HHS Public Access

Author manuscript

Knowl Based Syst. Author manuscript; available in PMC 2018 January 01.

Published in final edited form as:

Knowl Based Syst. 2017 January 1; 115: 15–26. doi:10.1016/j.knosys.2016.10.012.

Anonymizing 1:M microdata with high utility

Qiyuan Gong[a,*], **Junzhou Luo**[a], **Ming Yang**[a], **Weiwei Ni**[a], and **Xiao-Bai Li**[b]

[a]Southeast University, Nanjing, China

[b]University of Massachusetts Lowell, Massachusetts, USA

Abstract

Preserving privacy and utility during data publishing and data mining is essential for individuals, data providers and researchers. However, studies in this area typically assume that one individual has only one record in a dataset, which is unrealistic in many applications. Having multiple records for an individual leads to new privacy leakages. We call such a dataset a 1:M dataset. In this paper, we propose a novel privacy model called (*k, l*)-diversity that addresses disclosure risks in 1:M data publishing. Based on this model, we develop an efficient algorithm named 1:M-Generalization to preserve privacy and data utility, and compare it with alternative approaches. Extensive experiments on real-world data show that our approach outperforms the state-of-the-art technique, in terms of data utility and computational cost.

Keywords

Data anonymization; Data privacy; *k*-anonymity; *l*-diversity; 1:M microdata

1. Introduction

More and more organizations begin to publish microdata for research and analytics purposes. For example, a hospital may release patients' medical records so that researchers can study the characteristics of various diseases. The published microdata contains potentially identifiable sensitive information of individuals, which may lead to privacy disclosure. It is well-known that removing the personal identities from microdata is insufficient due to the possibility of *linking attacks*. An adversary can link some attributes, called quasi-identifier (QID), with external datasets to re-identify individuals. According to a study [1], approximately 87% of the population in the United States can be uniquely identified based on three QID attributes: gender, date of birth, and five-digit zip code. To preserve privacy during data publishing, a number of privacy models and algorithms have been proposed [2–4]. Organizations can anonymize their datasets with these algorithms to satisfy certain privacy models to protect individual privacy. For example, they can apply a *k*-anonymity based algorithm on micro-data to achieve *k*-anonymity such that each record will be indistinguishable from at least $k-1$ other records on QID attributes. So, the adversary cannot re-identify any individual in the dataset with probability higher than $1/k$.

*Corresponding author. gongqiyuan@seu.edu.cn (Q. Gong).

1.1. Motivation and challenges

Almost all previous works assume that one person has only one record in a dataset (called 1:1 dataset), making some data analysis tasks (e.g., health complication analysis, market basket analysis, etc.) not applicable. We call a microdata set that allows multiple records for the same person a 1:M dataset. In real-world database systems, 1:M dataset is more general than 1:1 dataset. For example, in social networking websites (e.g., Facebook, Twitter), a user may post multiple statuses or messages using the same account. A customer may have multiple purchase transactions in a supermarket, which can be identified by the credit card or membership card used. These scenarios have been largely overlooked by the previous works, limiting the applicability of the anonymity models and algorithms.

Consider a 1:M scenario, shown in Table 1(a), where a hospital wants to release patients' records for complication analysis. A patient may have multiple diagnosis records in the same dataset. If we apply *k*-anonymity to this dataset, we will get a sanitized dataset shown in Table 1(b), where the QID values are generalized based on generalization hierarchies shown later in Fig. 1. Note that a system generated personal identifier (PID) is kept to preserve the ownership of each record in order to perform the complication analysis. This *k*-anonymity dataset has two kinds of privacy leakage, as stated in Problems 1 and 2 below.

Problem 1 (Privacy model failure)—Directly applying existing 1:1 privacy models to 1:M datasets may cause privacy disclosure problems due to multiple occurrences of an individual in the dataset.

State-of-the-art privacy models are generally established on datasets with one record for each individual. Privacy models built on this kind of datasets no longer hold on 1:M datasets. In Table 1(b), the QIDs of Bob, David, Daisy and Alice are not well protected even though the dataset satisfies 2-anonymity requirement. For example, if an adversary knows Bob's QID values, i.e., < 18, *M*, 12000 >, he can directly get Bob's disease information, i.e., a_1, a_2 and b_2. Because only the first two records in Table 1(b) have such QID values, and both of them belongs to PID 1. So the adversary can determine that Bob must be PID 1 with 100% confidence. Worse yet, as all records of the same individual usually share the same QID values, the adversaries can infer more information using this knowledge. In our running example, PID 2 (David) and PID 1 (Bob) are assigned to group 2 in Table 1(b). An adversary may observe that PID1's QID values are generalized in group 2, while remaining unchanged in group 1. So he could infer PID2's QID values using PID1's QID information: PID 2 must be male and no older than 15 years old[1], living in a place with zip code from 10,001 to 15,000. PID1′s QID information becomes background knowledge for attacking individuals in his group. We call this *inconsistency attack*, which is caused by inconsistent generalization on the same QID values. Clearly, even if the data publisher applies *l*-diversity to this microdata, the privacy leakage caused by inconsistency attack still exists.

[1]Based on Table 1(b) and Fig. 1, PID2's age must be 15 years old or younger.

Problem 2 (SA fingerprint identification)—In a 1:M dataset, different sensitive attribute (SA) values of the same individual form a feature, which can uniquely identify individuals.

Multiple SA values of the same individual may form a fingerprint called *SA fingerprint*. Similar to unique set in set-valued data publishing [5], these fingerprints can be used to re-identify individuals. In Table 1(b), all patients' SA fingerprints are unique, making them vulnerable to adversaries. For example, if an adversary knows that Bob went to this hospital for treatment of a_1 and b_2, he can infer that tuples 1–3 belong to Bob with 100% confidence, and discover that Bob also has disease a_2. Herein a unique subset of SA fingerprint, i.e., $< a_1, b_2 >$, is enough to re-identify an individual. Different from linking attack on QID, the length of each SA fingerprint can be different, leading to a much higher entropy than QID. So, the more information the adversary gets about SA, the more likely he can identify the victims.

These two problems make 1:M data publishing much more complex than 1:1 data publishing. To achieve privacy, 1:M data anonymization needs more effort s and often causes more information loss for QID and SA attributes. Recent works [6,7] on 1:M problem, either causes privacy leakage [8] or leads to unacceptable information distortion [7]. It is necessary to develop an approach to address the privacy problem in the 1:M setting.

1.2. Contributions

The main contributions of this work are as follows:

1. We propose a new privacy model named (k, l)-diversity for the 1:M data publishing and provide analytical results for the proposed model. By enforcing k-anonymity on SA fingerprint and l-diversity on each equivalence class, (k, l)-diversity can protect QID and SA information during 1:M data publishing.
2. We propose an approach called 1:M-generalization based on existing algorithms. The proposed 1:M-generalization algorithm can efficiently compute anonymized 1:M dataset which satisfies (k, l)-diversity requirement with low information loss.
3. We evaluate our approach by conducting experiments on two real-world 1:M datasets, and compare our approach with existing state-of-art technique.

The rest of the paper is organized as follows. Section 2 reviews the related works. Section 3 formalizes the underlying concepts, analyzes 1:M problem, and introduces the proposed (k, l)-diversity models. Section 4 presents a new anonymization algorithm to achieve (k, l)-diversity. Section 5 describes our experiments that demonstrate the effectiveness of our algorithms. Section 6 concludes the paper with directions for future work.

2. Related work

Since the introduction of k-anonymity in [1,9,10], many privacy models [11,12] and anonymization algorithms [9,13] have been proposed to avoid privacy leakage during data publishing. Sweeney [1] first pointed out that removing identifying attributes (e.g., name and

email address) may not be sufficient to preserve privacy, as the QID attributes (e.g., zip code, gender and date of birth) may potentially identify the record owners. The author proposed *k*-anonymity to ensure that each record is indistinguishable with other $k-1$ records on QIDs. Machanavajjhala et al. [12] observed a drawback of *k*-anonymity that *k*-anonymized dataset is vulnerable to homogeneity attack and background knowledge attack. They proposed *l*-diversity model with SA diversity constraint to enhance privacy protection. Li et al. [11] discovered skewness attack and similarity attack on *l*-diversity, and further proposed t-closeness model with distribution constraint to preserve privacy. However, t-closeness cannot sufficiently protect the privacy of infrequent values, which are more vulnerable to privacy exposure. So Cao et al. [14] proposed β-likeness with strong constraint on relative confidence gain to achieve anonymity.

The goal of anonymization is to find a transformation that satisfies privacy model with minimal information loss. Since achieving this goal is NP-hard [15,16], all existing approaches achieve near-optimal anonymity with approximation algorithms. LeFevre et al. [17] employed the Apriori-like dynamic programming approach based on full-domain generalization. To reduce the information loss caused by global recoding, LeFevre et al. [18] developed a top-down greedy approximation multidimensional *k*-anonymization algorithm called *Mondrian*, based on local recoding. Xu et al. [19] proposed two clustering-based algorithms, which outperformed Mondrian on information loss. Ghinita et al. [20] mapped multi-dimensional anonymization problem to one-dimensional problem, and proposed two efficient algorithms named Hilb and iDist to solve the problem. Ni et al. [21] proposed a clustering-oriented method to keep nearest neighborhood structures of data points during anonymization. Guo et al. [22] developed a clustering-based anonymization approach to preserve the characteristics of data streams. Aggarwal et al. [23] found that when microdata contains a large number of attributes, any generalization necessarily losses considerable information in the microdata due to the curse of dimensionality. To overcome this drawback, Xiao et al. [24] proposed a method called Anatomy, which anonymizes microdata by breaking the correlation between QID and SA attributes. But Anatomy releases precise QID values, making it vulnerable to presence attacks. So Tao et al. [25] proposed AN-GEL to achieve better privacy and marginal publication. Recently, Wong et al. [26] found that *k*-anonymity can be achieved by non-homogeneous generalization, and proposed a technique named ring generalization to achieve higher utility while providing the same privacy guarantee. Xue et al. [27] adapted ring generalization for anonymizing sparse high-dimensional data, they proposed a nonreciprocal recoding anonymization scheme for such data. Doka et al. [28] formulated the optimal-utility *k*-anonymization problem as a network flow problem, and proposed freeform generalization for better utility.

All works discussed so far focus on anonymizing relational datasets. Recent studies have shown that anonymizing transactional set-valued data is quite different, due to the high dimensionality. Xu et al. [8] proposed item suppression based approach for publishing sensitive transaction data with high utility. Different from Xu et al., Terrovitis et al. [5] assumes that each item can be used in re-identification attacks, they proposed k^m-anonymous for transaction data publishing when adversary's background knowledge is limited to no more than m values. They developed Apriori-based anonymization algorithms [5] to achieve k^m-anonymous. He et al. [29] found that Apriori-based anonymization may

cause huge information loss. They proposed a local recoding algorithm named Partition under *k*-anonymity to preserve more data utility. Terrovitis et al. [30] proposed two local recoding algorithms based on Apriori-based anonymization, which outperform Partition on information loss.

All works above assume that one individual has exactly one record in a dataset. Only very few works [6,7] on data privacy consider 1:M datasets. Tao et al. [6] observed that existing models no longer hold when an individual has multiple records in the dataset. They proposed new privacy models based on *k*-anonymity and *l*-diversity, and developed new algorithms based on these models. But their approach does not address the problem caused by SA fingerprint, so an adversary who has SA information can re-identify individuals from their datasets. Poulis et al. [7] divided attributes into relational and transaction attributes and formulated the 1:M problem as a multi-objective optimization problem. They proposed (k, k^m)-anonymous with two constraints, i.e., *k*-anonymity and k^m -anonymous, on each EC. Then they showed that under their privacy model, minimizing information loss on either part would increase the information loss on the other. For instance, minimizing information loss on relational part will distort more information on transaction part. So they developed two algorithms to handle the tradeoff between the two aspects. However, their threat model is somewhat unrealistic. They assumed that the adversary knew both QID and SA information. If this is true, it is not clear why the adversary would still have the interest to launch an attack. On the other hand, such an assumption requires an extremely conservative privacy model, leading to extensive information loss. To reduce information loss, the privacy model we propose assumes that the adversary may have either QID or SA information of the target, but not both.

3. Problem setting and privacy model

Following the definitions in the literature, we classify attributes into four categories.

- **Personal identifier (PID)**: A system generated attribute that uniquely identify an individual in microdata. In a 1:M dataset, a PID is necessary to identify the ownership of different tuples.
- **Quasi-identifier (QID)**: Attributes that, in combination, can be linked with external information to re-identify individuals in microdata (e.g., age, gender and zip code).
- **Sensitive attributes (SA)**: Attributes that are confidential for an individual (e.g., disease, income). In a 1:M dataset, an individual may have multiple SA values.
- **Other attributes**: Attributes that do not fall into the previous three categories. These attributes are considered non-sensitive, and can be published directly.

Without loss of generality, we assume that different records of an individual have the same QID values. This assumption is reasonable because QIDs are typically demographic or geographic attributes, which remain unchanged at least for a certain period[2]. To simplify

[2]We surveyed existing online 1:M datasets, and found that QID values are relatively stable. Most changes on QID values are caused by mistake, e.g., changes on birthday and blood type.

discussion, we consider 1:M dataset with only one sensitive attribute, such that all sensitive values of an individual form a *SA fingerprint*, written as $< sa_1, sa_2, \ldots >$. In Table 1(a), all records of the same individual share the same QID values. So we can transform Tables 1(a) to 2(a), by merging the QID values of the same individual without losing any information. In this *transformed* dataset, a patient has only one record in the dataset, which consists of his/her QID values and SA fingerprint (also called RT-dataset in [7]). It satisfies the assumptions of 1:1 dataset, so existing privacy models and algorithms can be applied to it. This transformation also ensures that no inconsistency will occur after generalization, making *inconsistency attack* unavailable. As a result, aforementioned Problem 1 is no longer an issue with this transformation.

Let T be the original dataset, T' be the transformed dataset, and T^* be the anonymized dataset. Let n be the number of individuals in T. Note that the number of records in T' is also equal to n, but we will still call T' a 1:M dataset because of its context. Let d be the number of QID attributes. Since there is only one sensitive attribute, the attributes in T can be written as $\{A_1, A_2, \ldots, A_{d+1}\}$, where $\{A_1, A_2, \ldots, A_d\}$ are QID attributes and A_{d+1} is the sensitive attribute. Let t be a record in T, t_i be the ith record, $t[j]$ be the jth attribute value of t.

To address privacy breach on QID and SA fingerprint, we introduce two kinds of grouping methodologies:

Definition 1 (Equivalence Class (EC) [18])

For a 1:M dataset T', an equivalence class consists of all records with the same values on all the QID attributes in T'.

Definition 2 (Sensitive Attribute Fingerprint Bucket (SAFB))

For a 1:M dataset T', a sensitive attribute fingerprint bucket consists of all records with the same SA fingerprint in T'.

Both EC and SAFB are non-overlapping divisions on T', where $\cup EC_i = T'$, $\cup SAFB_j = T'$. For each $\forall i \neq j$, we have $EC_i \cap EC_j = \varnothing$, $SAFB_i \cap SAFB_j = \varnothing$. Note that EC is divided based on QID values, while SAFB is divided based on SA fingerprints. Records in EC are indistinguishable on QID, and records in SAFB are indistinguish-able on SA fingerprint. The adversary cannot re-identify individuals based on QID values if all the EC groups are sufficiently large. This is the basic idea of k-anonymity [1]. To deal with SA fingerprint identifications, we extend the notion of k-anonymity for SAFB.

Definition 3 ((k-anonymity for SA fingerprint))

A 1:M dataset T' is k-anonymity, if and only if every SAFB in T' has at least k records.

The k-anonymity defined above ensures that each record in T' is identical to at least $k - 1$ other records on SA fingerprint, such that the adversary cannot re-identify individuals with SA fingerprints or their fragments. The Partition algorithm in [29] is designed for this privacy model, and can anonymize SA fingerprint (set-valued data) with high efficiency. However, ECs in T' are not well protected, and they are vulnerable to linking attack,

homogeneity attack and background knowledge attack [12]. To guard against these attacks, we extend the *l*-diversity principle [12] to require that each EC contains at least *l* 'well-represented' SA fingerprints. Intuitively, *l*-diversity ensures privacy by increasing the 'diversity' in each EC, so that the adversary cannot reveal an individual's SA values without $l-1$ pieces of background knowledge. The *l*-diversity regarding SA fingerprints is defined below.

Definition 4 ((*l*-diversity for 1:M dataset))

A 1:M dataset T' satisfies *l*-diversity, if and only if any EC in T' contains at least *l* 'well-represented SA' fingerprints.

According to [12,20,24], 'well-represented' has many interpretations, leading to different kinds of *l*-diversity, e.g, (*c, l*)-diversity and entropy *l*-diversity. In this paper, we adopt the *l*-diversity formulation from [20,24], i.e., the probability of associating a record in EC with any SA fingerprint is at most $1/l$. Note that the above definition of *l*-diversity is in terms of SA fingerprints, not of SA values. We should point out that having *l* 'well-represented' SA values does not ensure the 'diversity' on SA fingerprints. Suppose, for instance, all records in an EC have the same SA fingerprint, that contains *l* 'well-represented SA values. Then, these records satisfy the traditional *l*-diversity for 1:1 dataset but do not satisfy our *l*-diversity requirement for 1:M dataset. If an adversary has the knowledge about the SA fingerprint of the target, the SA values of all individuals will be exposed. By applying *l*-diversity for 1:M dataset, we can ensure that such adversaries cannot determine any victim's SA values using SA fingerprints with a probability higher than $1/l$. However, *l*-diversity for 1:M dataset cannot prevent SA fingerprint identification as stated in Problem 2. If an adversary has a fragment of Bob's SA fingerprint, e.g., $< a_1, b_2 >$, which is unique in T', he can re-identify Bob and expose Bob's whole SA fingerprint.

To protect against privacy leakage in 1:M data publishing, we propose a new privacy model called (*k, l*)-diversity. The (*k, l*)-diversity can provide protection for both QID and SA fingerprint, reducing the risks of linking attack and attribute disclosure. The (*k, l*)-diversity model consists of two constraints: (1) *k*-anonymity constraint on SAFB to prevent re-identification caused by SA fingerprint (Problem 2). (2) *l*-diversity constraint on EC, which restricts the size of both QID attributes and SA fingerprint.

Definition 5 (((*k, l*)-diversity))

A 1:M dataset T' satisfies (*k, l*)-diversity, if and only if:

1. For any SAFB $\in T'$, there are at least *k* different individuals, and
2. For any EC $\in T'$, there are at least *l* 'well-represented' SA fingerprints.

Intuitively, constraint 1 is *k*-anonymity on SA fingerprints, which ensures that any SA fingerprint is associated with at least *k* individuals. Meanwhile, constraint 2 simultaneously ensures diversity on SA fingerprints and anonymity on QID values. To satisfy (*k, l*)-diversity, T' should satisfy constraints 1 and 2 at the same time. Unfortunately, satisfying both constraints with minimum information loss is NP-hard. Because each constraint is a

special case of the whole problem and both special cases are NP-hard. In Section 4, we propose a heuristic algorithm, which can achieve (k, l)-diversity with high efficiency.

Lemma 1

If a 1:M dataset T' satisfies (k, l)-diversity, then T' satisfies k-anonymity for SA fingerprints.

Proof—Let s be an arbitrary SAFB in T'. There must be at least k individuals in s. Since each individual has either 0 or 1 record in T', there must be at least k records in s. According to the definition of SAFB, these records are indistinguishable on SA fingerprint. So s satisfies the definition of k-anonymity.

Lemma 2

If a 1:M dataset T' satisfies (k, l)-diversity, then T' satisfies l-diversity for 1:M dataset.

Proof—Let ec be an arbitrary EC in T'. There must be at least l 'well-represented' SA fingerprints in ec. So ec satisfies the definition of l-diversity for 1:M dataset.

Theorem 1

If a 1:M dataset T' satisfies (k, l)-diversity, the adversary cannot re-identify any individual using SA or QID values with a probability higher than $max\{1/l, 1/k\}$.

Proof—Lemma 1 ensures that the probability of re-identifying SA fingerprint is at most $1/k$. Lemma 2 ensures that the adversary cannot re-identify any individual with probability larger than $1/l$ from QID values. So the probability that the adversary can re-identify any individual with SA or QID values from T' is $max\{1/l, 1/k\}$.

Table 2 (b) is a (2, 3)-diversity dataset for Table 1(a), obtained by using the 1:M-Generalization algorithm. That is, each SAFB has at least 2 records and each EC contains at least 3 'well-represented' SA fingerprints. So an adversary with QID or SA values cannot re-identify any individual in Table 2(b). For example, suppose an adversary knows some of the Bob's disease information, i.e., a_1 and b_2. Given Table 2(b), he will get two candidates, record 1 and 5, whose anonymized disease value $< A, b_2 >$ covers $< a_1, b_2 >$. Another adversary, who has Bob's QID values, will get three candidates, i.e., record 1, 2 and 3. Without additional information, none of these adversaries can re-identify Bob with probability higher than $max(1/2, 1/3) = 1 / 2$. Moreover, since each EC in Table 2(b) contains at least three 'well-represented' SA fingerprints, the adversary cannot reduce the scope of candidates using homogeneity attack or background knowledge attack.

Another advantage of (k, l)-diversity is that l and k can be chosen separately. Data publisher can choose these two parameters according to the characteristics of dataset and privacy requirements. Note that k-anonymity may reduce the diversity of SA fingerprint, which may increase the difficulty of achieving l-diversity. So a higher l is not recommended when k is very large.

At present, anonymization is usually realized by generalization. By generalizing original specific values to more general values, records in an EC are indistinguishable from each other with regard to QID values. So the adversary cannot re-identify victims from well generalized microdata with high confidence. Generalization hierarchies are needed to conduct a generalization. Examples of generalization hierarchies are shown in Fig. 1.

In 1:M data publishing, two kinds of generalization mechanisms can be utilized: relational generalization [9] and transaction (or set-valued data) generalization [29]. These two kinds of generalization are quite different in nature [5]. In relational generalization, different relational values are generalized by their lowest common ancestor on a relational generalization hierarchy, as shown in Fig. 2(a). In Table 1(b), Bob's age value 18 and David's age value 14 are generalized to interval [11, 20], which covers both Bob and David's age. So the adversary cannot identify David's record even if he knew David is 14 years old. On the other hand, each generalization of different transaction sets is based on lowest common cut on a transaction generalization hierarchy, as shown in Fig. 2(b). This common cut will cover all values in the transaction sets. For example, we can generalize < a_1, a_2, b_2 > and < a_2, b_2 > to < A, b_2 >. In this case, < A, b_2 > covers all values involved in this generalization, i.e., A covers a_1 and a_2, and b_2 covers b_2. In this paper, we will try to incorporate two generalization mechanisms into one algorithm to achieve privacy protection.

4. Anonymization algorithm

Existing anonymization algorithms are largely partition methodologies on microdata. Typically, a data partition is performed only once based on either QID or SA attributes. However, the one-time partition may be insufficient for 1:M dataset, because both SA and QID parts may leak privacy information. Therefore, we perform data partition twice in our proposed algorithm. However, directly applying different algorithms on EC and SAFB will not achieve anonymity on both SA and QID attributes, because the realization on one algorithm may affect the other. Specially, we find that different anonymization orders lead to different results. If we perform SA anonymization after QID generalization, the later operation may violate l-diversity requirement. In contrast, if we perform SA anonymization before QID generalization, the diversity will be preserved. Therefore, we perform SA anonymization first, followed by QID generalization. Algorithm 1 provides an outline of 1:M-Generalization procedure. We use the Partition [29] for SA anonymization and the Mondrian [18] for QID generalization. Both algorithms are designed in a top-down manner, and can be implemented straightforwardly. They can be replaced by more powerful algorithms, e.g., TopDown [19], Hilb and iDist [20], with limited modification. The main phases of 1:M-generalization are as follows:

Algorithm 1

1:M-generalization.

Input: T, k and l

Output: T^*

`1:M-Generalization(T)`

```
// step 1
T′ ← Transform T into 1:1 dataset ;
// step 2
IT ← Partition(T′, k) ;
// step 3
T* ← Mondrian(IT, l) ;
return T*;
```

Step 1: Transformation. We first transform 1:M microdata to 1:1 microdata. This transformation is essential for handling the privacy disclosure in Problem 1. After this transformation, we can use existing algorithms and models for 1:1 datasets to solve the remaining problems.

Step 2: SA fingerprint anonymization. As show in Section 3, transaction (SA fingerprint) generalization is somewhat different from relational (QID) anonymization. First, the set-valued data is usually high-dimensional. Second, only one generalization hierarchy is involved during anonymization. To anonymize SA fingerprint, we apply a local recoding, top-down anonymization algorithm called Partition [29]. To our knowledge, Partition is perhaps the fastest algorithm for set-valued data publishing. It can anonymize set-valued dataset with high efficiency, while minimizing set-valued information loss. Algorithm 2 shows the pseudocode for Partition. After some modification, Partition can efficiently group records in k-sized groups based on the SA fingerprints similarity. Then, we anonymize records in each group using transaction generalization. After generalization, each group becomes a SAFB, where each record is indistinguishable from at least $k - 1$ records on SA fingerprint, satisfying constraint 1 of (k, l)-diversity.

Algorithm 2

Partition for SA fingerprint.

```
Partition(partition, k)
    if partition cannot be split then
        Add partition to global return @@list ;
    else
        // pick a node with max information gain
        splitNode ← pick_node(partition);
        // distribute records to subPartitions
        subPartitions ← distribute_data(partition, splitNode);
        // handle subPartitions with less than k records
        balance_partitions(subPartitions) ;
        for subPartition in subPartitions do
            Partition(subPartition, k) ;
```

Step 3: QID anonymization and SA diversity. After Step 2, both Problem 1 and Problem 2 have been addressed. But the dataset may still be subject to homogeneity attack and background knowledge attack, caused by lack of diversity in EC. In this step, we anonymize QID to satisfy constrains 2 of (k, l)-diversity by applying the Mondrian algorithm on IT. Mondrian is a widely used algorithm in relational data anonymization. It can anonymize QID effectively in a top-down manner. Algorithm 3 shows the pseudocode for Mondrian. Note that in our Mondrian implementation we check allowable split on SA fingerprints rather than on SA values, such that all subPartitions in our algorithms satisfy l-diversity on 1:M during anonymization. So after Step 3, each EC contains at least l 'well-represented' SA fingerprints, making homogeneity attack and background knowledge attack ineffective.

Let n be the number of individuals in T and $c(1 \leq c < < n)$ be the average number of occurrences for each individual. So cn is the total number of records in T. Note that if $c = 1$, then T is a 1:1 dataset. During transformation, we need to traverse all tuples in T, so the time complexity of Step 1 is $O(cn)$. In Step 2, each iteration of the Partition is at least a recursive binary partition on existing partitioned data, and the size of the root is n. In worst case, the partition depth is n/k, where each iteration produce two groups with $(n - k)$ and k records in $O(n - ck)$ running time. So the worst-case running time of Step 2 is $O(n^2/k)$. In balanced scenario, the partition depth is nearly $\log n$, where each iteration is an even binary split on current partition. In this case, the running time of Step 2 is $O(n\log n)$. Step 3 is much faster. According to [18], the time complexity of Step 3 is $O(n\log n)$. So the expected running time of 1:M-generalization is $O(n\log n)$ or $O(n^2/k)$ in the worst case. Later in Section 5, we can see that the actual performance of 1:M-generalization is nearly linear on real-world datasets.

Algorithm 3

Mondrian for 1:M data.

```
Mondrian(partition, l)
    if partition cannot be split then
        Add partition to global return list ;
    else
        /* choose the attribute with the widest (normalized) range    */
        of values
        dim ← Choose_Attribute(partition) ;
        if dim is numeric then
                threshold ← Choose_Threshold(partition, dim);
                lhs ← {t ∈ partition: t[dim] ≤ threshold} ;
                rhs ← {t ∈ partition: t [dim] > threshold} ;
                subPartition ← {lhs} U {rhs} ;
        else
                splitNode ← split(partition, dim);
                subPartitions ← distribute_data(partition,
                splitNode);
        for subPartition in subPartitions do
```

Mondrian(subPartition, *l*) ;

Lemma 3

After Step 2, the intermediate table IT satisfies constraint 1 of (k, l)-diversity.

Proof—According to [29], the Partition algorithm in Step 2 will generate an intermediate table *IT*, in which each partition (SAFB) contains at least *k* different individuals, satisfying constraint 1 of (*k, l*)-diversity.

Lemma 4

After Step 3, the table T satisfies constraint 2 of (k, l)-diversity.*

Proof—As mentioned earlier, our Mondrian implementation ensures that all subPartitions satisfy *l*-diversity for 1:M dataset during anonymization. So after Step 3, each EC contains at least *l* 'well-represented' SA fingerprints, satisfying constraint 2 of (*k, l*)-diversity.

Theorem 2

Microdata anonymized by 1:M-generalization satisfies (k, l)-diversity.

Proof—As we have proven in Lemmas 3 and 4, T^* satisfies constraint 1 and 2 of (*k, l*)-diversity after Step 2 and 3. Meanwhile, the two algorithms will not affect each. So microdata anonymized by 1:M-generalization satisfies (*k, l*)-diversity.

In Fig. 3, we show how the illustrative example in Table 1(a) is anonymized using the proposed 1:M-Generalization. The original dataset T is first transformed to T', which satisfies 1:1 requirement. Then in step 2, T' is transformed to IT by the Partition algorithm, using SA fingerprint generalization. According to Lemma 3, IT satisfies *k*-anonymity for SA fingerprint, but does not satisfies (*k, l*)-diversity. For example, when an adversary knows Bob's SA fingerprint or its fragment, he cannot re-identify Bob from IT with a probability higher than 1/*k*. But if an adversary knows Bob's QID values, i.e., < 18, *M*, 12,000 >, he can re-identify Bob with 100% confidence. That is, the Partition algorithm alone cannot ensure privacy during 1:M data publishing. So in the final step, IT is further generalized to T^* by the Mondrian algorithm, as shown in Fig. 3(d) (same as Table 2(b)). Following the result of Lemma 4, T^* satisfies (*k, l*)-diversity. That is, any individual in T^* cannot be uniquely re-identified using SA or QID values.

5. Experiments and analysis

In this section, we evaluate our approach in terms of data utility and computational efficiency. Specifically, we measure anonymized data quality and execution time to compare 1:M-generalization with RMR (RMERGER) from [7]. According to Section 2, RMR is an alternative approach anonymize data in 1:M settings, although it is originally developed for (k, k^m)-anonymous. Both of the algorithms are implemented in Python[3].

We used two real-world datasets, called the INFORMS[4] and Youtube[5] datasets, for the experiments. Both datasets have been used in related works [7,31]. The entire Youtube dataset is too large for RMR, so we chose a subset of Youtube that contains 85,607 records with 117,752 videos. We configured both datasets in the same way as in [7]. During pre-processing, we removed the duplicate SA values and records with missing values. The characteristics of the datasets are shown in Table 3. We choose $k = 10$, and $l = 5$ as the default parameters, and set $\delta = 0.65$ and $m = 2$ for RMR.

One of the challenges for experimental evaluation is to create generalization hierarchies for the dataset, especially for the SA attribute, i.e., diagnosis codes and related_videos. We followed the idea from [29] and constructed an evenly distributed generalization hierarchy. We built income and disease hierarchies with node fan-out $f = 5$, where the node fan-out indicates how many items are generalized from one level to its parent level in the hierarchy tree. All experiments were run on a HP ProLiant DL580 G5 with 16 GB memory running Linux (Ubuntu Core 13.04).

5.1. Information loss and data utility

To measure the information loss caused by anonymization, we used NCP (Normalized Certainty Penalty) [19]. Specifically, we used two kinds of NCP to measure information loss on QID and SA.

Let ν be a value of attribute A in table T. The basic NCP is defined as

$$NCP(\nu)=\begin{cases} 0 & |\nu|=1 \\ |\nu|/|A| & \text{otherwise} \end{cases} \quad (1)$$

where $|\nu|$ is the number of leaf nodes covered by ν corresponding to generalization hierarchies and $|A|$ is the total number of leaf nodes in attribute A. The value range of NCP is from 0 to 1. The value 0 means no distortion, whereas 1 means the values is generalized to the root of generalization hierarchy. Considering record 1 in Table 2(b), for example, when b_2 is generalized to B, the information loss is $NCP(B) = 2/6 = 1/3$. Similarly, when age value 18 is generalized to [16,20][6], the information loss is $NCP([16, 20]) = 5/100 = 5\%$.

To measure information loss in QID values, let t be a record in T, where t_i denote the ith records in T. Let $t[1], t[2], \ldots, t[d]$ be the QID values of t. QID–NCP is defined as

$$QID-NCP(t)=\frac{\sum_{j=1}^{d} NCP(t[j])}{d} \quad (2)$$

[3]Source code of 1:M_Generalization and RMR
[4]https://sites.google.com/site/informsdataminingcontest
[5]http://netsg.cs.sfu.ca/youtubedata/
[6]We assume age range is [1,100], and zip code range is [10001,30000].

$$QID-NCP(T)=\frac{\sum_{i=1}^{n} QID-NCP(t_i)}{n} \quad (3)$$

where n is the number of records in T. So the QID–NCP of Table 2(b) is (3 * (10/100 + 5000/30000) + 3 * (10/100 + 5000/30000))/6 = 26.67%.

Different from QID–NCP, information loss in SA values is specified in terms of the SA fingerprint in t, denoted as $t[d+1]$. Let p be a SA value in $t[d+1]$. SA–NCP is defined as

$$SA-NCP(T)=\frac{\sum_{i=1}^{n}\sum_{j=1}^{C(t_i[d+1])} NCP(p_j)}{\sum_{i=1}^{n} C(t_i[d+1])} \quad (4)$$

where $C(t[d+1])$ is the number of distinct SA values in SA fingerprint. So the SA-NCP of Table 2(b) is (2/6 + 2/6 + 2/6 + 2/6 + 2/6 + 2/6)/10 = 20%.

We compute QID–NCP and SA–NCP with varying parameters, and present the results in Figs. 4, 5 and 6. Note that in all test cases, QID-NCP of RMR is almost 65%. This is because we set δ = 0.65. RMR sacrifices QID-NCP to achieve better SA–NCP. By setting δ, RMR will merge clusters until the QID–NCP is larger than δ. We set δ = 0.65, because only in this value SA–NCP of RMR is comparable to 1:M-generalization. Otherwise, the RMR's SA–NCP will be much higher.

As shown in Fig. 4, 1:M-generalization preserves more utility than RMR on both QID and SA, when $l = 5$ and $l = 10$. We can observe the sharp increase of information loss from $k = 10$ to $k = 100$ for RMR. That is because RMR enforces both privacy models on EC, which greatly reduces the utility of each EC. So when k increases, RMR need to distort more information on QID and SA to achieve anonymity. Specially, when QID–NCP of each cluster increases, less group merging will be performed, which further causes SA–NCP to increase. So SA–NCP of RMR increases quickly when k increases. Conversely, 1:M-generalization enforces k-anonymity on the whole table, and enforces l-diversity on each EC, which greatly increases the data utility. Note that both QID–NCP and SA–NCP of 1:M-generalization are sensitive to k. As k increases, both QID–NCP and SA–NCP of 1:M-generalization increase gradually. To explain this, recall that SA values are anonymized by the Partition algorithm, which is sensitive to k. When k increases, Partition needs to distort more information to achieve k-anonymity on SA fingerprint, which increases SA-NCP. For the Mondrian algorithm, a larger k implies fewer SAFBs after Step 2, which reduces the diversity on SA fingerprint, making it harder to achieve l-diversity. It is clear from Fig. 4 that RMR will distort more information on QID when k increases. In Fig. 4(a) and (c), we can observe that a larger l will increase the QID–NCP of 1:M-generalization. This is because a larger l requires a larger size for each EC, which increases information loss. The SA–NCP values do not depend on l because Partition is unrelated to l and Mondrian does not change SA values. So, the results of 1M:-generalization for $l = 5$ and $l = 10$ are the same in Fig. 4(b) and (d).

Note that RMR does not contain parameter l. So in Fig. 5, the curve of RMR is stable. We can see that 1:M-generalization preserve more utility than RMR on INFORMS and Youtube datasets, even when $k = 50$. Different from k, only QID–NCP is sensitive to l. Because Partition is unrelated to l, and Mondrian does not change SA fingerprint values during anonymization, the SA–NCP result does not change while varying l. On the other hand, QID values with Mondrian are directly relevant to l. A larger l requires a larger size for each EC, which implies a higher level of generalization for QID values. So QID–NCP increases when l grows. Note that QID–NCP of 1:M-generalization may be higher than RMR, if l is much larger than k (e.g., $l \geq 15$, $k = 10$). But in that case, the privacy guarantee of 1:M-generalization is much higher than RMR, because l-diversity requirement provides better protection than k-anonymity requirement, when l is larger than k. Meanwhile, both QID–NCP and SA–NCP of 1:M-generalization will increase slightly when k increases.

To evaluate if data utility is sensitive to the size of datasets (n), we generated a series of subsets of the data by randomly sampling 5K, 10K, …, ALL records from the full dataset. For each size, we generated 10 sample sets using different random number seeds. We then ran both algorithms with default parameters. The final result is the average of the results from the 10 sample sets. As shown in Fig. 6, 1:M-generalization outperforms RMR on all datasets, especially on QID–NCP. Note that the SA–NCP of RMR reduces quickly when the size of a dataset increases. This is because a larger dataset makes RMR easier to satisfy (k, k^m)-anonymous, which reduces both QID-NCP and SA–NCP. Specially, a small QID–NCP on clusters allows more group merging, which further reduces SA–NCP. On the other hand, both QID–NCP and SA–NCP of 1:M-generalization is reduced gradually when the size of dataset grows. As both Partition and Mondrian is sensitive to dataset size, a larger dataset makes SAFBs and ECs easier to reach k-anonymity for SA fingerprint and l-diversity for 1:M dataset, resulting in smaller information loss.

5.2. Efficiency

We evaluate the efficiency by the total execution time of 1:M-generalization (excluding the pre-processing). Specifically, we ran both algorithms with varying datasets, k and l. The results are shown in Fig. 7. It is clear that 1:M-generalization is more efficient than RMR. The average execution time of 1:M-generalization is less than 30 s, while RMR is up to 1 h. As shown in Fig. 7(a) and (d), the total execution time of 1:M-generalization reduce slightly when k grows. This is because a larger k implies fewer splits during Partition, which reduces the running time of step 2. On the other hand, running time does not appear to change much with parameter l, as shown in Fig. 7(b) and (e). A possible explanation is that a larger l reduces the split level of Mondrian, but also increases the running time of checking if EC satisfies l-diversity. The most apparent change on running time is observed when varying the size of dataset, as shown in Fig. 7(c) and (f). The running time grows about linearly with increased volumes of data, which is expected because the average running time of the 1:M-generalization is $O(n \log n)$.

6. Conclusion and further study

This paper presents a study on privacy preserving data publishing for 1:M microdata. We propose a new privacy model named (*k, l*)-diversity to address this problem. Based on this model, we develop an algorithm called 1:M-generalization for anonymization, and compare it with alternative approaches. Extensive experiments with real-world datasets show that our approach outperforms the state-of-the-art both in terms of execution time and information loss.

This work also initiates several directions for future work. Recall that we have focused on the case where there is a single sensitive attribute. Extending our work to multiple sensitive attributes is a challenging topic. Furthermore, in this paper, we have considered only 1:M microdata that all records of the same individual share the same QID values. In practice, QID values of some records may be changed, forming a strong QID fingerprint, e.g., < 18, *M*, < *single, married* >, 21, 000 >, which needs to be carefully anonymized. Extending our work to such a scenario is an exciting topic.

Acknowledgments

This work is supported by National Natural Science Foundation of China under Grants No. 61272054, 61572130, 61632008, 61320106007, 61502100, 61370077 and61402104, Jiangsu Provincial Natural Science Foundation under Grants BK20150628, BK20140648 and BK20150637, the Fundamental Research Funds for the Central Universities under Grant 2242014R30010, Jiangsu Provincial Key Technology R&D Program under Grant BE2014603, Qing Lan Project of Jiangsu Province, Jiangsu Provincial Key Laboratory of Network and Information Security under Grant BM2003201, and Key Laboratory of Computer Network and Information Integration of Ministry of Education of China under Grant 93K-9.

Xiao-Bai Li's research was supported in part by the National Library of Medicine of the National Institutes of Health, USA, under Grant Number R01LM010942. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Library of Medicine or the National Institutes of Health.

References

1. Sweeney L. K-Anonymity: a model for protecting privacy. Int. J. Uncertain. Fuzz. Knowl. Based Syst. 2002; 10(5):557–570.
2. Fung BCM, Wang K, Chen R, Yu PS. Privacy-preserving data publishing: a survey of recent developments. ACM Comput. Surv. 2010; 42(14):1–14. 53.
3. Komishani EG, Abadi M, Deldar F. Pptd: preserving personalized privacy in trajectory data publishing by sensitive attribute generalization and trajectory local suppression. Knowl. Based Syst. 2016; 94:43–59.
4. Zhou B, Pei J, Luk W. A brief survey on anonymization techniques for privacy preserving publishing of social network data. SIGKDD Explor. Newsl. 2008; 10(2):12–22.
5. Terrovitis M, Mamoulis N, Kalnis P. Privacy-preserving anonymization of set-valued data. Proceedings of the 34th International Conference on Very Large Data Bases(VLDB), VLDB Endowment. 2008
6. Tao Y, Tong Y, Tan S, Tang S, Yang D. Protecting the publishing identity in multiple tuples. Data Appl. Security. 2008; XXII:205–218.
7. Poulis G, Loukides G, Gkoulalas-Divanis A, Skiadopoulos S. Anonymizing data with relational and transaction attributes. Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases(ECML/PKDD). 2013
8. Xu Y, Fung B, Wang K, Fu A, Pei J. Publishing sensitive transactions for item-set utility. Proceedings of 8th IEEE International Conference on Data Mining(ICDM). 2008:1109–1114.

9. Sweeney L. Achieving k-anonymity privacy protection using generalization and suppression. Int. J. Uncertain. Fuzz. Knowl. Based Syst. 2002; 10(5):571–588.
10. Ciriani V, di Vimercati SDC, Foresti S, Samarati P. K-anonymity, Secure Data Management in Decentralized Systems. Springer-Verlag. 2007
11. Li N, Li T, Venkatasubramanian S. t-closeness: Privacy beyond k-anonymity and l-diversity. IEEE 23rd International Conference on Data Engineering(ICDE), IEEE. 2007:106–115.
12. Machanavajjhala, A., Gehrke, J., Kifer, D., Venkitasubramaniam, M. L-diversity: privacy beyond k-anonymity. Washington, DC, USA: IEEE Computer Society; 2006.
13. Nergiz ME, Clifton C. Thoughts on k-anonymization. Data Knowl. Eng. 2007; 63:622–645.
14. Cao J, Karras P. Publishing microdata with a robust privacy guarantee. Proceedings of the 38th International Conference on Very Large Data Bases(VLDB), VLDB Endowment. 2012
15. Meyerson A, Williams R. On the complexity of optimal k-anonymity. Proceedings of the Twenty-third ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems(PODS), ACM. 2004:223–228.
16. Xiao X, Yi K, Tao Y. The hardness and approximation algorithms for l-diversity. Proceedings of the 13th International Conference on Extending Database Technology(EDBT), ACM. 2010:135–146.
17. LeFevre K, DeWitt DJ, Ramakrishnan R. Incognito: efficient full-domain k-anonymity. Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data(SIGMOD), ACM. 2005:49–60.
18. LeFevre K, DeWitt DJ, Ramakrishnan R. Mondrian multidimensional k-anonymity. Proceedings of the 22nd International Conference on Data Engineering(ICDE), IEEE Computer Society. 2006:25.
19. Xu J, Wang W, Pei J, Wang X, Shi B, Fu AW-C. Utility-based anonymization using local recoding. Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(SIGKDD), ACM. 2006:785–790.
20. Ghinita G, Karras P, Kalnis P, Mamoulis N. Fast data anonymization with low information loss. Proceedings of the 33rd International Conference on Very Large Data Bases(VLDB), VLDB Endowment. 2007:758–769.
21. Ni W, Chong Z. Clustering-oriented privacy-preserving data publishing. Knowl. Based Syst. 2012; 35:264–270.
22. Guo K, Zhang Q. Fast clustering-based anonymization approaches with time constraints for data streams. Knowl. Based Syst. 2013; 46:95–108.
23. Aggarwal CC. On k-anonymity and the curse of dimensionality. Proceedings of the 31st International Conference on Very Large Data Bases(VLDB), VLDB Endowment. 2005:901–909.
24. Xiao X, Tao Y. Anatomy: simple and effective privacy preservation. Proceedings of the 32nd International Conference on Very Large Data Bases(VLDB), ACM. 2006:139–150.
25. Tao Y, Chen H, Xiao X, Zhou S, Zhang D. Angel: enhancing the utility of generalization for privacy preserving publication. IEEE Trans. Knowl. Data Eng. 2009; 21(7):1073–1087.
26. Wong WK, Mamoulis N, Cheung DWL. Non-homogeneous generalization in privacy preserving data publishing. Proceedings of the 2010 ACM SIG-MOD International Conference on Management of Data(SIGMOD), ACM. 2010:747–758.
27. Xue M, Karras P, Raïssi C, Vaidya J, Tan K-L. Anonymizing set-valued data by nonreciprocal recoding. Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(SIGKDD), ACM. 2012:1050–1058.
28. Doka K, Xue M, Tsoumakos D, Karras P. k-anonymization by freeform generalization. Proceedings of the 10th ACM Symposium on Information, Computer and Communications Security(ASIA CCS), ACM. 2015:519–530.
29. He Y, Naughton JF. Anonymization of set-valued data via top-down, local generalization. Proceedings of the 35th International Conference on Very Large Data Bases(VLDB), VLDB Endowment. 2009
30. Terrovitis M, Mamoulis N, Kalnis P. Local and global recoding methods for anonymizing set-valued data. VLDB J. 2011; 20(1):83–106.

31. Loukides G, Liagouris J, Gkoulalas-Divanis A, Terrovitis M. Disassociation for electronic health record privacy. J. Biomed. Inf. 2014; 50:46–61.

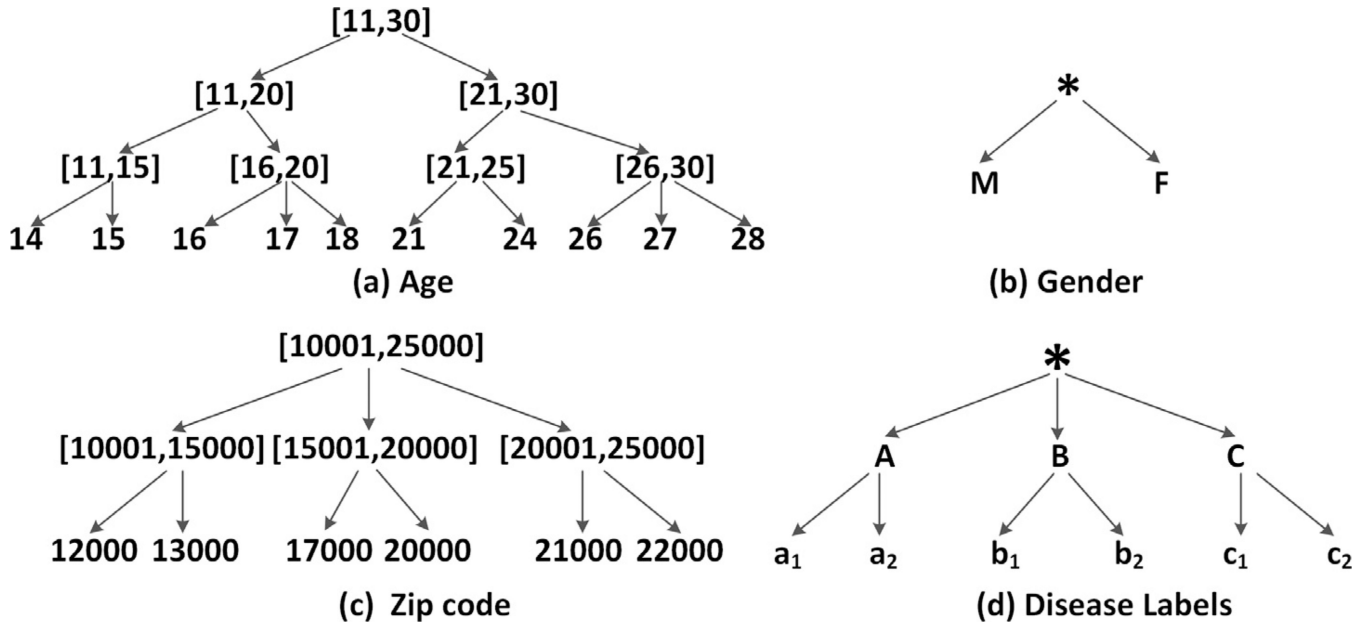


Fig. 1.
Generalization hierarchies.

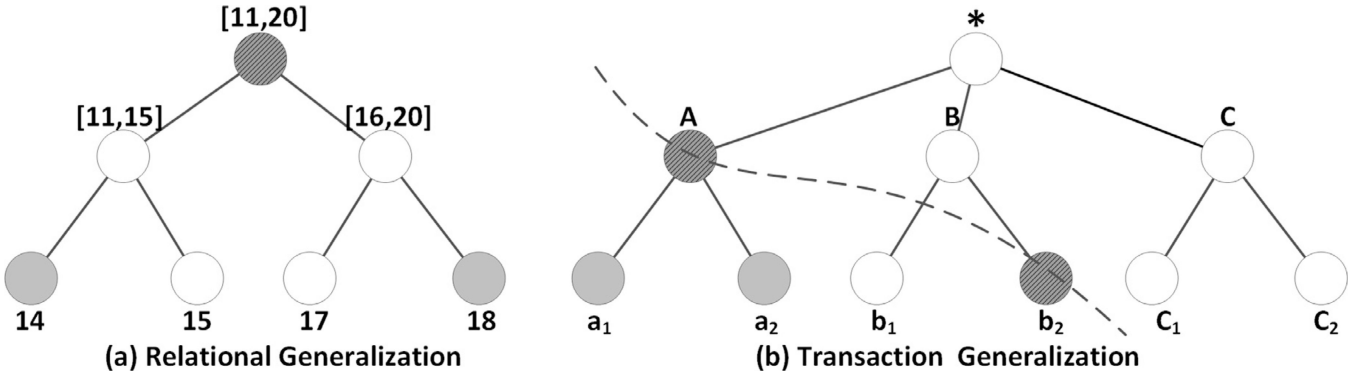


Fig. 2.
Relational generalization vs. transaction generalization.

tuple ID	PID	Age	Gender	Zip code	Disease
1(Bob)	1	18	M	12000	a_1
2(Bob)	1	18	M	12000	a_2
3(Bob)	1	18	M	12000	b_2
4(David)	2	14	M	13000	b_1
5(Tom)	3	21	F	21000	b_2
6(Simon)	4	16	M	14000	c_2
7(Daisy)	5	27	F	22000	a_2
8(Daisy)	5	27	F	22000	b_2
9(Alice)	6	28	F	21000	c_1
10(Alice)	6	28	F	21000	c_2

(a) T

Step 1
Transform

tuple ID	Age	Gender	Zip code	Disease
1-3(Bob)	18	M	12000	$< a_1, a_2, b_2 >$
4(David)	14	M	13000	$< b_1 >$
5(Tom)	21	F	21000	$< b_2 >$
6(Simon)	16	M	14000	$< c_2 >$
7-8(Daisy)	27	F	22000	$< a_2, b_2 >$
9-10(Alice)	28	F	21000	$< c_1, c_2 >$

(b) T'

Step 2
Partition

Age	Gender	Zip code	Disease
18	M	12000	$< A, b_2 >$
14	M	13000	$< B >$
21	F	21000	$< B >$
16	M	14000	$< C >$
27	F	22000	$< A, b_2 >$
28	F	21000	$< C >$

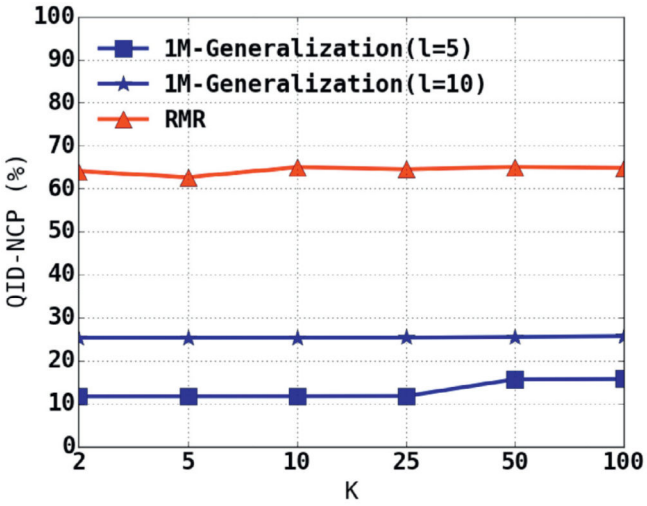
(c) IT

Step 3
Mondrian

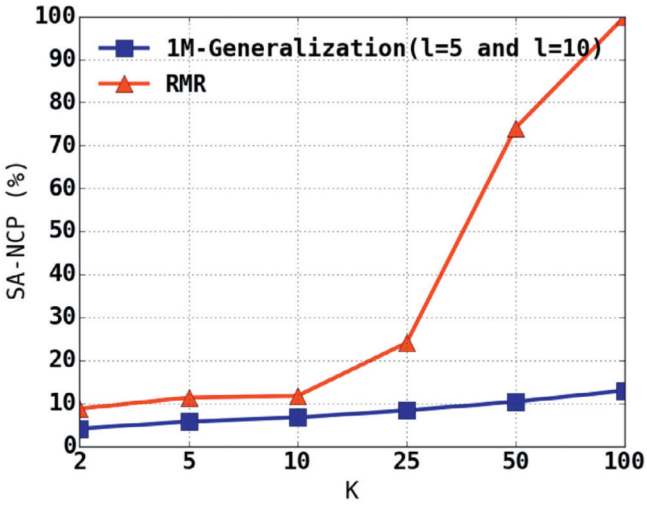
Age	Gender	Zip code	Disease
[11,20]	M	[10001,15000]	$< A, b_2 >$
[11,20]	M	[10001,15000]	$< B >$
[11,20]	M	[10001,15000]	$< C >$
[21,30]	F	[20001,25000]	$< B >$
[21,30]	F	[20001,25000]	$< A, b_2 >$
[21,30]	F	[20001,25000]	$< C >$

(d) T*

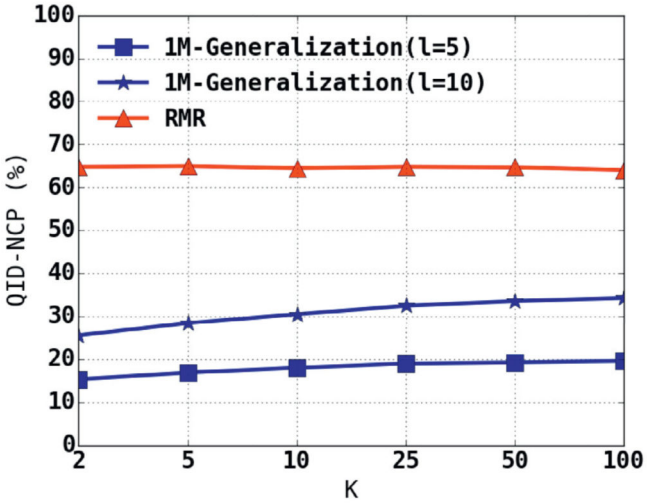
Fig. 3.
Anonymization 1:M dataset using 1:M-generalization.



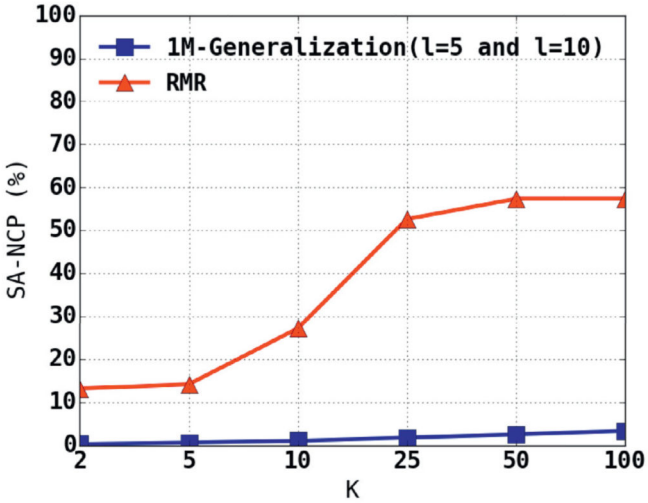
(a) QID-NCP on INFORMS



(b) SA-NCP on INFORMS

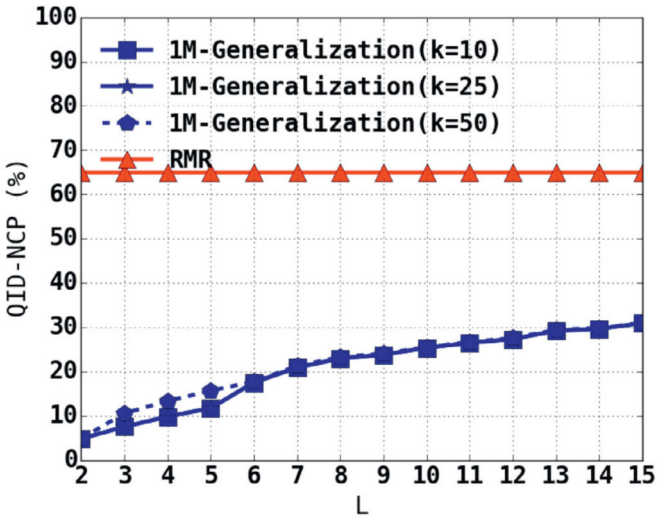


(c) QID-NCP on Youtube

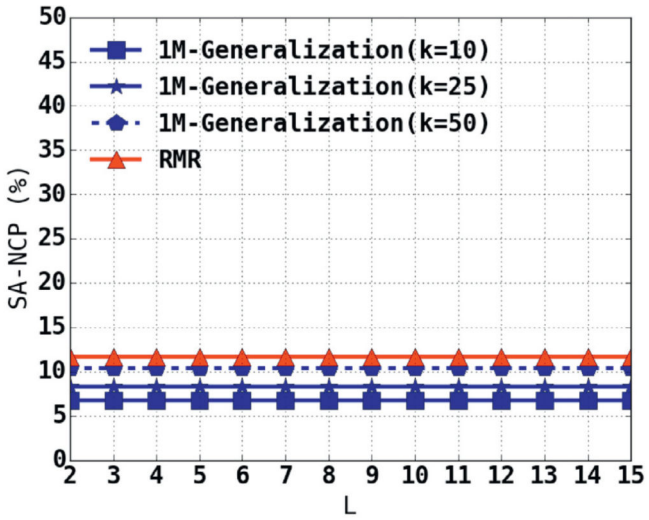


(d) SA-NCP on Youtube

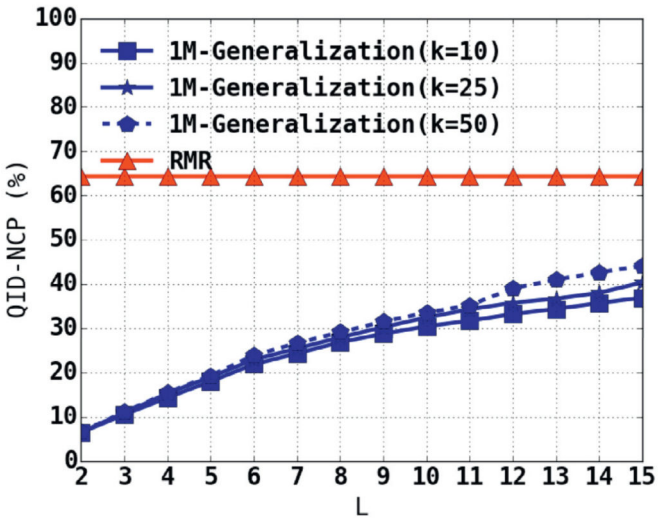
Fig. 4.
Information loss when varying *k*.



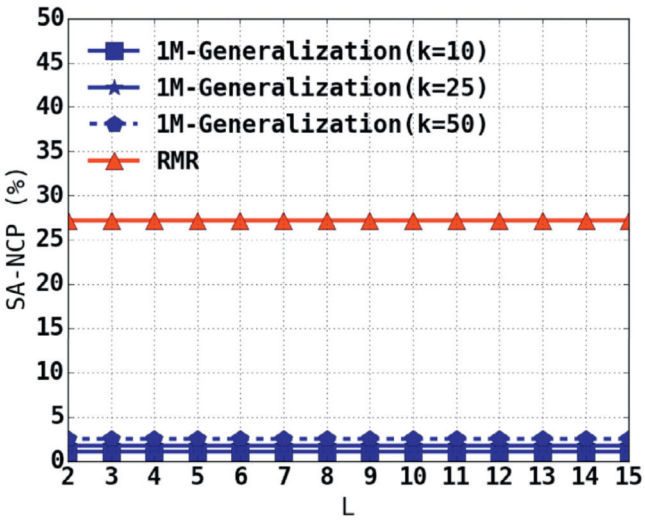
(a) QID-NCP on INFORMS



(b) SA-NCP on INFORMS

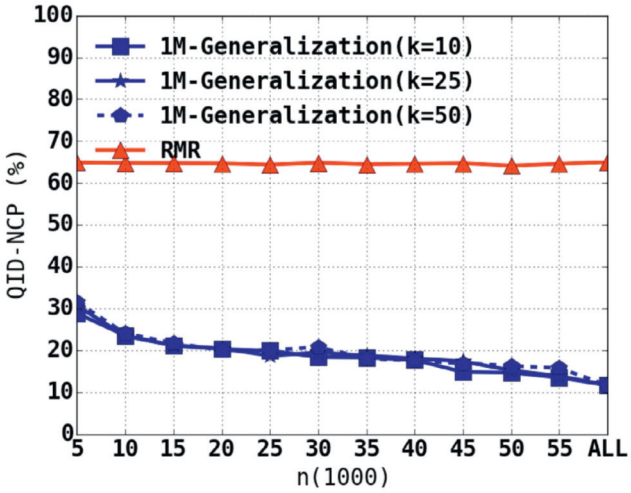


(c) QID-NCP on Youtube

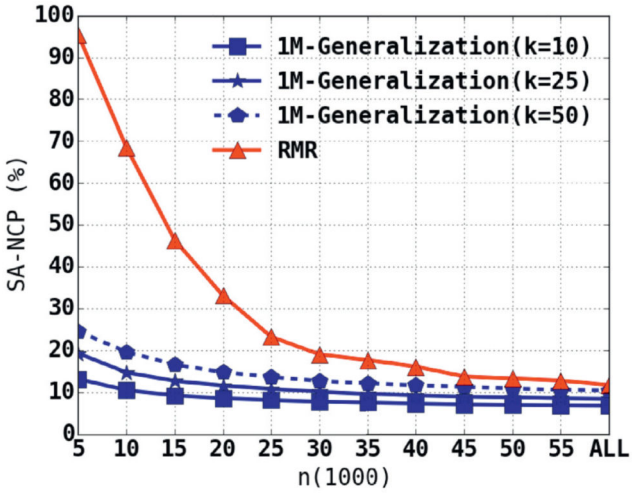


(d) SA-NCP on Youtube

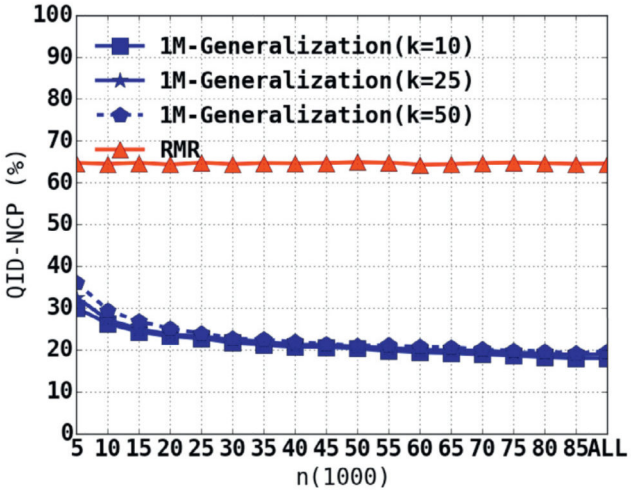
Fig. 5.
Information loss when varying *l*.



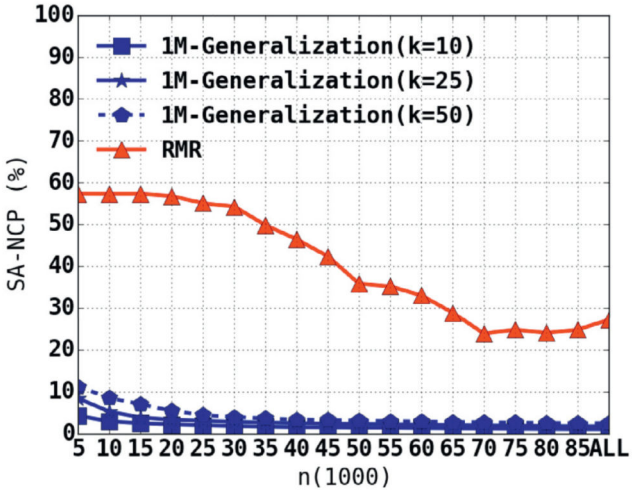
(a) QID-NCP on INFORMS



(b) SA-NCP on INFORMS

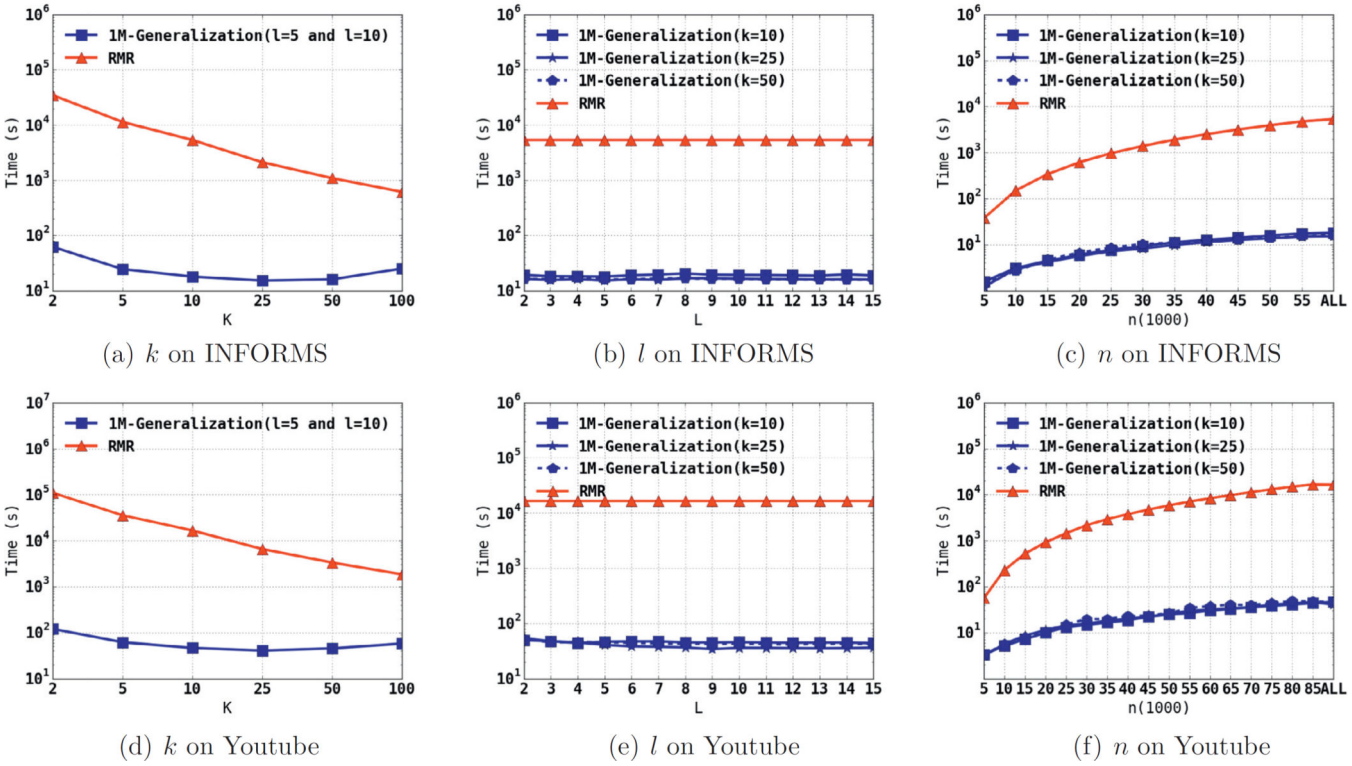


(c) QID-NCP on Youtube



(d) SA-NCP on Youtube

Fig. 6.
Information loss when varying *n*.



(a) k on INFORMS (b) l on INFORMS (c) n on INFORMS

(d) k on Youtube (e) l on Youtube (f) n on Youtube

Fig. 7.
Execution time.

Table 1

1:M Data publishing.

(a) Microdata for complication analysis						**(b) Published data using *k* -anonymity**				
tuple ID	**PID**	**Age**	**Gender**	**Zip code**	**Disease**	**PID**	**Age**	**Gender**	**Zip code**	**Disease**
1(Bob)	1	18	M	12,000	a_1	1	18	M	12,000	a_1
2(Bob)	1	18	M	12,000	a_2	1	18	M	12,000	a_2
3(Bob)	1	18	M	12,000	b_2	1	[11,20]	M	[10,001, 15,000]	b_2
4(David)	2	14	M	13,000	b_1	2	[11,20]	M	[10,001, 15,000]	b_1
5(Tom)	3	21	F	21,000	b_2	3	[11,30]	*	[10,001, 25,000]	b_2
6(Simon)	4	16	M	14,000	c_2	4	[11,30]	*	[10,001, 25,000]	c_2
7(Daisy)	5	27	F	22,000	a_2	5	27	F	22,000	a_2
8(Daisy)	5	27	F	22,000	b_2	5	27	F	22,000	b_2
9(Alice)	6	28	F	21,000	c_1	6	28	F	21,000	c_1
10(Alice)	6	28	F	21,000	c_2	6	28	F	21,000	c_2

Table 2

Transformed 1:M dataset and (*k*, *l*)-diversity.

(a) Transformed microdata					**(b) Published data by (*k*, *l*)-diversity**			
tuple ID	**Age**	**Gender**	**Zip code**	**Disease**	**Age**	**Gender**	**Zip code**	**Disease**
1–3(Bob)	18	M	12,000	< a_1, a_2, b_2 >	[11,20]	M	[10,001, 15,000]	< A, b_2 >
4(David)	14	M	13,000	< b_1 >	[11,20]	M	[10,001, 15,000]	< B >
5(Tom)	21	F	21,000	< b_2 >	[11,20]	M	[10,001, 15,000]	< C >
6(Simon)	16	M	14,000	< c_2 >	[21,30]	F	[20,001, 25,000]	< B >
7–8(Daisy)	27	F	22,000	< a_2, b_2 >	[21,30]	F	[20,001, 25,000]	< A, b_2 >
9–10(Alice)	28	F	21,000	< c_1, c_2 >	[21,30]	F	[20,001, 25,000]	< C >

Table 3

Description of the datasets.

Dataset	*n*	QID	SA	SA Domain
INFORMS	58,568	Month of birth	Diagnosis codes	632
		Year of birth		
		Race		
		Years of education		
		Income		
Youtube	85,607	Age	related_videos	117,752
		Category		
		Length		
		Rate		
		Ratings		
		Comments		