Journal of the Association for Information Systems (2020) **21**(6), 1461-1485
doi: 10.17705/1jais.00643

RESEARCH ARTICLE

ISSN 1536-9323

Protecting Privacy When Sharing and Releasing Data with Multiple Records per Person

Hasan B. Kartal[1], Xiao-Bai Li[2]

[1]University of Illinois at Springfield, USA, hasan.kartal@uis.edu
[2]University of Massachusetts Lowell, USA, xiaobai_li@uml.edu

Abstract

This study concerns the risks of privacy disclosure when sharing and releasing a dataset in which each individual may be associated with multiple records. Existing data privacy approaches and policies typically assume that each individual in a shared dataset corresponds to a single record, leading to an underestimation of the disclosure risks in multiple records per person scenarios. We propose two novel measures of privacy disclosure to arrive at a more appropriate assessment of disclosure risks. The first measure assesses individual-record disclosure risk based upon the frequency distribution of individuals' occurrences. The second measure assesses sensitive-attribute disclosure risk based upon the number of individuals affiliated with a sensitive value. We show that the two proposed disclosure measures generalize the well-known *k*-anonymity and *l*-diversity measures, respectively, and work for scenarios with either a single record or multiple records per person. We have developed an efficient computational procedure that integrates the two proposed measures and a data quality measure to anonymize the data with multiple records per person when sharing and releasing the data for research and analytics. The results of the experimental evaluation using real-world data demonstrate the advantage of the proposed approach over existing techniques for protecting privacy while preserving data quality.

Keywords: Data Privacy, *k*-Anonymity, *l*-Diversity, Gini Index, kd-Trees

Roger Chiang was the accepting senior editor. This research article was submitted on March 30, 2019 and underwent three revisions.

1 Introduction

Recent developments in business analytics and big data technologies enable organizations to share and analyze large amounts of various types of personal data (Abbasi, Sarker, & Chiang, 2016; Chen, Chiang, & Storey, 2012; Maass et al.; 2018). This has caused growing concerns about individual privacy, leading to tightened privacy laws and regulations, such as the General Data Protection Regulation (GDPR) recently introduced by the European Union (EU 2016). Privacy disclosure risk arises in many business analytics and big data applications, including the sharing of patient records among different healthcare providers, the distribution of online review data for product recommendations and personalized services, and the sharing of online purchase behavior data by e-commerce companies (Cavusoglu et al. 2016; Kordzadeh & Warren, 2017; Menon & Sarkar 2016). These applications typically involve combining data records from different sources or integrating records of individuals from different temporal and/or geographic points within the same data source. An essential aspect in these applications is that each individual typically corresponds to multiple records, a situation referred to as *multiple records per person* (MRPP) in this paper. MRPP scenarios are very common in big data applications where data are often stored in different

types of databases. While the use of MRPP enhances the value of the data, it also increases the level of disclosure risk. This work examines privacy disclosure risk in MRPP scenarios, a problem that has not been adequately addressed in the literature and in practice.

Multiple records per person is very common in health information systems. A representative example is the Rochester Epidemiology Project (REP), led by the Mayo Clinic (Rocca et al., 2012; Sauver et al., 2012). The REP is a collaboration of clinics, hospitals, and other medical facilities in Minnesota and Wisconsin in the US for sharing medical records for research and public health service. The REP system connects patient medical records across all participating medical providers. Specifically, multiple records of a patient collected by different medical providers are linked with a patient ID in the system. With linked records, the system can be used to improve continuity of care and to study hospital readmission problems. When the system is applied to a large group of patients over an extended time period, it also supports more general research into disease trends in the community. The REP data are available to both participating health organizations and external medical researchers and have been used to support more than 2,000 publications across a wide range of diseases. There are also several other systems similar to REP in other regions of the US (Rocca et al., 2012).

Figure 1 shows a screenshot of the REP display (provided by Sauver et al., 2012). It is clear that a patient often has multiple records in the system. To protect patient privacy when releasing the data to a third party, REP follows the policies specified in the Health Insurance Portability and Accountability Act (HIPAA) by de-identifying the records before releasing them (DHHS, 2000). To comply with HIPAA policy, all personally identifiable fields in Figure 1 are removed for public access. The fields that remain available include patient demographics (e.g., gender, birth year, etc.) and medical information (e.g., disease, treatment, etc.). HIPAA de-identification policies, however, may be insufficient in protecting patient privacy (Sweeney, 2002), particularly when each patient has multiple records in the released data (to be discussed later).

The MRPP setting also appears in many other data releasing and sharing applications. In a widely publicized incident, Netflix awarded one million dollars to a 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 attracted thousands of participants and lasted for three years, was considered a big success in data mining and business analytics. Once the winners were declared, Netflix immediately announced plans for another contest. A few months later, however, the company canceled the plans when it was found that the de-identified data released for the contest, which included movie names and ratings associated with customers, could, in fact, be used to reidentify the customers. The cancellation was necessary in order to settle a class-action lawsuit on privacy violations (Lohr, 2010). In the data provided by Netflix, each movie viewer typically has many movie-viewing records that are linked by a viewer ID. This record linkage is necessary because the recommendation system needs to analyze the associations of movies viewed and rated by a viewer in a certain time period. With multiple linked records, however, disclosure risk increases significantly.

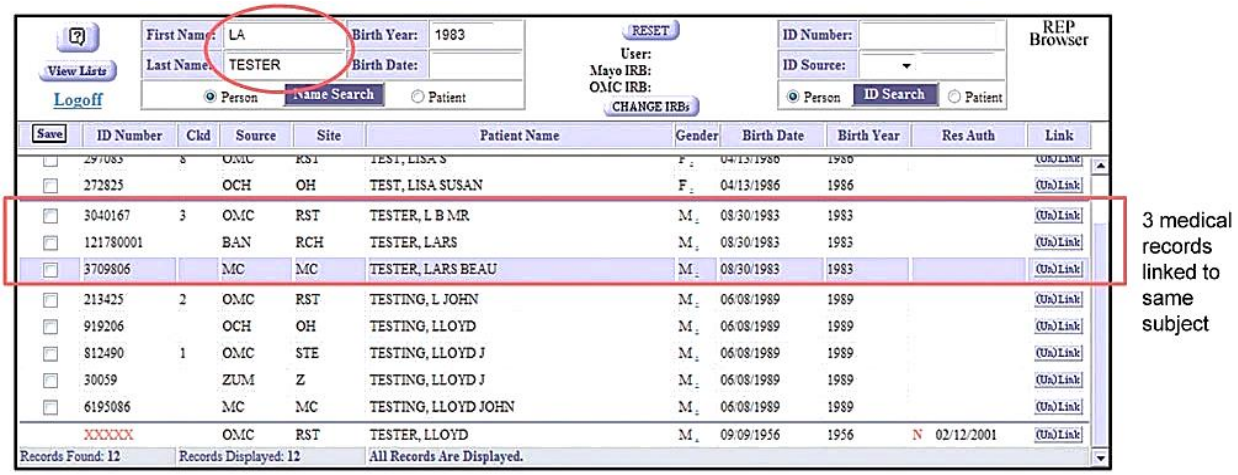

Figure 1. A Screenshot of the REP Display (Source: Sauver et al., 2012, p. 1621)

In the above examples, there are three parties involved from a privacy perspective: (1) the data-owner organization(s) (e.g., REP and Netflix) who wants to make the data available to third-party users while protecting the privacy of the individuals involved; (2) individuals (e.g., REP patients and Netflix customers) whose personal data were collected by the data-owner organization and who want their private information protected; and (3) third-party data users (e.g., medical researchers outside REP and the Netflix data recipients) who want to use the data acquired from the data owner to perform data analysis and research. When a data user attempts to use the acquired data to reveal individuals' private information, the user is called an adversary. This study focuses on how the data-owner organization can release useful data to third-party users while preserving the privacy of the individuals involved.

When released data include personal information, a common practice to address privacy concerns is to de-identify the data before their release. De-identification removes direct identifiers such as individuals' names, phone numbers, and addresses. It is a primary approach in HIPAA's privacy rule (DHHS, 2000). However, it has been shown that de-identification alone does not sufficiently protect against identity disclosure (Samarati & Sweeney, 1998; Sweeney, 2002). Some combinations of demographic attributes, such as age, gender, and zip code, can be used to reidentify individuals from a de-identified dataset (Xu, 2007). In fact, Sweeney (2002) found out that 87% of the population in the United States can be uniquely identified with three demographic attributes—gender, date of birth, and five-digit zip code—which are accessible from some publicly available data sources, such as voter registration records. These publicly available or easily accessible attributes are called quasi-identifiers (QIs). Often, QI attributes are useful for data analysis and need to be included in the released data. To prevent privacy disclosure, a well-known technique called *k*-anonymity (Samarati & Sweeney, 1998; Sweeney, 2002) generalizes the QI attribute values so that each record in a released dataset cannot be distinguished among at least *k* records based on the QI attribute values. A group of records sharing the same QI values is referred to as a QI-group.

This study seeks to assess and mitigate disclosure risks when releasing data with multiple records per person. Following the convention in literature, the released dataset is typically de-identified and includes two types of attributes: (1) quasi-identifier (QI) attributes, which are normally not considered as confidential by individuals, such as age, gender, and zip code. However, the values of the QI attributes can often be obtained from public sources that also contain identifying attributes. So, these QI attributes can be used by an adversary to reidentify the individuals in the de-identified data released, resulting in identity disclosure. (2) Sensitive attributes, which contain private information that an individual typically does not want revealed, such as income, disease, and sexual orientation. The QI attributes, such as age, gender, and zip code, can be obtained (along with identifying attributes) from many data sources, including public sources (e.g., voter registration records) and commercial sources (e.g., data vendors that sell consumer data). In some cases, the adversary knows the QI attributes of the target because they are colleagues, friends, or neighbors. Because of these realistic scenarios, in the data privacy literature (Samarati & Sweeney, 1998; Machanavajjhala et al., 2006; Fung et al., 2010; Li & Sarkar, 2011; Li & Sarkar, 2013; El Emam et al., 2013), it is normally assumed that the adversary knows the QI attribute values of the target individuals but not the sensitive attribute values. The adversary then attempts to disclose the sensitive values based on the information in the QI attributes. We adopt the same assumption in this study.

For the purposes of analyzing privacy-disclosure risk, the literature describes two types of disclosure (Duncan & Lambert, 1989; Li & Sarkar, 2014): (1) identity disclosure, or reidentification, in which an adversary is able to match a record in a dataset to an individual, and (2) sensitive-attribute disclosure, in which an adversary is able to deduce the sensitive-attribute value(s) of an individual record, even without knowing the identity of the individual. The *k*-anonymity model considers the reidentification risk but not the attribute disclosure risk. Therefore, even when the reidentification risk of an individual is sufficiently limited in a QI-group, attribute disclosure may still occur when there is little diversity in the values for a sensitive attribute. To address this problem, the *l*-diversity principle was proposed (Machanavajjhala et al. 2006), which requires that each QI-group contains at least one well-represented (relatively balanced) sensitive value so that sensitive values are sufficiently diversified.

k-anonymity, *l*-diversity, and other existing methods for privacy-preserving data release, including official privacy policy like HIPAA, all assume that each individual corresponds to a single record (Fung et al., 2010). When multiple records in a dataset are associated with the same individual, a QI-group of *k* records may contain fewer than *k* individuals. Also, sensitive values in a QI-group in an MRPP setting may not be distributed as diversified, as in a single record per person setting, even if the group satisfies *l*-diversity. As a result, *k*-anonymity and *l*-diversity do not provide intended privacy protections in MRPP cases, as they do in single-record cases.

This study is designed to address the limitations of the existing well-known privacy techniques, namely *k*-

anonymity and *l*-diversity, when an individual has multiple records in a dataset. We propose two novel disclosure-risk measures for the MRPP setting. The first measure, called *g*-balance, assesses the identity disclosure risk of an individual record, while the second measure, called *h*-affiliation, assesses sensitive-attribute disclosure risk. Based on these two measures, we develop an efficient algorithm for protecting against both identity and attribute disclosures in MRPP cases. Our work follows the same assumption in the literature; that is, the adversary knows the QI attribute values of the target individual who is in the released dataset, and attempts to disclose the sensitive values of the target based on the known QI information. In terms of data anonymization, it is a common practice to protect against identity disclosure by applying anonymization techniques to QI attributes while keeping sensitive attributes unchanged (DHHS, 2000; Fung et al., 2010). We follow the same practice in this study.

This work makes a contribution to data privacy research and practice in several ways. First, we investigate an important problem that has largely been overlooked in the literature. While it is common to see multiple records per person in many data-sharing applications, little attention has been devoted to the study of related privacy-disclosure problems. Our study fills this gap in the literature. Second, we propose two novel metrics for measuring individual-record disclosure and attribute-disclosure risks, respectively, in MRPP cases. We show that the two proposed measures generalize the well-known *k*-anonymity and *l*-diversity measures, and work for cases involving either a single record or multiple records per person. We develop an efficient algorithm that integrates the two proposed metrics and a data-utility metric for anonymizing data in MRPP scenarios. Third, we validate the effectiveness of the proposed approach using real-world data and demonstrate that the proposed approach is superior to existing techniques for protecting privacy while preserving data quality for releasing data with multiple records per person.

2 An Illustrative Example

Consider the hypothetical examples shown in Tables 1 and 2. Table 1 shows the data stored in the electronic medical records (EMR) system of a medical provider, Franklin Center for Lung Diseases, and Table 2 shows the data stored in the EMR system of another provider, Lexington Gastroenterology Clinic. We can see that many patients have multiple records, either within a system or across different systems. Notice that because the records were collected at different times, a patient may have different age or ZIP values in different records. Some of these attributes, such as age, gender, ZIP, and disease, are to be included in a data repository for data sharing or public access. Because of various technical, organizational, and policy issues, there are also many attributes or items that might not be shared, including patient and physician names, detailed dates, and addresses. Some medical information, such as clinical narrative, lab test report, radiology images, and treatment details associated with each patient visit, may or may not be shared depending on the sharing agreement. The aggregated data are shown in Table 3a, where patient names are listed for easy illustration only.

Using *k*-anonymity, the dataset will be divided into some QI-groups, each containing at least *k* records. For practical reasons, it is also required that the same person's records should be grouped into the same QI-group. There are many approaches to grouping data, but the common idea is to group the data such that the records within a QI-group are as close to each other as possible by some distance measure calculated based on the QI-attribute values. After the grouping, the QI values of all records within a group are generalized into the same value, using the group's value domain or range to make the records within a group indistinguishable.

Table 1. Data from Franklin Center for Lung Diseases

Admission no.	Name	Age	Gender	ZIP	Disease	Physician	Other data in Franklin Center EMR System
1001	Ashley	86	Female	20375	Asthma	Dr. Cox	Example data include clinical narratives, lab test reports, radiology images, and/or treatment details associated with each patient visit
1002	Charlie	69	Male	20048	Pneumonia	Dr. Khan	
1003	Harry	74	Male	20400	Asthma	Dr. Cox	
1004	Harry	75	Male	20400	Bronchitis	Dr. Cox	
1005	Charlie	70	Male	20048	Pneumonia	Dr. Khan	
1006	Charlie	71	Male	20048	Pneumonia	Dr. Khan	
1007	Edward	84	Male	20090	Pneumonia	Dr. Smith	
1008	Fred	78	Male	20400	Pneumonia	Dr. Smith	
1009	Harry	76	Male	20400	Asthma	Dr. Patel	

Table 2. Data from Lexington Gastroenterology Clinic

Admission no.	Name	Age	Gender	ZIP	Disease	Physician	Other data in Lexington Clinic EMR System
2001	Ashley	86	Female	20375	Reflux	Dr. Jones	Example data include clinical narratives, lab test reports, radiology images, and/or treatment details associated with each patient visit
2002	Bob	85	Male	20375	Reflux	Dr. Jones	
2003	Charlie	71	Male	20048	Gastritis	Dr. Moore	
2004	Diana	84	Female	20090	Ulcer	Dr. Taylor	
2005	Charlie	71	Male	20048	Gastritis	Dr. Moore	
2006	Diana	84	Female	20090	Gastritis	Dr. Moore	
2007	Edward	84	Male	20090	Gastritis	Dr. Moore	
2008	Greg	78	Male	20420	Ulcer	Dr. Taylor	
2009	Harry	74	Male	20400	Ulcer	Dr. Brown	
2010	Harry	76	Male	20400	Ulcer	Dr. Brown	

Table 3. Aggregated Data

a. Original data

Admission no.	Name	Age	Gender	ZIP	Disease
1001	Ashley	86	Female	20375	Asthma
2001	Ashley	86	Female	20375	Reflux
2002	Bob	85	Male	20375	Reflux
1002	Charlie	69	Male	20048	Pneumonia
1005	Charlie	70	Male	20048	Pneumonia
1006	Charlie	71	Male	20048	Pneumonia
2003	Charlie	71	Male	20048	Gastritis
2005	Charlie	71	Male	20048	Gastritis
2004	Diana	84	Female	20090	Ulcer
2006	Diana	84	Female	20090	Gastritis
1007	Edward	84	Male	20090	Pneumonia
2007	Edward	84	Male	20090	Gastritis
1008	Fred	78	Male	20400	Pneumonia
2008	Greg	78	Male	20420	Ulcer
1003	Harry	74	Male	20400	Asthma
1004	Harry	75	Male	20400	Bronchitis
1009	Harry	76	Male	20400	Asthma
2009	Harry	74	Male	20400	Ulcer
2010	Harry	76	Male	20400	Ulcer

b. Anonymized data

QI-Group	Age	Gender	ZIP	Disease
1	85-86	*	20375	Asthma
1	85-86	*	20375	Reflux
1	85-86	*	20375	Reflux
2	69-71	Male	20048	Pneumonia
2	69-71	Male	20048	Pneumonia
2	69-71	Male	20048	Pneumonia
2	69-71	Male	20048	Gastritis
2	69-71	Male	20048	Gastritis
3	84	*	20090	Ulcer
3	84	*	20090	Gastritis
3	84	*	20090	Pneumonia
3	84	*	20090	Gastritis
4	74-78	Male	20400-20420	Pneumonia
4	74-78	Male	20400-20420	Ulcer
4	74-78	Male	20400-20420	Asthma
4	74-78	Male	20400-20420	Bronchitis
4	74-78	Male	20400-20420	Asthma
4	74-78	Male	20400-20420	Ulcer
4	74-78	Male	20400-20420	Ulcer

If generalization is impossible or inappropriate, the QI values will be suppressed. Table 3b is a *k*-anonymized version of Table 3a with $k = 3$. The records within a group are very close to each other in terms of the values of the QI attributes of age, gender, and ZIP. For example, Diana and Edward's four records are grouped together (in Group 3) because they share the same age and ZIP values. However, their genders are different and must be suppressed because there is no meaningful way to generalize them. Ashley and Bob are grouped together in a similar manner. Note that Fred and Greg are extremely close, but they cannot form a group because of the 3-anonymity requirement, so they are grouped together with Harry.

Although Table 3b satisfies the 3-anonymity requirement, some records in the anonymized data can be easily reidentified. Consider an adversary who knows Charlie's QI attribute values (i.e., age, gender and ZIP code values in Group 2 of Table 3b). This adversary also knows that Charlie is in the dataset (which is easy to know for a system like REP because the system covers all the residents in the region). If the adversary randomly selects one of the five records in Group 2, he would successfully identify one of Charlie's records. In this study, we do not assume that the adversary knows the number of records a target individual has. However, disclosure is deemed to occur when any one of an individual's records is identified. We can also apply this principle to assess the disclosure risks of each of the other individuals in the dataset. For example, in the last group, Harry is associated with five out of seven records while Fred and Greg have only one record each. Assume that the adversary knew Harry's QI attribute values, and also Fred's and Greg's values. Then, by random guessing, the adversary would have a much higher probability of successfully matching one of Harry's records versus Fred's or Greg's record. Thus, through random guessing, Harry has a higher disclosure risk than Fred or Greg, even if the adversary does not know that Harry has more records than Fred or Greg.

When each individual corresponds to a single record, the probability of linking a target to a specific individual using QI values is, at most, $1/k$ in a *k*-anonymized table. When an individual may have multiple records, the individual-record disclosure risk can be assessed based on the individual's record frequency. In a QI-group containing *k* individuals, let f_i be the number of records associated with the ith individual, the individual-record disclosure risk can be assessed by $f_i / \sum_{i=1}^{k} f_i$. The dataset in Table 3b satisfies 3-anonymity, but the probability of successfully matching a record in the dataset to an individual may be higher than 1/3. For example, the probability of matching a record in the second group to Charlie is 100%; for Harry's records, the probability of matching is 5/7.

Next, we consider attribute-disclosure risk in MRPP cases. As indicated, the *l*-diversity assumes a single record per person; it is thus not effective for reducing attribute-disclosure risk in MRPP scenarios. For Group 2 in Table 3b, for example, even though the group is 2-diverse, the sensitive values, pneumonia, and gastritis, can be disclosed individually or together. This is because both values are affiliated with the same patient, Charlie, whose records can thus be easily reidentified.

As explained above, in applications with multiple records per person, data users often want to observe how conditions or preferences of a person change over time, or they may want to analyze how different behaviors or outcomes co-occur for the same individual. In such cases, it is necessary to link the multiple records of the same person with a person identifier (PID). A PID is a system-generated number or label that uniquely (but anonymously) determines a person.

To illustrate the sensitive-attribute disclosure problem across different individuals in an MRPP scenario, consider Table 4, taken from Group 3 in Table 3b. For illustration purposes, we use the first letter of the patient's name as the PID value (which is unlikely to be the case in real applications). Because the group contains at least three distinct sensitive values, it satisfies the basic requirement of *l*-diversity where *l*=3. However, an adversary who finds that his or her target (Diana or Edward) is in this group will know that the target has gastritis even though the adversary does not know which PID corresponds to the target. This is because both patients in the group have gastritis.

Table 4. An *l*-Diverse QI-Group Vulnerable to Sensitive-Attribute Disclosure

PID	Age	Gender	ZIP	Disease
D	84	*	20090	Ulcer
D	84	*	20090	Gastritis
E	84	*	20090	Pneumonia
E	84	*	20090	Gastritis

A similar problem can be seen in Group 4 of Table 3b. Although four different diseases exist in the sensitive attribute, two patients (i.e., Greg and Harry) have ulcers. As two out of three patients are affiliated with the same disease, the likelihood of a patient in the group having the disease is 66.7%, which is higher than any commonly acceptable risk level. Therefore, when individuals have multiple records with multiple sensitive values, even if the values in a QI-group appear to be diverse, an adversary may still infer the individuals' sensitive values with a high probability because of the broad affiliation of a sensitive value with different individuals in the group. In short, *l*-diversity is not an appropriate criterion for assessing attribute-disclosure risk in an MRPP scenario.

It appears that some of the MRPP problems illustrated above may be addressed by way of database decomposition. For example, Table 3 could be decomposed into two relational tables. The first table would contain the PID and QI attributes (age, gender, and ZIP code) and the second table would contain the attributes of PID and disease. The two tables could be joined by the PID to create Table 3. Applying *k*-anonymity to the first table would ensure that each QI-group has at least *k* individuals. This way, Charlie would not appear alone in a QI-group. However, this post-decomposition *k*-anonymity method would not reduce the individual-record disclosure risk for the other groups in Table 3, because the final released dataset would be in a multiple record per person format. Similarly, applying *l*-diversity to the second table would not address the sensitive-attribute disclosure problem illustrated in Table 4. Furthermore, as mentioned at the beginning of the paper, this study considers applications where data are not necessarily stored in relational databases. In this situation, decomposition may not be practical because of the lack of well-defined database schema.

3 Related Work

k-anonymity is a privacy model designed to prevent or mitigate the reidentification problem based on QI attributes (Sweeney, 2002). With *k*-anonymity, when an adversary attempts to identify an individual in a dataset using QI values, the individual cannot be linked to a particular record with a probability higher than 1/*k*. However, individuals in a *k*-anonymized group can still be subject to high attribute disclosure risk if their sensitive attribute values are the same or similar. In this case, the adversary can disclose the sensitive information of the target individual with certainty or high probability, even though the adversary cannot tell which record in the dataset corresponds to the target individual. To address this issue, the *l*-diversity model has been proposed (Machanavajjhala et al., 2006), which requires that a sensitive attribute includes at least *l* well-represented values in each group of anonymized data. Further details and developments with respect to anonymization techniques can be found in Fung et al. (2010). Essentially all of the existing approaches to anonymization assume that each individual corresponds to a single record.

Many real-world datasets, such as patient visitation records, account transactions, and online reviews and ratings, often consist of multiple records for the same individual (El Emam et al., 2009, El Emam et al., 2013). In these cases, *k*-anonymity and *l*-diversity approaches are not appropriate for assessing or mitigating the disclosure risk. El Emam et al. (2009) conducted a case study to evaluate the reidentification risks of patients using a real pharmacy prescription dataset containing individuals with multiple records. They reported that reidentification risks for the individuals in the dataset were quite high. The study, however, does not propose a privacy model for handling MRPP problems.

There have been a few studies concerning problems related to privacy disclosure in MRPP applications. Wang and Fung (2006) address the privacy disclosure problem in sequentially released multiple datasets. They assume that each dataset is a different projection of the same underlying database. The privacy problem considered depends on the presence of a sensitive attribute and the study only concerns attribute disclosure. Our study addresses problems related to both identity and attribute disclosure, as discussed above. We investigate the record-identification problem based on the chance of finding an individual's record in a dataset, which is unrelated to the presence of a sensitive attribute. In addition, we do not assume that multiple data releases include different projections of the same database.

Nergiz, Clifton, and Nergiz (2007) discuss anonymization issues with multiple relational tables. Their approach assumes a restrictive relational database schema. The privacy problem is also contingent upon the presence of sensitive attributes. In addition, the approach assumes that a domain generalization hierarchy can be defined for the values of the QI attributes. Our work does not assume a relational database schema, and, as explained above, the problem we study cannot be addressed by decomposing relational tables and then applying *k*-anonymity and *l*-diversity principles to the decomposed tables. Also, our approach does not rely on a known domain generalization hierarchy.

To address the MRPP disclosure problem, Tao et al. (2008) propose an approach that ensures that every QI-group contains at least *K* individuals or PIDs, each having one or more records. We call this approach "PID-based *K*-anonymity" (with a capital letter *K*). Although this is a reasonable improvement over traditional *k*-anonymity models, further investigation reveals that PID-based *K*-anonymity also does not

provide an adequate level of protection against MRPP disclosures. For example, the last QI-group in Table 3b is considered the most secure group in the dataset using PID-based *K*-anonymity because it contains three people, implying a maximum disclosure risk of *1*/3; but it fails to protect Harry at this security level (probability of matching a record in that group to Harry is 5/7 = 71%). Such high disclosure risks are caused by the unbalanced frequency distribution of the individuals in their QI-groups. In the third QI-group, both Diana and Edward are protected from privacy disclosure with a probability of 2/4 = 0.50. In other words, a smaller *K* may provide better protection than a larger *K*, indicating that the PID-based *K* parameter does not adequately represent the level of protection or risk in the unbalanced frequency distribution case in the MRPP scenario.

In short, existing approaches make different assumptions and have several limitations when it comes to assessing privacy risks in MRPP scenarios. Our proposed approach overcomes these limitations and effectively extends the single-record-based *k*-anonymity and *l*-diversity approaches to MRPP problems.

4 Disclosure Risk and Data Quality Measures

It is clear that in MRPP scenarios, disclosure risk for an individual is closely related to the individual's occurrence frequency. In general, the more unbalanced the occurrence distribution of individuals in a QI-group, the greater the disclosure risk. To limit this risk, individuals' occurrence distribution in QI-groups in the released dataset should be well balanced. Therefore, the basic idea of our approach for reducing individual-record disclosure risk is to create QI-groups that contain a sufficient number of individuals with relatively balanced occurrence distributions.

Given the initial dataset *D*, we can divide *D* into individual-based subsets such that for each individual in *D*, all the records of this individual must be in one and only one subset; i.e., two subsets of *D* cannot contain different records of the same individual. All subsets mentioned in this paper refer to such individual-based subsets; thus, we will omit the term "individual-based." We now define our first proposed measure, called *g*-balance, based on the classical Gini index in economics and machine learning (Breiman et al., 1984), which we adopt to measure disclosure risk.

Definition 1 (*g*-balance): Let *t* be the dataset *D* or a subset of *D* and n_t be the number of individuals in *t*, and c_i be the number of occurrences of the *i*th individual in *t*. The *g*-balance of *t* is defined as

$$g(t) = 1 - \sum_{i=1}^{n_t} \left(\frac{c_i}{\sum_{j=1}^{n_t} c_j} \right)^2 \qquad (1)$$

The *g*-balance measure achieves the maximum when individuals in *t* are evenly distributed, i.e., all c_i's are equal (Breiman et al., 1984). It achieves the minimum of zero when *t* consists of records of a single individual, i.e., $n_t = 1$ (with any number of occurrences of the individual). A larger *g* value indicates a more balanced occurrence distribution in *t*, which suggests better protection against disclosure after the QI values are generalized. With this observation, we say that a QI-group *t* satisfies *g*-balance requirement for a specified g^* value if $g(t) \geq g^*$. The *g* value is related to the number of individuals in a QI-group, as stated in Theorem 1 below.

Theorem 1: If a QI-group *t* satisfies the *g*-balance requirement for a specified *g* value, then the QI-group has at least $1/(1-g)$ individuals; i.e.,

$$n_t \geq \frac{1}{1-g} \qquad (2)$$

The proofs of Theorem 1 and all other mathematical results are provided in the Appendix. Based on Theorem 1, in forming QI-groups for MRPP problems, we can control the number of individuals in a group by specifying an appropriate *g* threshold value. When each individual in the group corresponds to a single record, there is a direct relationship between the *g* value and the *k* value in *k*-anonymity, as stated in Corollary 1 below.

Corollary 1: If a QI-group with *k* individuals satisfies the *g*-balance requirement and each individual in the group corresponds to a single record, then

$$k = \frac{1}{1-g} \qquad (3)$$

It is clear from Theorem 1 and Corollary 1 that the *g*-balance measure generalizes the *k*-anonymity measure. Furthermore, it is straightforward to see that Equation (3) also holds for PID-based *K*-anonymity (i.e., when *k* is replaced by *K*). We note that the balance/skewness of the occurrence distribution can be quantified by some other statistical dispersion measures such as entropy. We chose to use the Gini index because it allows us to efficiently derive Theorem 1 and Corollary 1.

Our proposed method uses binary partitioning to split the dataset into two smaller subsets recursively to form

QI-groups. After the partitioning is completed, the QI values in the final subsets are generalized similarly to *k*-anonymity. We denote the parent set for a split by t_p and the two child subsets of t_p by t_1 and t_2. It can be shown that the *g* value before a split is always greater than or equal to the weighted average *g* value after the split (Breiman et al., 1984). Such a decrease in *g*-balance value implies an increase in disclosure risk. To measure this difference, we define *g*-balance change below.

Definition 2 (*g*-balance change): Let g_p, g_1 and g_2 be the *g*-balance values for t_p, t_1 and t_2, respectively. Let c_{i_p}, c_{i_1} and c_{i_2} be the number of occurrences of the i_pth individual in t_p, the i_1th individual in t_1 and the i_2 th individual in t_2, respectively, where $\sum c_{i_p} = \sum c_{i_1} + \sum c_{i_2}$. The *g*-balance change from splitting t_p into t_1 and t_2 is:

$$\Delta g(t_p) = g_p - \frac{\sum c_{i_1}}{\sum c_{i_p}} - \frac{\sum c_{i_2}}{\sum c_{i_p}} g_2 \qquad (4)$$

Next, we consider the sensitive-attribute disclosure risk. As discussed in the introduction, the traditional *l*-diversity principle is not appropriate for MRPP scenarios. Different individuals in a QI-group may have very diverse sensitive values. However, because each individual may have multiple sensitive values, it is possible that a certain sensitive value is shared by many or even all the individuals in the group (while the other sensitive values may be diversified). If a sensitive value is affiliated with all individuals, then this value is disclosed with certainty (e.g., Gastritis associated with both Diana and Edward in Table 4). In general, the larger the proportion of individuals with which a sensitive value is affiliated, the higher the disclosure risk. Following the convention in data privacy literature (Fung et al., 2010; Machanavajjhala et al., 2006), we assume all sensitive values are equally important. So, the attribute disclosure risk of a QI-group can be determined by the sensitive value that is affiliated with the largest proportion of the individuals in the group.

Definition 3 (*h*-affiliation): Let *t* be the dataset *D* or a subset of *D*, n_t be the number of individuals in *t*, and n_j be the number of individuals in *t* affiliated with the *j*th sensitive value. The *h*-affiliation of *t* is defined as

$$h(t) = \max_j \frac{n_j}{n_t} \qquad (5)$$

The *h*-affiliation measure achieves the maximum of one when all individuals in *t* are affiliated with a common sensitive value. It achieves the minimum of $1/n_t$ when no individuals in *t* share any common sensitive value. Clearly, a larger *h* value suggests a higher sensitive-attribute disclosure risk. With this observation, we say that a QI-group *t* satisfies *h*-affiliation requirement for a specified h^* value if $h(t) \le h^*$.

A common yet conservative interpretation of the "*l* well-represented values" in *l*-diversity is that the relative frequency of the most frequent sensitive value in a QI-group cannot be greater than $1/l$ (Machanavajjhala et al., 2006; Xiao & Tao, 2006). This *l*-diversity requirement can be specified as $\max(k_j)/k \le 1/l$, where k_j is the number of records in the QI-group having the *j*th sensitive value, and *k* is the total number of records in the QI-group. When each individual corresponds to a single record only, it is easy to see that $n_j = k_j, \forall j$, and so $\max(n_j) = \max(k_j)$. Then, the *h*-affiliation requirement is equivalent to the *l*-diversity requirement and the h^* value is simply the reciprocal of the *l* value:

$$h^* = \frac{1}{l} \qquad (6)$$

So, the *h*-affiliation measure generalizes the *l*-diversity measure. The *h*-affiliation has the following property related to data partitioning.

Lemma 1. When a dataset is partitioned into subsets, the *h*-affiliation for at least one subset will be greater than or equal to the *h*-affiliation of the dataset before the partitioning.

Lemma 1 suggests that splitting data into subsets generally increases the attribute disclosure risk. We mentioned earlier that splitting data also increases the individual-record disclosure risk because of a change in *g*-balance value. These properties provide a theoretical basis for our proposed recursive partitioning algorithm in assessing both disclosure risks.

Our method keeps track of the number of individuals affiliated with each sensitive value in the group. Figure 2 shows the partitioning process for the dataset in Table 3. Figure 2a ($t_p = D$) shows the dataset *D* in Table 3a. The final QI-groups include Figure 2b (t_1), Figure 2d (t_{21}) and Figure 2e (t_{22}), all having a well-balanced frequency distribution of the individuals. Within each group (subset), no disease is affiliated with more than 50% of the patients. We describe how to compute *g* and *h* values here and will explain how the splits are determined in the next section.

First, the *g*-balance of *D* given in Figure 2a (i.e., t_p) is computed by substituting the c_i values from the table into Equation (1) as follows:

$$g(t_p) = 1 - \left(\frac{2}{19}\right)^2 - \left(\frac{1}{19}\right)^2 - \left(\frac{5}{19}\right)^2 - \left(\frac{2}{19}\right)^2 - \left(\frac{2}{19}\right)^2 - \left(\frac{1}{19}\right)^2 - \left(\frac{1}{19}\right)^2 - \left(\frac{5}{19}\right)^2 = 0.82\,.$$

For example, the first 2/19 applies to the first individual Ashley, whose number of occurrences (c_1) is 2, and the total number of occurrences of all individuals is $\sum_j c_j = 19$. The h-affiliation of t_p is computed by substituting n_j values into Equation (5) as below:

$$h = \max\left(\frac{2}{8}, \frac{1}{8}, \frac{3}{8}, \frac{3}{8}, \frac{2}{8}, \frac{3}{8}\right) = 0.375,$$

where each number inside the parentheses is the fraction of individuals affiliated with each of the six diseases in the table. For example, $\frac{2}{8}$ represents that 2 out of 8 patients have Asthma. When t_p is split into t_1 (Figure 2b) and t_2 (Figure 2c) based on gender, g values for t_1 and t_2 are respectively 0.5 and 0.747; h values are both 0.5. Then, the g-balance change by splitting t_p into t_1 and t_2 is computed using Equation (4):

$$\Delta g(t_p) = 0.82 - (\frac{4}{19})0.5 - (\frac{15}{19})0.747 = 0.1252.$$

Similarly, if the dataset is split based on the median of Age or ZIP, the corresponding Δg value will be 0.1707 or 0.1802, respectively.

Our recursive partitioning method adopts the idea of the well-known kd-tree technique (Friedman, Bentley, & Finkel, 1977), where each split is determined based on the variance of the QI attributes. Typically, the QI attribute with the largest variance at each iteration is used to split the data, as this will result in the most significant reduction in variance in the partitioned data. A lower variance in a QI-group leads to a better data utility because it causes a smaller information loss (i.e., loss in variation) after QI values within the partitioned group are generalized. In other words, with a smaller within-group variance, the generalized values will be closer to the original values. Thus, we use variance to measure the quality of anonymized data. Variance is calculated by considering multiple records per person since the released dataset will be in multiple record per person format.

Name	Age	Gender	ZIP	Disease
A	86	Female	20375	Asthma
A	86	Female	20375	Reflux
B	85	Male	20375	Reflux
C	69	Male	20048	Pneumonia
C	70	Male	20048	Pneumonia
C	71	Male	20048	Pneumonia
C	71	Male	20048	Gastritis
C	71	Male	20048	Gastritis
D	84	Female	20090	Ulcer
D	84	Female	20090	Gastritis
E	84	Male	20090	Pneumonia
E	84	Male	20090	Gastritis
F	78	Male	20400	Pneumonia
G	78	Male	20420	Ulcer
H	74	Male	20400	Asthma
H	75	Male	20400	Bronchitis
H	76	Male	20400	Asthma
H	74	Male	20400	Ulcer
H	76	Male	20400	Ulcer

$g = 0.82$ $h = 0.375$

a. $t_p = D$

Name	Age	Gender	ZIP	Disease
A	86	Female	20375	Asthma
A	86	Female	20375	Reflux
D	84	Female	20090	Ulcer
D	84	Female	20090	Gastritis

$g = 0.50$ $h = 0.50$

b. t_1

Name	Age	Gender	ZIP	Disease
B	85	Male	20375	Reflux
C	69	Male	20048	Pneumonia
C	70	Male	20048	Pneumonia
C	71	Male	20048	Pneumonia
C	71	Male	20048	Gastritis
C	71	Male	20048	Gastritis
E	84	Male	20090	Pneumonia
E	84	Male	20090	Gastritis
F	78	Male	20400	Pneumonia
G	78	Male	20420	Ulcer
H	74	Male	20400	Asthma
H	75	Male	20400	Bronchitis
H	76	Male	20400	Asthma
H	74	Male	20400	Ulcer
H	76	Male	20400	Ulcer

$g = 0.747$ $h = 0.50$

c. t_2

Name	Age	Gender	ZIP	Disease
B	85	Male	20375	Reflux
E	84	Male	20090	Pneumonia
E	84	Male	20090	Gastritis
F	78	Male	20400	Pneumonia
G	78	Male	20420	Ulcer

$g = 0.72$ $h = 0.50$

d. t_{21}

Name	Age	Gender	ZIP	Disease
C	69	Male	20048	Pneumonia
C	70	Male	20048	Pneumonia
C	71	Male	20048	Pneumonia
C	71	Male	20048	Gastritis
C	71	Male	20048	Gastritis
H	74	Male	20400	Asthma
H	75	Male	20400	Bronchitis
H	76	Male	20400	Asthma
H	74	Male	20400	Ulcer
H	76	Male	20400	Ulcer

$g = 0.50$ $h = 0.50$

e. t_{22}

Figure 2. Data Partitioning Process for Dataset in Table 3a

In calculating variance, we transform categorical QI values into numeric or ordered values based on coding methods suggested in LeFevre, DeWitt, and Ramakrishnan (2006), and normalize all numeric values (original or transformed) to the unit scale of range [0, 1]. For example, for gender, we assign zero for female and one for male. If the QI attribute has more than two unordered categories, additional binary attributes are created to handle multiple categories one by one.

5 The Proposed Algorithm

There are two objectives in our partitioning process for anonymizing data: (1) to minimize disclosure risks, which means keeping change in *g*-balance and increase in *h*-affiliation as small as possible, and (2) to minimize information loss after generalization by reducing the variance of the partitioned data as much as possible before generalization. We note again that *g*-balance is used to achieve a more balanced frequency distribution in QI-groups, while *h*-affiliation is used to ensure that no sensitive values occur too frequently in any given QI-group.

We indicated that the g value will decrease during the data partitioning process and proposed using the Δg measure to quantify this change in g value. While Lemma 1 suggests that *h*-affiliation generally increases with data partitioning, it is difficult to construct a composite measure to quantify the changes in both *g*-balance and *h*-affiliation simultaneously. Therefore, our strategy is to first use *g*-balance to determine how to split the data and then use *h*-affiliation as a constraint to check whether the partitioned QI-groups satisfy the sensitive attribute protection requirement. This idea of handling two different disclosure risk criteria in computation is similar to that of Machanavajjhala et al. (2006), where the *l*-diversity requirement is checked after a group is formed based on a *k*-anonymity algorithm. The QI values of all individuals are generalized after the entire partitioning process is completed.

On the other hand, there is a nice way to integrate *g*-balance (as a disclosure risk measure) and variance (as an information loss measure) into a single combined measure. It is clear that a split with a small *g*-balance change $\Delta g(t)$ is preferred since it implies a small increase in disclosure risk after splitting the data. In terms of information loss, it is preferred that the generalized value for each QI-group is as close to the original individual values as possible. In other words, the variance in a group after data partitioning should be as small as possible. Therefore, the attribute with the largest variance should be used to split the data so that the partitioned groups will have their within-group variance reduced most significantly. We define a *g*-balance/variance ratio measure below to represent this trade-off between disclosure protection and data quality. It is used as the splitting criterion in the data partitioning process.

Definition 4 (balance-variance ratio): Let *t* be the dataset *D* or a subset of *D* and $v_j(t)$ be the variance of the *j*th QI attribute in *t*. The balance-variance ratio for splitting *t* on the *j*th QI attribute is defined as

$$r_j(t) = \frac{\Delta g(t)}{v_j(t)} \qquad (7)$$

The balance-variance ratio represents the marginal decrease in *g*-balance per unit variance of a QI attribute. Because a small *g*-balance change and a large variance are preferred for a candidate split, the QI attribute that has the minimum balance-variance ratio should be selected for partitioning the data at each iteration. The proposed algorithm recursively splits data into two subsets at the median of the QI attribute having the minimum balance-variance ratio. If the QI attribute is of the ordered categorical type, the split is made at the between-category point closest to the median among all between-category points.

Table 5 describes the steps of the proposed algorithm, where two user-specified privacy requirement parameters are used: minimum *g*-balance value, *g**, and maximum *h*-affiliation value, *h**. The computational time complexity of the algorithm is equivalent to that of a kd-tree, which is of $O(N \log N)$ for a dataset of *N* records (Friedman et al., 1977). This is very efficient for handling large datasets.

Table 5. The Proposed Algorithm

	Input: Dataset *D*, threshold values *g and *h**.**
Step 1	For the current dataset *t*, compute $r_j(t)$ for each QI attribute *j*. Let *j** be the QI with minimum $r_j(t)$.
Step 2	(i) Split *t* into two subsets at the median of attribute *j**. (ii) If the *g*-balance value of any subset of *t* is smaller than *g** or *h*-affiliation value of any subset of *t* is greater than *h**, undo split and set *j** to the QI attribute with the next smallest $r_j(t)$ and go to (i). Stop splitting if no QI attribute can be assigned to *j**.
Step 3	Repeat Steps 1 and 2 for each subset until no further split can be made.
Step 4	Generalize the QI values in each subset.

Continuing with the illustrative example in Figure 2a, suppose $g^* = 0.5$ and $h^* = 0.5$. For gender, we thus have $v_2 = 0.1662$ and $r_2 = \Delta g / v_2 = 0.1252 / 0.1662 = 0.753$. Similarly, for age, $v_1 = 0.1202$ (with normalized values) and $r_1 = 0.1707/0.1202 = 1.420$; and for ZIP, $v_3 = 0.1959$ and $r_3 = 0.1802/0.1959 = 0.920$. So, the second QI attribute, gender, is selected for the first split since r_2 is the smallest. The two subsets are shown in Figure 2b (t_1) and Figure 2c (t_2). Subsequently, Figure 2c (t_2) can be further split into Figure 2d (t_{21}) and Figure 2e (t_{22}) based on the first QI attribute age. Figure 2b (t_1), Figure 2d (t_{21}), and Figure 2e (t_{22}), cannot be split further since splitting any of these tables causes a g value to be smaller than 0.5 and/or an h value to be greater than 0.5 for at least one of the child subsets.

The QI-groups of the anonymized dataset using the proposed algorithm is shown in Table 6b. For illustration, we also provide Table 6a, which is the same as Table 3a except that records are reordered to match the records in the anonymized dataset. It can be seen that the anonymized dataset has multiple individuals in each QI-group with well-balanced frequency distributions and no disease occurs too frequently relative to the number of individuals in each QI-group. The g and h values in each QI-group all satisfy the threshold requirements; i.e., $g \geq g^* = 0.5$ and $h \leq h^* = 0.5$ (in the first group, $g = 0.5, h = 0.5$; in the second group, $g = 0.5, h = 0.5$; and in the third group, $g = 0.72, h = 0.5$).

Table 6. The Original and Anonymized Datasets (g* = 0.50; h* = 0.50)

a. Original dataset

Name	Age	Gender	ZIP	Disease
Ashley	86	Female	20375	Asthma
Ashley	86	Female	20375	Reflux
Diana	84	Female	20090	Ulcer
Diana	84	Female	20090	Gastritis
Bob	85	Male	20375	Reflux
Edward	84	Male	20090	Pneumonia
Edward	84	Male	20090	Gastritis
Fred	78	Male	20400	Pneumonia
Greg	78	Male	20420	Ulcer
Charlie	69	Male	20048	Pneumonia
Charlie	70	Male	20048	Pneumonia
Charlie	71	Male	20048	Pneumonia
Charlie	71	Male	20048	Gastritis
Charlie	71	Male	20048	Gastritis
Harry	74	Male	20400	Asthma
Harry	75	Male	20400	Bronchitis
Harry	76	Male	20400	Asthma
Harry	74	Male	20400	Ulcer
Harry	76	Male	20400	Ulcer

b. Anonymized dataset

PID	Age	Gender	ZIP	Disease
A	84-86	Female	20090-20375	Asthma
A	84-86	Female	20090-20375	Reflux
D	84-86	Female	20090-20375	Ulcer
D	84-86	Female	20090-20375	Gastritis
B	78-85	Male	20090-20420	Reflux
E	78-85	Male	20090-20420	Pneumonia
E	78-85	Male	20090-20420	Gastritis
F	78-85	Male	20090-20420	Pneumonia
G	78-85	Male	20090-20420	Ulcer
C	69-76	Male	20048-20400	Pneumonia
C	69-76	Male	20048-20400	Pneumonia
C	69-76	Male	20048-20400	Pneumonia
C	69-76	Male	20048-20400	Gastritis
C	69-76	Male	20048-20400	Gastritis
H	69-76	Male	20048-20400	Asthma
H	69-76	Male	20048-20400	Bronchitis
H	69-76	Male	20048-20400	Asthma
H	69-76	Male	20048-20400	Ulcer
H	69-76	Male	20048-20400	Ulcer

6 Experimental Evaluation

To evaluate the performance of the proposed method, we conducted an experimental evaluation study using three real-world databases. The first is a healthcare database provided by the INFORMS Data Mining Section (2008) for its first data mining contest. The database consists of four related datasets, two of which are closely related to MRPP problems: patient demographics and patient conditions. The patient demographics dataset includes attributes such as patient ID, year of birth, gender, race, years of education, marital status, income, and poverty level. The patient conditions dataset includes attributes such as patient ID, ICD-9 diagnosis code, and year. The two datasets were joined into one set according to patient ID. From the patient demographics dataset, we chose year of birth, years of education, income, and poverty level for the QI attributes; from the patient conditions set, we chose ICD-9 diagnosis code as the sensitive attribute. We obtained count values based on patient ID in the patient conditions dataset. After removing records with missing information, 117,307 medical condition records for 29,531 patients remained. Approximately 25,000 patients out of 29,531 had multiple visits.

The second database contains movie rating data collected by Harper and Konstan (2016). This dataset is somewhat similar to the Netflix dataset discussed in the introduction, but the Netflix dataset is no longer available. The database consists of three related datasets, two of which are closely related to MRPP problems: user and rating datasets. The user dataset provides demographic information for 943 users, including user ID, age, gender, occupation, and ZIP code. We selected age, gender, and ZIP code for the QI attributes. The rating dataset contains 100,000 user ratings (1-5 scale) for 1,682 movies, with attributes of user ID, movie ID, rating, and a time stamp. Initially, we considered using movie ID as the sensitive attribute. However, it turned out that some popular movies were very common to a large number of users, making it difficult to consider it sensitive. In order to make sensitive attribute values more meaningful, we considered the attributes movie ID, rating, and month/year together as the sensitive attribute; that is, we combined the values of the attributes of movie ID, rating, and month/year to create the sensitive attribute value. Again, the user and rating datasets were joined into one set according to user ID.

The third database contains financial data from a bank about their clients, accounts, and transactions (PKDD, 1999). The database consists of eight related datasets, four of which are closely related to MRPP problems: client, account, demographics, and transactions. The four datasets were joined into one set using client ID and account ID. We chose date of birth, salary level, and account open date as the QI attributes. We defined the sensitive attribute as the transaction amount per period, which was rounded to the nearest hundred dollars. After removing records with missing values, the aggregated dataset contained 273,508 transaction records from 3,674 clients.

7 Evaluation of Individual-Recorded Disclosure Risk

We first compared our *g*-balance based method with *k*-anonymity and PID-based *K*-anonymity in terms of individual-record disclosure risk and data quality for the MRPP scenario. As discussed above, the individual-record disclosure risk (*IDR*) of the *i*th person in a QI-group of size *k* can be defined by $IDR_i = f_i / \sum_{i=1}^{k} f_i$, where f_i is the number of records associated with the *i*th person in the QI-group. In general, individuals in a QI-group may have different *IDR* values. 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 & Tao, 2006). So, we defined the individual-record disclosure risk of a QI-group *q* as the maximum *IDR* value in the group, written as $GIDR_q = \max(IDR_i)$. To evaluate individual-record disclosure risk for an anonymized dataset with *m* QI-groups, we use the maximum and average *GIDR* measures, defined below:

$$MaxGIDR = \max_{q=1,\ldots,m} GIDR_q$$

$$AvgGIDR = \frac{1}{m}\sum_{q=1}^{m} GIDR_q$$

We ran a *k*-anonymity algorithm (LeFevre et al., 2006) using seven different *k* values: $k = 2, 3, 5, 7, 10, 20, 50$, and also used the same seven values for PID-based *K*-anonymity, $K = 2, 3, 5, 7, 10, 20, 50$. Based on Equation (3), which provides a direct relationship between the *k* (or *K*) and *g* values in the one record per person case scenario, we then selected seven corresponding *g* values: $g = 0.50, 0.67, 0.80, 0.86, 0.90, 0.95, 0.98$. Some individuals in the dataset had only one record per person and it is possible that these individuals were assigned to the same QI-group. So, the selection of corresponding *g* values ensures that the dataset anonymized with the *g*-balance based method satisfies respective *k*-anonymity and PID-based *K*-anonymity requirements in this situation. Tables 7, 8, and 9 show the maximum and average individual-record disclosure risks for the three datasets anonymized based on the chosen *k*, *K*, and *g* values, respectively.

Table 7. Comparison of Individual-Record Disclosure Risks in Patient Data

a. Risk with *k*-anonymity		
k	***MaxGIDR* (%)**	***AvgGIDR* (%)**
2	100.00	88.55
3	100.00	82.07
5	100.00	70.63
7	100.00	62.57
10	100.00	53.28
20	100.00	36.63
50	76.67	20.75

b. Risk with PID-based *K*-anonymity		
K	***MaxGIDR* (%)**	***AvgGIDR* (%)**
2	96.67	59.26
3	94.29	46.48
5	77.14	34.09
7	76.74	27.40
10	59.02	21.56
20	45.59	13.62
50	17.46	7.03

c. Risk with *g*-balance		
G	***MaxGIDR* (%)**	***AvgGIDR* (%)**
0.50	68.18	44.11
0.67	54.17	33.52
0.80	40.00	23.54
0.86	33.33	18.33
0.90	25.58	14.16
0.95	16.45	8.55
0.98	7.33	4.25

Table 8. Comparison of Individual-Record Disclosure Risks in Movie Data

a. Risk with *k*-anonymity		
k	***MaxGIDR* (%)**	***AvgGIDR* (%)**
2	100.00	99.27
3	100.00	99.27
5	100.00	99.27
7	100.00	99.27
10	100.00	99.27
20	100.00	99.27
50	100.00	85.94

b. Risk with PID-based *K*-anonymity		
K	***MaxGIDR* (%)**	***AvgGIDR* (%)**
2	96.07	63.00
3	90.21	50.07
5	81.35	35.37
7	81.02	28.65
10	50.74	21.87
20	35.90	13.33
50	9.87	6.96

c. Risk with *g*-balance		
G	***MaxGIDR* (%)**	***AvgGIDR* (%)**
0.50	67.37	45.39
0.67	53.89	34.67
0.80	35.46	24.01
0.86	30.16	18.83
0.90	23.54	13.18
0.95	11.98	7.65
0.98	5.88	4.41

Table 9. Comparison of Individual-Record Disclosure Risks in Bank Data

a. Risk with *k*-anonymity		
k	***MaxGIDR* (%)**	***AvgGIDR* (%)**
2	100.00	99.84
3	100.00	99.84
5	100.00	99.82
7	100.00	99.77
10	100.00	99.47
20	100.00	97.59
50	100.00	87.99

b. Risk with PID-based *K*-anonymity		
K	***MaxGIDR* (%)**	***AvgGIDR* (%)**
2	98.14	57.11
3	82.58	41.36
5	56.61	28.07
7	41.64	20.52
10	33.43	16.01
20	17.19	8.58
50	6.08	3.98

c. Risk with *g*-balance		
g	***MaxGIDR* (%)**	***AvgGIDR* (%)**
0.50	66.49	40.04
0.67	51.06	30.65
0.80	34.24	21.03
0.86	25.69	15.19
0.90	19.10	11.92
0.95	10.30	6.39
0.98	5.12	3.09

It can be observed that *k*-anonymity is very ineffective against MRPP disclosure. In Tables 7a, 8a, and 9a, the maximum disclosure risks are 100% for $k = 2$ through 20, meaning that some QI-groups consist of multiple records of only one individual, which can be uniquely reidentified. Even with $k = 50$, the maximum risks are still very high (76.67%, 100%, and 100% for the patient, movie, and bank data, respectively). While PID-based *K*-anonymity does a better job than *k*-anonymity, the *g*-balance based method clearly outperforms both *k*-anonymity and PID-based *K*-anonymity in every comparison category. The individual-record disclosure risks using *g*-balance are significantly lower than those using the other two methods in every risk assessment scenario. This is because of the balanced frequency distribution of individuals within a QI-group in the proposed method, which is designed to limit the disclosure risk of individuals' records with multiple occurrences. In addition, because of the direct relationship between the *k* (or *K*) and *g* values when each individual in a QI-group has only one record, the dataset anonymized with the *g*-balance based method satisfies corresponding *k*-anonymity and PID-based *K*-anonymity requirements in the one record per person case scenario.

Next, we evaluate data quality by measuring information loss because of generalization. Let *D* be the original dataset with *N* individuals and *d* QI attributes, and D^* be the anonymized version of *D*, where its QI values are generalized using the means of the QI attributes in each QI-group. Let x_{ij} and x_{ij}^* be the normalized values of the *j*th QI attribute of the *i*th individual in *D* and D^*, respectively. Information loss because of generalization can be measured using the average normalized error (*ANE*), computed by

$$ANE = \frac{1}{d * N} \sum_{j=1}^{d} \sum_{i=1}^{N} |x_{ij} - x_{ij}^*|$$

ANE measures the average normalized distances between the original and generalized QI values. A small *ANE* suggests a small information loss and thus is desirable for higher data quality. In order to compare the *ANE*s between the two methods, it is necessary to "control" the disclosure risk at the same level for all methods. Thus, we gradually adjusted the *k* and *K* values in *k*- and *K*-anonymity and *g* values in *g*-balance such that the resulting *MaxGIDR* values are comparable at six target levels: 1%, 2%, 5%, 10%, 15%, and 20% for the patient data (in data privacy research and practice, disclosure risk is more often measured using the maximum risk instead of average risk). For the movie and bank datasets, because some individuals had a very high number of occurrences, it was not possible to get the target levels of 1% and 2% for *MaxGIDR*. So, we set five target levels: 5%, 10%, 15%, 20% and 25%. To be conservative, we made the *MaxGIDR* value from *g*-balance slightly smaller than that from *k*-anonymity at each level. We can then compare the related *ANE* values.

The results of this experiment are given in Tables 10, 11, and 12. The *ANE* values with *g*-balance are considerably smaller than those with *k*- and *K*-anonymity at all levels while the *MaxGIDR* values with *g*-balance are about the same as (or slightly smaller than) those with *k*- and *K*-anonymity at all levels. This suggests that the *g*-balance based method results in smaller information loss than *k*- and *K*-anonymity, given about the same individual-record disclosure risk. One explanation is that *k*- and *K*-anonymity reduce the risk only by increasing the group size, which directly causes information loss when the QI values within a group are generalized. The *g*-balance method focuses on the occurrences of each individual and assigns individuals with similar occurrence frequencies into the same group, which does not necessarily require increasing the group size. A second explanation is that our proposed algorithm partitions data into QI-group using the balance-variance ratio, which can achieve a superior tradeoff between disclosure risk and information loss. In summary, the results from Tables 7 through 12 indicate that the proposed *g*-balance method outperforms *k*- and *K*-anonymity in terms of both privacy protection and data quality.

Table 10. Comparison of Information Loss Given Individual-Record Disclosure Risks in Patient Data

Target *MaxGIDR* (%)	*k*-Anonymity		PID-based *K*-anonymity		*g*-Balance	
	MaxGIDR (%)	*ANE*	*MaxGIDR* (%)	*ANE*	*MaxGIDR* (%)	*ANE*
20	19.41	1.479	19.19	1.118	18.79	0.469
15	16.64	2.951	16.27	1.814	14.45	0.923
10	9.60	4.031	9.14	3.160	9.03	2.477
5	5.36	10.505	4.64	9.229	4.49	8.962
2	2.44	34.815	2.48	34.571	2.43	22.789
1	1.43	81.068	1.93	78.809	1.35	55.298

Table 11. Comparison of Information Loss Given Individual-Record Disclosure Risks in Movie Data

Target *MaxGIDR* (%)	*k*-Anonymity		PID-based *K*-anonymity		*g*-Balance	
	MaxGIDR (%)	*ANE*	*MaxGIDR* (%)	*ANE*	*MaxGIDR* (%)	*ANE*
25	25.58	0.908	26.97	0.902	25.00	0.506
20	20.87	1.022	21.38	0.984	20.40	0.867
15	15.23	2.133	15.78	2.107	14.57	1.723
10	12.07	2.785	12.01	2.918	11.98	2.205
5	5.92	8.551	6.02	8.976	5.88	7.825

Table 12. Comparison of Information Loss Given Individual-Record Disclosure Risks in Bank Data

Target *MaxGIDR* (%)	*k*-Anonymity		PID-based *K*-Anonymity		*g*-Balance	
	MaxGIDR (%)	*ANE*	*MaxGIDR* (%)	*ANE*	*MaxGIDR* (%)	*ANE*
25	24.85	1.263	23.55	0.610	23.53	0.498
20	22.05	1.341	21.64	0.806	21.39	0.557
15	15.93	1.811	14.42	1.431	14.38	0.947
10	10.64	3.298	10.85	2.804	10.30	1.786
5	5.55	8.230	5.68	7.781	5.48	4.608

8 Evaluation of Sensitive-Attribute Disclosure Risk

We now compare *h*-affiliation with *l*-diversity in the effectiveness of measuring sensitive-attribute disclosure risk. While QI values are generalized to satisfy *k*-anonymity or *g*-balance requirements, we note that sensitive attributes are usually not subject to change in most data privacy approaches (DHHS, 2000; Fung et al., 2010). Indeed, *l*-diversity and *h*-affiliation mitigate sensitive-attribute disclosure risk by forming the QI-groups and adjusting the group sizes to include diverse sensitive values rather than changing the sensitive values.

As discussed earlier, sensitive-attribute disclosure risk (*SAR*) in a QI-group depends on how many individuals in the group are associated with a sensitive value. For a QI-group of size *k*, *SAR* for the *j*th sensitive value can be defined by $SAR_j = n_j/k$, where n_j is the number of people having the *j*th sensitive value in the QI-group. Furthermore, we define the sensitive-attribute disclosure risk of a QI-group *q* to be the maximum *SAR* among all sensitive values, written as $GSAR_q = \max(SAR_j)$. To evaluate sensitive-attribute disclosure risk for an anonymized dataset with *m* QI-groups, we use the maximum and average *GSAR* measures, defined below:

$$MaxGSAR = \max_{q=1,\dots,m} GSAR_q$$

$$AvgGSAR = \frac{1}{m}\sum_{q=1}^{m} GSAR_q$$

The results of the previous section show that *g*-balance is more effective than both traditional *k*-anonymity and PID-based *K*-anonymity in reducing disclosure risk and information loss. Therefore, we compared *h*-affiliation with *l*-diversity based on the QI-groups formed using *g*-balance. We set three *g*-balance threshold levels, $g^* = 0.50, 0.67$ and 0.80 for the experiment. For each g level, we applied several threshold values for *l*-diversity and *h*-affiliation to examine the maximum and average sensitive-attribute disclosure risks.

We ran our algorithm with *l*-diversity measure using five different l values: $l = 2, 3, 5, 10, 20$. Based on Equation (6), we have $h = 1/l$ in the one-record-per-person case. So, we selected five corresponding h values: $h = 0.50,\ 0.33,\ 0.20,\ 0.10,\ 0.05$. Tables 13, 14, and 15 show the maximum and average sensitive-attribute disclosure risks for the two datasets anonymized based on those *l* and *h* values.

It can be observed that *l*-diversity is very ineffective in controlling sensitive-attribute disclosure risk in the MRPP problem. For example, with the patient data when $g^* = 0.5$, the *MaxGSAR* values are 100% for *l* =

2 through 10, which means that in some QI-groups all individuals have a common sensitive value, causing the sensitive value to be disclosed with certainty. As *l*-diversity increases, disclosure risk generally decreases. However, there is not an intuitive connection between *l* values and actual risk levels. We note that for the movie dataset, *MaxGSAR* and *AvgGSAR* values with *l*-diversity do not change when *l* values are increased from 2 through 20. This is because in the original dataset each individual has at least 20 occurrences (lower frequency individuals were removed from the original data to balance the frequency distributions). It is practically impossible to tell, based on the results of either dataset, what risk a QI-group has for a given *l*-diversity value. On the other hand, it is clear that *h*-affiliation is more effective in controlling sensitive-attribute disclosure risk in MRPP problems. For example, when $g^* = 0.5$, the *MaxGSAR* values in the patient data decreases from 50% to around 4.4% for *h* = 0.50 through 0.05. Similarly, in the movie and bank cases for the same g^* and *h* values, *MaxGSAR* decreases from 50% to 4.8% and 6.21%, respectively. Clearly, *h* value closely represents the actual maximum sensitive-attribute disclosure risk in the MRPP problem and is more effective in controlling the risk than *l*-diversity. Also, it can be seen that as the *h*-affiliation value gets smaller, it performs much better than the corresponding *l*-diversity measure. In summary, for each corresponding *l* and *h* value at each of the three g^* levels, *h*-affiliation results in substantially lower maximum and average sensitive-attribute disclosure risks than *l*-diversity.

For *h*-affiliation in the patient and bank data (Tables 13 and 15), the *MaxGSAR* values with the same *h* value are the same for all three different *g* threshold values. This is because the groups formed by the proposed algorithm are constrained by the *h* threshold values, instead of the *g* threshold values. For the movie data (Table 14), some of the *MaxGSAR* values with *h*-affiliation for $g^* = 0.8$ are different from the corresponding *MaxGSAR* values for $g^* = 0.5$ (and $g^* = 0.67$). For example, when $g^* = 0.8$ and $h = 0.5$, $MaxGSAR = 30\%$, whereas when $g^* = 0.5$ and $h = 0.5$, $MaxGSAR = 50\%$. This is because the groups formed by the proposed algorithm are constrained by $g^* = 0.8$ and $g^* = 0.5$ (instead of $h = 0.5$), respectively.

Table 13. Comparison of Sensitive-Attribute Disclosure Risk in Patient Data

Threshold g*	***l*-Diversity**			***h*-Affiliation**		
	l	***MaxGSAR* (%)**	***AvgGSAR* (%)**	***h***	***MaxGSAR* (%)**	***AvgGSAR* (%)**
0.50	2	100.00	35.39	0.50	50.00	32.26
	3	100.00	33.70	0.33	31.58	20.91
	5	100.00	31.24	0.20	20.00	15.35
	10	100.00	22.85	0.10	10.00	8.26
	20	27.78	12.78	0.05	4.39	3.88
0.67	2	80.00	27.01	0.50	50.00	25.38
	3	80.00	26.83	0.33	31.58	20.58
	5	80.00	26.23	0.20	20.00	15.31
	10	80.00	21.49	0.10	10.00	8.25
	20	27.78	12.83	0.05	4.39	3.88
0.80	2	62.50	20.88	0.50	50.00	20.84
	3	62.50	20.86	0.33	31.58	19.13
	5	55.56	20.75	0.20	20.00	14.91
	10	55.56	19.08	0.10	10.00	8.29
	20	27.78	12.83	0.05	4.39	3.88

Table 14. Comparison of Sensitive-Attribute Disclosure Risk in Movie Data

Threshold g^*	*l*-Diversity			*h*-Affiliation		
	l	***MaxGSAR* (%)**	***AvgGSAR* (%)**	***h***	***MaxGSAR* (%)**	***AvgGSAR* (%)**
0.50	2	66.67	27.40	0.50	50.00	26.96
	3	66.67	27.40	0.33	28.57	18.68
	5	66.67	27.40	0.20	20.00	15.01
	10	66.67	27.40	0.10	10.00	7.28
	20	66.67	27.40	0.05	4.84	3.82
0.67	2	50.00	19.70	0.50	50.00	19.70
	3	50.00	19.70	0.33	28.57	18.15
	5	50.00	19.70	0.20	20.00	14.91
	10	50.00	19.70	0.10	10.00	7.28
	20	50.00	19.70	0.05	4.84	3.82
0.80	2	30.00	13.95	0.50	30.00	13.95
	3	30.00	13.95	0.33	30.00	13.95
	5	30.00	13.95	0.20	20.00	12.96
	10	30.00	13.95	0.10	10.00	7.28
	20	30.00	13.95	0.05	4.84	3.82

Next, we evaluate the performance of *h*-affiliation in terms of data quality. We again first used *g*-balance to form the QI-groups with three thresholds, $g^* = 0.50, 0.67$ and 0.80. We then applied *l*-diversity and *h*-affiliation for anonymizing data and compare their data quality by measuring information loss because of generalization. While identity disclosure reveals both the identity and sensitive values of an individual, attribute disclosure does not necessarily lead to the unique identification of an individual. So, the minimum threshold values used for attribute disclosure risk are usually larger than those for identity disclosure (Duncan & Lambert, 1989; Machanavajjhala et al., 2006; Fung et al., 2010). For each *g* level, we compared the *ANE* values of both methods by controlling the maximum sensitive-attribute disclosure risk (*MaxGSAR)* at three target levels: 10%, 15%, and 20%. To be conservative, we kept the *MaxGSAR* values from *h*-affiliation slightly smaller than those from *l*-diversity and then compared the corresponding *ANE* values.

The results of this experiment are given in Tables 16, 17, and 18. The *ANE* values with *h*-affiliation are substantially smaller than those with *l*-diversity at all levels while the *MaxGSAR* values with *h*-affiliation are about the same as (or slightly smaller than) those with *l*-diversity at all levels. This suggests that the *h*-affiliation results in smaller information loss than *l*-diversity, given the same sensitive-attribute disclosure risk. In the patient data, some diseases are very common across all individuals, as well as for those within a QI-group. *l*-Diversity does not have a built-in mechanism to deal with this problem when a patient has multiple records with multiple diseases. It relies solely on increasing the group size to satisfy a target *MaxGSAR* level, which results in much larger group sizes than those with *h*-affiliation. This issue is not so significant for the movie and bank data because no sensitive values are common across a large number of individuals. Because of this different characteristic in the data, the differences in *ANE* values between *l*-diversity and *h*-affiliation in the patient data are substantially larger than those in the movie and bank data.

In summary, the findings from Tables 13 through 18 indicate that the proposed *h*-affiliation method provides a more intuitive representation of sensitive-attribute disclosure risk, and it outperforms *l*-diversity in terms of both privacy protection and data quality in data with multiple records per person.

Table 15. Comparison of Sensitive-Attribute Disclosure Risk in Bank Data

Threshold g^*	*l*-Diversity			*h*-Affiliation		
	l	*MaxGSAR* (%)	*AvgGSAR* (%)	h	*MaxGSAR* (%)	*AvgGSAR* (%)
0.50	2	100.00	34.19	0.50	50.00	31.59
	3	75.00	33.05	0.33	30.77	21.20
	5	66.67	26.21	0.20	20.00	15.37
	10	27.38	14.49	0.10	10.00	7.92
	20	20.00	10.08	0.05	6.21	6.21
0.67	2	75.00	27.94	0.50	50.00	27.62
	3	75.00	27.78	0.33	30.77	21.20
	5	50.00	25.72	0.20	20.00	15.42
	10	40.00	17.19	0.10	10.00	8.63
	20	20.00	10.08	0.05	6.21	6.21
0.80	2	57.14	22.21	0.50	50.00	22.13
	3	57.14	22.21	0.33	30.77	19.99
	5	50.00	21.75	0.20	20.00	15.42
	10	33.33	16.86	0.10	10.00	8.63
	20	20.00	10.08	0.05	6.21	6.21

Table 16. Comparison of Information Loss Given Attribute Disclosure Risks in Patient Data

Threshold g^*	Target *MaxGSAR* (%)	*l*-Diversity		*h*-Affiliation	
		MaxGSAR (%)	*ANE*	*MaxGSAR* (%)	*ANE*
0.50	20	20.12	14.989	18.37	0.404
	15	15.96	27.249	14.29	1.536
	10	10.80	259.512	9.47	28.317
0.67	20	20.10	15.012	18.08	0.435
	15	14.67	31.880	13.99	1.752
	10	10.43	309.400	9.21	82.312
0.80	20	20.10	16.295	17.97	0.575
	15	14.81	28.871	14.29	1.550
	10	8.92	467.803	8.89	139.537

Table 17. Comparison of Information Loss Given Attribute Disclosure Risks in Movie Data

Threshold g^*	Target *MaxGSAR* (%)	*l*-Diversity		*h*-Affiliation	
		MaxGSAR (%)	*ANE*	*MaxGSAR* (%)	*ANE*
0.50	20	20.00	0.383	19.29	0.198
	15	15.34	0.899	14.29	0.404
	10	10.59	1.959	10.00	0.789
0.67	20	21.11	0.354	20.00	0.176
	15	14.63	0.900	14.56	0.412
	10	9.95	2.005	9.77	0.831
0.80	20	19.09	0.711	18.90	0.277
	15	15.34	0.975	13.90	0.446
	10	9.95	2.139	8.70	1.082

Table 18. Comparison of Information Loss Given Attribute Disclosure Risks in Bank Data

Threshold *g**	Target *MaxGSAR* (%)	*l*-Diversity		*h*-Affiliation	
		MaxGSAR (%)	*ANE*	*MaxGSAR* (%)	*ANE*
0.50	20	20.00	1.795	19.81	0.614
	15	15.79	2.222	15.69	1.208
	10	10.20	8.646	10.17	5.509
0.67	20	21.05	2.132	20.00	0.656
	15	20.25	7.581	15.00	1.759
	10	10.29	21.036	10.00	12.240
0.80	20	21.05	2.955	19.52	0.732
	15	15.13	8.402	14.64	1.922
	10	10.20	40.637	9.70	13.715

9 Discussion

MRPP is an essential aspect of many business analytics and big data applications. Existing data privacy approaches typically assume that each individual corresponds to a single record, which may be inadequate for protecting privacy in MRPP scenarios. The proposed approach overcomes the limitations of existing well-known approaches, effectively reducing the risk of individual-record disclosure and attribute disclosure in MRPP scenarios. Therefore, this research has significant managerial, organizational, and societal implications. The proposed approach should alleviate individuals' concerns about loss of privacy and confidentiality and increase their willingness to allow their data to be shared for secondary uses, such as medical research that benefits society or personalized services that benefit the users themselves. It should also reduce organizations' concerns about possible privacy violations, enabling organizations to share high-quality data safely for legitimate research and analytics purposes in a big data environment.

The proposed approach reduces the disclosure risks in MRPP problems by considering individuals' frequency distribution in a dataset. It balances out the risks of an unbalanced frequency distribution by assigning individuals with the same or similar occurrence frequencies to a QI-group. Thus, the advantage of this approach over traditional methods should be more visible when individuals' frequency distribution is more unbalanced, as was observed in the evaluation study. The frequency distribution in the patient dataset was more unbalanced than that of the movie dataset because individuals with fewer than 20 occurrences had been removed from the original movie dataset before it was released. As a result, the proposed approach performed better with the patient dataset than with the movie dataset. While removing low-frequency individuals reduces risk caused by unbalanced distributions, it also results in information loss. Using our proposed approach, the removal of these low-frequency individuals from the movie dataset would have been unnecessary. Instead, they would be assigned to low-frequency QI-groups, and their disclosure risk could be controlled by the use of *g*-balance and *h*-affiliation measures.

The proposed *g*-balance and *h*-affiliation measures are easy to use because of their relationships with widely used *k*-anonymity and *l*-diversity, respectively. In practice, *k* and *l* values are typically chosen between 5 and 20 (El Emam et al. 2009, 2013; LeFevre et al., 2006; Machanavajjhala et al., 2006; Sweeney, 2002). To set the threshold values *g** and *h** in MRPP cases, the user can first consider these commonly used *k* and *l* values and then calculates the corresponding *g* and *h* values based on Equations (3) and (6) for the thresholds. This ensures that the dataset anonymized with the *g** and *h** thresholds satisfies respective *k*-anonymity and *l*-diversity requirements in case there is only a single record for an individual.

We have considered only one sensitive attribute in this study, but our idea can be easily extended to cases with multiple sensitive attributes. When there are multiple sensitive attributes, *h*-affiliation criteria can be applied to each sensitive attribute in a QI-group separately. In the proposed algorithm, the *h*-affiliation is used as a constraint to check whether the partitioned QI-groups satisfy the sensitive-attribute protection requirement for one sensitive attribute. When there are multiple sensitive attributes, the *h*-affiliation constraint must be satisfied for each of those sensitive attributes. Computationally, this involves checking *h*-affiliation conditions multiple times in Step 2(ii) of the proposed algorithm (Table 5), which is easy to implement.

10 Conclusion

In this study, we investigate the MRPP disclosure problem that is largely overlooked in the data privacy literature. We propose a novel approach to protect data against MRPP-based individual-record disclosure and sensitive-attribute disclosure. We demonstrate that the proposed approach provides significantly better privacy protection against MRPP disclosures than traditional approaches while maintaining greater data quality. Using the proposed approach, organizations can effectively evaluate and mitigate privacy risks with their data when an individual in the dataset has multiple records.

Acknowledgments

The authors are grateful to the senior editor and the five anonymous reviewers for their insightful comments and suggestions that have improved the paper considerably. Xiao-Bai Li's research was supported in part by the National Library of Medicine of the National Institutes of Health under Grant No. R01LM010942. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Library of Medicine or the National Institutes of Health.

References

Abbasi, A., Sarker, S., & Chiang, R. H. (2016). Big data research in information systems: Toward an inclusive research agenda. *Journal of the Association for Information Systems, 17*(2), i-xxxii.

Breiman, L., Friedman, J., Olshen, R. A., Stone, C. J. (1984). *Classification and regression trees.* Wadsworth.

Cavusoglu, H., Phan, T. Q., Cavusoglu, H, Airoldi, E. M. (2016). Assessing the impact of granular privacy controls on content sharing and disclosure on Facebook. *Information Systems Research, 27*(4), 848-879.

Chen, H., Chiang, R. H. L., Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS Quarterly, 36*(4), 1165-1188.

Department of Health and Human Services (DHHS) (2000). Standards for privacy of individually identifiable health information. *Federal Register, 65*(250), 82462-82829.

Duncan, G. T., & Lambert, D. (1989). The risk of disclosure for microdata. *Journal of Business and Economic Statistics, 7*(2), 201-217.

El Emam, K., Dankar, F. K., Neisa, A., Jonker, E. (2013). Evaluating the risk of patient re-identification from adverse drug event reports. *BMC Medical Informatics and Decision Making, 13*(1), Article 114.

El Emam, K., Dankar, F. K., Vaillancourt, R., Roffey, T., Lysyk, M. (2009). Evaluating the risk of re-identification of patients from hospital prescription records. *The Canadian Journal of Hospital Pharmacy, 62*(4), 307-319.

European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD). (1999). PKDD'99 Discovery Challenge. http://lisp.vse.cz/pkdd99/ Challenge/chall.htm

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

Friedman, J. H., Bentley, J. L., & Finkel, R. A. (1977). An algorithm for finding best matches in logarithmic expected time. *ACM Transactions on Mathematical Software, 3*(3), 209-226.

Fung, B., Wang, K., Chen, R., & Yu, P. S. (2010). Privacy-preserving data publishing: A survey of recent developments. *ACM Computing Surveys, 42*(4), Article 14.

Harper, F. M., Konstan, J. A. (2016). The movieLens datasets: History and context. ACM *Transactions on Interactive Intelligent Systems, 5*(4), Article 19.

INFORMS Data Mining Section (2008). INFORMS Data Mining Contest 2008. https://sites.google.com/site/informsdataminingcontest/

Kordzadeh, N., & Warren, J. (2017). Communicating personal health information in virtual health communities: An integration of privacy calculus model and affective commitment. *Journal of the Association for Information Systems, 18*(1), 45-81.

Krumm, J. (2009). A survey of computational location privacy. *Personal and Ubiquitous Computing, 13*(6), 391-399.

LeFevre, K., DeWitt, D. J., & Ramakrishnan, R. (2006). Mondrian multidimensional k-anonymity. *Proceedings of the 22nd International Conference on Data Engineering*.

Li, X-B. & Sarkar, S. (2011). Protecting privacy against record linkage disclosure: A bounded swapping approach for numeric data. *Information Systems Research, 22*(4), 774-789.

Li, X-B. & Sarkar, S. (2013). Class restricted clustering and micro-perturbation for data privacy. *Management Science, 59*(4), 796-812.

Li, X-B. & Sarkar, S. (2014). Digression and value concatenation to enable privacy-preserving regression. *MIS Quarterly, 38*(3), 679-698.

Lohr, S. (2010, March 13). Netflix cancels contest after concerns are raised about privacy. *New York Times*.

Maass, W., Parsons, J., Purao, S., Storey, V. C., & Woo, C. (2018). Data-driven meets theory-driven research in the era of big data: Opportunities and challenges for information systems research. Journal of the *Association for Information Systems, 19*(12), 1253-1273.

Machanavajjhala, A., Gehrke, J., Kifer, D., & Venkitasubramaniam, M. (2006). l-Diversity: Privacy beyond k-anonymity. *Proceedings of the 22nd IEEE International Conference on Data Engineering*.

Menon, S., & Sarkar S. (2016). Privacy and big data: Scalable approaches to sanitize large transactional databases for sharing. *MIS Quarterly, 40*(4), 963-981.

Nergiz, M. E., Clifton, C., & Nergiz, A. E. (2007). Multirelational k-anonymity. *Proceedings of the 23rd International Conference on Data Engineering*.

Rocca, W. A., Yawn, B. P., Sauver, J. L. S., Grossardt, B.R., & Melton, L.J. (2012). History of the Rochester epidemiology project: Half a century of medical records linkage in a US population. *Mayo Clinic Proceedings, 87*(12), 1202-1213.

Samarati, P., & Sweeney, L. (1998). Protecting privacy when disclosing information: k-Anonymity and its enforcement through generalization and suppression. *Proceedings of the IEEE Symposium on Research in Security and Privacy*.

St Sauver, J. L., Grossardt, B. R., Yawn, B. P., Melton III, L. J., Pankratz, J. J., Brue, S. M., Rocca, W. A. (2012). Data resource profile: The Rochester epidemiology project (REP) medical records-linkage system. *International Journal of Epidemiology, 41*(6), 1614-1624.

Sweeney, L. (2002). k-Anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10*(5), 557-570.

Tao, Y., Tong, Y., Tan, S., Tang, S., & Yang, D. (2008). Protecting the publishing identity in multiple tuples. *Proceedings of 22nd Annual IFIP WG 11.3 Working Conference on Data and Applications Security*.

Wang, K., & Fung, B. (2006). Anonymizing sequential releases. Proceedings of the 12th ACM SIGKDD *International Conference on Knowledge Discovery and Data Mining (pp. 414-423).*

Xiao, X., & Tao, Y. (2006). Anatomy: Simple and effective privacy preservation. *Proceedings of the 32nd Conference on Very Large Data Bases*.

Xu, J., Wang, G. A., Li, J., & Chau, M. (2007). Complex problem solving: Identity matching based on social contextual information. *Journal of the Association for Information Systems, 8*(10), 525-545.

Appendix

Proofs of Theorem 1 and Corollary 1

The g-balance function, $g(t)$, achieves the maximum when individuals in t are evenly distributed; i.e., when all the c_i's in equation (1) are equal to the same value c. In this case, we can write equation (1) as follows:

$$g_{\max} = 1 - \sum_{i=1}^{n_t} \left(\frac{c}{n_t c}\right)^2 = 1 - \sum_{i=1}^{n_t} \left(\frac{1}{n_t}\right)^2 = 1 - \frac{1}{n_t} \tag{A1}$$

When each individual in t corresponds to a single record, $c_i = c = 1, \forall i$, and $g = g_{\max}$. Substituting n_t in (A1) by k, we obtain Equation (3) in Corollary 1.

Theorem 1 can be proven by using proof by contradiction. Suppose Equation (2) in Theorem 1 is incorrect; that is, it is possible that

$$n_t < \frac{1}{1-g}$$

Then, rearranging the inequality, we have

$$g > 1 - \frac{1}{n_t}$$

It follows from Equation (A1) that the right-hand side is $g_{\max}$. So, we have $g > g_{\max}$, which is a contradiction. This completes the proof.

Proof of Lemma 1

Let t_p, t_1 and t_2 be the parent dataset and its two subsets with n_{t_p}, n_{t_1} and n_{t_2} the number of individuals in each set, respectively. Let h_p, h_1 and h_2 be the corresponding h-affiliation value, and j_p, j_1 and j_2 be the index of the corresponding sensitive value defined by equation (5), respectively. Let n_{j_p}, n_{j_1} and n_{j_2} be the number of individuals affiliated with j_p, j_1 and j_2, respectively. Then,

$$h_p = n_{j_p}/n_{t_p},\ h_1 = n_{j_1}/n_{t_1},\ h_2 = n_{j_2}/n_{t_2},\ \text{and}\ n_{t_1} + n_{t_2} = n_{t_p}.$$

We show that

$$h_p \le \max\{h_1, h_2\}. \tag{A2}$$

It follows from Equation (5) that $n_{j_1} + n_{j_2} \ge n_{j_p}$. Without loss of generality, assume $h_1 \le h_2$. Then, if $h_p < h_1$, (A2) is obtained immediately. Now, consider $h_p \ge h_1$. If this is true, then

$$\begin{aligned}
&\frac{n_{j_1}+n_{j_2}}{n_{t_1}+n_{t_2}} \ge \frac{n_{j_p}}{n_{t_p}} \ge \frac{n_{j_1}}{n_{t_1}}, \ \Rightarrow\ n_{t_1}n_{j_1} + n_{t_1}n_{j_2} \ge n_{t_1}n_{j_1} + n_{t_2}n_{j_1}, \ \Rightarrow\ n_{t_1}n_{j_2} \ge n_{t_2}n_{j_1},\\
&\Rightarrow \quad n_{t_1}n_{j_2} + n_{t_2}n_{j_2} \ge n_{t_2}n_{j_1} + n_{t_2}n_{j_2} \ \Rightarrow\ (n_{t_1} + n_{t_2})n_{j_2} \ge (n_{j_1} + n_{j_2})n_{t_2},\\
&\Rightarrow \quad \frac{n_{j_2}}{n_{t_2}} \ge \frac{n_{j_1}+n_{j_2}}{n_{t_1}+n_{t_2}} \ge \frac{n_{j_p}}{n_{t_p}}.
\end{aligned}$$

That is, $h_p \le h_2$, and (A2) is obtained. This completes the proof.

About the Authors

Hasan Kartal is an assistant professor in the Department of Management Information Systems at the University of Illinois at Springfield. He earned his PhD at the University of Massachusetts Lowell in Management Information Systems in 2017. His principal research interests are in data science, business analytics, medical and health informatics, data privacy, information economics, big data, text mining, and social media analytics. He is currently focused on information privacy and privacy-preserving data sharing. His works have been published in high-quality peer-reviewed journals such as the *Journal of the Association for Information Systems*, *International Journal of Information Management*, *Information Systems Frontiers*, *Computers & Industrial Engineering, Journal of Global Information Management,* and *ACM Transactions on Management Information Systems*. He participates in and presents at international conferences such as ICIS, INFORMS, and HICSS with a Best Paper Award in the 2015 Workshop on Information Technologies and Systems (WITS 2015).

Xiao-Bai Li is a professor of information systems in the Department of Operations and Information Systems at the University of Massachusetts Lowell, USA. He received his PhD in management science from the University of South Carolina. His research focuses on data science, business analytics, data privacy, and information economics. He has received funding for his research from the National Institutes of Health and National Science Foundation, USA. His work has appeared in *Information Systems Research*, *MIS Quarterly*, *Management Science*, *Operations Research*, *Journal of the Association for Information Systems*, *INFORMS Journal on Computing*, *IEEE Transactions on Knowledge and Data Engineering*, *IEEE Transactions on Systems, Man, and Cybernetics*, *Decision Support Systems*, *Communications of the ACM*, and *European Journal of Operational Research*, among others. He currently serves as an associate editor for *Information Systems Research*, *Decision Support Systems*, *ACM Journal of Data and Information Quality*, and *Information Technology and Management*.