ICIS 2019 Proceedings

Analytics and Data Science

A Decision Tree Approach for Assessing and Mitigating Background and Identity Disclosure Risks

Haifang Yang
Dalian University of Technology, yhf19880108@126.com

Mingzheng Wang
Zhejiang University, wangmzh@zju.edu.cn

Xiangpei Hu
Dalian University of Technology, drhxp@dlut.edu.cn

Xiaobai Li
University of Massachusetts Lowell, xiaobai_li@uml.edu

Yang, Haifang; Wang, Mingzheng; Hu, Xiangpei; and Li, Xiaobai, "A Decision Tree Approach for Assessing and Mitigating Background and Identity Disclosure Risks" (2019). *ICIS 2019 Proceedings*. 17.
https://aisel.aisnet.org/icis2019/data_science/data_science/17

A Decision Tree Approach for Assessing and Mitigating Background and Identity Disclosure Risks

Completed Research Paper

Haifang Yang
School of Economics and Management
Dalian University of Technology
Dalian, Liaoning, China
drhfyang@126.com

Mingzheng Wang
Department of Data Science and
Engineering Management
Zhejiang University
Hangzhou, Zhejiang, China
wangmzh@zju.edu.cn

Xiangpei Hu
School of Economics and Management
Dalian University of Technology
Dalian, Liaoning, China
drhxp@dlut.edu.cn

Xiao-Bai Li
Manning School of Business
University of Massachusetts Lowell
Lowell, MA 01854, U.S.A.
xiaobai_li@uml.edu

Abstract

The Facebook/Cambridge Analytica data scandal shows a type of privacy threat where an adversary attacks on a massive number of people without prior knowledge about their background information. Existing studies typically assume that the adversary knew the background information of the target individuals. This study examines the disclosure risk issue in privacy breaches without such an assumption. We define the background disclosure risk and re-identification risk based on the notion of prior and conditional probabilities respectively, and integrate the two risk measures into a composite measure using the Minimum Description Length principle. We then develop a decision-tree pruning algorithm to find an appropriate group size considering the tradeoff between disclosure risk and data utility. Furthermore, we propose a novel tiered generalization method for anonymizing data at the group level. An experimental study has been conducted to demonstrate the effectiveness of our approach.

Keywords: data privacy, decision trees, background disclosure, identity disclosure

Introduction

The recent developments in big data technologies have facilitated the sharing of personal data in numerous application areas, such as database marketing, healthcare data sharing, social media data analysis, and data sharing among supply-chain partners. However, sharing personal data for business use often raises concerns about personal privacy. Governments around the world have increasingly toughened the laws for privacy protection. European Union has recently introduced the General Data Protection Regulation (GDPR) (EU 2016), which requires data owner organizations obtain consent from individual consumers when their identifiable data are to be shared with a third party. Despite the tightened privacy laws, repeated privacy invasion cases continue to cause outrage among the public. Recently, the social media giant Facebook acknowledged that over 87 million users were affected in a data privacy scandal carried out by Cambridge Analytica, a political consulting firm. According to news reports (Frier 2018;

Romm 2019), Cambridge Analytica had accessed personal information of these 87 million users without their consent, using Facebook's friends connections from about 270,000 users who participated in a study conducted by the firm. Based on the data, Cambridge Analytica was able to build the voting-related profiles for those users and create personalized ads for them, which could potentially influence their voting decisions. This unprecedented data harvesting and abuse has raised urgent privacy concern and caused backlash from the public. As a result, Cambridge Analytica has filed for insolvency proceedings and closed its operations, and Facebook is facing possible multi-billion dollars fine for its privacy violation (Romm 2019).

Two aspects in the Facebook/Cambridge Analytica case are worth noting. First, the privacy adversary attempted to obtain information from a massive number of people, rather than from a few focused subjects. Second, the adversary did not have much background information about the individuals to be targeted; instead, the background information was acquired as part of privacy disclosure process. The background information here refers to non-sensitive information (e.g., Facebook users' demographic and social-economic attributes), which is different from identity information, as well as sensitive information that the adversary ultimately pursues (e.g., political ideology or voting stance). The two aspects above are also common in many marketing applications with privacy concerns, where a large number of consumers are targeted and their background information is acquired during the marketing process (Duhigg 2012). These aspects have not been specifically studied in the data privacy literature. Often, research efforts are focused on relatively small number of data subjects (i.e., the subjects having highest identity disclosure risks). Also, many data privacy approaches assume that the adversary already knew the background information about the target individuals.

A large body of research in data privacy has focused on the risk of *re-identification*, which concerns how likely an adversary can match a record in a de-identified dataset to a target individual using some combination of the individual's non-sensitive attributes, called *quasi-identifier* (QI) attributes, such as age, gender and zip code (Sweeney 2002). To prevent re-identification, Sweeney (2002) and Samarati (2001) propose the k-anonymity model, which requires each record should be indistinguishable from at least $k-1$ other records with respect to the QI values in the released dataset. However, k-anonymity model and many other related privacy models assume that the adversary knew the QI values of the target individuals. This may not be true in practice. Indeed, Dwork (2006, 2011) criticize that most of anonymization approaches need assumptions on the adversary's auxiliary information about the target individuals. To address this problem, Dwork (2006, 2011) propose the differential privacy model, which does not make any assumption about an adversary's prior knowledge. However, differential privacy is typically used for database query applications or aggregate data analysis. It is not designed for releasing or sharing an entire dataset of individual records, because there would likely be huge information loss in the released dataset in order to satisfy the differential privacy requirements (Machanavajjhala et al. 2008).

Our work uses anonymization framework to address privacy disclosure problem in the context of data sharing. Different from k-anonymity, we do not assume that an adversary knew the background information of the target individuals. We consider a privacy attack situation where an adversary has access to a de-identified dataset and attempts to acquire personal information about a large number of individuals in the dataset. We study the problem by considering two types of privacy disclosures. The first type concerns how feasible or likely for an adversary to find the background information about the individuals in the dataset, assuming the adversary does not have the background information. We call this a *background disclosure* problem. To the best of our knowledge, the background disclosure problem has not been formally investigated in literature. After background disclosure, the adversary proceeds to pursue the second type of disclosure, which is to identify individual targets based on their background information. This second type of disclosure is commonly referred to as *re-identification* in literature.

In this paper, we propose a novel approach to assess and mitigate background disclosure and re-identification risks in a released dataset. We use a decision tree technique to partition the dataset into groups/subsets. The background disclosure risk and re-identification risk for the individuals in a group are measured based on the attribute values that define the group and the size of the group, respectively. Accordingly, the two risk measures can be described with prior probability and conditional probability, respectively. We integrate the two risk measures into a composite measure based on the notion of Minimum Description Length (MDL) (Rissanen 1978). We then develop a decision-tree pruning algorithm to find an appropriate group size considering the tradeoff between disclosure risk and data utility. Furthermore, we propose a novel tiered generalization method for anonymizing data at the group level, which takes advantage of the tiered structure of the decision tree to effectively limit background disclosure and also reduce information loss due to generalization.

This work makes contributions to the data privacy literature in several aspects:

1. We consider privacy attacks targeting a large number of people, where the adversary does not have prior knowledge about these people. We introduce the notion of background disclosure inherent in this problem.
2. We propose a novel privacy measure that integrates background disclosure and re-identification risks based on the MDL principle. We develop a decision-tree algorithm that uses this measure in partitioning data into groups for anonymization.
3. We propose a new tiered generalization method that takes advantage of the tiered structure of the decision tree to effectively mitigate the background disclosure and also reduce information loss due to generalization.

Related Work

A main stream of data privacy research is based on a popular privacy model called k-anonymity (Samarati 2001; Sweeney 2002). Given a dataset to be shared, the k-anonymity approach first divides the data into a number of groups, called *QI groups*, with at least k records in each group. The QI values are then anonymized using the same generalized values within a QI group so that a record is indistinguishable from at least $k-1$ other records with respect to the QI values. Assuming that the adversary knew the QI values of the target individuals but not the other attribute values, the re-identification risk for any individual in a k-anonymized dataset is $1/k$ at most. A series of refined models, such as l-diversity (Machanavajjhala et al. 2007) and t-closeness (Li et al. 2007), are proposed to overcome k-anonymity's limitations in handling disclosures caused by sensitive attributes. These popular models typically assume that the adversary knew the target individual's record is in the released data and thus focus on how to reduce the probability of uniquely identifying the right record. On the other hand, another privacy model called δ-presence considers the scenario where knowing the presence of target individuals in the released data already constitutes a privacy disclosure; it thus focuses on how to protect against this presence disclosure (Nergiz et al. 2007; Nergiz and Clifton 2010). This stream of studies all assume that the adversary knew the QI values of the target individuals. They do not address the background disclosure problem concerned in this study.

Another stream of studies represented by Dwork (2006, 2011) criticize that it is hard or even unrealistic to make assumption on the adversary's prior knowledge about the target individuals. Dwork and her colleagues propose the differential privacy principal that does not require the assumption. Differential privacy is appealing in that it guarantees that addition or deletion of any individual's record does not significantly affect the result of data analysis. However, differential privacy is mainly applicable to database query or aggregate data analysis, and does not work well for dataset release. Differential privacy treats data utility as a secondary objective, considered only after privacy criteria are rigorously satisfied. As a result, it causes huge information loss when applied to releasing a dataset at the individual record level (Machanavajjhala et al. 2008). Our work is in the context of dataset sharing, where differential privacy is not very effective. But in line with differential privacy, we do not make assumption on the adversary's prior knowledge about the background information of the target individuals.

We should point out the two differences between the term "quasi-identifier" (QI) in literature and the term "background information" used in this paper. First, the QI attributes are predefined by the data owner organization while the attributes representing background information are not predefined. Second, quasi-identifier is assumed known to the adversary but background information is not; the adversary needs to acquire the background information in the privacy disclosure process. Usually, the background information can be obtained in publically available data (e.g., voter registration list) or acquired from commercial data vendors (FTC 2014), where the identities of the individuals are also included. We call this type of data *reference data* in this paper. Note that the reference data do not include the sensitive attributes (e.g., purchase transaction data or financial information) that the adversary attempts to find.

Instead, the sensitive information is contained in the de-identified data released by the data owner organization.

Several studies have exploited tree-based partitioning methods to generate QI groups (LeFevre et al. 2006; LeFevre et al. 2008; Fu et al. 2010; Li and Sakar 2009; Li and Sakar 2014). LeFevre et al. (2006) used kd-trees to cluster data into groups to reduce the information loss caused by generalization on QI values. LeFevre et al. (2008) adopt decision trees to split data into groups for privacy protection while ensuring the prediction accuracy. Fu et al. (2010) focus on the conditions to preserve decision tree models while protecting data privacy. Li and Sarkar (2009, 2014) investigate the sensitive attribute disclosure problems when classification and regression trees are used for privacy breach, respectively. None of these studies, however, has considered the background disclosure problem. Li and Sarkar (2009) observe that tree-based approaches can be used to identify a large number of vulnerable individuals. We have mentioned earlier that the problem we study involves disclosure of private information on a large number of people. Hence, our approach addresses the issue using decision trees too.

Generalization is one of the most widely used methods for anonymizing data. In traditional approaches such as *k*-anonymity, *l*-diversity and *t*-closeness, individual QI attribute values of the records in a QI group are generalized into the same value to make them indistinguishable. While this uniform generalization method anonymizes the records well, it also causes considerable information loss. Various alternative approaches for anonymizing data have been proposed to reduce information loss, including Anatomy (Xiao and Tao 2006), *k*-type anonymization (Gionis et al. 2008), and ring generalization (Wong et al. 2010). With respect to the problem we study, it can be shown that uniform generalization also makes background disclosure easier. In order to mitigate both background disclosure risk and information loss, we propose a tiered generalization approach.

An Illustrative Example

Consider an online retail company that has large sets of customer and transaction data. The company shares the data with some consulting firms based on some data sharing agreements. To comply with GDPR, the identities of the customers are removed but background and transaction data, such as age, gender, marital status, and purchase transactions, are included. The consulting firms may use the data to develop analytical models for serving the retail company's business. Some of these firms may have or are able to obtain the reference data that include these customers' identity and background information, but not transaction information. These consulting firms can then match the customers in the de-identified dataset provided by the retail company with those in the reference data using the background information common in both data sources, producing individually identifiable data that may be private and sensitive. The consulting firms can then provide these identifiable data to their other clients for mass marketing and promotion. Below, we use a hypothetical dataset to illustrate how the decision tree technique can be used by a third party like the consulting firms to determine the background information needed to re-identify the de-identified data, as well as the chance for re-identifying the individuals.

Consider a dataset that contains 14 records of customers with three background attributes, Age, Gender (female or male), Marital Status (married or not married), and a class attribute (yes or no) indicating the customer has bought a certain medical product or not. Figure 1 shows a decision tree built on this dataset. The records in the dataset are partitioned into five leaves (in rounded rectangle). In each node, the first row shows the node ID, followed by the distribution of class values at this node, expressed as {number of Yes', number of No's}. The second row indicates the classification decision and the number of errors (i.e., misclassified records) with the decision. The classification decision at each node is determined by the majority class value occurred in this node, and the classification errors equals to the number of records with minority class values. For example, Node 4 contains two records, both having a Yes for the class value; so the classification decision is Yes with zero error. At each edge, the splitting criterion, which is calculated by some metrics such as information gain or Gini index in most existing tree building models, indicates the background attribute values used to assign the records to the child nodes. The background attribute values of the records within a leaf will be generalized so that each leaf corresponds to a QI group in the *k*-anonymity model.

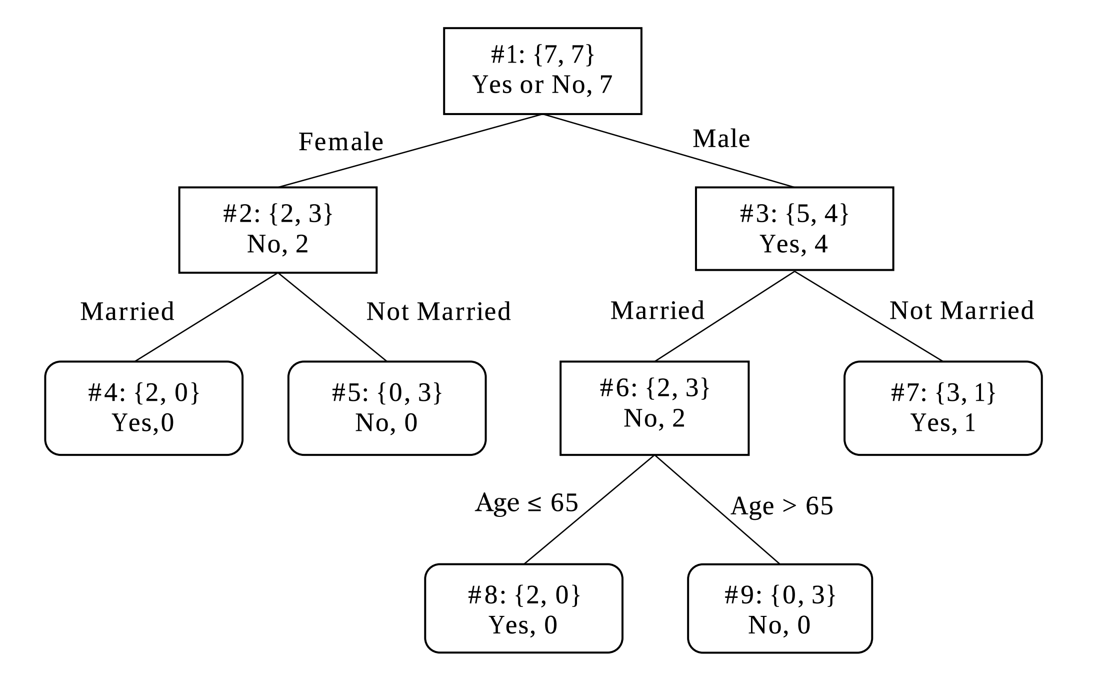

Figure 1. An Example of Decision Tree

Using this tree, an adversary who has no predefined targets can easily determine which individuals' background information is easier or harder to obtain. For example, in order to re-identify the two records in node 4, the adversary needs to acquire a reference dataset that contains gender and marital status attributes. To re-identify the two records in node 8, however, the adversary needs to know one more attribute, age, in addition to gender and marital status. Therefore, the records in node 4 have a greater background disclosure risk than those in node 8. On the other hand, both nodes 4 and 8 include two records. Based on the *k*-anonymity principle, the re-identification risk for the records in these two nodes are the same, both equaling 1/2. Considering both background disclosure and re-identification, it is easy to see that the records in node 4 have a higher disclosure risk than those in node 8.

Background disclosure risk depends not only on the number of attributes in the background information, but also on the size or range of the attribute values. For example, nodes 8 and 9 involve the same three attributes (Gender, Marital Status, and Age). The gender and marital status for the two nodes are the same. The only difference between them is that the records in node 8 are aged 65 years or younger while those in node 9 are older than 65 years. Given the age distribution in any typical population, it would be easier to get reference data to cover people aged 65 years or younger than that to cover adult people older than 65 years. That is, background disclosure for those in node 8 is easier (i.e., risk is higher) than those in node 9. On the other hand, re-identification risk for those in node 8 is also higher (1/2) than those in node 9 (1/3). Therefore, the records in node 8 have higher combined disclosure risk than those in node 9.

Disclosure Risk Measures and Utility Measure

The above example shows that the background information necessary for an adversary to re-identify an individual in a group/subset can be represented by the attributes used to split the data to form the group. In general, the more attributes required to specify the background information for the group, the more difficult to acquire the required background information. In a decision tree, this is illustrated as a longer path from the root to the leaf. In addition to the number of attributes, the background information is also related to the values of the attribute involved. In general, the larger the range or size of the attribute values, the easier to obtain the reference data covering these values. Note that smaller range attribute values may result in higher re-identification risk. We will discuss this issue later when analyzing the re-identification risk.

The second aspect above is related to the domain of an attribute, which refers to all the values the attribute contain. The domain size for a categorical attribute is the number of distinct values in the dataset; for a numeric attribute it is the value range of the attribute in the dataset. For example, the domain size for Gender is two in the root node and becomes one in the other node (either Female or Male). The domain size for Age in the root node is the entire Age value range in the dataset and it reduces to "> 65" in node 9. For an attribute j, let V_j and $V_j(t)$ be its domain size in the full dataset and in a node t, respectively. Assume that data values follow a uniform distribution, which is an often used assumption in the data privacy literature (Machanavajjhala et al. 2007; Xiao and Tao 2006). If based on attribute j only, the probability that an individual is in node t is $P_j(t) = V_j(t)/V_j$. If node t is obtained by splitting m attributes, then the probability an individual is in node t is $\prod_{j=1}^{m} P_j(t)$. This probability can be used to measure the degree of difficulty or easiness for an adversary to obtain the relevant background information for the individuals in node t. Note that the probability product above assumes the attributes are mutually independent, which may not be the case for some attributes (e.g., age is independent of gender, but not of marital status). However, the measure will not be used to actually estimate the probabilities; rather, it is to be used to compare the degree of difficulty to obtain the background information in different nodes. So, the assumption is reasonable in this sense.

Another, perhaps more usual, way to describe the background information is to use the information measure in the information theory literature (Shannon 1948). In our context, the information content for a node t with background information probability $P_j(t)$ is $-\log_2 P_j(t)$. Note that the larger the $P_j(t)$ value, the smaller the $-\log_2 P_j(t)$ value. Thus, a smaller value in information content means the information is more general while a larger value means the information is more specific. For the m attribute case, the information content corresponding to $\prod_{j=1}^{m} P_j(t)$ is $-\sum_{j=1}^{m} \log_2 P_j(t)$. We define background information disclosure risk based on this measure.

Definition 1. For a node t obtained using m attributes, the *background information granularity* (BIG) is defined by

$$BIG(t) = -\sum_{j=1}^{m} \log_2 P_j(t) \tag{1}$$

Note that $BIG(t)$ is a positive number because $\log_2 P_j(t)$ is negative. It follows from (1) that the larger the m value and/or the larger the $-\log_2 P_j(t)$ value, the larger the $BIG(t)$ value. In other words, the more attributes and/or more detailed attribute information required to identify node t, the larger the $BIG(t)$ value. This means a larger $BIG(t)$ value indicates a smaller background disclosure risk for the individuals in node t. Lemma 1 below states that the background disclosure risk of node t is always larger than that of its child node.

Lemma 1. Let t_c be a child node of node t. Then

$$BIG(t) < BIG(t_c) \quad \forall t \tag{2}$$

The proof is straightforward because a child node is formed by splitting on the parent node using either an additional attribute or a smaller domain size of the same attribute.

After the background information in a node t is acquired, the adversary proceeds to re-identify the individuals in node t. Based on the k-anonymity principle (Sweeney 2002), the values of each background attribute for all individuals in node t will be generalized into the same value. Therefore, the probability for correctly re-identifying an individual i in node t ($i \in t$) is $P(i|t) = 1/n_t$, where n_t is the number of records in node t. Since this probability is conditioned on getting the background information for node t, we rename this measure as *conditional re-identification risk*. The corresponding information measure is $-\log_2 P(i|t) = -\log_2\left(\frac{1}{n_t}\right) = \log_2 n_t$, formally defined below.

Definition 2. Given a node t with n_t records, the *information for conditional re-identification* (ICR) of a record in node t is

$$ICR(t) = \log_2 n_t \tag{3}$$

Clearly, $ICR(t)$ increases with n_t. So, the larger the number of records in a node, the more information needed for conditional re-identification, and the smaller the conditional re-identification risk.

Lemma 2. Let t_c be a child node of node t. Then

$$ICR(t) > ICR(t_c) \ \forall t \tag{4}$$

The proof is straightforward because a child node has a smaller number of records and thus a smaller ICR value than its parent node.

Statistically, the background disclosure risk $P(t)$ can be regarded as the prior probability for an adversary to find the background information for the individuals in node t, while the re-identification risk $P(i|t)$ can be viewed as the conditional probability for the adversary to correctly identify a record in node t. As we pointed out earlier, existing related work assumes that background information is known and thus the prior probability for background disclosure is 100%. We do not make this assumption. We have also defined two information-based disclosure risk measures, BIG and ICR, corresponding to the prior probability and conditional probability respectively. To consider the two disclosure risks jointly, one approach is to construct a posterior probability measure by joining the prior probability and conditional probability. However, it is computationally more convenient to combine the two information-based the measures based on the well-known Minimum Description Length (MDL) principle (Rissanen 1978).

Given a set of hypotheses $\mathcal{H}$ for describing data D, let $L(H)$ be the length or amount of information required to describe a hypothesis $H (H \in \mathcal{H})$ and $L(D|H)$ be the length or amount of information required to describe the data D given the hypothesis H. The MDL principle states that the best hypothesis is the one that minimizes the total description length $L(H) + L(D|H)$. In our problem context, H represents the background information and $D|H$ represents re-identification given the information in H. Based on this interpretation, we introduce a measure for assessing the combined background disclosure and re-identification risks.

Definition 3. The *total information for disclosing individuals* (TIDI) in node t is the sum of node t's background information granularity (BIG) and its information for conditional re-identification (ICR), written as

$$R(t) = -\sum_{j=1}^{m} \log_2 P(t) + [-\log_2 P(i|t)] = -\sum_{j=1}^{m} \log_2 \frac{V_j(t)}{V_j} + \log_2 n_t \tag{5}$$

The TIDI measure considers the combined information required for both background and re-identification disclosures. The smaller the TIDI value, the smaller the required combined information and the higher the combined disclosure risk. In Figure 1, for example, $R(2) = -\log_2 \frac{1}{2} + \log_2 5 = 3.322$, and $R(6) = \left(-\log_2 \frac{1}{2} - \log_2 \frac{1}{2}\right) + \log_2 5 = 4.322$. So, the five records at node 2 have higher combined disclosure risk than the five records at node 6. This is intuitively understandable because the background information for node 2 is simpler than that for node 6 while the re-identification risks for both nodes are the same.

As the value of $R(t)$ decreases, the amount of information required for identifying the individuals in node t becomes smaller, which means the disclosure risk increases for the individuals in node t. As discussed earlier, when a decision tree grows, the BIG component of the $R(t)$ measure tends to be lager while the ICR component tends to be smaller. Therefore, it is interesting to analyze how $R(t)$ measure changes as a decision tree grows; i.e., when a node is split into a branch.

Definition 4. A *branch* B_t is a subsection of a decision tree that includes an internal node t and all its leaves and non-leaf descendant nodes. If t is the root of the tree, then the branch is the entire tree.

In Figure 1, for example, B_2 includes nodes 2, 4 and 5; and B_3 includes nodes 3, 6, 7, 8 and 9.

In data privacy research and practice, it is a common practice to measure disclosure risk based on the maximum risk instead of average risk (Sweeney 2002; Fung et al. 2010; Xiao and Tao 2006). Thus, in a branch, the leaf with the smallest TIDI value (i.e., the maximum disclosure risk) should have the highest priority for protection. The TIDI measure for a branch is defined accordingly.

Definition 5. Let t_l $(l = 1, \dots, g)$ be the g leaves in branch B_t. The TIDI measure of branch B_t is defined by the leaf having the minimum TIDI value, i.e.,

$$R(B_t) = \min_{l=1,\dots,g} R(t_l) \tag{6}$$

Take B_2 in Figure 1 as an example. The two child nodes, #4 and #5, have the same value for the first (BIG) component of the TIDI measure, but node 4 has a smaller value for the second (ICR) component of TIDI than node 5 because node 4 has a fewer number of records than node 5. Therefore, $R(B_2)$ is defined by $R(4)$.

Lemma 3. The TIDI value of node t is always larger than or equal to that of branch B_t. That is,

$$R(t) \geq R(B_t) \ \ \forall t \tag{7}$$

Proof. Continuing the notation used earlier, let $\beta_l = \prod_{j=1}^{m} P_j(t_l) \,/\, \prod_{j=1}^{m} P_j(t)$ where t_l $(l = 1, \dots, g)$ are the g leaves in branch B_t. It is easy to see that $\sum_{l=1}^{g} \beta_l = 1$ and $n_t = \sum_{l=1}^{g} n_{t_l}$, where n_{t_l} is the number of records in t_l. Using proof by contradiction, suppose Equation (7) does not hold; that is, there exist a node t such that $R(t) < R(B_t)$. Then

$$\begin{aligned}
R(B_t) - R(t) &= \min_{l=1,\dots,g} R(t_l) - R(t) \\
&= \min_{l=1,\dots,g} \left(-\sum_{j=1}^{m} \log_2 P_j(t_l) + \log_2 n_{t_l} \right) - \left(-\log_2 \prod_{j=1}^{m} P_j(t) + \log_2 n_t \right) \\
&= \min_{l=1,\dots,g} \left(-\log_2 \left[\beta_l \prod_{j=1}^{m} P_j(t) \right] + \log_2 n_{t_l} + \log_2 \prod_{j=1}^{m} P_j(t) - \log_2 n_t \right) \\
&= \min_{l=1,\dots,g} \left(-\log_2 \beta_l - \log_2 \prod_{j=1}^{m} P_j(t) + \log_2 \prod_{j=1}^{m} P_j(t) + \log_2 n_{t_l} - \log_2 n_t \right) \\
&= \min_{l=1,\dots,g} \left(-\log_2 \beta_l + \log_2 \frac{n_{t_l}}{n_t} \right) = \min_{l=1,\dots,g} \log_2 \left(\frac{n_{t_l}}{n_t \beta_l} \right) > 0.
\end{aligned}$$

In other words, $\forall l \in \{1, \dots, g\},\ n_{t_l} > n_t \beta_l$. Then,

$$n_t = \sum_{l=1}^{g} n_{t_l} > \sum_{l=1}^{g} (n_t \beta_l) = n_t \sum_{l=1}^{g} \beta_l = n_t$$

This simplifies to $n_t > n_t$, which is a contradiction. Therefore, Equation (7) holds. □

Lemma 3 suggests that growing a decision tree will cause the TIDI value to decrease and the combined disclosure risk to increase. On the other hand, pruning a decision tree will cause the combined disclosure risk to decrease. Therefore, from a disclosure-control perspective, it is desired to have a smaller tree, which results in generalizing attribute values over larger groups.

Decision trees are typically used for classification analysis, and an overly small tree will cause larger classification errors. Given a decision tree, its classification accuracy is measured by the sum of the numbers of majority class (i.e., the class whose value occurs most frequently) at each leaf, while the classification error is determined by the total number of all minority classes. To discuss the tradeoff between classification accuracy and disclosure control, we define classification errors for a node and a branch, respectively.

Definition 6A. Let n_t and n_t^{major} be the number of the total records and the number of the records with the majority class in node t, respectively. The *node error* of t is equal to the number of all the misclassified records at node t, i.e.,

$$E(t) = n_t - n_t^{major} \tag{8}$$

Definition 6B. The *branch error* of branch B_t is the sum of the node errors from all leaves of B_t $(t_l, l = 1,2, \dots, g)$; that is,

$$E(B_t) = \sum_{l=1}^{g} E(t_l) \tag{9}$$

For example, in Figure 1, $E(2) = 5 - 3 = 2$, and $E(B_2) = E(4) + E(5) = 0 + 0 = 0$. Similarly, $E(3) = 4$, and $E(B_3) = 1 + 0 + 0 = 1$.

It is well-known from the decision tree literature (Breiman et al. 1984; Quinlan 1993; Witten et al. 2017) that a split always reduces training errors. In other words, the branch error of B_t are always smaller than or equal to the node error of t, i.e., $E(B_t) \leq E(t)$.

Next, we propose a measure, called error-risk ratio, that considers the tradeoff between disclosure risks (including background disclosure and conditional re-identification risks) and data utility (as represented by classification errors).

Definition 7 (Error-Risk Ratio). The *error-risk* (ER) *ratio* for an internal node t is defined as

$$\omega_t = \frac{R(t) - R(B_t)}{E(t) - E(B_t)} \tag{10}$$

Here, ω_t (≥ 0) represents the marginal decrease in disclosure risk per unit training error increment due to pruning off branch B_t into leaf node t. A larger ω_t value suggests a larger reduction in disclosure risk given the same increment in training errors, or a smaller increase in errors for the same reduction in disclosure risk. So, a larger ω_t value is preferred in selecting a branch for pruning. We describe next how this tradeoff measure is used in the proposed tree pruning algorithm.

Proposed Tree Pruning Algorithm

Our goal is to find a best tradeoff among the disclosure risks and classification accuracy by constructing a decision tree with an appropriate size. We approach this goal through a tree pruning algorithm based on the maximum ER ratio criterion. The steps of the algorithm proceed in an iterative process. At each iteration, it calculates the value of ω_t for every branch in the current tree. The branch that has the maximum value of ω_t is pruned to a leaf. The process continues until a prespecified minimum leaf-size criterion is satisfied. The parameter k sets the minimum number of records required in a leaf, which is used to control the re-identification risk of a group. It also has an impact on background information granularity as well as classification error. The proposed algorithm, called error-risk pruning (ERP) algorithm, is described in Figure 2.

```
Input: $T$, an unpruned classification tree; $k$, the minimum number of records in a leaf.
Output: A pruned tree $T'$.
1. For every internal node $t$, calculate the $\omega_t$ based on Equation (10).
2. Select the node having the largest $\omega_t$ value and denote it by $t^*$. Let $\ell^*$ be the leaf of branch $B_{t^*}$
   having the minimum number of records and $n_{\ell^*}$ be the number of records in $\ell^*$.
3. (i) If $n_{\ell^*} < k$, then prune branch $B_{t^*}$ into a leaf;
   (ii) Otherwise, let $t^*$ be the internal node with the next largest $\omega_t$ and go to (i). Stop pruning if
   no node can be assigned to $t^*$.
4. Repeat steps 1-3 for the last pruned tree until all leaves satisfy the minimum size criterion.
```

Figure 2. The Error-Risk Pruning (ERP) Algorithm

Like usual decision tree algorithms, the ERP algorithm is computationally quite efficient, as stated in Lemma 4 below.

Lemma 4. The worst-case time complexity of the ERP algorithm is of order $O(M^2)$, where M is the number of internal nodes of the tree.

Proof. Consider an unpruned tree T consisting of M internal nodes. The initial computations on ER ratios for all internal nodes are of order $O(M)$. Identifying the node with the largest ER ratio to prune off also takes a time of order $O(M)$. In the worst-case scenario, to prune off all internal nodes in the tree, there will be at most $M-1$ iterations for these operations (each subsequent iterations involving fewer number of calculations and comparisons). Therefore, the overall time complexity of the ERP algorithm is $O(M)+\mathrm{O}(M-1)+\cdots+O(1)$, which is $O(M^2)$. □

Next, we explain the ERP procedure using the example in Figure 1 (by considering the tree as a full tree for illustration purpose). Suppose that the range of Age values is from 22 to 80. The calculations for the node and branch TIDI values, and the ER ratio at internal node 6 are provided as follows:

$R(6)=\left(-\log_2\frac{1}{2}-\log_2\frac{1}{2}\right)+\log_2 5=4.322,$

$R(B_6)=\min\left\{\left(-\log_2\frac{1}{2}-\log_2\frac{1}{2}-\log_2\frac{65-22}{80-22}\right)+\log_2 2\,,\left(-\log_2\frac{1}{2}-\log_2\frac{1}{2}-\log_2\frac{80-65}{80-22}\right)+\log_2 3\right\}=3.432,$

and $\omega_6=\frac{4.322-3.432}{2-0}=0.445.$

Similarly, we have the following error-risk ratios for all the other internal nodes:

$\omega_1=0.135, \omega_2=0.161, \omega_3=0.246.$

Suppose that $k=3$. Since branch B_6 has the largest ER ratio and meets the condition in step 3(i), it is pruned first. Note that branches B_2 and B_6 have the same classification errors. Our algorithm selects B_6 for pruning because the reduction in combined background disclosure and re-identification risk by pruning off B_6 is larger than that by pruning off B_2.

After pruning off B_6, the process continues for this pruned tree until all leaves of the tree satisfy the minimum size criterion.

Tiered Generalization

After pruning is completed, the data are partitioned into several groups, each corresponding to a leaf. Then, the background attribute values within each group are generalized. Usually, generalization requires that the QI attribute values of all records within a group should be the same (hereafter called *uniform generalization*). This uniform generalization could cause large amount of information loss (Xiao and Tao 2006; Li and Sarkar 2014). As far as background disclosure is concerned, uniform generalization also makes background information easier to acquire because generalized data tends to contain fewer background attributes and larger domain size of the attributes than the original data. To address this problem, we propose a new *tiered generalization* method that takes advantage of the tiered structure of a decision tree and generalizes the records in a group with different attribute values while still satisfying the principle of k-anonymity.

Record No.	Original data		Tiered Generalization		Uninform Generalization		Class Attribute
	Age	Marital Status	Age	Marital Status	Age	Marital Status	
1	57	Married	[42-61]	{Married, Not Married}	[29-61]	*	Yes
2	61	Married	[42-61]	{Married, Not Married}	[29-61]	*	Yes
3	42	Not Married	[29-61]	{Married, Not Married}	[29-61]	*	No
4	29	Not Married	[29-42]	Not Married	[29-61]	*	No
5	38	Not Married	[29-42]	Not Married	[29-61]	*	No

Table 1. Example of Tiered Generalization and Uniform Generalization with *k* = 3

Table 1 shows the original data and the anonymized data using tiered and uniform generalization methods, with $k = 3$, on the five records in nodes 2 in Figure 1. We add the Age value for each record that was not shown in Figure 1, but omit the Gender attribute since it is the same for all the records. Clearly, the Age and Marital Status values of any record in the original data can be matched to those of at least three records in the data anonymized by the tiered generalization. Therefore, the records with tiered generalization satisfy 3-anonymity. Next, we use this example to illustrate the computational steps of the tiered generalization.

The proposed method contains three main steps for a group formed by the ERP pruning:

Step 1: *Subgrouping*. The records in a group t formed by the ERP pruning are partitioned into subsets $t_i (i = 1, \ldots, l)$ based on their groups/leaves before pruning. In Table 1, the five records are divided into two subsets based on their member groups before pruning, with records 1 and 2 (in node 4 originally) in one subset and records 3, 4 and 5 in the other subset (in node 5 originally).

Step 2: *Tiered nearest neighbor searching*. Let n_i be the number of records in a subset t_i. If $n_i < k$, the tiered generalization needs to find $k - n_i$ records from the tiered neighboring subsets of t_i. The tiered neighbors of t_i is determined based on the tier structure of the decision tree. The algorithm first searches the records from the nodes that share the same parent node with t_i. If the number of records in the neighboring nodes is smaller than $k - n_i$, then the search goes to the nodes that share the same grandparent node of t_i. This process continues until the neighboring nodes have at least $k - n_i$ records. At this time, if the neighboring nodes have more than $k - n_i$ records, then the $k - n_i$ records that are closest to the centroid of t_i are selected based on the Euclidean distance measure. These $k - n_i$ records are called *shared records*, while the other records are called *non-shared records*. The shared records need to be found for each t_i having $n_i < k$.

In table 1, the first subset (based on node 4) has only two records. In order to satisfy 3-anonymity, the subset needs one shared record from the tiered nearest neighboring subset (node 5). Based on the Euclidean distance, record 3 is selected as the shared record.

Step 3: *Generalizing with sharing*. The background attribute values for the non-shared records will be generalized using the value domain of the subset. The value domain for a numeric attribute is the value range of the attribute in the subset. For a categorical attribute, the value domain includes all distinct values of the attribute in the subset. The background attribute values for the shared records will be generalized using the combined value domain of the subset itself and all of the related neighboring subsets. Continuing with Table 1, the Age and Marital Status values of the non-shared records 1 and 2 are generalized using their value domains in the first subset. The non-shared records 4 and 5 are generalized similarly using their value domains in the second subset. The shared record 3 is generalized using the combined value domains from both subsets.

Similar to uniform generalization, the tiered generalization replaces numeric attribute values using the range value in the relevant subset(s). For a categorical attribute, the tiered generalization replaces the values of the records in a subset by a set-valued form that enumerates all the distinct values of the attribute in the relevant subset(s). For example, in Table 1, the values of Marital Status for the records in the first subset are generalized into {Married, Not Married}. For another example, suppose the Age values are also considered categorical or discrete, then the generalized values for the above five records would be in the set-valued form as follows: {42, 57, 61} for records 1 and 2, {29, 38, 42} for records 4 and 5, and {29, 38, 42, 57, 61} for record 3.

It is observed that using the uniform generalization, the five records have to be generalized into the same values for Age and Marital Status in order to satisfy 3-anonymity. Clearly, the tiered generalization causes much smaller information loss in the data than the uniform generalization. In general, given the same k-anonymity requirement, the tiered generalization typically results in better data quality than the uniform generalization.

The tiered generalization procedure is computationally very efficient. The most time-consuming operation is Step 2 (tiered nearest neighbor searching) while the time for the other two steps are negligible. For a leaf node t with n_t records, finding the nearest records takes $O(kn_t)$ time. This search operation needs to be done for all L leaves. So, the total complexity for the proposed method is of $O(kn_tL) = O(kN)$, where N is the total number of records in the entire dataset.

Experimental Evaluation

We have conducted an experimental study on two real-world datasets that are commonly used as benchmark datasets in data-mining research and also well-documented as having some privacy implications (Dua and Graff 2019). The performances are evaluated on re-identification and background disclosure risks, as well as data utility for classification analysis. The datasets are described as below.

The first dataset, *Contra*, is taken from a National Indonesia Contraceptive Prevalence Survey (Lerman et al. 1991). It consisted of 1,473 married women's records with 10 attributes, five categorical and five numeric, including contraceptive method choice (class attribute, with three categories), wife's age, wife's education, husband's education, number of children, occupation, religion, etc.

The second dataset, *Adult*, is a census dataset extracted from a census bureau database (Kohavi 1996). After removing the instances with too many unknown values, the dataset contains 45,222 records, with seven categorical and seven numeric attributes such as salary, gender, age, race, education, occupation, work class, capital gain, and so on. This data has been used to build classification models to predict whether a person makes over $50,000 a year, where the salary is the class attribute with binary categories.

We compare our proposed method with the Mondrian method proposed by LeFevre et al. (2008), which is, to our knowledge, the only existing tree-based data privacy research that considers identity disclosure as well as the classification utility. Most of recent studies on privacy problems using decision trees are quite different from our research and thus are not suitable for comparison purposes. For example, Li and Sarkar (2009) consider attribute disclosure and alter the data by swapping sensitive values, whereas Li and Sarkar (2014) addresses a regression problem. Neither of them considers identity disclosure problem. There are two critical differences between the proposed method and Mondrian: (1) the proposed method considers both background disclosure and re-identification risks and integrates the two risk factors into a combined measure for partitioning and anonymizing data, while Mondrian only focuses on re-identification risk by requiring the anonymized data to satisfy k-anonymity; and (2) the proposed method adopts the tiered generalization method while Mondrian uses the uniform generalization method. To separate the performance difference due to the proposed ERP pruning method and that due to the tiered generalization, we also included an "ERP only" variation for the experiment, which implements the ERP pruning but not the tiered generalization (the uniform generalization is used instead).

For simplicity, we assume that all non-class attributes are background attributes and subject to generalization. The class attribute is considered as the sensitive attribute. Following the common practice in the framework of data anonymization, the values of the sensitive (class) attribute are not changed.

In data anonymization studies, a prespecified parameter k for the minimum group size is often used as a control measure to facilitate comparisons on the other risk and utility measures. This is a common practice in previous experimental studies (Machanavajjhala et al. 2007; Li et al. 2007; LeFevre et al. 2008; Li and Sarkar 2014). Different k values result in different group sizes, and lead to different performances on disclosure risks and utility of the anonymized data. We used three typical k values for all three methods: $k = 10$, 20, and 30. As discussed earlier, parameter k measures re-identification risk. With the same k value, we evaluate the performances on background disclosure risk and data utility based on the following performance measures.

To assess the background disclosure risk related to the required background information, we adopt an information granularity metric called Global Certainty Penalty (GCP) (Ghinita et al. 2007). To measure the GCP of the entire anonymized dataset, we first introduce the information granularity measure (also called the normalized certainty penalty or NCP) (Xu et al. 2006) for a generalized value g of an attribute A, as follows:

$$NCP(g) = \begin{cases} 0, & \text{if } |g| = 1 \\ |g| \;/ |A|\,, & \text{otherwise} \end{cases} \tag{11}$$

where $|g|$ and $|A|$ are the domain sizes of attribute A in g and in the entire data, respectively. For the generalized value g, the domain size for numeric attribute is the range size from its lower bound to its upper bound, and for categorical attribute it is the number of the distinct values in this generalized value. For example, for generalized value "Age = [29-42]" the domain size is 13; and for "Marital Status = *", the domain size is 2, because Marital Status has only two values (married, not married). The value of NCP is ranged from 0 to 1, of which 0 means no generalization, whereas 1 means the value is generalized to the domain size of the attribute in the entire dataset. Finally, for the released dataset T with n records and m background attributes, let g_{ij} be the generalized value of attribute j in record i. The GCP measure is defined as

$$GCP(T) = \frac{\sum_{i=1}^{n}\left[\sum_{j=1}^{m} NCP(g_{ij})\right]}{m * n} \tag{12}$$

The value of GCP is also ranged from 0 to 1. A larger GCP indicates that it is easier to acquire the related background information for the individuals in the released dataset, which means a higher background disclosure risk.

With respect to the utility of anonymized data, we use the classification error rate resulting from the tree model that is built on the anonymized data. For comparison, the classification performances on the original datasets are also reported as a benchmark, which were computed using C4.5 decision tree system implemented in the Weka software package (Witten et al. 2017).

For all the three methods, we run a 10-fold cross-validation experimental scheme on the two datasets, which is similar to the experimental methodology used by LeFevre et al. (2008), briefly described next. First, a dataset was randomly divided into 10 equally sized blocks for 10 test runs. Each time one block was used as the test set while the remaining 9 blocks together were the training set. The training set is then anonymized by applying one of the three anonymization methods to generate an anonymized training set. The anonymized testing set is formed by using the decision tree built from the original data to partition the original testing set into groups and then generalizing the groups. Finally, the classification model is trained using the anonymized training set and tested using the anonymized testing set. This procedure is repeated for 10 runs, each using a different training/testing combination. The reported result is computed by averaging the results of the 10 cross-validation runs on each performance measure.

Data set	Method	Group Size (k)	Time (Second)	GCP	Classification Error Rate
Contra	Original				47.79%
	Mondrian ERP-only Proposed method	10	0 0 3	0.4174 0.4093 0.4071	58.48% 49.02% 48.27%
	Mondrian ERP-only Proposed method	20	0 0 3	0.5136 0.5131 0.5127	59.89% 59.36% 56.55%
	Mondrian ERP-only Proposed method	30	0 0 3	0.5771 0.5728 0.5602	60.28% 60.29% 59.06%
Adult	Original				14.62%
	Mondrian ERP-only Proposed method	10	49 54 75	0.4711 0.4251 0.3648	24.81% 21.09% 16.63%
	Mondrian ERP-only Proposed method	20	33 48 66	0.5010 0.4262 0.3696	25.09% 24.48% 17.08%
	Mondrian ERP-only Proposed method	30	28 39 62	0.5195 0.4695 0.3737	27.76% 24.93% 17.44%

Table 2. Results of Primary Experiments

The results of the experiments are shown in Table 2. It is observed that, for the same group size k, the GCP values with the ERP-only method are smaller than those with Mondrian in both datasets. This indicates that given the same re-identification risk, the proposed ERP pruning method produces anonymized data with lower background disclosure risk than Mondrian does. This is due to the use of the integrated background and identity disclosure risk measure in the ERP pruning method for reducing the background disclosure risks. Moreover, the GCP values with the proposed method including both ERP pruning and tiered generalization are further smaller than those with the ERP-only method. This is owing to the tiered generalization that preserves more detailed background information, making background disclosure more difficult. The results also show that the GCP value increases with the parameter k, this is reasonable because with a larger k value the group size becomes larger and attribute values in a group need to be generalized to cover larger domains.

With respect to the data utility, the classification error rates resulting from the ERP-only method are generally lower than those from Mondrian for both datasets. A possible explanation is that ERP is better than Mondrian in *grouping* the data by considering background information granularity, so the classification models based on the data produced from ERP are more accurate than those from Mondrian. Moreover, the classification error rates from the proposed method with ERP pruning plus tiered generalization are further lower than those from the ERP-only method. This is due to the capability of the tiered generalization to generalize background information at more detailed levels than the uniform generalization, preserving more accurate relationship between background attributes and the class attribute. The results on both datasets also indicate that the tiered generalization can work for both numeric and categorical attributes. In summary, given the same re-identification risk, the proposed methods outperforms Mondrian in terms of both background disclosure risk and data utility.

As for the runtime, all three methods can be completed within seconds or minutes. The proposed method ran slightly slower than ERP-only, which in turn ran slightly slower than Mondrian. This is understandable because ERP pruning and tiered generalization each requires additional time to compute. Nevertheless, the runtimes for the proposed method are reasonable and practical. The results demonstrate that the proposed algorithm can be well-suited for large data applications.

Discussion and Conclusion

The recent cases of privacy invasion have presented a new privacy problem in which an adversary attempts to attack on massive number of people. This massive privacy breach problem is different from conventional data privacy problem, as the adversary does not have prior knowledge about the background information of the target individuals. The adversary needs to first find the background information about the target individuals, and then identify which specific record matches a target individual. To the best of our knowledge, none of the existing privacy models is appropriate for this data privacy problem. The proposed method can not only protect against background and identity disclosures, but also effectively enhance the utility of the data, compared with current state-of-the-art benchmark method.

This research has significant practical values and business implications. As data-sharing is being increasingly adopted in practice, there is a rising sentiment that consumer privacy is being severely eroded. The proposed approach enables organizations to alleviate consumers' concerns about loss of privacy and confidentiality, increasing their willingness to allow their data to be shared for secondary uses, which in turn helps organizations to better comply with privacy laws and regulations such as GDPR and HIPAA (Health Insurance Portability and Accountability Act). Furthermore, GDRP and HIPAA allow anonymized personal data to be shared without individuals' consent. Therefore, organizations can implement the proposed approach to transform the original data into high-quality anonymized data and share them for legitimate business analytics and consumer research, without violating the privacy laws and regulations.

When using the proposed approach, an important parameter that should be carefully considered is the group size k. As discussed earlier, when generalization is used for each group, the re-identification risk for any individual in an anonymized dataset is $1/k$ at most. So, the larger the k value, the smaller the re-identification risk. However, a larger k value also causes a larger information loss after individual values in a group are generalized into the same value. So, the k value has an impact on both disclosure risk and data utility. In our experiment, we have chosen k values to be between 10 and 30 based on the choices from prior studies (Machanavajjhala et al. 2007; Li and Sarkar 2014). In practice where the size of the dataset is typically large, *k* can be set to a relatively large value to get a sufficiently small re-identification risk. In this case, the generalized data can still have sufficiently high data quality because the number of groups obtained from a huge dataset can still be large.

There are several possible future research directions. The current work focuses on privacy-preserving method for classification analysis with a single target workload. Future studies could extend the method to other application domains such as regression analysis or multiple target workload. In addition, alternative generalization or anonymization techniques, aiming to further improve the data utility while preserving privacy, deserve further study. This work focuses on the data science and analytics aspect of the privacy disclosure problem. An action research following this work will be another interesting agenda for future research.

Acknowledgements

Mingzheng Wang's research is partly supported by the National Science Foundation of China (NSFC) under Grants 71931009 and 71671023, the Foundation for Innovative Research Groups of NSFC under Grant 71421001, and the Development Foundation of Teaching and Scientific Research for Teachers of Liberal Arts in Zhejiang University. Xiao-Bai Li's research was supported in part by the National Library of Medicine of the National Institutes of Health, USA, 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.

References

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

Dua, D., and Graff, C. 2019. UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science (http://archive.ics.uci.edu/ml).

Duhigg, C. 2012. "How Companies Learn Your Secrets," *The New York Times Magazine*, February 16, 2012, p. 10.

Dwork, C. 2006. "Differential Privacy," *in Proceedings of 33rd International Colloquium on Automata, Languages and Programming, Part I*, Lecture Notes in Computer Science (4052), Berlin, Heidelberg: Springer, pp. 1-12.

Dwork, C. 2011 "A Firm Foundation for Private Data Analysis," *Communications of the ACM* (54:1), pp. 86-95.

European Parliament and Council of the European Union (EU) 2016. "General Data Protection Regulation," *Official Journal of the European Union*, May 4, 2016 (http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679&from=EN).

Federal Trade Commission (FTC). 2014. Data Brokers: A Call for Transparency and Accountability (http://www.ftc.gov/system/files/documents/reports/data-brokers-call-transparency-accountability-report-federal-trade-commission-may-2014/140527databrokerreport.pdf).

Frier, S. 2018. "Facebook Says Data on Most of Its 2 Billion Users Is Vulnerable," *Bloomberg*, April 4, 2018 (https://www.bloomberg.com/news/articles/2018-04-04/facebook-says-data-on-87-million-people-may-have-been-shared).

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. 7:1-7:20.

Fung, B., Wang, K., Chen, R., and Yu, P.S. 2010. "Privacy-Preserving Data Publishing: A Survey of Recent Developments," *ACM Computing Surveys* (42:4), pp.14.

Ghinita, G., Karras, P., Kalnis, P., and Mamoulis, N. 2007. "Fast data anonymization with Low information loss," in *Proceedings of. 33rd International Conference on Very Large Data Bases,* New York: ACM Press, pp.758-769.

Gionis, A., Mazza, A., and Tassa, T.2008. "*k*-anonymization revisited," *in Proceedings of 24th IEEE International Conference on Data Engineering*, IEEE Computer Society, Washington, DC, pp. 744-753.

Kohavi, R. 1996. "Scaling Up the Accuracy of Naive-Bayes Classifiers: a Decision-Tree Hybrid," *in Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, AAAI Press, Palo Alto, CA, pp.202-207.

LeFevre, K., DeWitt, D.J., and Ramakrishnan, R. 2006. "Mondrian multidimensional k-Anonymity," *in Proceedings of 22nd IEEE International Conference on Data Engineering*, , IEEE Computer Society, Washington, DC, pp. 25-35.

LeFevre, K., DeWitt, D.J., 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.

Lerman, C., Molyneaux, J.W., Pangemanan, S., and Iswarati. 1991. "The determinants of contraceptive method and service point choice," *in secondary Analysis of the 1987 National Indonesia Contraceptive Prevalence Survey, Volume 1: Fertility and Family Planning. Honolulu, HI: East-West Population Institute.*

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*, EEE Computer Society, Washington, DC, 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.

Li, X.B., and Sarkar, S. 2014. "Digression and Value Concatenation to Enable Privacy-Preserving Regression," *MIS Quarterly* (38: 6), pp.79-698.

Machanavajjhala, A., Gehrke, J., Kifer, D., and Venkitasubramaniam, M. 2007. "L-diversity: Privacy beyond K-anonymity," *ACM Transactions on Knowledge Discovery from Data* (1:1), Article 3, pp.3:1-3:52.

Machanavajjhala, A., Kifer, D., Abowd, J., Gehrke, J., and Vilhuber, L. 2008. "Privacy: Theory Meets Practice on the Map," in *Proceedings of the 24th IEEE International Conference on Data Engineering*, Washington, DC: IEEE Computer Society, pp. 277-286.

Nergiz, M.E., Atzori, M., and Clifton, C. 2007. "Hiding the presence of individuals from shared databases," *in Proceedings of 2007 ACM SIGMOD International Conference on Management of data*, New York: ACM Press, pp.665-676.

Nergiz, M.E., and Clifton, C. 2010. "δ-Presence without Complete World Knowledge," *IEEE Transactions on Knowledge and Data Engineering* (22:6), pp. 868-883.

Quinlan, J.R. 1993 *C4.5: Programs for Machine Learning*, San Mateo, CA: Morgan Kaufmanns Publishers.

Rissanen, J. 1978. "Modeling by shortest data description," *Automatica* (14:5), pp.465-658.

Romm, T. 2019. "The U.S. Government and Facebook Are Negotiating a Record, Multibillion-Dollar Fine for the Company's Privacy Lapses," *The Washington Post*, February 14, 2019 (https://www.washingtonpost.com/technology/2019/02/14/us-government-facebook-are-negotiating-record-multi-billion-dollar-fine-companys-privacy-lapses/?utm_term=.b0ce9c6169d2).

Samarati, P. 2001. "Protecting Respondent's Privacy in Microdata Release," *IEEE Trans. Knowledge and Data Engineering (*13:6), pp.1010-1027.

Shannon, C.E. 1948. "A mathematical theory of communication," *The Bell System Technical Journal* (27:3), pp. 379-423.

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

Witten, I.H., Frank, E., Hall, M.A., and Pal, C.J. 2017. *Data Mining: Practical Machine Learning Tools and Techniques*, Cambridge, MA: Morgan Kaufmann Publishers.

Wong, W. K., Mamoulis, N., and Cheung, D. W. L. 2010. "Non-homogeneous generalization in privacy preserving data publishing," in Proceedings of 2010 ACM SIGMOD International Conference on Management of Data, New York: ACM Press, pp.747-758.

Xiao, X., and Tao, Y. 2006. "Anatomy: Simple and Effective Privacy Preservation," *in Proceedings of International Conference on Very Large Data Bases*, New York: ACM Press, pp.139-150.

Xu. J., Wang, W., Pei, J., Wang, X., Shi, B., and Fu, A.W.C. 2006. "Utility-based anonymization using local recoding," *in Proceedings of 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York: ACM Press, pp. 785-790.