PROTECTING PRIVACY AGAINST REGRESSION ATTACKS IN PREDICTIVE DATA MINING

Completed Research Paper

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

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

Abstract

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-mining technique, can be used to effectively reveal individuals' sensitive data. This problem, which we call a "regression attack," has been overlooked in the literature. 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 adopts a novel measure which considers the tradeoff between disclosure risk and data utility in a regression tree pruning process. We also propose a dynamic value-concatenation method, which overcomes the limitation of requiring a user-defined generalization hierarchy in traditional k-anonymity approaches. Our approach can be used for anonymizing both numeric and categorical data. An experimental study is conducted to demonstrate the effectiveness of the proposed approach.

Keywords: Data privacy, data mining, regression trees, anonymization

Introduction

Predictive data-mining techniques, such as regression and classification, have been widely used by organizations to build knowledge-discovery and business-intelligence solutions. They have been applied to a variety of domains that involve using personal data, including database marketing, healthcare study, and financial analysis. While these techniques are used by organizations to better understand and serve their customers, and thus gain competitive advantages, there are growing concerns about invasions to privacy by these techniques. In a widely-publicized incident, Netflix had recently awarded $1 million to a research team led by two AT&T employees for winning a contest to improve the predictive accuracy of the company's movie recommendation system by over 10%. The contest, which lasted for three years, was considered by many to be a great research and business success. However, Netflix had to cancel plans for a sequel when it was discovered that the de-identified data released for the contest, which included movie recommendations and choices made by customers, could in fact be used to re-identify the customers (Lohr 2010). Concerns about privacy have also caused data quality and integrity to deteriorate. According to Teltzrow and Kobsa (2004), 82% of online users have refused to give personal information and 34% have lied when asked about their personal habits and preferences.

From a privacy viewpoint, the attributes of data on individuals can be classified into three categories: (1) *explicit identifiers*, which can be used to directly identify an individual, including name, social security number, phone number, and driver's license 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*, which are normally not considered as sensitive by individuals, such as age, gender, race, education, and occupation. However, the values of some of these attributes can often be used to identify individuals by matching data from different sources, resulting in identity disclosure. Such attributes are 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 privacy-preserving data mining research, it is typically assumed that the explicit identifiers have already been removed from the data. Data masking is applied to QI attributes to prevent or limit the re-identification risk (and the sensitive attributes are typically released in their original values). We assume the same setting in this study.

This research investigates a privacy disclosure problem that occurs when a *k*-anonymity based approach is used to mask data that is used for regression purposes. 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 (Sweeney 2002). When the k-anonymized data is intended for regression analysis, a regression tree technique is often used (LeFevre et al. 2008; Fu et al. 2010). Regression trees, introduced by Breiman et al. (1984), 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. As such, regression trees are widely used in predictive data mining. However, regression trees can be used as a tool to effectively reveal sensitive information about individuals. This can be accomplished by setting the sensitive attributes as the response attributes, and building a regression tree to help reveal the individuals' sensitive values. We call this use of regression trees for "mining" personal information a *regression attack*.

To illustrate the problem, consider an example dataset containing 14 individuals, as shown in Table 1. There are two numeric QI attributes (Age and YearsEdu), one categorical QI attribute (Occupation, with four categories), and two numeric sensitive attributes (Income and Asset). Given this dataset, a privacy intruder can set the two sensitive attribute as responses and build a regression tree based on the methods of Breiman et al. (1984) and De'ath (2002, for multiple 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). A split criterion is specified along with the edge representing the split. With this tree, it is very 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 who is a 'professional' is included in the dataset, then he can find the ranges of the Income and Asset values for the individual, which are very narrow.

Table 1. An 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	45	18	technical	66	116
9	48	16	technical	68	129
10	62	16	unskilled	60	110
11	30	18	managerial	69	124
12	56	17	managerial	72	133
13	42	18	professional	74	137
14	51	20	professional	77	143

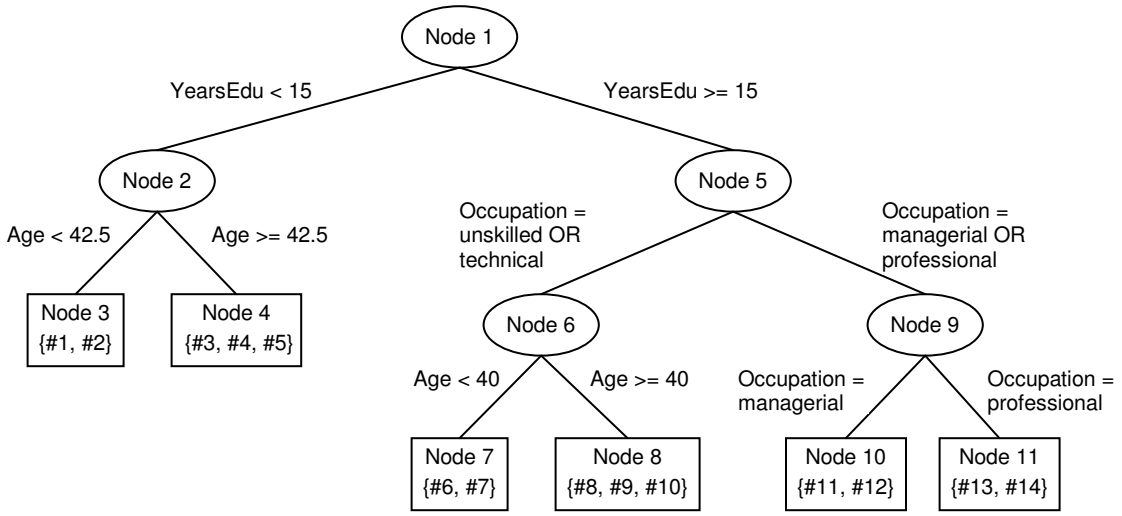

Figure 1. A Regression Tree Built on Data in Table 1

Similar results can be achieved using other methods as well. For example, if the intruder knew some attribute values of his target subjects, he can issue an ad-hoc query to directly search for these targets. A regression tree technique, however, can compromise data privacy in a systematic way. It is more effective than an ad-hoc query for disclosure in several aspects. First, for a numeric QI attribute, only a value range, instead of the exact value, is needed to determine the targets. Second, sensitive values can often be revealed with regression trees using only a small subset of the QI attributes (whereas an ad-hoc query often requires more attributes, depending on the sequence in which the attributes are considered). Third, a regression tree shows which targets are more vulnerable to attacks (i.e., sensitive values can be determined fairly accurately) and which QI attributes are the key attributes for disclosure. This can help an intruder identify new targets, or gather additional data for those targets whose QI attributes are partially available to the intruder. Finally, a regression attack can simultaneously identify a large number of target individuals.

Table 2. An Illustrative Example: *k*-Anonymized Data

No.	*k* = 2			*k* = 4			Income ($000)	Asset ($000)
	Age	YearsEdu	Occupation	Age	YearsEdu	Occupation		
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
8	[45-62]	[16-18]	*	[33-62]	[16-18]	*	66	116
9	[45-62]	[16-18]	*	[33-62]	[16-18]	*	68	129
10	[45-62]	[16-18]	*	[33-62]	[16-18]	*	60	110
11	[30-56]	[17-18]	managerial	[30-56]	[17-20]	skilled	69	124
12	[30-56]	[17-18]	managerial	[30-56]	[17-20]	skilled	72	133
13	[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

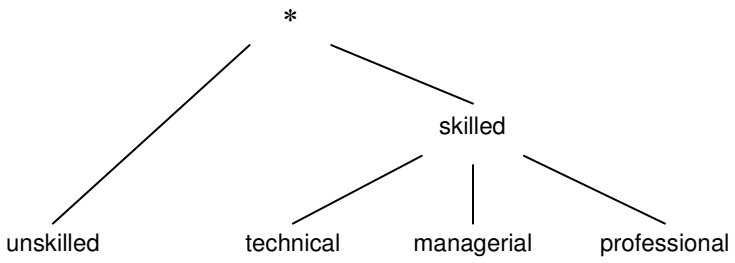

Figure 2. Generalization Hierarchy for Occupation Attribute

Existing *k*-anonymity-based techniques cannot effectively deal with such a regression attack. The basic idea behind *k*-anonymity is to partition a dataset into groups with at least *k* records in each group, and then anonymize the QI attribute values with the same generalized value within a group, so that the records in a group are indistinguishable. For a numeric attribute, *k*-anonymity replaces the original values in a group with the group range. For a categorical attribute, it generalizes the values based on a user-defined hierarchy. LeFevre et al. (2008) propose a *k*-anonymity-based method called *Regression Mondrian* for data to be used for regression analysis. 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). The generalization hierarchy for the Occupation attribute is given in Figure 2 (where a symbol * represents suppression of values). 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. So, the intruder can still obtain the sensitive information fairly accurately for the individuals in these groups, even though he cannot positively identify the individuals. For example, if he knew that an individual who is a 'professional' with '18-20' years of education, he can find that the individual's income is between $74,000 and $77,000. This kind of situation also occurs for some of the 4-anonymized groups (e.g., the group with record numbers 11, 12, 13 and 14).

Another 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 values, the values will have to be suppressed, 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 suppression symbol (∗) based on the hierarchy. These suppressed values cause the utility of the released data to deteriorate.

In this study, we address the privacy disclosure and data utility problems discussed above. To protect against privacy disclosure from regression attacks, we propose a regression-tree-based approach, which adopts a novel measure that considers the tradeoff between disclosure risk and data utility in the regression-tree pruning process. To overcome the limitation due to pre-defined generalization hierarchies, we propose a dynamic value-concatenation method that merges categorical values based on the hierarchical structure of the regression trees. We call the proposed technique MART (for Multivariate Anonymization with Regression Trees). The main contributions of this research are summarized as follows.

- *The originality of the problem*. The regression attack problem has not been formally studied in the literature. Existing privacy-preserving data-mining techniques are not appropriate in dealing with this problem. Particularly, we demonstrate that using a *k*-anonymity technique without caution can increase the disclosure risk for the sensitive data.
- *The novelty of the approach*. We propose a novel measure that considers the tradeoff between disclosure risk and data utility for multiple numeric sensitive attributes, which is used in constructing regression trees for data partitioning. For data anonymization, we propose a dynamic value-concatenation method that merges categorical values based on the hierarchy of the regression trees. The two components of the proposed approach are both new to the literature.
- *The practicality of the technique*. The proposed MART technique can be used for anonymizing both numeric and categorical QI attributes. MART is computationally very efficient and is much faster than the traditional *k*-anonymity algorithms. It is therefore well-suited for large-scale data-mining applications.

The rest of the paper is organized as follows. In the next section, we discuss prior research related to our problem. In the follow-up section, 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 datasets. The final section concludes the paper and provides directions for future research.

Related Work

There is a large body of research on privacy-preserving data publishing and mining (Aggarwal and Yu 2008). A significant part of the literature is related to the *k*-anonymity framework, proposed by Sweeney (2002) and Samarati (2001). 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 the same dataset. In this way, the identity of an individual is expected to be better protected. *K*-anonymity is a general-purpose technique for privacy-preserving data publishing. It may be ineffective when the anonymized data is used for data mining, because it is not designed specifically to preserve the relationships between the sensitive attributes and the QI attributes.

In the data-mining area, privacy issues have been studied by Agrawal and Srikant (2000) and Lindell and Pinkas (2002), among others. A number of studies develop privacy-preserving data-mining approaches under the *k*-anonymity framework. Fung et al. (2007) propose a top-down refinement method for classification problem that satisfies the *k*-anonymity principle. Friedman et al. (2008) developed a set of *k*-anonymity-based algorithms for various data-mining tasks, including classification, clustering and association rules mining (but not regression). LeFevre et al. (2008) and Fu et al. (2010) propose *k*-anonymity based approaches for classification and regression problems, both using classification and regression trees. Neither of these two studies, however, addresses the regression attack problem. 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. This study, however, considers sensitive numeric data and the approach we propose is intended for regression application.

The *k*-anonymity approach focuses on re-identification risk only and does not consider sensitive value disclosure. It generalizes different but similar QI attribute values into the same value within a group. The new values produced by the generalization operation are still correct with respect to the generalized categories. 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 disclosure risk.

To address this issue, Machanavajjhala et al. (2006) propose a privacy principle called *l*-diversity that applies to categorical sensitive attributes. The *l*-diversity principle requires that a sensitive attribute should include at least *l* well-diversified values in the *k*-anonymized data. 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 either unnecessary or 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 this problem, Li et al. (2007) propose another privacy principle called *t*-closeness, which requires that, for each group, the distance between the distribution 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 an *l*-diversity requirement. The *t*-closeness measure is applicable to a single numeric attribute, but is not appropriate for correlated multiple numeric attributes. Furthermore, the *t*-closeness approach does not consider prediction errors because it is intended for general-purpose data publishing. As a result, the anonymized data might not be suitable for regression analysis.

In summary, there is a lack of research in the data privacy literature that addresses the regression attack problem. Therefore, 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 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 $y_i(t)$ $(i = 1,...,n_t)$ be the value of the response attribute in the *i*th record in node *t*, and $\bar{y}(t)$ be the mean of the response attribute values in node *t*. The univariate *SSE* at node *t* is defined as

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

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, $t_1,...,t_m$. The reduction in *SSE*, $e(t) - [e(t_1) + \cdots + e(t_m)]$, 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 for multivariate regression trees in the literature. The splitting criteria proposed were 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 $\mathbf{y}_i(t) = [y_{i1}(t),...,y_{ir}(t)]'$ be the values of the response attributes in the ith record in node t, and $\bar{\mathbf{y}}(t)$ 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

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

With this measure, a multivariate regression tree can be built similarly 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 closer in the Income and Asset values.

The method 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. Fu et al. (2010) employ a similar idea (limited to univariate regression trees), but impose additional constraints in the tree-growing process to preserve the tree structure on the anonymized data. As mentioned earlier, neither method has considered sensitive value disclosure that is vulnerable to a regression attack.

The approach we propose involves not only growing a regression tree but also pruning the tree. For a traditional regression tree, the purpose of pruning is to avoid over-fitting problem. Therefore, the usual pruning method in regression trees aim at minimizing prediction error. In our problem, however, both disclosure risk and prediction error should be considered in selection of 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, since the overall distributions are usually public information. In other words, when anonymized data is released, the data recipient can expect the overall parameters, such as the means and covariances of the response attributes for the entire dataset, to be reasonably close to the original parameters. Indeed, in many cases, such original parameters are released with the data.

To measure the disclosure risk for records at a node with the above property, 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 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 *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 benchmark.

Definition 1. Let $\mathbf{S}$ be the scatter matrix of the response attributes on the entire dataset and S_{jk} be the (j,k) element of $\mathbf{S}$. 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 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. The node digression has the following property:

Lemma 1. *If there is no linear dependence between response attributes, then,*

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

The proofs of this lemma and all other mathematical properties are provided in the Appendix. 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 weighted digression of the child nodes (and their descendants). To describe this property, we first define some terms.

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, for instance, 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. The *branch digression* of B_t is defined as the sum of its leaf node digressions, i.e.,

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

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 branch root node t. Hence, the branch digression for* B_t *is always greater than the node digression for t; that is,*

$$\Delta(\ell) > \Delta(t), \ \forall \ell, t \ \Rightarrow \ \Delta(B_t) > \Delta(t), \forall t \ . \tag{6}$$

Lemma 2 states that a split of a node always increases digression. In other word, Δ-digression increases monotonically in the depth of the node (with respect to its ancestor nodes). So, pruning of a branch into a leaf always reduces digression. Next, we define the error for a node *t* and a branch B_t.

Definition 4. The *node error e*(*t*) is the *SSE* of node *t* defined in Equations (1) and (2). The *branch error* $e(B_t)$ is defined as the sum of its leaf node errors:

$$e(B_t) = \sum_{\ell=1}^{m} e(\ell) \, , \tag{7}$$

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:

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

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

Error-Digression Pruning

During the pruning process, we want the reduction in error 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 node 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's 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 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 reidentification 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 below:

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

where $\tilde{\boldsymbol{\Sigma}}$ and $\tilde{\boldsymbol{\Sigma}}_t$ are the sample covariance matrix for the entire dataset and node *t* subset respectively, and *r* is the number of the responses. The L_t statistic follows a chi-squared distribution with $r(r+1)/2$ degrees of freedom. As such, 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 3. 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.

0. Let *k* be the minimum number of records in a leaf, and α be the significance level for the likelihood ratio test.
1. For each internal node *t*, calculate the q_t value based on Equation (8) and L_t value based on Equation (9).
2. Select the node *t** having the smallest q_t value. If $n_{t^*} < k$ and the *p*-value for L_t is smaller than α, then prune the corresponding branch into a leaf; other wise, go to Step 1.
3. Repeat Steps 1 and 2 until all nodes satisfy the minimum size and significance level criteria.

Figure 3. The Error-Digression Pruning (EDP) Algorithm

We explain the EDP procedure using the example in Figure 1 and Table 1. We provide below for node 9 the results of the node error and digression, the branch error and digression, the error-digression ratio and the *p*-value of the likelihood ratio test statistic (denoted p_9).

$$e(9) = 0.0537, \quad e(B_9) = 0.0155, \quad \Delta(9) = 0.0494, \quad \Delta(B_9) = 0.1017, \quad q_9 = 0.7297, \quad p_9 = 0.0089.$$

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

$$q_2 = 2.0392, \quad q_6 = 2.9492, \quad q_5 = 12.5078; \text{ and } \quad p_2 = 0.0365, \quad p_6 = 0.1065, \quad p_5 = 0.1310.$$

Suppose $k = 2$ and $\alpha = 0.05$. Then, node 9 will be pruned off 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 same minimum 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 reidentification. 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. LeFevre et al. (2008) also suggest alternative values

such as mean and median for replacement. Fu et al. (2010) propose a somewhat complex scheme to anonymize numeric QI values. In this study, we focus on anonymizing categorical QI attributes. Numeric QI attribute values are simply replaced by subset means.

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 a pre-defined hierarchy shown in Figure 2, many categorical values in the anonymized data (Table 2) are suppressed. 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 a 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 divides each category into a branch. This method is not effective for our purpose. For the illustrative example, if such a multi-way split is made on the Occupation attribute, then a generalization of this attribute will force the suppression of its values. Binary splits, on the other hand, allow more flexibility for generalization.

For an attribute with c categories, there are $2^c - 1$ binary partitions of these categories (e.g., there are 15 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 (and classification trees with only two classes), Breiman et al. (1984) show that there is a way to order the categories in a certain sequence so that the best partition is one of the cuts of the sequence. As a result, there is only $c - 1$ (instead of $2^c - 1$) possible partitions. This method is implemented in our splitting algorithm.

Table 3. An Illustrative Example: Anonymized Data Using Value-Concatenation

No.	*k* = 2			*k* = 4			Income ($000)	Asset ($000)
	Age	YearsEdu	Occupation	Age	YearsEdu	Occupation		
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
8	[45-62]	[16-18]	unskilled+technical	[33-62]	[16-18]	unskilled+technical	66	116
9	[45-62]	[16-18]	unskilled+technical	[33-62]	[16-18]	unskilled+technical	68	129
10	[45-62]	[16-18]	unskilled+technical	[33-62]	[16-18]	unskilled+technical	60	110
11	[30-56]	[17-18]	managerial	[30-56]	[17-20]	managerial+professional	69	124
12	[30-56]	[17-18]	managerial	[30-56]	[17-20]	managerial+professional	72	133
13	[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

The method of value-concatenation 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; note that our proposed method may partition the data differently, depending on the α parameter). 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.

There can be some alternative schemes to the value-concatenation method. One way is to include proportion information into the concatenated categories. For example, for 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 weighted concatenation may present a challenge to regression analysis because there will be significantly more concatenated categories than the original ones. Note that the value-concatenation method can be applied not only for data with regression trees, but also for data with classification trees.

Experiments

We conducted experiments on two real-world financial and healthcare datasets to evaluate the proposed method. We first describe the datasets 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). They include salary offered, position, course load, number of years teaching, region, year indicator, etc. Salary offered and number of years teaching were considered as the sensitive attributes.

Alcohol. This dataset was taken from Kenkel and Terza (2001), which study factors affecting individuals' drinking behaviors. It includes data on 2,467 male individuals, each with 17 attributes (3 numeric and 14 categorical). There are demographic attributes such as age, race, education, marital status, and employment type, as well as several health and health-insurance related attributes. The attributes income and drinking frequency (biweekly) were considered as the sensitive attributes.

The experiment was conducted on regression analysis. We set the sensitive attributes as responses for regression. This is the most appropriate for assessing the tradeoff between protecting against regression attacks and preserving data quality for regression analysis. Suppose a non-sensitive attribute is set as the response. If the final regression model shows that the relationship between the sensitive attribute (which is a non-response variable) and the non-sensitive response is insignificant, then it will be easy for a data anonymization method to alter the sensitive attribute values while keeping the regression model practically unchanged. This situation is unlikely to occur when the sensitive attribute is set as the response. For simplicity, we assume all non-sensitive attributes are QI attributes and thus are anonymized. The values of sensitive attributes are not changed, following the k-anonymity protocol.

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 uses multivariate regression trees. As discussed earlier, there are three key differences between MART and RM: (1) MART considers sensitive value disclosure while RM does not; (2) For categorical QI attributes, MART uses dynamic value-concatenation while RM uses generalization that requires a user-defined hierarchy; (3) MART uses binary split while RM uses multi-way split for categorical attributes. We defined a generalization hierarchy for each categorical attributes in a dataset, based on the ideas provided by LeFevre et al. (2008). For numeric QI attributes, we replace the original values by the group averages for both MART and RM.

In the k-anonymity studies, reidentification risk is measured by minimum group size k. Our approach could allow a smaller group size as long as the sensitive value disclosure risks in the group is sufficiently low. Therefore, we use the average (instead of minimum) group size for measuring re-identification risk. To facilitate comparisons over multiple criteria, we adjusted group size parameters in RM and MART, as well as the α parameter in MART to produce regression trees with about the same average group size. The performances of the two techniques are then evaluated on the sensitive value disclosure risk and data utility measures, 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:

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

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 identified a target group t, he 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.

The classical linear regression model was used for the experiment. We applied a 10-fold cross-validation procedure, which used 90% of the data as the training set and the remaining 10% as the test set, and repeated the process 10 times, each using different training and test sets (LeFevre et al. 2008). 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, 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. For multiple responses, the *MAPE* measure 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. Since a 10-fold cross validation procedure was used, we report the average results over the 10 runs for *MAPE*.

Table 4. Results of Experiments

Data	Method	Average Group Size	RSD (%)	MAPE (%)
Offer	RM	10	29.31	5.68
	MART	10	33.34	4.59
	RM	20	31.36	7.61
	MART	20	36.33	5.03
Alcohol	RM	25	78.97	14.85
	MART	25	92.08	9.89
	RM	40	86.98	15.29
	MART	40	93.92	10.33

The results of the experiments are shown in Table 4. For each dataset, we compared RM and MART with two different sets of group sizes. It is clear that, for the same re-identification risk (i.e., the same group size), the *RSD* values with MART is larger than those with RM for all group sizes in both datasets, which indicates lower disclosure risk for the sensitive attribute values. This, we believe, can be explained by the use of the Δ-digression measure in MART for reducing the risk. For regression analysis, MART outperforms RM in all cases. This is likely due to the use of dynamic value-concatenation method in generalizing categorical QI attribute values, which is better in preserving data quality than the pre-defined generalization hierarchies.

Both algorithms ran very fast, completed procedures within a few seconds. They are much faster than the traditional k-anonymity algorithms (Sweeney 2002; and Samarati 2001). The runtimes for the two

algorithms were almost the same, which is expected because they use similar basic regression tree algorithms. The runtime results are thus not reported.

Conclusion and Extensions

Regression techniques have been widely used not only as a tool for analytics in business and public domains, but also as a research method in management, science and social science studies. Therefore, the regression attack problem we investigate is vitally important. We have presented a novel approach for protecting against sensitive value disclosure by such an attack. We have also proposed the value-concatenation method to limit identity disclosure risk. We have shown analytically that the proposed Δ-digression measure has some important properties that will serve to evaluate disclosure risk when multiple numeric sensitive attributes are targeted. Our experimental study demonstrates the effectiveness of the proposed approach.

In the proposed approach, the disclosure risk is considered only at the tree-pruning stage. However, the error-digression measure may also be used for splitting the data so that the disclosure risk can be considered at the tree-growing stage. A possible outcome of this strategy is the reduced prediction power of the regression trees built based on the data anonymized in this way. It would be interesting to investigate the potential of this alternative approach and compare it with the present approach from a risk-utility tradeoff perspective.

We plan to investigate alternative methods to mask the partitioned data in future. For example, we anonymize all QI attributes for simplicity. This may overly weaken the predictive power of the model built on the anonymized data. A better approach would be to apply more stringent anonymization for those QI attributes that do not appear in a regression tree (i.e., less useful in prediction) than those that appear.

We also plan to explore the idea of weighted value-concatenation method mentioned earlier. The simple weighted coding method suggested earlier would work well for data released for simple publishing purposes such as reporting summary statistics. It will be difficult to use for more sophisticated data mining and analysis such as regression and classification. One possible approach is to break up a record with concatenated attribute values into multiple weighted-records and then assign a weight accordingly to each "fractional" record. Some decision tree techniques use weighting methods to deal with missing values, which could be adopted for weighted concatenation.

The relative squared distance (*RSD*) defined in Equation (10), which is used to evaluate the sensitive-value disclosure risk, is not very intuitive. We will explore more intuitive measures for evaluating the risk. A possible alternative is to quantify the risk based on the notion of sensitive-value matching, on lines similar to record linkage.

Appendix

Proof of Lemma 1. Let M be the total number of subsets partitioned by the tree, and $n_t\ (t=1,...,M)$ be the number of records in node t. Consider any two responses Y_j and Y_k. Let $y_{ij}^t\ (i=1,...,n_t)$ be the value of Y_j in the ith 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:

$$\begin{aligned}&(y_{ij}^t - \overline{Y}_j)(y_{ik}^t - \overline{Y}_k) = \\ &(\overline{y}_j^t - \overline{Y}_j)(\overline{y}_k^t - \overline{Y}_k) + (\overline{y}_j^t - \overline{Y}_j)(y_{ik}^t - \overline{y}_k^t) + (y_{ij}^t - \overline{y}_j^t)(\overline{y}_{ik}^t - \overline{Y}_k) + (y_{ij}^t - \overline{y}_j^t)(y_{ik}^t - \overline{y}_k^t).\end{aligned} \tag{A1}$$

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:

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

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 $y_{ik}^{t'}$ $(i=1,...,n_{t'})$ 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 $(\ell=1,...,m)$ be the number of records in leaf ℓ. Let y_{ij}^{ℓ} $(i=1,...,n_\ell)$ be the value of Y_j in the *i*th 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 $\sum_{\ell=1}^{m}s_{jk}(\ell)$. Write 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'. Rearrange the terms in (A6), we have

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

So,

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

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,

$$\mathbf{D}(\ell')-\mathbf{D}(t)=\mathbf{b}, \quad \text{(A9)}$$

It follows from the same argument as in the proof of Lemma 1 that **b** is a form of scatter matrix and thus $|\mathbf{b}| > 0$. On the other side, $\mathbf{D}(\ell')$ and $\mathbf{D}(t)$ are both scatter matrices, which can be diagonalized into $\mathbf{X}\mathbf{\Lambda}_{\ell'}\mathbf{X}'$ and $\mathbf{X}\mathbf{\Lambda}_{t}\mathbf{X}'$, with determinant equal to $|\mathbf{\Lambda}_{\ell'}|$ and $|\mathbf{\Lambda}_{t}|$ respectively. Therefore,

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

References

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

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

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

Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. 1984. *Classification and Regression Trees*, Belmont, CA: Wadsworth.

De'ath, G. 2002. "Multivariate Regression Trees: A New Technique for Modeling Species–Environmental Relationships," *Ecology* (83:4), pp. 1105-1117.

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

Fu, Y., Chen, Z., Koru, G., and Gangopadhyay, A. 2010. "A Privacy Protection Technique for Publishing Data Mining Models and Research Data," *ACM Transactions on Management Information Systems* (1:1), Article 7, pp. 7:1-17:20.

Fung, B.C.M., Wang, K., and Yu, P.S. 2007. "Anonymizing Classification Data for Privacy Preservation," *IEEE Transactions on Knowledge and Data Engineering* (19:5), pp. 711-725.

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

Johnson, R.A., and Wichern, D.W. 2002. *Applied Multivariate Statistical Analysis*. Upper Saddle River, NJ: Prentice Hall.

Kenkel, D.S., and Terza, J.V. 2001. "The Effect of Physician Advice on Alcohol Consumption: Count Regression with an Endogenous Treatment Effect," *Journal of Applied Econometrics* (16:2) , pp. 165-184.

LeFevre, K., DeWitt, D.J., and Ramakrishnan, R. 2008. "Workload-Aware Anonymization Techniques for Large-Scale Datasets," *ACM Transactions on Database Systems* (33:3), Article 17, pp. 17:1-17:47.

Li, N., Li, T., and Venkatasubramanian, S. 2007. "*t*-Closeness: Privacy Beyond *k*-Anonymity and *l*-Diversity," in *Proceedings of the 23rd IEEE International Conference on Data Engineering*, Istanbul, Turkey, pp. 106-115.

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

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

Lindell, Y., and Pinkas, B. 2002. "Privacy Preserving Data Mining," *Journal of Cryptology* (15:3), pp. 177-206.

Lohr, S. 2010. "Netflix Cancels Contest after Concerns Are Raised about Privacy," *New York Times*, March 13, B3.

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

Morrison, D.F., 1990. *Multivariate Statistical Methods*, New York: McGraw-Hill.

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

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

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