerial+professional	77	143

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 4

Results of Primary Experiments

Data	Method	Group Size	RSD	Linear Regression MAPE	Regression Tree MAPE
Offer	Original			0.1520	0.1528
	RM	10	0.5861	0.1678	0.1674
	MART		0.6258	0.1670	0.1657
	RM	20	0.6493	0.1700	0.1703*
	MART		0.7098	0.1673	0.1670*
	RM	30	0.6919	0.1745**	0.1751**
	MART		0.7598	0.1673**	0.1672**
Alcohol	Original			5.7562	6.3045
	RM	10	0.6121	6.3465	7.0129**
	MART		0.7160	6.2149	6.4172**
	RM	20	0.6705	6.5732	7.2324**
	MART		0.7533	6.3593	6.8005**
	RM	30	0.6725	6.9313**	7.9977**
	MART		0.8768	6.5498**	6.8842**
Credit	Original			0.3710	0.3873
	RM	10	0.5209	0.4664**	0.4662
	MART		0.5894	0.4595**	0.4657
	RM	20	0.6301	0.4647**	0.4671
	MART		0.6664	0.4604**	0.4665

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Data	Method	Group Size	RSD	Linear Regression MAPE	Regression Tree MAPE
	RM	30	0.6402	0.4664**	0.4719*
	MART		0.7080	0.4624**	0.4683*

** The results of the two methods are statistically significantly different at α = 0.05.

* The results of the two methods are statistically significantly different at α = 0.1.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 5

Results of Experiment on Census Data

Method	Time (second)	RSD	Linear Regression MAPE	Regression Tree MAPE
Original			0.8135	0.8161
Baseline	8345.0	0.8098	0.8904**	0.8876**
RM	10.1	0.5492	0.8286**	0.8267**
MART	10.5	0.8234	0.8192**	0.8236**

** The results of the pairwise comparisons across the three methods are all statistically significantly different at $\alpha = 0.05$.