WP. 19
ENGLISH ONLY

UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE (UNECE)

CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION

STATISTICAL OFFICE OF THE EUROPEAN UNION (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Tarragona, Spain, 26-28 October 2011)

Topic (iv): Balancing data quality and data confidentiality

On Balancing Disclosure Risk and Data Utility in Transaction Data Sharing Using R-U Confidentiality Map

Prepared by Grigorios Loukides and Jianhua Shao, Cardiff University and
Aris Gkoulalas-Divanis, IBM Research-Zurich, Switzerland

On balancing disclosure risk and data utility in transaction data sharing using R-U confidentiality map

Grigorios Loukides*, Aris Gkoulalas-Divanis**, Jianhua Shao*

* School of Computer Science, Cardiff University, Cardiff, UK, {g.loukides, j.shao}@cs.cf.ac.uk

** Information Analytics Lab, IBM Research - Zurich, Rüschlikon, Switzerland, agd@zurich.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 achieving 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 traced, which allows the production of high-quality anonymization solutions.

1 Introduction

Transaction datasets about individuals are increasingly collected and shared by organizations and businesses to support a wide spectrum of applications, including e-commerce [20] and biomedicine [11]. These datasets are comprised of records, called *transactions*, which consist of sets of items, such as the products purchased by customers from a supermarket, or the diagnosis codes contained in patients' electronic medical records.

Publishing transaction data needs to be performed in a way that prevents *re-identification* (i.e., the association between an individual and their transaction) to adhere to data sharing policies and regulations [1, 2, 5]. 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 [3]. For instance, releasing Table 1(a) after removing individuals' names would still allow an attacker, who knows that *Anne* is

diagnosed with a, b, 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 [7, 12, 18]. These methods anonymize data using item *generalization* (i.e., they replace items with more general/abstract ones) and/or suppression (i.e., they eliminate 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. Table 1(b), for example, is produced from Table 1(a) when the method of [18] 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 Table 1(b), is no more than $\frac{1}{6}$.

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 Apriori

Unfortunately, maximizing both the privacy protection and utility of anonymized data is computationally infeasible [12], and these two properties can only be traded-off. While producing data with a "desired" trade-off is essential in practice, how this may be achieved is still not addressed. In this paper, we evaluate the privacy/utility trade-off offered by three popular transaction data anonymization algorithms [7, 12, 18], when they are applied to e-commerce [20] and electronic medical record datasets from the Vanderbilt University Medical Center [13]. We examine how an R-U confidentiality map [8] can be constructed for anonymized transaction data. We also show how an R-U confidentiality map can be used to assist data publishers in balancing protection from re-identification with data utility and in comparing different anonymization methods.

The remainder of the paper is organized as follows. Section 2 provides the necessary background and Section 3 discusses the concept of R-U confidentiality map and its use in transaction data publishing. In Section 4, we provide experimental results and, in Section 5, we conclude the paper.

2 Background

In this section, we review the techniques of generalization and suppression for anonymizing transaction data. We also discuss popular anonymization principles and algorithms to guard against re-identification, as well as a measure to capture

data utility based on aggregate query answering accuracy.

2.1 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. Each *transaction* T_n, $n = 1, ..., N$, corresponds to a unique individual and is a pair $T_n = \langle tid, I \rangle$, where *tid* is a unique identifier and I is the 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.

2.2 Generalization and suppression

Producing data that prevents re-identification can be achieved by generalization and suppression, which, contrary to perturbative methods [15], 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 [12]. Thus, generalization often incurs less information loss than suppression [11].

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 can be considered as a mapping function Φ from $\mathcal{I}$ to the space of generalized items $\tilde{\mathcal{I}}$, which is constructed by assigning each item $i \in \mathcal{I}$ to a unique generalized item $\Phi(i) = \tilde{i}$ in $\tilde{\mathcal{I}}$ that contains i.

2.3 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 $i \in \mathcal{D}$, 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 that $sup(i, \mathcal{D}) \leq sup(\Phi(i), \tilde{\mathcal{D}})$, 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 data of Table 1(a) and by 6 transactions in the anonymized version of this table, shown in Table 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 [10]. Since privacy principles originally developed for relational data, such as k-anonymity [16], have shown to cause excessive information loss when applied to protect transaction data, alternative privacy principles have been developed. For example, Terrovitis et al. [18] argued that it may be difficult for an attacker to acquire knowledge about all items of a transaction and proposed the k^m-anonymity principle, which is defined as follows.

Definition 2.1 (k^m-anonymity)**.** *Given parameters k and m, a dataset $\mathcal{D}$ satisfies k^m-anonymity when $sup(I, \mathcal{D}) \geq k$, for each m-itemset I in $\mathcal{D}$.*

A k^m-anonymous dataset provides protection from attackers who know up to 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 released dataset. To enforce k^m-anonymity, Terrovitis et al. [18] designed the Apriori Anonymization, henceforth referred to as Apriori. Apriori operates in a bottom-up fashion, beginning with 1-itemsets (items) and subsequently considering incrementally larger itemsets. In each iteration, the algorithm enforces k^m-anonymity using the full-subtree, global generalization model [9].

Motivated by applications, including biomedical data sharing, in which the potentially linkable itemsets are known, Loukides et al. [12] 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 contained in the constraint, as explained below.

Definition 2.2 (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 $sup(p, \tilde{\mathcal{D}}) \geq k$ or $sup(p, \tilde{\mathcal{D}}) = 0$.*

The authors of [12] also proposed limiting the amount of allowable generalization for each item by introducing the concept of *utility constraint*. A utility constraint is a set of items that are allowed to be generalized together, and it models an analysis requirement. A set of utility constraints is specified by data publishers and given as input to an anonymization algorithm. When the generalized dataset produced by the algorithm satisfies the specified utility constraints (i.e., no item in a utility constraint is generalized together with an item not contained in the constraint), it is guaranteed that the generalized dataset remains useful for analysis. An algorithm that can anonymize data, based on privacy and utility constraints, called COnstrained-based Anonymization of Transactions (COAT), was introduced in [12]. This algorithm operates in a greedy fashion and employs global generalization and

suppression. The choice of the items generalized by COAT is governed by utility constraints. 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 utility constraint to ensure privacy.

Another approach to prevent re-identification when adversarial knowledge is expressed as a set of privacy constraints was recently proposed by Gkoulalas et al. [7]. This approach models transaction data anonymization as a constrained clustering problem, where the objective is to construct a clustering comprised of generalized items, such that it satisfies privacy constraints and incurs minimal information loss. This problem is shown to be NP-hard and is dealt by with a heuristic algorithm, called Privacy-constrained Clustering-based Transaction Anonymization (PCTA). PCTA iteratively selects the privacy constraint p that is most likely to require a small amount of generalization in order to be satisfied. Then, it examines all possible ways to generalize items in p and applies the one that leads to the minimum amount of information loss. The process continues until the privacy constraint is satisfied, at which point the next non-satisfied privacy constraint is examined.

2.4 Capturing utility using ARE

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, criteria that measure the information loss incurred by anonymization have been proposed [12, 18]. Examples of such criteria are the Normalized Certainty Penalty (NCP), which is used as an objective measure in the Apriori algorithm, and the Utility Loss (UL), which guides the anonymization performed by COAT and PCTA.

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

SELECT COUNT($\tilde{T}_n$) FROM $\tilde{\mathcal{D}}$
WHERE $\tilde{I}$ supports $\tilde{T}_n$ in $\tilde{\mathcal{D}}$
(b)

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

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. Given a workload of queries, ARE reflects the average number of transactions that are retrieved incorrectly as part of query answers [12]. 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 [11], 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.

3 R-U Confidentiality map

As maximizing both privacy protection and utility offered by anonymized data is computationally infeasible, 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, based on the concept of R-U confidentiality map [8]. The R-U confidentiality map was originally proposed for additive noise [8], but has been applied to different privacy-preserving techniques, including topcoding [6], as well as k-anonymization and randomization [17]. Using an R-U confidentiality map, data publishers are able to select an anonymization with a "good" balance between data utility and privacy, which is beneficial when they do not have specific requirements for privacy protection and data utility. In addition, R-U confidentiality maps enable a comparison of the effectiveness of different anonymization algorithms, which is not trivial when the algorithms are based on different privacy principles (e.g., Apriori and COAT) or optimization strategies (e.g., COAT and PCTA).

In our context, an R-U is a curve that illustrates the effectiveness of an anonymization method in terms of the level of privacy protection from re-identification (henceforth referred to as *Risk*) and data utility for aggregate query answering (henceforth referred to as *Utility*). To construct 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. An example of an R-U confidentiality map constructed based on solutions derived by COAT can be seen in Fig. 3(a).

Utility for an anonymized dataset $\tilde{\mathcal{D}}$ and a workload of queries $\mathcal{W}$ is measured as $\frac{1}{ARE}$, and *Risk* as the upper bound for the probability of performing re-identification using $\tilde{\mathcal{D}}$, which is computed as $1/min_{\forall p \in \mathcal{P}} sup(p, \tilde{\mathcal{D}})$, where $\mathcal{P}$ is the set of the specified privacy constraints. To demonstrate the feasibility of using an R-U map, we opted for simple measures, assuming that ARE and $sup(\bigcup_{\forall i \in p} \Phi(i), \tilde{\mathcal{D}})$ are non-zero. However, we acknowledge the fact that data publishers may want to consider other measures, such as *NCP* for *Utility* and top q-percentiles of *Risk* [17].

4 Experimental results

To allow a direct comparison between the tested algorithms, we configured all of them as in [7] 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 Apriori provided by the authors of [18] 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 [11,18]. 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 [13]. These datasets are referred to as *VNEC* and $VNEC_{KC}$ and were introduced in [11]. The datasets we used have different characteristics, shown in Table 1. To measure *Utility*, we used the query workloads of [7].

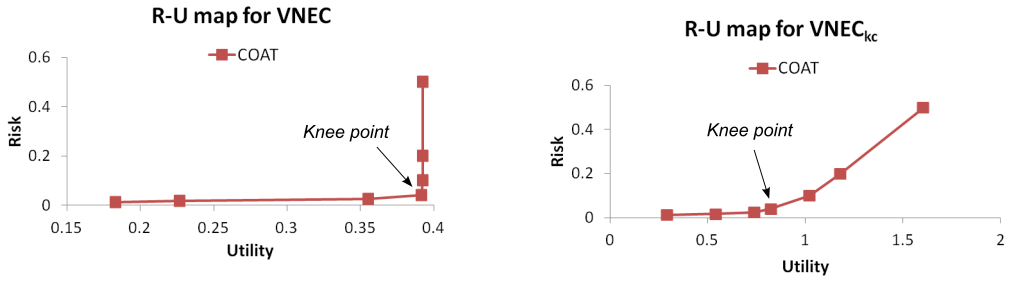
Dataset	N	$\|\mathcal{I}\|$	**Max. size of** T	**Avg. size of** T
BMS2	77512	3340	161	5.0
VNEC	2762	5830	25	3.1
$VNEC_{KC}$	1335	305	3.1	5.0

Table 1: Description of used datasets

First, we applied COAT to the *VNEC* and $VNEC_{KC}$ datasets using different k values ranging from 2 to 80 and setting all other parameters as in the single-visit case described in [12]. This configuration yielded *Risk* values that vary from 1 (when data are published intact) to 0.0125. The R-U maps for *VNEC* and *VNEC* are shown in Figs. 3(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 map, locating the knee point can be performed using various methods [14, 19], such as the angle-based method [19]. The latter method resulted in finding the Knee points shown in Figs. 3(a) and (b) below.

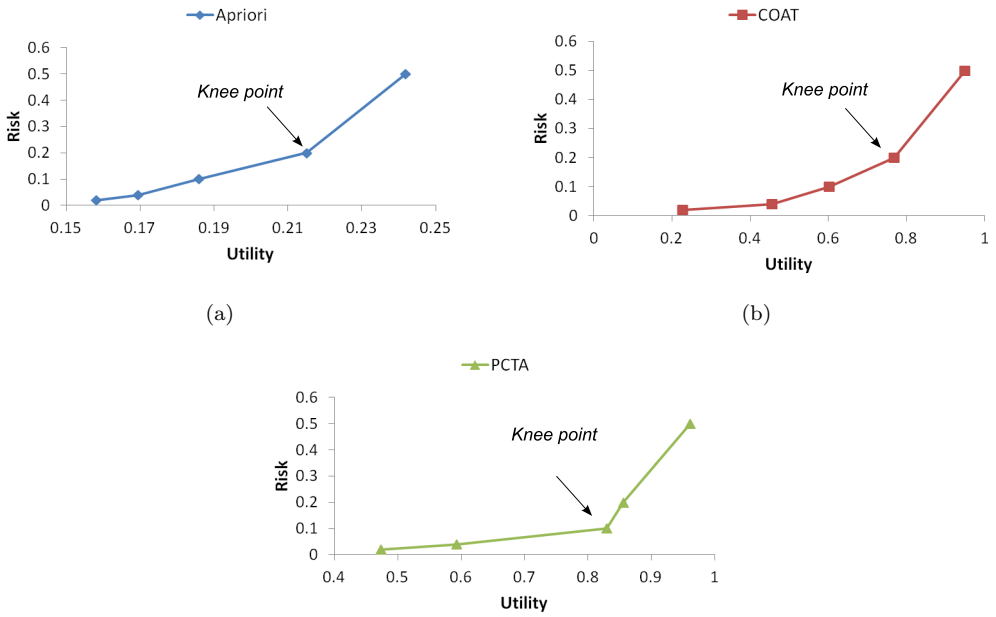
Then, we applied Apriori, COAT, and PCTA on *BMS2* using different k values ranging from 2 to 100. The R-U maps for these algorithms are illustrated in Figs. 4(a), (b), and (c). In this experiment, all algorithms were configured to achieve k^2-anonymity and COAT ran with a single utility constraint, effectively allowing any possible item generalization. Observe that the shape of the curves for the three algorithms differs significantly and that the construction of R-U confidentiality map allows data publishers to find Knee points, i.e., to release anonymizations with a "good" utility/privacy trade-off using either of these algorithms.

Finally, we considered a scenario in which data publishers have a certain maximum acceptable level of *Risk* and want to release the anonymized version of their



(a) (b)

Figure 3: R-U maps for (a) *VNEC* and (b) $VNEC_{kc}$.



(a) (b)

(c)

Figure 4: R-U maps for (a)Apriori, (b)COAT, and (c) PCTA (*BMS2*).

dataset that offers the maximum *Utility* for this level of *Risk*. This scenario is common, for example, in biomedical data sharing, where the typical maximum acceptable level of *Risk* is 0.2 [4, 12]. Assuming that data publishers can use one of the Apriori, COAT, and PCTA agorithms, the R-U confidentiality map, shown in Fig. 5, can help them identify PCTA as the algorithm to use, since it offers better *Utility* than both Apriori and COAT, across all tested levels of *Risk*.

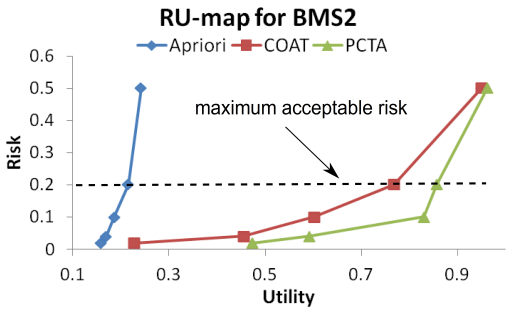


Figure 5: Comparison of anonymization algorithms using R-U maps for *BMS2*.

5 Conclusions

Several transaction data anonymization methods were developed recently, but how they can be used to derive anonymizations with a "desired" utility/privacy trade-off has not been considered. In this paper, we address this issue by applying the concept of R-U confidentiality map. We explain how R-U maps can be constructed and used in transaction data anonymization and, through experiments using real data, we demonstrate the feasibility of our methodology.

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] Health Insurance Portability and Accountability Act of 1996 United States Public Law.

[3] M. Barbaro and T. Zeller. A face is exposed for aol searcher no. 4417749. New York Times, Aug 2006.

[4] K. El Emam and F. K. Dankar. Protecting privacy using k-anonymity. *Journal of the American Medical Informatics Association*, 15(5):627–637, 2008.

[5] Council European Parliament. EU Directive on privacy and electronic communications.

[6] G. Duncan G and S.L. Stokes. Disclosure risk vs. data utility: the r-u confidentiality map as applied to topcoding. *Chance*, 17:1620, 2004.

[7] A. Gkoulalas-Divanis and G. Loukides. PCTA: Privacy-constrained Clustering-based Transaction Data Anonymization. In *EDBT PAIS (to appear)*, 2011.

[8] G.T.Duncan, S.A.Keller-McNulty, and S.L. Stokes. Disclosure risk vs. data utility: The R-U confidentiality map. los alamos national laboratory technical report, LA-UR-01-6428. 2001.

[9] V. S. Iyengar. Transforming data to satisfy privacy constraints. In *KDD*, pages 279–288, 2002.

[10] G. Loukides, J.C. Denny, and B. Malin. The disclosure of diagnosis codes can breach research participants' privacy. *Journal of the American Medical Informatics Association*, 17:322–327, 2010.

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

[12] G. Loukides, A. Gkoulalas-Divanis, and B. Malin. COAT: Constraint-based anonymization of transactions. *KAIS*, 28(2):251–282, 2011.

[13] D. Roden, J. Pulley, M. Basford, G.R. Bernard, E.W. Clayton, J.R. Balser, and D.R. Masys. Development of a large scale de-identified dna biobank to enable personalized medicine. *Clinical Pharmacology and Therapeutics*, 84(3):362–369, 2008.

[14] V. Satopaa, J. Albrecht, D. Irwin, and B. Raghavan. Finding a "kneedle" in a haystack: Detecting knee points in system behavior. In *International Conference on Distributed Computing Systems Workshops (ICDCSW)*, pages 166 –171, 2011.

[15] Francesc Sebé, Josep Domingo-Ferrer, Josep Maria Mateo-Sanz, and Vicenç Torra. Post-masking optimization of the tradeoff between information loss and disclosure risk in masked microdata sets. In *Inference Control in Statistical Databases, From Theory to Practice*, pages 163–171, 2002.

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

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

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

[19] Q. Zhao, V. Hautamaki, and P. Frnti. Knee point detection in bic for detecting the number of clusters. In *Advanced Concepts for Intelligent Vision Systems*, pages 664–673, 2008.

[20] Z. Zheng, R. Kohavi, and L. Mason. Real world performance of association rule algorithms. In *KDD*, pages 401–406, 2001.