Int J Software Informatics, Volume 6, Issue 3 (2012), pp. 399–417
International Journal of Software and Informatics, ISSN 1673-7288
©2012 by ISCAS. All rights reserved.
E-mail: ijsi@iscas.ac.cn
http://www.ijsi.org
Tel: +86-10-62661040

Assessing Disclosure Risk and Data Utility Trade-off in Transaction Data Anonymization

Grigorios Loukides[1], Aris Gkoulalas-Divanis[2], and Jianhua Shao[1]

[1] (School of Computer Science, Cardiff University, Cardiff, UK)
[2] (IBM Smarter Cities Technology Centre, IBM Research-Ireland, Ireland)
Email: {g.loukides,j.shao}@cs.cf.ac.uk, arisdiva@ie.ibm.com

Abstract Organizations and businesses, including financial institutions and healthcare providers, are increasingly collecting and disseminating information about individuals in the form of transactions. A transaction associates an individual with a set of items, each representing a potentially confidential activity, such as the purchase of a stock or the diagnosis of a disease. Thus, transaction data need to be shared in a way that preserves individuals' privacy, while remaining useful in intended tasks. While algorithms for anonymizing transaction data have been developed, the issue of how to achieve a "desired" balance between disclosure risk and data utility has not been investigated. In this paper, we assess the balance offered by popular algorithms using the R-U confidentiality map. Our analysis and experiments shed light on how the joint impact on disclosure risk and data utility can be examined, which allows the production of high-quality anonymization solutions.
Key words: anonymization; transaction data; R-U Map; trade-off

Loukides G, Gkoulalas-Divanis A, Shao JH. Assessing disclosure risk and data utility trade-off in transaction data anonymization. *Int J Software Informatics*, **Vol.6, No.3 (2012): 399–417. http://www.ijsi.org/1673-7288/6/i137.htm**

1 Introduction

Privacy has long been held as a basic human value that needs protection[60]. In recent years, however, protection for privacy is becoming increasingly threatened, mainly as a result of widespread use of computer and communication technologies. An increasing amount of sensitive data about individuals is entering into computer systems everyday, either out of necessity (e.g. in healthcare) or through lack of careful control (e.g. on social networks). It is important therefore that we consider how sensitive and personal information is protected in today's data collection, management and sharing activities.

Protecting data privacy requires us to ensure (i) security - stored data are not lost or accessed by unauthorized users; (ii) secrecy - no one is able to eavesdrop the data while they are transmitted between authorized users; and (iii) anonymity - private and sensitive information about individuals is not disclosed when data are released. Note that in many cases, releasing data about individuals is not an option, but a necessity. For example, data relating to drug side effects may be collected from

---

Corresponding author: Jianhua Shao, Email: j.shao@cs.cardiff.ac.uk

Received 2011-11-14; Revised 2012-08-22; Accepted 2012-09-04.

www.ijsi.org
Copyright by Institute of Software, Chinese Academy of Sciences. All rights reserved

Name	*Diagnosis codes*
Anne	a b c d e f
Greg	a b e g
Jack	a e
Tom	b f g
Mary	a b
Jim	c f

(a)

Diagnosis codes
(a, b, c, d, e, f, g)
(a, b, c, d, e, f, g)
(a, b, c, d, e, f, g)
(a, b, c, d, e, f, g)
(a, b, c, d, e, f, g)
(a, b, c, d, e, f, g)

(b)

Figure 1. (a) Original dataset, and (b) output of AA

patients through General Practitioners, and such data will need to be released to drug researchers for analysis.

In this paper we consider privacy protection through data anonymity. More specifically, we consider the protection of *transaction data* which are increasingly collected by and shared among organizations and businesses to support a variety of applications, including e-commerce[73] and biomedicine[40]. These datasets are comprised of records, called *transactions*, which consist of sets of items (also called *itemsets*), such as the products purchased by customers from a supermarket, or the diagnosis codes contained in patients' electronic medical records. For example, the table in Fig. 1(a) shows a set of medical records (transactions), each containing the diagnosis codes (items) associated with a specific patient.

Since transaction data can contain private information about individuals, their release and sharing need to be performed in a way that prevents *re-identification* (i.e., the association between an individual and their transaction), so that data sharing policies and regulations are observed[3,1,15]. Note that re-identification is possible even when no explicit identifiers are contained in the released data, as shown in the AOL search data incident[9]. For instance, releasing the table in Fig. 1(a) after removing individuals' names would still allow an attacker, who knows that *Anne* is diagnosed with a and c, to associate her with the first transaction in the table and infer all of her diagnoses.

Several methods that protect transaction data by limiting the probability of re-identification have been proposed[20,41,64]. These methods anonymize data using item *generalization* (i.e., replacing items with more general/abstract ones) and/or *suppression* (i.e., eliminating some items from the data), until the aforementioned probability becomes $\frac{1}{k}$ or less, where k is a parameter that is specified by the data publisher. The table in Fig. 1(b), for example, is produced from the table in Fig. 1(a) when the method of Ref. [64] is applied with $k = 6$. Observe that all diagnosis codes are replaced by (a, b, c, d, e, f, g), which is a *generalized item* interpreted as representing any non-empty subset of $\{a, b, c, d, e, f, g\}$, and that the probability of re-identifying an individual, using the table in Fig. 1(b), is no more than $\frac{1}{6}$.

Protecting transaction data through anonymity has a penalty to pay: the anonymized data will incur some information loss, thereby affecting the utility of the anonymized data. For example, using the table in Fig. 1(b), we will no longer be able to answer accurately how many patients have diagnosis codes a, b and c. Unfor-

www.ijsi.org

www.ijsi.org

Copyright by Institute of Software, Chinese Academy of Sciences. All rights reserved

tunately, achieving both maximum privacy protection and utility of anonymized data is not feasible[41], and these two properties can only be traded-off. This is because making data more anonymous (better protection) often means that data will be less useful (worse utility) in analytic studies[45]. Thus, it is important to consider how a "desired" balance may be achieved in transaction data anonymization. The level of protection and utility that an anonymized transaction dataset will have, depends on which specific method is used to anonymize the data, and within that method, how certain parameters are configured. For example, for the method used to produce the table in Fig. 1(b), setting $k = 4$ could result in an anonymous dataset that has quite different protection and utility properties from those of the table shown in Fig. 1(b). In addition, anonymization methods are based on heuristics, which do not guarantee maximizing data utility.

Consequently, it may be difficult for data publishers to use the proposed anonymization methods effectively in practice to (i) decide which method will deliver the best result in terms of data utility or privacy protection, or (ii) produce an anonymous dataset that has the required or desired balance between these two properties. At the same time, addressing this issue is important, because applying anonymization methods is becoming a requirement in several application domains[1,2].

In this paper, we study how the disclosure risk and the data utility of an anonymous transaction dataset can be assessed jointly. More specifically, we show how an R-U confidentiality map[27] can be constructed for anonymized transaction data, and how this map can be used to examine the following:

- *How the qualities of solutions produced by different methods may be compared.* We evaluate the privacy/utility trade-off offered by three popular transaction data anonymization algorithms[20,41,64], when they are applied to e-commerce[73] and electronic medical record datasets from the Vanderbilt University Medical Center[55]. We show how an R-U confidentiality map can be used to compare different anonymization methods meaningfully.

- *How the parameters of a method may be configured to achieve a required trade-off.* We consider two scenarios here. In the first one, the data publisher wishes to understand what configuration of the parameters would result in an anonymization that has a best privacy risk and data utility trade-off. In the second scenario, the publisher has an upper bound on acceptable privacy risk in mind, and wants to select a configuration that would maximize data utility within this bound. We show how an R-U Confidentiality map can be used to assist data publishers in analyzing the trade-off between disclosure risk and data utility in both of these cases.

Our analysis and experiments shed light on how the joint impact on disclosure risk and data utility can be examined, thereby allowing the production of high-quality anonymization solutions.

The remainder of the paper is organized as follows. Section 2 discusses techniques for transaction data anonymization, while Section 3 provides the necessary background for the techniques that are used in our work. Section 4 discusses the concept of R-U confidentiality map and its use in transaction data publishing. In Section 5, we provide experimental results and, in Section 6, we conclude the paper.

www.jos.org

Copyright by Institute of Software, Chinese Academy of Sciences. All rights reserved

2 Techniques for Anonymizing Transaction Data

Data about individuals need to be published to support a growing number of applications. This has led to the development of methods for anonymizing different types of data, such as relational[7,10,30,32−34,36,43,44,46,47,52,56,68], transaction[64,69,29,41,19,20,65,42,11], sequential[54], trajectory[4,26,25,71,31], and graph data[37,28,39]. Our work is related to transaction data anonymization, however, it employs many concepts that were introduced in the context of relational data anonymization. Thus, we first discuss relational data anonymization methods, in Section 2.1. Subsequently, we review techniques for anonymizing transaction data, in Section 2.2. We note that our intention is to provide a brief overview of the relevant anonymization techniques, and not an extensive survey of different transaction data anonymization methods. For such a survey, we refer the reader to Ref. [17].

2.1 Relational data anonymization

Re-identification concerns via seemingly innocuous attributes were raised by Sweeney in Ref. [62]. In her work, Sweeney showed that the publication of a relational table that contains de-identified data about individuals still allows an attacker to associate individuals with their records. This is because publicly available datasets, such as voter registration lists, can be linked to the published data, based on potentially linkable attributes, called *quasi-identifiers*, such as date of birth, gender, and zip code. A principle to prevent re-identification, called k-anonymity, was also proposed in Ref. [62]. This principle requires each record of the published table to contain the same values in all quasi-identifiers with at least $k-1$ other records in the table. Satisfying k-anonymity offers protection against re-identification, because the probability of linking an individual to their true record, based on quasi-identifiers, is no more than $\frac{1}{k}$. The parameter k controls the level of offered privacy and is set by data publishers.

The process of enforcing k-anonymity is called k-anonymization, and it can be thought of as a two-step process: (i) finding groups of at least k records that minimize the level of data transformation required to satisfy k-anonymity, and (ii) transforming these records to make the table k-anonymous. The first step can be achieved by employing various strategies, such as binary search[56], data partitioning[33], or clustering[43,6,68], to search for a “good” anonymization, as finding a solution with minimal information loss is NP-hard[49]. For instance, LeFevre et al.[33] proposed Mondrian, a partitioning-based algorithm that is reminiscent of the kd-tree construction. Mondrian starts by considering the entire dataset, and then iteratively partitions it into smaller sets of at least k records. Within each iteration, the partitioning is performed by splitting the data across the median of the quasi-identifier with the largest domain.

To transform the data in order to k-anonymize them, the techniques of suppression[70,67] and generalization[56,62,30] have been proposed. In the context of relational data, suppression involves the removal of values in quasi-identifiers, while generalization involves the replacement of values in quasi-identifiers with more general, but semantically consistent values. Generalization can be performed using several different models (also referred to as *recoding* models), which are classified into *global* and *local*. Global generalization models require all records that have the same set of

Copyright by Institute of Software, Chinese Academy of Sciences. All rights reserved

values over quasi-identifiers to be generalized in exactly the same way, whereas local models lift this requirement. Iyengar, for example, proposed the *full-subtree*, global generalization model[30]. This generalization model assumes the existence of a generalization hierarchy, which organizes the different ways in which specific values may be replaced by more general ones, and requires replacing entire subtrees of values with their closest common ancestor in the generalization hierarchy. Local generalization models have been shown to incur lower information loss than global models[68], but they can cause problems for data mining algorithms to work effectively on anonymized data[16].

The anonymization of relational data can be performed based on several privacy principles, such as l-diversity[48], t-closeness[35] and tuple-diversity[43]. Unlike k-anonymity, these principles do not attempt to prevent re-identification, but the association of individuals with sensitive information. To forestall this threat, the works of Refs. [48,35,43] model the sensitive information using a non quasi-identifier attribute, called *sensitive* attribute, and attempt to enhance the protection of published relational data by requiring the values of the sensitive attribute in each anonymous group to follow a certain distribution. For instance, t-closeness[35] measures privacy based on the distance between two distributions; the distribution of the sensitive values in an anonymous group of records in the published table, and the distribution of all sensitive values in the table. When these two distributions are "close" to one another, it is assumed that strong privacy is achieved, as no more sensitive information will be disclosed from the group, than from the table itself. Machanavajjhala et al.[48] and Li et al.[35] extended the Incognito k-anonymization algorithm[32] to enforce l-diversity and t-closeness, respectively, while a bottom-up, clustering-based algorithm for enforcing tuple-diversity has been proposed in Ref. [43].

However, transaction data cannot be anonymized using the aforementioned principles and algorithms without incurring excessive information loss[5,69], for two reasons. First, only a small number out of thousands of possible items are contained in a transaction. Second, transactions have a varying number of items. These characteristics of transaction data make it difficult to find values that are sufficiently similar to allow the type of anonymization described above, with "low" information loss.

2.2 Transaction data anonymization

Guarding against re-identification in transaction data publishing has attracted significant research interest, and several privacy principles and algorithms to achieve this have been proposed[64,29,41,40]. The goal of the privacy principles against re-identification is to prevent an attacker who knows all or some of the items of an individual, from associating the individual with a "small" number of transactions. He et al.[29], for example, assumed that attackers have knowledge of all the items in an individual's transaction and, to guard against such attackers, they introduced a k-anonymity-based principle, called *complete k-anonymity*. This principle requires each transaction in the published dataset to be indistinguishable from at least $k-1$ other transactions in the dataset. Terrovitis et al.[64,65] considered attackers, who know up to m items of an individual and proposed k^m-anonymity to protect against them. More recently, Loukides et al.[4] introduced the principle of *privacy-constrained anonymity*, which assumes that only certain items in an individual's transaction can

Copyright by Institute of Software, Chinese Academy of Sciences. All rights reserved

be used for re-identification. These items are specified by the data publishers, based on their domain knowledge, or automatically, by assuming a worst-case scenario[41]. Privacy-constrained anonymity is more general than both complete k-anonymity and k^m-anonymity, and it has been shown to allow the publication of patient data that remain useful for biomedical analysis tasks[40,22].

Anonymizing data based on any of the above principles, while minimizing the level of information loss incurred by anonymization, is NP-hard[41]. Thus, several heuristic algorithms have been proposed. Specifically, He et al.[29] and Terrovitis et al.[65] considered anonymization algorithms that perform local generalization, based on the full-subtree generalization model. The latter model is similar to that of Ref. [30], but it is applied to items and allows transactions containing the same items to be anonymized differently. The algorithm developed by He et al.[29] is called *Partition* and enforces complete k-anonymity. Partition starts by generalizing all items to the most generalized item (i.e., the item lying in the root of the generalization hierarchy) and then replaces this item with its immediate descendants in the hierarchy, if complete k-anonymity is satisfied. Then, iteratively, it replaces generalized items with less general items (one at a time, starting with the one that incurs the least amount of data distortion), as long as complete k-anonymity is satisfied, or the generalized items are replaced by leaf-level items in the hierarchy. Terrovitis et al.[65] introduced the Local Recoding Anonymization (LRA) algorithm, which enforces k^m-anonymity. LRA partitions a dataset horizontally into sets in a way that would result in low information loss when the data is anonymized, and then generalizes items in each set separately. Due to the local generalization models they adopt, the Partition and LRA algorithms produce data that may not be mined effectively.

Thus, in this work, we consider anonymization algorithms that employ global generalization[64,41,20]. These algorithms employ different privacy models and are based on different heuristics. Consequently, it is not straightforward for data publishers to use these algorithms to construct anonymizations with a desired trade-off between data utility and privacy protection. Our approach attempts to address this specific issue. The first algorithm we consider is called Apriori Anonymization (AA) and has been proposed by Terrovitis et al.[64] to enforce k^m-anonymity. AA works in an iterative, bottom-up fashion. Since an itemset appears in no more transactions than any of its subsets does, it is possible for itemsets that need protection to be examined in a progressive fashion; from single items to sets of m items. Thus, AA generalizes larger itemsets, based on the way their subsets have been generalized. Generalization is performed by traversing the generalization hierarchy in a bottom-up, breadth-first way, using the full-subtree, global generalization model that was proposed by Iyengar[30]. The replacement of the items in an itemset with more general items can increase its support. This helps the enforcement of k^m-anonymity, but increases the level of information loss. Thus, AA starts from leaf-level nodes in the hierarchy and then examines the immediate ascendants of these items, one at a time. This is reminiscent to the strategy followed by the Apriori association rule mining algorithm[8].

Loukides et al.[41] developed COnstraint-based Anonymization of Transactions (COAT), a greedy, bottom-up algorithm to enforce privacy-constrained anonymity. COAT employs a *set-based*, global generalization model, which allows any set of items

www.ijsi.org

Copyright by Institute of Software, Chinese Academy of Sciences. All rights reserved

to be generalized together and does not require generalization hierarchies. The choice of the items generalized by COAT is governed by utility constraints, which are specified by data publishers. These constraints correspond to the most generalized items that can be used to replace a set of items, thus they limit the generalizations to those that are acceptable for intended applications. Specifically, given a set of utility constraints, COAT attempts to construct a generalized item that is not more general than its corresponding utility constraint. When such an item is not found, COAT selectively suppresses a minimum number of items from the corresponding privacy constraint to ensure that privacy-constrained anonymity is satisfied.

Recently, the Privacy-constrained Clustering-based Transaction Anonymization (PCTA) algorithm was proposed by Gkoulalas-Divanis et al.[20]. This algorithm aims to satisfy the privacy-constrained anonymity principle and employs the set-based, global generalization model[41]. PCTA adopts a bottom-up approach that iteratively merges clusters formed by the items of the original dataset. Each original item initially forms a singleton cluster and, subsequently, singleton clusters are merged, in a way that is reminiscent of hierarchical agglomerative clustering algorithms, to satisfy privacy constraints with low information loss.

Another privacy threat in transaction data publishing is *sensitive information disclosure*, which involves the association of individuals with their sensitive information. Sensitive information disclosure has been considered in several works[69,11,42]. Xu et al.[64], for example, introduced the principle of (h, k, p)-coherence, which prevents both re-identification and sensitive information disclosure. The (h, k, p)-coherence treats items that can lead to identity disclosure, called *public items*, similarly to k^m-anonymity (the function of parameter p is the same as m in k^m-anonymity), while limiting the probability an attacker infers any non-public item using a threshold h. The authors of Ref. [64] also developed *Greedy*, a suppression-based algorithm to enforce (h, k, p)-coherence. Greedy works by discovering all unprotected itemsets of minimal size and protects them by iteratively suppressing the item contained in the greatest number of those itemsets. In another line of work, Cao et al.[11] introduced a novel privacy principle against sensitive information disclosure, called ρ-uncertainty. This principle guards against attackers who can use any combination of items, either public or non-public, to infer an individual's sensitive information, and it can be enforced by an algorithm that first iteratively suppresses non-public itemsets, and then generalizes public items using the full-subtree, global generalization model[30]. Loukides et al.[42] proposed a model that prevents both re-identification and the inference of sensitive information, while allowing detailed privacy requirements to be specified and enforced by anonymization. In Ref. [42], the privacy requirements are expressed using implications, called *PS-rules*, each between a set of public items and a set of sensitive items (i.e., items that model the sensitive information of an individual). In this work, we examine anonymization methods that prevent re-identification and leave a study of methods that prevent the disclosure of sensitive information for future work.

Last, we note that there are methods for preventing the inference of sensitive knowledge patterns, which are specified by data publishers, when transaction data are mined after their release. This threat is beyond re-identification and can be addressed by *knowledge hiding* methods[53,50,58,61,23,12,51,24,21]. These methods work by

Copyright by Institute of Software, Chinese Academy of Sciences. All rights reserved

transforming the original data so that sensitive knowledge patterns cannot be mined after their release, whereas interesting, non-sensitive patterns can still be discovered by mining algorithms. Several methods work by hiding association rules[66] based on heuristics[50,53,58,61] or exact techniques[23,24]. Oliveira et al.[53], for example, proposed a heuristic algorithm, which allows different privacy levels to be exercised for rules and performs hiding with a single dataset scan. Methods that focus on hiding classification rules based on suppression[13] or reconstruction[51] have also been proposed.

3 Background

In this section, we present the concepts that are used by the transaction data anonymization algorithms we consider in our work. After introducing some notation, we discuss the data transformation techniques employed by these algorithms, in Section 3.2. Subsequently, we define the anonymization principles that are used by these algorithms to guard against re-identification, in Section 3.3. Last, we present the measures that the anonymization algorithms considered in this work employ to capture data utility, in Section 3.4.

3.1 Notation

Let $\mathcal{I} = \{i_1, ..., i_M\}$ be a finite set of literals, called *items*. Any subset $I \subseteq \mathcal{I}$ is called an *itemset* over $\mathcal{I}$, and is represented as the concatenation of the items it contains. An itemset that has m items, or equivalently a *size* of m, is called an m-itemset and its size is denoted with $|I|$. A dataset $\mathcal{D} = \{T_1, ..., T_N\}$ is a set of N transactions. For instance, Fig. 1(a) illustrates a transaction dataset that contains 6 transactions, where $\mathcal{I} = \{a, b, c, d, e, f, g\}$ and ae is a 2-itemset.

Each *transaction* T_n, $n = 1, ..., N$, is associated with a unique individual and is a pair $T_n = \langle tid, I\rangle$, where tid is a unique identifier and I is an itemset. A transaction $T_n = \langle tid, J\rangle$ *supports* an itemset I, if $I \subseteq J$. Given an itemset I in $\mathcal{D}$, we use $sup(I, \mathcal{D})$ to represent the number of transactions $T_n \in \mathcal{D}$ that support I. For example, the itemset ae is supported by 3 transactions in the dataset shown in Fig. 1(a).

3.2 Generalization and suppression

Anonymizing transaction data against re-identification can be achieved by generalization and suppression, as mentioned in Section 2.2. These methods differ from perturbative methods[59], such as noise addition or data swapping, in that they allow data semantics to be preserved (i.e., an individual will not be associated with false information). Applying suppression results in publishing an anonymized version $\tilde{\mathcal{D}}$ of $\mathcal{D}$ from which one or more items contained in $\mathcal{D}$ have been removed. On the other hand, generalization transforms an original dataset $\mathcal{D}$ to an anonymized dataset $\tilde{\mathcal{D}}$ by mapping items in $\mathcal{D}$ to generalized items[41]. Thus, generalization often incurs less information loss than suppression40].

Suppression and generalization can be applied *globally*, when each occurrence of an item i in $\mathcal{D}$ is suppressed or replaced by the same generalized item $\tilde{i}$ in $\tilde{\mathcal{D}}$, respectively, or *locally*, when this restriction is lifted. Global generalization is defined as follows.

www.ijsi.org

Copyright by Institute of Software, Chinese Academy of Sciences. All rights reserved

Definition 3.1 (Global generalization). A global generalization is a partition $\tilde{\mathcal{I}}$ of $\mathcal{I}$ in which each item i in $\mathcal{I}$ is mapped to a generalized item $\tilde{i}$ in $\tilde{\mathcal{I}}$ that contains i, using a generalization function $\Phi : \mathcal{I} \rightarrow \tilde{\mathcal{I}}$.

Consider $\mathcal{I} = \{a, b, c, d\}$ for example. $\tilde{\mathcal{I}} = \{(a), (b), (c, d)\}$ is a global generalization of $\mathcal{I}$. In this generalization, items a and b are mapped to themselves (i.e., $\Phi(a) = \tilde{i_1} = (a)$ and $\Phi(b) = \tilde{i_2} = (b)$), while c and d are mapped to a generalized item (i.e., $\Phi(c) = \Phi(d) = \tilde{i_3} = (c, d)$ which is interpreted as representing c or d or c and d).

3.3 Anonymization principles and algorithms

To see how generalization can be used to prevent re-identification, observe that, given an anonymized dataset $\tilde{\mathcal{D}}$, an attacker, who knows that an individual is associated with an item i that is supported by $\mathcal{D}$, can link this individual to their transaction with a probability of at most $\frac{1}{sup(\Phi(i),\tilde{\mathcal{D}})}$. It is also easy to see from Definition 3.1 that in global generalization it holds that $sup(i, \mathcal{D}) \leqslant sup(\Phi(i), \tilde{\mathcal{D}})$, because $\Phi(i)$ in $\tilde{\mathcal{D}}$ is supported by all transactions that support i in $\mathcal{D}$, as well as by transactions that support any other item in $\mathcal{D}$ that is mapped to $\Phi(i)$ in $\tilde{\mathcal{D}}$. For example, b is supported by 4 transactions in the original table in Fig. 1(a) and by 6 transactions in the anonymized version of this table, shown in Fig. 1(b). Thus, generalizing i can lead to reducing the probability of re-identifying an individual. On the other hand, a globally suppressed item is not supported by any transactions in $\tilde{\mathcal{D}}$, hence the probability of re-identifying an individual based on this item is zero.

Suppression and generalization, however, need to be used in a principled manner, as otherwise it is possible for either unprotected or practically useless data to be produced[38]. Consider, for example, that the dataset shown in Fig. 1(a) is published after suppressing all items except c and f. The published data does not prevent re-identification, because cf will be supported by a single transaction in the published data. Thus, an attacker who knows that *Jim* is associated with cf and is represented in the published dataset, can uniquely associate *Jim* with his transaction.

As discussed in Section 2.1, the privacy principles that were developed for anonymizing relational data, such as k-anonymity[62], would cause excessive information loss if they were applied to protect transaction data[5], and alternative privacy principles have been developed. In what follows, we formally define the privacy principles that are employed by the anonymization algorithms we consider. The first of these principles is k^m-anonymity, which has been proposed by Terrovitis et al.[64] and is defined as follows.

Definition (k^m-anonymity). Given parameters k and m, a dataset $\mathcal{D}$ satisfies k^m-anonymity when $sup(I, \mathcal{D}) \geqslant k$, for each m-itemset I in $\mathcal{D}$.

A k^m-anonymous dataset provides protection from attackers who know up to any m items of an individual, because it ensures that any combination of these items cannot be used to associate this individual with less than k transactions of the published dataset.

Motivated by applications, including biomedical data sharing, in which the potentially linkable itemsets are known, Loukides et al.[41] proposed the concept of privacy constraint, which is defined as a set of potentially linkable items from $\mathcal{I}$. Satisfying a privacy constraint imposes a lower bound of k to the support of itemsets that need to be protected, and thus limits the probability of re-identification based on the items

Copyright by Institute of Software, Chinese Academy of Sciences. All rights reserved

contained in the constraint, as explained below.

Definition (Privacy constraint). A *privacy constraint* $p = \{i_1, ..., i_r\}$ is a set of potentially linkable items in $\mathcal{I}$. Given a parameter k of anonymity, p is *satisfied* in $\tilde{\mathcal{D}}$ when either $sup(p, \tilde{\mathcal{D}}) \geqslant k$ or $sup(p, \tilde{\mathcal{D}}) = 0$.

A set of privacy constraints is provided as input in both the COAT and PCTA algorithms, and these algorithms ensure that all privacy constraints in the set will be satisfied in the anonymized data produced by them.

3.4 Capturing data utility

A transaction dataset can be anonymized in many different ways, but the one that harms data utility the least, is typically preferred. To capture data utility, the AA algorithm[64] employs the Normalized Certainty Penalty (NCP) measure, which was originally proposed in the context of relational data anonymization by Xu et al.[68]. Before providing the definition of the NCP measure, let us consider a generalization hierarchy $\mathcal{H}$ and define a function LD, which is applied to an item i that has been mapped to $\tilde{i}$ when $\mathcal{D}$ is anonymized to $\tilde{\mathcal{D}}$. The output of LD is the fraction of the number of leaf-level descendants of the subtree rooted at $\tilde{i}$ in $\mathcal{H}$, over the total number of leaves in $\mathcal{H}$. Based on this definition, the NCP is computed as explained below.

Definition 3.4 (Normalized Certainty Penalty (NCP)). Given a generalization hierarchy $\mathcal{H}$, and an original dataset $\mathcal{D}$ that has been anonymized to $\tilde{\mathcal{D}}$ using generalization, the Normalized Certainty Penalty for $\tilde{\mathcal{D}}$ is defined as

$$NCP(\tilde{\mathcal{D}}) = \frac{\sum_{\forall i \in \mathcal{I}} sup(i, \mathcal{D}) \times LD(i, \tilde{\mathcal{D}})}{\sum_{\forall i \in \mathcal{I}} sup(i, \mathcal{D})}$$

Thus, NCP is expressed as the weighted average of the information loss of all generalized items, which are penalized based on the number of leaf-level descendants they have in the generalization hierarchy. Consider, for example, a two-level generalization hierarchy that has the generalized item (a, b, c, d, e, f, g) at the root level, and the items a to g at the leaf-level. The LD scores for all items in the anonymized dataset shown in Fig. 1(b) are 1. This is because, all these items are mapped to (a, b, c, d, e, f, g), and both the subtree rooted at (a, b, c, d, e, f, g) and the generalization hierarchy have 7 leaf-level nodes. Also, a and b are supported by 4 transactions in the original dataset shown in Fig. 1(a), e and f are supported by 3 transactions, c and g by 2 transactions, and d by 1 transaction. Thus, the NCP score for the anonymized dataset of Fig. 1(b) is $\frac{4\times1+4\times1+3\times1+3\times1+2\times1+2\times1+1\times1}{4+4+3+3+2+2+1} = 1$.

The COAT and PCTA algorithms use the *Utility Loss* (UL) measure, which was proposed in Ref. [41] and can be used to capture the information loss incurred by both generalization and suppression. The following definitions explain the computation of the *Utility Loss* (UL) measure.

Definition 3.5 (Utility loss for a generalized item). The Utility Loss (UL) for a generalized item $\tilde{i}$ is defined as

$$UL(\tilde{i}) = \frac{2^{|\tilde{i}|} - 1}{2^{|\mathcal{I}|} - 1} \times w(\tilde{i}) \times \frac{sup(\tilde{i}, \tilde{\mathcal{D}})}{N}$$

where $|\tilde{i}|$ denotes the number of items in $\mathcal{I}$ that are mapped to $\tilde{i}$, and $w : \tilde{\mathcal{I}} \to [0, 1]$ is a function assigning a weight according to the perceived usefulness of $\tilde{i}$ in analysis.

Copyright by Institute of Software, Chinese Academy of Sciences. All rights reserved

www.ijsi.org

Definition 3.6 (Utility loss for an anonymized dataset). The Utility Loss (UL) for an anonymized dataset $\tilde{\mathcal{D}}$ is defined as

$$UL(\tilde{\mathcal{D}}) = \sum_{\forall \tilde{i} \in \tilde{\mathcal{I}}} UL(\tilde{i}) + \sum_{\forall suppressed\ item\ i_m \in \mathcal{I}} \mathcal{Y}(i_m)$$

where $\mathcal{Y} : \mathcal{I} \rightarrow \Re$ is a function that assigns a penalty, which is specified by data owners, to each suppressed item.

UL quantifies information loss based on the size, weight and support of generalized items, imposing a "large" penalty on generalized items that are comprised of a large number of "important" items that appear in many transactions. The denominators $(2^{|\mathcal{I}|} - 1)$ and N in Definition 3.5 are used for normalization purposes, so that the scores for UL are in $[0, 1]$. Moreover, a weight w is used to penalize generalizations exercised on more "important" items. This weight is specified by the data publishers, based on the perceived importance of the items to the subsequent analysis tasks. For instance, assuming that the weight of the generalized item (a, b, c, d, e, f, g) is 1, the UL of the anonymized dataset shown in Fig. 1(b) is computed as $\frac{2^7-1}{2^7-1} \times 1 \times \frac{6}{6} = 1$.

Another way to quantify data utility is to assume that anonymized data are intended for a specific task and measure how accurately they support this task compared to the original data. Average Relative Error (ARE) is a criterion that captures data utility, based on the accuracy of performing query answering on anonymized data. This criterion has been used to evaluate the AA, COAT, and PCTA algorithms, and is also employed by our approach, as we will discuss later in the paper. Given a workload of queries, ARE reflects the average number of transactions that are retrieved incorrectly as part of query answers[41]. Consider, for example, the COUNT query illustrated in Fig. 2(a). Assuming that $I = a$ and $\mathcal{D}$ is the dataset of Fig. 1(a), we can derive an answer of 4 for this query. However, we cannot do the same when this query is applied to the anonymized dataset shown in Fig. 1(b), and an estimated answer needs to be derived. Based on the method of Ref. [40], for example, the estimated answer for this query is 3, and the Relative Error is $\frac{|4-3|}{4} = 0.25$. Given a number of such queries, ARE is computed by averaging their Relative Error scores.

`SELECT COUNT(`T_n`) FROM` $\mathcal{D}$
`WHERE` T_n supports I in $\mathcal{D}$

(a)

`SELECT COUNT(`$\tilde{T_n}$`) FROM` $\tilde{\mathcal{D}}$
`WHERE` $\tilde{T_n}$ supports I in $\tilde{\mathcal{D}}$

(b)

Figure 2. COUNT query applied to (a) original, and (b) anonymized data

4 R-U Confidentiality Map

As maximizing both privacy protection and utility offered by anonymized data is not feasible, the goal of data publishers becomes to produce anonymized data with a "desired" trade-off between these two properties. This calls for a study of the relationship between disclosure risk and data utility, which can be conducted empirically,

www.jos.org.cn

Copyright by Institute of Software, Chinese Academy of Sciences. All rights reserved

based on the concept of R-U confidentiality map[27]. The R-U confidentiality map was originally proposed for additive noise by Duncan et al.[27], but it has been applied to different privacy-preserving techniques, such as topcoding[18], k-anonymization, and randomization[63].

In our context, an R-U confidentiality map is used to study the effectiveness of an anonymization method in terms of the *joint impact* on privacy protection and data utility it produces, for a given set of data under different parameter settings. This is illustrated in Fig. 3, where *Utility* is measured as $\frac{1}{ARE}$ and *Risk* is calculated as $1/min_{\forall p \in \mathcal{P}} sup(p, \tilde{\mathcal{D}})$, with $\mathcal{P}$ being the set of the specified privacy constraints, and ARE and $sup(\bigcup_{\forall i \in p} \Phi(i), \tilde{\mathcal{D}})$ assumed to be non-zero. That is, the utility is a measure of average query answering accuracy of running a workload of queries $\mathcal{W}$ on an anonymized dataset $\tilde{\mathcal{D}}$, and the risk is the upper bound on the probability of re-identification occurring using $\tilde{\mathcal{D}}$. Note that to demonstrate the feasibility of using an R-U confidentiality map, we have opted for measures based on ARE and $sup(\bigcup_{\forall i \in p} \Phi(i), \tilde{\mathcal{D}})$. However, an R-U confidentiality map is not limited to these. We acknowledge the fact that data publishers may want to consider other measures, such as *NCP* for *Utility* and top q-percentiles of *Risk*[63], in studying and comparing the quality of different anonymizations.

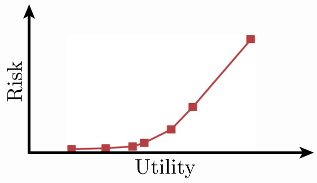


Figure 3. An R-U confidentiality map

To construct a curve in an R-U confidentiality map, we map a set of anonymization solutions, which are produced by applying the same method using different parameters, to a set of two-dimensional points. The x and y coordinates of each point correspond to the level of *Utility* and *Risk* offered by the anonymization solution, respectively. Using an R-U confidentiality map, data publishers are then able to select an anonymization with a "desired" balance between data utility and privacy. For example, as shown in Fig. 4, if the data publisher has an upper bound on risk (represented by the horizontal line) and a minimum requirement on utility (represented by the vertical line), then only the solutions in the lower-right quadrant are acceptable. In addition, R-U confidentiality maps enable a comparison of the effectiveness of different anonymization algorithms, in which case multiple R-U curves will be plotted in a single map, one for each method under comparison. Note that transaction data anoanymization algorithms are designed based on different privacy principles (e.g., AA and COAT) or optimization strategies (e.g., COAT and PCTA), so their comparison is not straightforward. We will discuss these issues further in the next section.

www.ijsi.org
www.ijsi.org

Copyright by Institute of Software, Chinese Academy of Sciences. All rights reserved

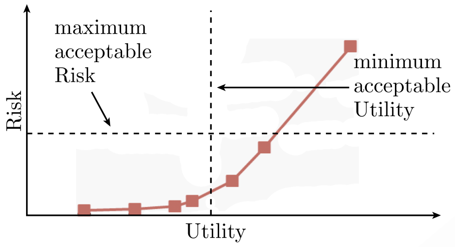


Figure 4. Selecting desired anonymizations

5 Experimental Evaluation

To allow a direct comparison between the tested algorithms, we configured all of them as in Ref. [20] and transformed the resultant anonymized datasets by replacing each generalized item with the set of items it contains. In our experiments, no items were suppressed. We used a C++ implementation of AA provided by the authors of Ref. [64] and implemented COAT and PCTA also in C++. All methods were executed on an Intel 2.0GHz machine with 4GB of RAM and tested using a common framework to measure data utility.

In our experiments, we used the *BMS-WebView-2* dataset (referred to as *BMS2*), which contains click-stream data from an e-commerce site and has been used in Refs. [64,40]. In addition, we used 2 real datasets that contain de-identified patient records derived from the Electronic Medical Record (EMR) system of Vanderbilt University Medical Center[55]1). These datasets are referred to as *VNEC* and $VNEC_{KC}$ and were introduced in Ref. [40]. The datasets we used have different characteristics, shown in Table 1. To measure *Utility*, we used the query workloads of Ref. [20].

Table 1 Description of used datasets

Dataset	N	$\lvert\mathcal{I}\rvert$	**Max. size of** T	**Avg. size of** T
BMS2	77512	3340	161	5.0
VNEC	2762	5830	25	3.1
$VNEC_{KC}$	1335	305	3.1	5.0

In the following, we consider two scenarios in which a desired trade-off between disclosure risk and data utility needs to be achieved. In the first scenario, a user seeks to find the parameters of a specific anonymization algorithm that result in the best trade-off. This case models a practical scenario, when a data publisher has decided to use a specific algorithm and has little technical expertise. The second scenario we consider involves a data publisher who seeks to limit the acceptable privacy risk by selecting a configuration of an algorithm that would maximize data utility within this bound. This type of scenario is realistic and common, for example, in biomedical data sharing where the typical maximum acceptable level of *Risk* is 0.2[14,41].

We started by considering the first scenario. We applied COAT to the *VNEC* and $VNEC_{KC}$ datasets using different k values ranging from 2 to 80 and setting all

---

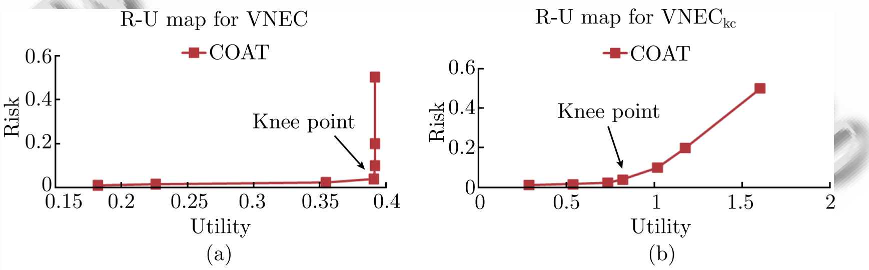
1) These datasets are proprietary and were made available to the first author, while he was working at Vanderbilt University.

www.ijsi.org
www.ijsi.org

Copyright by Institute of Software, Chinese Academy of Sciences. All rights reserved

other parameters as in the single-visit case described in Ref. [41]. This configuration yielded *Risk* values that vary from 1 (when data are published intact) to 0.0125. The R-U confidentiality maps for *VNEC* and $VNEC_{KC}$ are shown in Figs. 5(a) and (b), respectively. Observe that, in these two graphs, both the *Utility* scores for the same *Risk* level and the shape of the curves are different. This makes finding a "desired" trade-off between utility and privacy difficult and justifies the need for using an R-U confidentiality map. Using the latter, data publishers, who do not have specific requirements for data privacy and utility, can release the anonymization corresponding to the *Knee* point on the graph, i.e., the point where there exists the most significant local change in the curve. Given the coordinates of the points of the R-U confidentiality map, locating the knee point can be performed using various methods[72,57]. In this paper, we used the angle-based method[72] to find the Knee points shown in Figs. 5(a) and (b).

Observe that the knee-point in Fig. 5(a) corresponds to an anonymized dataset that has more than 2 times better utility than the dataset with the worst utility has (i.e., the dataset corresponding to the leftmost point in Fig. 5) and 12.5 times better protection than the least protected dataset has (i.e., the dataset corresponding to the rightmost point in Fig. 5). A similar observation can be made from Fig. 5(b). These results show that the use of R-U map in transaction data publishing is extremely beneficial in practice, as it allows data publishers to release anonymized datasets with a desired utility/privacy trade-off.



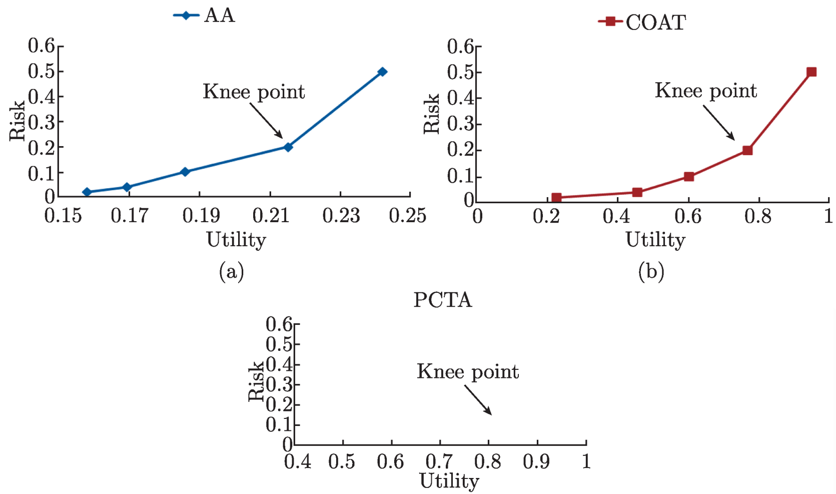
(a) (b)

Figure 5. R-U confidentiality maps for (a) *VNEC* and (b) $VNEC_{kc}$

Then, we applied AA, COAT, and PCTA on *BMS2* using different k values ranging from 2 to 100. The R-U confidentiality maps for these algorithms are illustrated in Figs. 6(a), (b), and (c). In this experiment, all algorithms were configured to achieve k^2-anonymity and COAT ran with a single utility constraint, effectively allowing any possible item generalization. Observe that the shape of the curves for the three algorithms differs significantly, and again the construction of R-U confidentiality map allows data publishers to determine which anonymization is preferred, for example, those at Knee points would represent the anonymizations produced by each of the algorithms with a "good" utility/privacy trade-off. These results demonstrate the effectiveness of our approach when it is applied to many different anonymization algorithms.

www.ijsi.org

Copyright by Institute of Software, Chinese Academy of Sciences. All rights reserved



(a) (b) (c)

Figure 6. R-U confidentiality maps for (a)AA, (b)COAT, and (c) PCTA (*BMS2*)

Finally, we considered the second scenario, in which a data publisher has a maximum acceptable level of *Risk* in mind and wants to release the anonymized version of the dataset that offers the maximum *Utility* within this level of *Risk*. So, if the data publisher in our case was asked to use one of the AA, COAT, and PCTA algorithms to anonymize the *BMS2* dataset with an upper bound disclosure risk of 0.2, then the R-U confidentiality map, shown in Fig. 7, would suggest that PCTA algorithm is preferred, since within this *Risk* bound, it offers better *Utility* than either AA or COAT does. A quantitatively similar result to the one shown in Fig. 7 was observed when our approach was applied to the *VNEC* and $VNEC_{KC}$ datasets, but it was omitted for brevity.

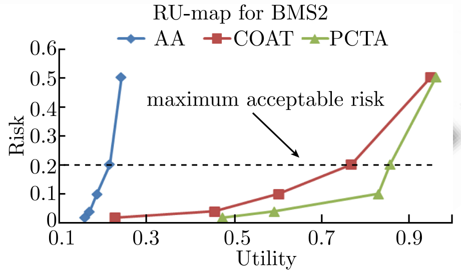


Figure 7. Comparing algorithms using R-U confidentiality map for *BMS2*

Thus, it can be seen that our approach allows a data publisher to select the anonymization algorithm that maximizes *Utility* without violating the acceptable level of *Risk*. Moreover, this can be performed by simply selecting the anonymization algorithm that produces the rightmost point below the horizontal line in the R-U map and configuring it with the parameters corresponding to this point.

www.jos.org.cn

Copyright by Institute of Software, Chinese Academy of Sciences. All rights reserved

6 Conclusions

Several transaction data anonymization methods have been developed recently, but how they may be used to derive anonymizations with a “desired” utility/privacy trade-off has not been considered. In this paper, we addressed this issue by applying the concept of R-U confidentiality map. We explained how R-U maps can be constructed and demonstrated how they may be used in assessing the disclosure risk and data utility trade-off offered by transaction data anonymization solutions.

Through experiments using real data, we have shown the feasibility of our proposed methodology in this paper. However, some further work on this is still necessary. In particular, different risk and utility measures need to be considered and experimented, in order to fully understand how R-U confidentiality maps may be used in assessing and balancing the quality of anonymization in more complex settings, for example, by considering attackers’s background knowledge and data analysis requirements as part of *Risk* and *Utility* assessment.

Acknowledgement

The authors would like to thank Manolis Terrovitis, Nikos Mamoulis, and Panos Kalnis for providing the implementation of the AA algorithm, as well as Bradley Malin (Vanderbilt University), for helpful discussions. The research of the first author is supported by a Research Fellowship from the Royal Academy of Engineering.

References

[1] National Institutes of Health. Policy for sharing of data obtained in NIH supported or conducted genome-wide association studies. NOT-OD-07-088. 2007.
[2] HM Government. Open data white paper - unleashing the potential. 2012.
[3] Health Insurance Portability and Accountability Act of 1996 United States Public Law.
[4] Abul O, Bonchi F, Nanni M. Never walk alone: Uncertainty for anonymity in moving objects databases. ICDE. 2008. 376–385.
[5] Aggarwal CC. On k-anonymity and the curse of dimensionality. VLDB. 2005. 901–909.
[6] Aggarwal G, Feder T, Kenthapadi K, Khuller S, Panigrahy R, Thomas D, Zhu A. Achieving anonymity via clustering. PODS’06. 2006. 153–162.
[7] Aggarwal G, Kenthapadi F, Motwani K, Panigrahy R, Thomasand D, Zhu A. Approximation algorithms for K-anonymity. Journal of Privacy Technology. 2005.
[8] Agrawal R, Srikant R. Fast algorithms for mining association rules in large databases. VLDB. 1994. 487–499.
[9] Barbaro M, Zeller T. A face is exposed for AOL searcher No. 4417749. New York Times, Aug 2006.
[10] Bayardo RJ, Agrawal R. Data privacy through optimal k-anonymization. 21st ICDE. 2005. 217–228.
[11] Cao J, Karras P, Raïssi C, Tan K. *rho*-uncertainty: iference-proof transaction anonymization, PVLDB, 2010, 3(1): 1033-1044.
[12] Chen K, Liu L. Privacy preserving data classification with rotation perturbation. ICDM. 2005. 589–592.
[13] Clifton C. Using sample size to limit exposure to data mining. Journal of Computer Security, 2000, 8(4): 281–307.
[14] El Emam K, Dankar FK. Protecting privacy using k-anonymity. Journal of the American Medical Informatics Association, 2008, 15(5): 627–637.
[15] European Parliament, Council. EU Directive on privacy and electronic communications.
[16] Fung BCM, Wang K, Chen R, Yu PS. Privacy-preserving data publishing: a survey on recent

Copyright by Institute of Software, Chinese Academy of Sciences. All rights reserved

developments. ACM Comput. Surv., 2010, 42.

[17] Fung BCM, Wang K, Fu AW, Yu PS. Introduction to privacy-preserving data publishing: concepts and techniques. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series, 2010.

[18] Duncan GG, Stokes SL. Disclosure risk vs. data utility: the R-U confidentiality map as applied to topcoding. Chance, 2004, 17: 1620.

[19] Ghinita G, Tao Y, Kalnis P. On the anonymization of sparse high-dimensional data. ICDE. 2008. 715–724.

[20] Gkoulalas-Divanis A, Loukides G. PCTA: Privacy-constrained clustering-based transaction data anonymization. EDBT PAIS (to appear), 2011.

[21] Gkoulalas-Divanis A, Loukides G. Revisiting sequential pattern hiding to enhance utility. KDD, 2011. 1316–1324.

[22] Gkoulalas-Divanis A, Loukides G. Utility-guided clustering-based transaction data anonymization. Transactions on Data Privacy, 2012, 5(1): 223–251.

[23] Gkoulalas-Divanis A, Verykios VS. Exact knowledge hiding through database extension. TKDE, 2009, 21(5): 699–713.

[24] Gkoulalas-Divanis A, Verykios VS. Hiding sensitive knowledge without side effects. KAIS, 2009, 20(3): 263–299.

[25] Gkoulalas-Divanis A, Verykios VS, Bozanis P. A network aware privacy model for online requests in trajectory data. DKE, 2009, 68(4): 431–452.

[26] Gkoulalas-Divanis A, Verykios VS. A free terrain model for trajectory k-anonymity. DEXA. 2008. 49–56.

[27] Duncan GT, Keller-McNulty SA, Stokes SL. Disclosure risk vs. data utility: The R-U confidentiality map. Los Alamos National Laboratory Technical Report, LA-UR-01-6428. 2001.

[28] Hay M, Miklau G, Jensen D, Towsley D, Weis P. Resisting structural re-identification in anonymized social networks. PVLDB, 2008, 1(1): 102–114.

[29] He Y, Naughton JF. Anonymization of set-valued data via top-down, local generalization. PVLDB, 2009, 2(1): 934–945.

[30] Iyengar VS. Transforming data to satisfy privacy constraints. KDD. 2002. 279–288.

[31] Kaplan E, Pedersen TB, Savas E, Saygin Y. Discovering private trajectories using background information. DKE, 2010, 69(7): 723–736.

[32] LeFevre K, DeWitt DJ, Ramakrishnan R. Incognito: efficient full-domain k-anonymity. SIGMOD. 2005. 49–60.

[33] LeFevre K, DeWitt DJ, Ramakrishnan R. Mondrian multidimensional k-anonymity. ICDE. 2006. 25.

[34] Li J, Tao Y, Xiao X. Preservation of proximity privacy in publishing numerical sensitive data. SIGMOD'08. 2008. 473–486.

[35] Li N, Li T, Venkatasubramanian S. t-Closeness: privacy beyond k-anonymity and l-diversity. ICDE. 2007. 106–115.

[36] Li T, Li N. Injector: mining background knowledge for data anonymization. ICDE, 2008. 446–455.

[37] Liu K, Terzi E. Towards identity anonymization on graphs. 2008 SIGMOD. 2008. 93–106.

[38] Loukides G, Denny JC, Malin B. The disclosure of diagnosis codes can breach research participants' privacy. Journal of the American Medical Informatics Association, 2010, 17(3): 322–327.

[39] Loukides G, Gkoulalas-Divanis A. Privacy challenges and solutions in the social web. ACM Crossroads, 2009, 16(2): 14–18.

[40] Loukides G, Gkoulalas-Divanis A, Malin B. Anonymization of electronic medical records for validating genome-wide association studies. Proceedings of the National Academy of Sciences, 2010, 17: 7898–7903.

[41] Loukides G, Gkoulalas-Divanis A, Malin B. COAT: Constraint-based Anonymization of Transactions. KAIS, 2011, 28(2): 251-282.

[42] Loukides G, Gkoulalas-Divanis A, Shao J. Anonymizing transaction data to eliminate sensitive inferences. DEXA. 2010. 400–415.

[43] Loukides G, Shao J. Capturing data usefulness and privacy protection in k-anonymisation. SAC.

www.jos.org.cn

Copyright by Institute of Software, Chinese Academy of Sciences. All rights reserved

2007. 370–374.

[44] Loukides G, Shao J. Clustering-based k-anonymisation algorithms. DEXA. 2007. 761–771.

[45] Loukides G, Shao J. Data utility and privacy protection trade-off in k-anonymisation. PAIS. 2008. 36–45.

[46] Loukides G, Shao J. An Efficient Clustering Algorithm for -Anonymisation. Journal of Computer Science and Technology, 2008, 23(2): 188–202 .

[47] Loukides G, Tziatzios A, Shao J. Towards Preference-Constrained-Anonymisation. DASFAA International Workshop on Privacy- Preserving Data Analysis (PPDA). 2009. 231–245.

[48] Machanavajjhala A, Gehrke J, Kifer D, Venkitasubramaniam M. l-Diversity: privacy beyond k-anonymity. ICDE. 2006. 24.

[49] Meyerson A, Williams R. On the complexity of optimal k-anonymity. PODS. 2004. 223–228.

[50] Moustakides GV, Verykios VS. A Max-Min approach for hiding frequent itemsets. ICDM Workshops. 2006. 502–506.

[51] Natwichai J, Li X, Orlowska M. Hiding classification rules for data sharing with privacy preservation. DAWAK. 2005. 468–467.

[52] Nergiz ME, Clifton C. Thoughts on k-anonymization. DKE, 2007, 63(3): 622–645.

[53] Oliveira SRM, Zaïane OR. Protecting sensitive knowledge by data sanitization. ICDM. 2003. 613–616.

[54] Pensa RG, Monreale A, Pinelli F, Pedreschi D. Pattern-preserving k-anonymization of sequences and its application to mobility data mining. The 1st International Workshop on Privacy in Location-Based Applications. 2008.

[55] Roden D, Pulley J, Basford M, Bernard GR, Clayton EW, Balser JR, Masys DR. Development of a large scale de-identified DNA biobank to enable personalized medicine. Clinical Pharmacology and Therapeutics, 2008, 84(3): 362–369.

[56] Samarati P. Protecting respondents identities in microdata release. TKDE, 2001, 13(9): 1010–1027.

[57] Satopaa V, Albrecht J, Irwin D, Raghavan B. Finding a “Kneedle” in a haystack: detecting knee points in system behavior. International Conference on Distributed Computing Systems Workshops (ICDCSW). 2011. 166–171.

[58] Saygin Y, Verykios VS, Clifton C. Using unknowns to prevent discovery of association rules. SIGMOD Record, 2001, 30(4): 45–54.

[59] Sebé F, Domingo-Ferrer J, Mateo-Sanz J, Torra V. Post-masking optimization of the tradeoff between information loss and disclosure risk in masked microdata sets. Inference Control in Statistical Databases, From Theory to Practice. 2002. 163–171.

[60] Smith R, Shao J. Privacy and e-commerce: a consumer-centric perspective. Electronic Commerce Research, 2007, 7(2): 89–116.

[61] Sun X, Yu PS. A border-based approach for hiding sensitive frequent itemsets. 5th IEEE International Conference on Data Mining, 2005. 8 .

[62] Sweeney L. K-anonymity: a model for protecting privacy. IJUFKS, 2002, 10: 557–570.

[63] Teng Z, Du W. Comparisons of k-anonymization and randomization schemes under linking attacks. Proc. of the Sixth International Conference on Data Mining. 2006. 1091–1096.

[64] Terrovitis M, Mamoulis N, Kalnis P. Privacy-preserving anonymization of set-valued data. PVLDB. 2008, 1(1): 115–125.

[65] Terrovitis M, Mamoulis N, Kalnis P. Local and global recoding methods for anonymizing set-valued data. VLDB J, 2011, 20(1): 83–106.

[66] Verykios VS, Gkoulalas–Divanis A. A Survey of Association Rule Hiding Methods for Privacy. chapter 11, pages 267–289. Privacy Preserving Data Mining: Models and Algorithms. Springer, 2008.

[67] Wang SJ, Wong SK, Prusinkiewicz P. An algorithm for multidimensional data clustering. ACM Trans. on Mathematical Software, 1988, 14(2): 153-162.

[68] Xu J, Wang W, Pei J, Wang X, Shi B, Fu AW-C. Utility-based anonymization using local recoding. KDD. 2006. 785–790.

[69] Xu Y, Wang K, Fu AW-C, Yu PS. Anonymizing transaction databases for publication. KDD. 2008. 767–775.

www.ijsi.org

www.ijsi.org

Copyright by Institute of Software, Chinese Academy of Sciences. All rights reserved

[70] Sung SY, Liu Y, Xiong H, Ng PA. Privacy preservation for data cubes. Knowledge Information Systems, 2006, 9(1): 38–61.

[71] Yarovoy R, Bonchi F, Lakshmanan LVS, Wang WH. Anonymizing moving objects: how to hide a MOB in a crowd?. EDBT. 2009. 72–83.

[72] Zhao Q, Hautamaki V, Frnti P. Knee point detection in BIC for detecting the number of clusters. Advanced Concepts for Intelligent Vision Systems, 2008. 664–673.

[73] Zheng Z, Kohavi R, Mason L. Real world performance of association rule algorithms. KDD. 2001. 401–406.

www.ijsi.org

www.ijsi.org

Copyright by Institute of Software, Chinese Academy of Sciences. All rights reserved