Data Min Knowl Disc (2011) 23:379–406
DOI 10.1007/s10618-010-0208-4

Publishing anonymous survey rating data

Xiaoxun Sun · Hua Wang · Jiuyong Li · Jian Pei

Received: 22 July 2009 / Accepted: 15 November 2010 / Published online: 26 November 2010

Abstract We study the challenges of protecting privacy of individuals in the large public survey rating data in this paper. Recent study shows that personal information in supposedly anonymous movie rating records are de-identified. The survey rating data usually contains both ratings of sensitive and non-sensitive issues. The ratings of sensitive issues involve personal privacy. Even though the survey participants do not reveal any of their ratings, their survey records are potentially identifiable by using information from other public sources. None of the existing anonymisation principles (e.g., k-anonymity, l-diversity, etc.) can effectively prevent such breaches in large survey rating data sets. We tackle the problem by defining a principle called (k, ϵ)-anonymity model to protect privacy. Intuitively, the principle requires that, for each transaction t in the given survey rating data T, at least $(k - 1)$ other transactions in T must have ratings similar to t, where the similarity is controlled by ϵ. The

Responsible editor: M.J. Zaki.

X. Sun (✉)
Australian Council for Educational Research, 19 Prospect Hill Road, Camberwell, VIC, Australia
e-mail: sun@acer.edu.au; xiaoxun.sun@gmail.com

H. Wang
Department of Mathematics Computing, University of Southern Queensland,
Toowoomba, QLD, Australia

J. Li
School of Computer and Information Science, University of South Australia,
Adelaide, SA, Australia
e-mail: jiuyong.li@unisa.edu.au

J. Pei
School of Computing Science, Simon Fraser University, Burnaby, BC, Canada
e-mail: jpei@cs.sfu.ca

(k, ϵ)-anonymity model is formulated by its graphical representation and a specific graph-anonymisation problem is studied by adopting graph modification with graph theory. Various cases are analyzed and methods are developed to make the updated graph meet (k, ϵ) requirements. The methods are applied to two real-life data sets to demonstrate their efficiency and practical utility.

Keywords $(k\epsilon)$-anonymity · Survey rating data · Graphical representation

1 Introduction

The problem of privacy-preserving data publishing has received a lot of attention in recent years (Sweeney 1997; Hansell 2006; Narayanan and Shmatikov 2008). Privacy preservation on relational data has been studied extensively. A major type of privacy attack on relational data includes re-identifying individuals by joining a published data set containing sensitive information with the external data sets modeling background knowledge of attackers (Li et al. 2009; Samarati and Sweeney 1998a; Machanavajjhala et al. 2006). Most of the existing work is formulated in contexts of several organizations, such as hospitals, publishing detailed data (also called microdata) about individuals (e.g. medical records) for research or statistical purposes.

Privacy risks of publishing microdata are well-known (LeFevre et al. 2006b; Wang and Fung 2006; Kifer and Gehrke 2006; Zhang et al. 2007). Famous attacks include de-anonymisation of the Massachusetts hospital discharge database by joining it with a public voter database (Sweeney 1997) and privacy breaches caused by AOL search data (Hansell 2006). Even if identifiers such as names and social security numbers have been removed, the adversary can use linking (Sweeney 2002), homogeneity and background attacks (Machanavajjhala et al. 2006) to re-identify individual data records or sensitive information of individuals. To overcome the re-identification attacks, the mechanism of k-anonymity was proposed (Samarati and Sweeney 1998a; Sweeney 2002). Specifically, a data set is said to be k-anonymous if, on the quasi-identifier (QID) attributes (the maximal set of join attributes to re-identify individual records), each record is identical with at least $(k - 1)$ other records. The larger the value of k, the better the privacy protection is. Although k-anonymity has been well adopted, Machanavajjhala et al. (2006) showed that a k-anonymous data set may still have some subtle but severe privacy problems due to the lack of diversity in sensitive attributes. Particularly, a large body of research contributes to transforming a data set to meet a privacy principle [k-anonymity (Sweeney 1997; Samarati 2001), l-diversity (Machanavajjhala et al. 2006), (α, k)-anonymity (Wong et al. 2006), t-closeness (Li and Li 2007)] using techniques such as generalization, suppression (removal), permutation and swapping of certain data values while minimizing certain cost metrics (Iyengar 2002; Wang et al. 2004; Meyerson and Williams 2004; Bayardo and Agrawal 2005; Fung et al. 2005; LeFevre et al. 2006a; Li and Li 2009).

Recently, a new privacy concern has emerged in privacy preservation research: how to protect the privacy of individuals in published large survey rating data. For example, movie rating data, supposedly to be anonymized, is de-identified by linking un-anonymized data from another source (Frankowski et al. 2006). On October 2, 2006, Netflix,

Table 1 (a) A published survey rating data set containing ratings of survey participants on both sensitive and non-sensitive issues (b) Public comments on some non-sensitive issues of some participants of the survey. By matching the ratings on non-sensitive issues with public available preferences, t_1 is linked to Alice, and her sensitive rating is revealed

ID	Non-sensitive			Sensitive
	Issue 1	Issue 2	Issue 3	Issue 4
t_1	6	1	*null*	6
t_2	1	6	*null*	1
t_3	2	5	*null*	1
t_4	1	*null*	5	1
t_5	2	*null*	6	5

Name	Non-sensitive issues		
	Issue 1	Issue 2	Issue 3
Alice	Excellent	So bad	–
Bob	Awful	Top	–
Jack	Bad	–	Good

the world's largest online DVD rental service, announced a $1-million Netflix Prize for improving their movie recommendation service (Hafner 2006). To aid contestants, Netflix publicly released a data set containing 100,480,507 movie ratings, created by 480,189 Netflix subscribers between December 1999 and December 2005. Narayanan and Shmatikov (2008) have shown that an attacker only needs a little bit information of an individual to identify the anonymized movie rating transaction of the individual in the data set. They re-identified Netflix movie ratings using the Internet Movie Database (IMDb)[1] as a source of auxiliary information and successfully identified the Netflix records of known users, uncovering their political preferences and other potentially sensitive information. In this paper, we will refer to two types of data as "survey rating data" and "relational data".

1.1 Motivation

The structure of large survey rating data is different from relational data, since it does not have fixed personal identifiable attributes. The lack of a clear set of personal identifiable attributes makes the anonymisation challenging (Zhou et al. 2008; Xu et al. 2008; Ghinita et al. 2008). In addition, survey rating data contains many attributes, each of which corresponds to the response to a survey question, but not all participants need to rate all issues (or answer all questions), which means a lot of cells in a data set are empty. For instance, Table 1a is a published survey rating data set containing ratings of survey participants on both sensitive and non-sensitive issues. The higher the rating is, the more preferred the participant is towards the issue. "*null*" means the

[1] http://www.imdb.com/.

participant did not rate the issue. Table 1b contains comments on non-sensitive issues of some survey participants, which might be obtained from public information sources such as personal weblogs or social network.

However, individuals in the anonymous survey rating data set are potentially identifiable based on their public comments from other sources (Narayanan and Shmatikov 2008). By matching the ratings of non-sensitive issues with publicly available preferences, an adversary can identify a small number of candidate groups that contain the record of the victim. It is unfortunate if there is only one record in the candidate group. For example, Alice is at risk of being identified in Table 1a, since t_1 is unique and could be linked to Alice's comments in Table 1b. This simple example motivates the first challenge:

How to preserve individual's privacy through identity protection in a large survey rating data set?

Though several models and algorithms have been proposed to preserve privacy in relational data, most of the existing studies can deal with relational data only (Sweeney 1997; Machanavajjhala et al. 2006; Li and Li 2007; Wong et al. 2006). Divide-and-conquer methods are applied to anonymize relational data sets due to the fact that tuples in a relational data set are separable during anonymisation. In other words, anonymizing a group of tuples does not affect other tuples in the data set. However, anonymizing a survey rating data set is much more difficult since changing one record may cause a domino effect on the neighborhoods of other records, as well as affecting the properties of the whole data set (details in Sect. 4.3). Hence, previous methods can not be applied to deal with survey rating data and it is much more challenging to devise anonymisation methods for large survey rating data than for relational data. Therefore, the second arising challenge is:

How to anonymize a large survey rating data while maintaining the least amount of distortion?

1.2 Contributions

Faced with these challenges, in this paper we study privacy preserving techniques for large survey rating data sets, and propose a new model and methods to preserve privacy in published large survey rating data sets.

This paper presents a systematic study towards the identity protection in large survey rating data sets. Firstly, we propose a privacy principle called (k, ϵ)-anonymity, which demands that for each transaction in a given survey rating data set, there are at least other $(k-1)$ similar transactions, where similarity is measured by ϵ. (k, ϵ)-anonymity guarantees that no individual is identifiable with confidence up to a function of ϵ with probability greater than $1/k$. Both k and ϵ define the degree of identity protection from different perspectives. The parameter ϵ specifies, for each transaction t, the length of ϵ-proximate neighborhood, whereas $1/k$ limits the probability that an adversary realizes t falling in that ϵ-proximate neighborhood.

Secondly, we formulate the (k, ϵ)-anonymity model using a graphical representation, design a metric to quantify graph modification operations and formally define the graph-anonymisation problem that, given a graphical representation G, asking

for the k-decomposable graph stemmed from G with the minimum number of graph modification operations. Given a survey rating data set T and ϵ, we prove that if the graphical representation G of T is k-decomposable, then T is (k, ϵ)-anonymous. This interpretation of anonymity prevents the re-identification of individuals by adversaries with a priori knowledge of the degree of certain nodes. Then, we make a thorough analysis of the modification strategies and prove the correctness and completeness of the proposed modification strategies. Finally, we apply the approaches to real-world rating data sets and demonstrate that the utility of the anonymous rating as well as the statistical properties are well preserved, and our methods are efficient.

The rest of the paper is organized as follows. Section 2 surveys related work. Section 3 discusses fundamental concepts and proposes the novel (k, ϵ)-anonymity principle for identity protection in a large survey rating data set. Section 4 introduces the graphical representation of (k, ϵ)-anonymity models and develops the anonymisation method for large survey rating data sets. Section 5 includes the results of experimental evaluations on two real-life data sets. Finally, Sect. 6 concludes the paper with directions for future work.

2 Related work

Privacy preserving data publishing has received considerable attention in recent years, especially in the context of relational data (Aggarwal 2005; Samarati and Sweeney 1998b; Samarati 2001; Machanavajjhala et al. 2006; Li and Li 2007). All these works assume a given set of attributes QID on which an individual is identified, and anonymize data records on the QID. Aggarwal (2005) presents a study on the relationship between the dimensionality of QID and information loss, and concludes that, as the dimensionality of QID increases, information loss increases quickly. Large survey rating data sets present a worst case scenario for existing anonymisation approaches because of the high dimensionality of QID and sparseness of the data sets. To our best knowledge, all existing solutions in the context of k-anonymity (Samarati and Sweeney 1998b; Samarati 2001), l-diversity (Machanavajjhala et al. 2006) and t-closeness (Li and Li 2007) assume a relational table, which typically has a low dimensional QID. Survey rating data sets, on the other hand, are characterized by sparseness and high dimensionality, which makes the current state-of-art principles incapable handling the anonymisation of large survey rating data sets.

There are few previous works considering the privacy of large rating data. In collaboration with MovieLens recommendation service, Frankowski et al. correlated public mentions of movies in the MovieLens discussion forum with the users' movie rating histories in the internal Netflix data set (Frankowski et al. 2006). Recent study reveals a new type of attack on anonymized MovieLens data (Narayanan and Shmatikov 2008). The supposedly anonymized movie rating data is re-identified by linking non-anonymized data from other sources. To our best knowledge, no anonymisation models and methods exist for preserving privacy for large survey rating data sets.

Privacy-preservation of transactional data has been acknowledged as an important problem in the data mining literature (Atzori et al. 2005a,b; Verykios et al. 2004; Ghinita et al. 2008; Xu et al. 2008; He and Naughton 2009). The privacy threats

caused by publishing data mining results such as frequent item sets and association rules is addressed in Atzori et al. (2005a,b). The work in Atzori et al. (2008), Verykios et al. (2004) focus on publishing anonymous patterns, where the patterns are mined from the original data, and the resulting set of rules is sanitized to present privacy breaches. In contrast, our work addresses the privacy threats caused by publishing a large survey rating data. Recent work (Ghinita et al. 2008; Xu et al. 2008; He and Naughton 2009) targets anonymisation of transaction data. Our work aims to prevent individual identity disclosure in a large survey rating data set.

Graph approaches have been applied in solving anonymization problems (Zhou et al. 2008; Liu and Terzi 2008). Liu and Terzi (2008) study a specific graph-anonymization problem. A graph is called k-degree anonymous if for every node v, there exists at least $k-1$ other nodes in the graph with the same degree as v. This definition of anonymity prevents the re-identification of individuals by adversaries with a priori knowledge of the degree of certain nodes. The anonymization problem we consider in this paper is partially related to it but different. We not only study how to modify the graph to make it k-decomposable, but also analyze how to anonymize the underlying data set, which is beyond the study of Liu and Terzi (2008). Pei and Zhou in Zhou et al. (2008) consider yet another definition of graph anonymity—a graph is k-anonymous if for every node there exist at least $k-1$ other nodes that share isomorphic neighborhoods; in this case the neighborhood of a node is defied by its immediate neighbors and the connections between them. This definition of anonymity in graphs is different from ours. In a sense it is a more strict one. Given the difference in the definition, the corresponding algorithmic problems arising in Zhou et al. (2008) are also different from the problems we consider in this paper.

3 (k, ϵ)-anonymity

In this section, we formally define the (k, ϵ)-anonymity model for protecting privacy in a large survey rating data set.

We assume that survey rating data publishes people's ratings on a range of issues. Some issues are sensitive, such as income level and sexuality frequency, while some are non-sensitive, such as the opinion of a book, a movie or a kind of food. Each survey participant is cautious about his/her privacy and does not reveal his/her ratings. However, an attacker can use auxiliary information to identify an individual's sensitive ratings in supposedly anonymous survey rating data. The auxiliary information of an attacker includes: (i) knowledge that a victim is in the survey rating data and; (ii) preferences of the victims on some non-sensitive issues. For instance, an attacker may find a victim's preference (not exact rating scores) by personal familiarity or by reading the victim's comments on some issues from personal weblogs or social networks. We assume that attackers know preferences of non-sensitive issues of a victim but do not know exact ratings and want to find out the victim's ratings on some sensitive issues. Our objective is to design an effective model to protect privacy of people's sensitive ratings in published survey rating data.

Given a survey rating data set T, each transaction contains a set of numbers indicating the ratings on some issues. Let $(o_1, o_2, \ldots, o_p, s_1, s_2, \ldots, s_q)$ be a transaction,

$o_i \in \{1:r, null\},\ i = 1, 2, \ldots, p$ and $s_j \in \{1:r, null\},\ j = 1, 2, \ldots, q$, where r is the maximum rating and *null* indicates that a survey participant did not rate. $o_1, \ldots, o_p$ stand for non-sensitive ratings and $s_1, \ldots, s_q$ denote sensitive ratings. Each transaction belongs to a survey participant. Let $T_A = \{o_{A_1}, o_{A_2}, \ldots, o_{A_p}, s_{A_1}, s_{A_2}, \ldots, s_{A_q}\}$ be the ratings for a survey participant A and $T_B = \{o_{B_1}, o_{B_2}, \ldots, o_{B_p}, s_{B_1}, s_{B_2}, \ldots, s_{B_q}\}$ be the ratings for a participant B. We define the dissimilarity between two non-sensitive rating scores as follows.

$$Dis(o_{A_i}, o_{B_i}) = \begin{cases} |o_{A_i} - o_{B_i}| & \text{if } o_{A_i}, o_{B_i} \in \{1:r\} \\ 0 & \text{if} o_{A_i} = o_{B_i} = \text{null} \\ r & \text{otherwise} \end{cases} \quad (1)$$

Definition 1 (ϵ-*proximate*) Given a small positive number ϵ, if for $1 \le i \le p,\ Dis(o_{A_i}, o_{B_i}) \le \epsilon$, then transactions T_A and T_B are ϵ-proximate.

If two transactions are ϵ-proximate, the dissimilarity between their non-sensitive ratings is bound by ϵ. In Table 1a, if $\epsilon = 1$, ratings 5 and 6 may have no difference in interpretation, so t_4 and t_5 are 1-proximate based on their non-sensitive rating.

Definition 2 ((k, ϵ)-*anonymity*) A survey rating data set is (k, ϵ)-anonymous if every transaction in the survey rating data set has at least $(k-1)\epsilon$-proximate neighbors.

The idea behind (k, ϵ)-anonymity is to make each transaction in a survey rating data set similar with at least other $(k-1)$ transactions in order to avoid linking to individual's sensitive ratings. (k, ϵ)-anonymity can well protect identity privacy, since it guarantees that no individual is identifiable with confidence up to a function of ϵ with probability greater than $1/k$. Both parameters k and ϵ are intuitive and operable in real-world applications. By varying the values of k or ϵ, we strengthen the protection from different perspectives. Specifically, the parameter ϵ captures the protection proximate neighborhood of each survey participant in that raising ϵ enlarges the protection range of each sensitive value. The purpose of elevating k is to lower an adversary's chance of beating that protection.

Given a survey rating data set T and the values of k, ϵ, the objective of (k, ϵ)-anonymisation is to modify T to make it satisfy the k, ϵ requirements. Generally speaking, if T has already met this privacy requirement, we can publish it without any modifications; otherwise, we need develop modification techniques for satisfying the (k, ϵ) requirements. Next, we discuss this problem.

4 Anonymize survey rating data

In this section, we describe our modification strategies through the graphical representation of the (k, ϵ)-anonymity model. Firstly, we introduce some preliminaries and quantify the distortion caused by anonymization. Secondly, we present the (k, ϵ)-anonymity model with graphs. Finally, we describe the modification strategies in detail.

Table 2 Sample survey rating data (I)

ID	Non-sensitive			Sensitive
	Issue 1	Issue 2	Issue 3	Issue 4
t_1	3	6	*null*	6
t_2	2	5	*null*	1
t_3	4	7	*null*	4
t_4	5	6	*null*	1
t_5	1	*null*	5	1
t_6	2	*null*	6	5

4.1 Preliminaries

Given a survey rating data set T, we define a binary flag matrix $F(T)$ to record if there is a rating or not for each non-sensitive issue (column). $F(T)_{ij} = 1$ if the ith participant rates the jth issue and $F(T)_{ij} = 0$ otherwise. For instance, the flag matrix associated with the rating data of Table 2 is

$$\mathbf{F} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix} \tag{2}$$

in which each row corresponds to survey participants and each column corresponds to non-sensitive issues. In order to measure the distance between two vectors in the flag matrix, we borrow the concept of Hamming distance (Hamming 1980).

Definition 3 (*Hamming Distance*) Hamming distance between two vectors in the flag matrix of equal length is the number of positions for which the corresponding symbols are different. We denote the Hamming distance between two vectors v_1 and v_2 as $H(v_1, v_2)$.

In other words, Hamming distance measures the minimum number of substitutions required to change one vector into the other, or the number of errors that transformed one vector into the other. For example, if $v_1 = (1, 1, 0)$ and $v_2 = (1, 0, 1)$, then $H(v_1, v_2) = 2$. If the Hamming distance between two vectors is zero, then these two vectors are identical. In order to categorize identical vectors in the flag matrix, we introduce the concept of Hamming group.

Definition 4 (*Hamming Group*) Hamming group is the set of vectors in which the Hamming distance between any two vectors of the flag matrix is zero. The maximal Hamming group is a Hamming group that is not a subset of any other Hamming group.

For example, there are two maximal Hamming groups in the flag matrix (2) made up of vectors {(1, 1, 0), (1, 1, 0), (1, 1, 0), (1, 1, 0)} and {(1, 0, 1), (1, 0, 1)} and they correspond to groups $\{t_1, t_2, t_3, t_4\}$ and $\{t_5, t_6\}$ of T.

4.2 Distortion metrics

In this section, we define a measure of information loss.

Definition 5 (*Tuple distortion by edge addition*) Let $t = (t_1, t_2, \ldots, t_m)$ be a tuple and $t' = (t'_1, t'_2, \ldots, t'_m)$ be an anonymized tuple of t. Then, the distortion of this anonymisation is defined as:

$$\text{Distortion\_additon}(t, t') = \sum_{i=1}^{m} |t_i - t'_i|$$

For example, if the tuple $t = (5, 6, 0)$ is generalized to $t' = (5, 5, 0)$, then the distortion of this anonymisation is $|5 - 5| + |6 - 5| + |0 - 0| = 1$.

Definition 6 (*Data set total distortion*) Let $T' = (t'_1, t'_2, \ldots, t'_n)$ be the anonymized data set from $T = (t_1, t_2, \ldots, t_n)$. Then, the total distortion of this anonymisation is defined as:

$$\text{Distortion}(T, T') = \sum_{i=1}^{n} \text{Distortion\_addition}(t_i, t'_i)$$

For example, let $T = (t_1, t_2, t_3, t_4)$, where $t_1 = (5, 6, 0)$, $t_2 = (2, 5, 0)$, $t_3 = (4, 7, 0)$ and $t_4 = (5, 6, 0)$. Let the anonymized view be $T' = (t'_1, t'_2, t'_3, t'_4)$, where $t'_1 = (5, 5, 0)$, $t'_2 = (3, 5, 0)$, $t'_3 = (3, 7, 0)$ and $t'_4 = (5, 7, 0)$. Then, the distortion between the two data sets is $1 + 1 + 1 + 1 = 4$.

4.3 Graphical representation

Given a survey rating data set $T = \{t_1, t_2, \ldots, t_n\}$, its graphical representation is the graph $G = (V, E)$, where V is a set of nodes, and each node in V corresponds to a record t_i $(i = 1, 2, \ldots, n)$ in T, and E is the set of edges, where two nodes are connected by an edge if and only if the distance between two records is bounded by ϵ with respect to the non-sensitive ratings (Eq. (1)).

Two nodes t_i and t_j are called connected if G contains a path from t_i to t_j $(1 \leq i, j \leq n)$. The graph G is called connected if every pair of distinct nodes in the graph can be connected through some paths. A connected component is a maximal connected subgraph of G. Each node belongs to exactly one connected component, as does each edge. The degree of the node t_i is the number of edges incident to t_i $(1 \leq i \leq n)$.

Theorem 1 *Given the survey rating data set T with its graphical representation G, T is (k, ϵ)-anonymous if and only if the degree of each node of G is at least $(k - 1)$.*

Proof "⇐": Without loss of generality, we assume that G is a connected graph. If for every node v in G, the degree of v is greater than $(k-1)$, which means there are at least $(k-1)$ other nodes connecting with v, then according to the construction of the graph, two nodes have an edge connection if and only if their distance is bounded by ϵ. Therefore, T satisfies (k, ϵ)-anonymity property.

"⇒": If T is (k, ϵ)-anonymous, then according to the definition of (k, ϵ)-anonymity, each record in T is ϵ-proximate with at least $(k-1)$ other records, and then in the graphical representation G of T, the degree of each node should be at least $(k-1)$. □

With the equivalent condition proven in Theorem 1, we see that in order to make $T(k, \epsilon)$-anonymous, we need to modify its graphical representation G to ensure that each node in G has degree of at least $(k-1)$. Next, we introduce the general graph anonymization problem. The input to the problem is a simple graph $G = (V, E)$ and an integer k. The requirement is to use a set of graph-modification operations on G in order to construct a graph $G' = (V', E')$ with the degree of each node in G' is at least $k-1$. The graph modification operation considered in this paper is edge addition (adding edges is by modifying values of transactions represented as nodes). We require that the output graph be over the same set of nodes as the original graph, that is, $V' = V$. Given T and ϵ, we denote the graphical representation of T as G. In order to meet k and ϵ requirements, we modify G to G', and the underlying data set T is changed into T'. We capture the distortion between T and T' ($Distortion(T, T')$) as the distortion of anonymizing G to G' denoted by $D(G)$, i.e., $D(G) = Distortion(T, T')$.

Problem 1 *Given a graph $G = (V, E)$ and an integer k, find a graph $G' = (V, E')$ with $E' \bigcap E = E$ by modifying values of some tuples so that the degree of each node of the corresponding graph is at least $(k-1)$ and the distortion $D(G)$ is minimized.*

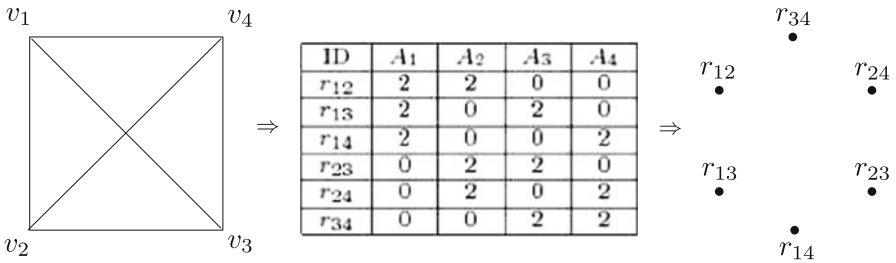
Theorem 2 *Problem 1 is NP-hard.*

Proof The NP-hardness proof of the Problem 1 is transformed from the problem of Edge Partition into 4-Cliques (Garey and Johnson 1979).

Edge Partition Into 4-Cliques: Given a simple graph $G = (V, E)$, with $|E| = 6m$ for some integer m, can the edges of G be partitioned into m edge-disjoint 4-cliques?

Given an instance of Edge Partition into 4-Cliques. We first construct a rating data set T as follows. For each vertex $v_i \in V$, construct an issue A_i. For each edge $e \in E$, where $e = (v_1, v_2)$, create a pair of records $r_{v1,v2}$, where the record has the ratings of both issues A_1 and A_2 equal to 2 and all other issues equal to 0. We then construct the graphical representation G' of T by setting $k = 6$, $\epsilon = 1$. The objective here is to add the edges to make the degree of each node in G' at least $(k-1)$, and we apply the cost metrics defined in Sect. 4.2. We show that the cost of making the degree of each node in G' at least $(k-1)$ is at most $12m$ if and only if E can be partitioned into a collection of m edge-disjoint 4-cliques.

"⇐" Suppose E can be partitioned into a collection of m disjoint 4-cliques. Consider one 4-clique C with vertices v_1, v_2, v_3 and v_4 Fig. 1a. Then, the rating data set T constructed from C is shown in Fig. 1b and the graphical representation G' of T is Fig. 1c.

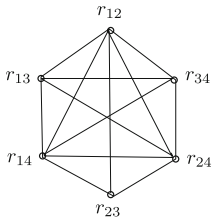


ID	A_1	A_2	A_3	A_4
r_{12}	2	2	0	0
r_{13}	2	0	2	0
r_{14}	2	0	0	2
r_{23}	0	2	2	0
r_{24}	0	2	0	2
r_{34}	0	0	2	2

(a) **(b)** **(c)**

Fig. 1 **a** one 4-clique C; **b** a rating data set T constructed from C; **c** graphical representation G' of T with $\epsilon = 1$

ID	A_1	A_2	A_3	A_4
r_{12}	1	1	0	0
r_{13}	1	0	1	0
r_{14}	1	0	0	1
r_{23}	0	1	1	0
r_{24}	0	1	0	1
r_{34}	0	0	1	1



ID	A_1	A_2	A_3	A_4
r_{12}	2	2	1	1
r_{13}	2	1	2	1
r_{14}	2	1	1	2
r_{23}	1	2	2	1
r_{24}	1	2	1	2
r_{34}	1	1	2	2

(a) **(c)** **(b)**

Fig. 2 Two possible modifications of the rating data set T with $k = 6, \epsilon = 1$

Since there are three 2's and three 0's for each issue in T, with the privacy requirement $k = 6$ and $\epsilon = 1$, the distance of any pair of nodes is bounded by 2, which is greater than the given ϵ. To satisfy the requirements, we can either change all the 2's or 0's in T to 1's, which has the cost of $3 \times 4 \times m = 12m$ (shown in Fig. 2).

"$\Rightarrow$" Suppose the cost of making the degree of each node in G' is at most $12m$. As G is a simple graph, any record only has two ratings of 2 and any six records should have at least four issues whose distances are greater than the given ϵ. The modification can be made by either changing 2 or 0 to 1. So, each record should have at least two 1's in T when its graphical representation G' satisfies the condition that each node in G' has the degree of at least 5. Then, the cost of making the degree of each node in G' is at least $6 \times 2 \times m = 12m$. Combining with the proposition that the cost is at most $12m$, we obtain the cost is exactly equal to $12m$ and thus each record should have exactly two 1's in the solution. Each group should have exactly 6 records. Suppose the six modified records contain 2 1's in issues A_1, A_2, A_3 and A_4. This corresponds to a 4-clique with vertices v_1, v_2, v_3 and v_4. Thus, we conclude that the solution corresponds to a partition into a collection of m edge-disjoint 4-cliques. □

Even though we can present the equivalent connection between the problem of anonymizing survey rating data and Problem 1, it is not easy to solve Problem 1. The difficulties are occurred in two main aspects. The first difficulty comes from the NP-hardness results of Problem 1, which makes no polynomial time algorithms for solving the problem ($P \neq NP$) and the only practical methods are heuristic. The second, but not the least difficulty is the domino effects. If the degree of a node is less

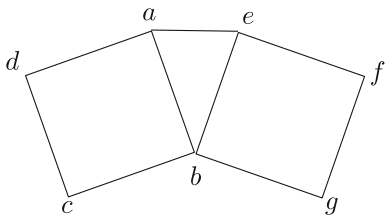


Fig. 3 An example of domino effects

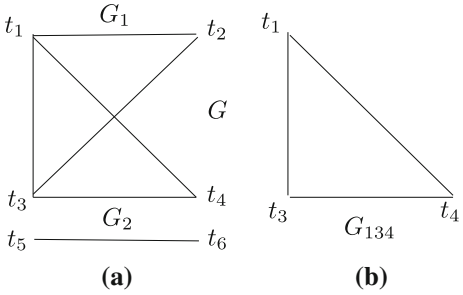


Fig. 4 Graphical representation example

than $(k-1)$, we need to add some edges to make its degree $(k-1)$. However, this simple operation could cause the domino effects to other nodes. The domino effect is a chain reaction that occurs when a small change causes a similar change nearby, which then will cause another similar change, and so on. In the graphical representation of the survey rating data set, if we add an edge to two nodes that are originally not connected, then the distance between these two nodes should be bounded by ϵ. Since the distance between these two nodes are changed, it is mostly likely that the distance between these two nodes and other nodes are affected as well. If this happens, it is hard to regulate the modification either on the graphical representation or on the survey rating data set. Take Fig. 3 as an example. Since node b is connected with nodes a, c, e, g, if we are going to change the degree of b, all the nodes are subject to this change, and the whole structure of the graph would be different. To avoid this domino effect, we further reduce the anonymization problem to ensure that the change of one node's degree has no effects on other nodes. In this paper, we adopt the concept of k-clique for the reduction.

We say G is a clique if every pair of distinct nodes is connected by an edge. The k-clique is a clique with at least k nodes. The maximal k-clique is the a k-clique that is not a subset of any other k-clique. We say the connected component $G = (V, E)$ is k-decomposable if G can be decomposed into several k-cliques $G_i = (V_i, E_i)$ $(i = 1, 2, \ldots, m)$, and satisfies $V_i \bigcap V_j = \emptyset$ for $(i \neq j)$, $\bigcup_{i=1}^{m} V_i = V$, and $\bigcup_{i=1}^{m} E_i \subseteq E$. The graph is k-decomposable if all its connected components are k-decomposable. The decomposability of the graph has the following monotonicity property.

Proposition 1 *If a graph $G = (V, E)$ is k_1-decomposable, then it is also k_2-decomposable, for every $k_2 \leq k_1$.*

For instance, the graphical representation of the survey rating data in Table 2 with $\epsilon = 2$ is shown in Fig. 4a. In Fig. 4a, there are two connected components, G_1 and

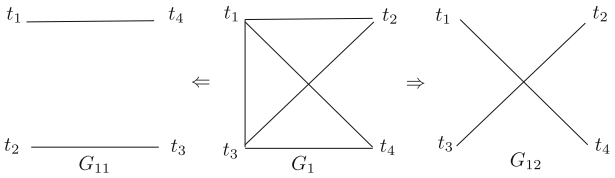
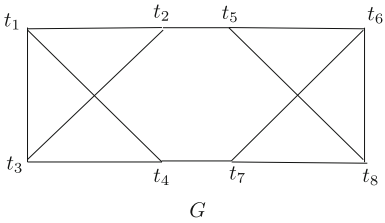


Fig. 5 Two possible 2-decompose of G_1

Fig. 6 A counter example



G_2, where G_2 is the 2-clique. G_{134} is a maximal 3-clique in G_1 (shown in Fig. 4b). G is 2-decomposable, since both G_1 and G_2 are 2-decomposable. Two possible 2-decompositions of G_1, G_{11} and G_{12} are shown in Fig. 5.

Note that if G is k-decomposable, then the degree of each node is at least $(k-1)$. However, on the other hand, if the degree of every node in G is at least $(k-1)$, G is not necessarily k-decomposable. A counterexample is shown in Fig. 6. For each node of G, the degree is at least 3, but G is not 4-decomposable. Although k-decomposability of G is a stronger condition than requiring the degree of the nodes in G to be at least $(k-1)$, it can avoid the domino effect through edge addition operations. From Theorem 1, we have the following corollary.

Corollary 1 *Given the survey rating data set T with its graphical representation G, if G is k-decomposable, then T is (k, ϵ)-anonymous.*

For instance, the survey rating data shown in Table 2 is (2,2)-anonymous since its graphical representation (Fig. 4a) is 2-decomposable.

Problem 2 *Given a graph $G = (V, E)$ and an integer k, modify values of some tuples to make the corresponding graph $G' = (V, E')$ k-decomposable with $E' \bigcap E = E$ such that the distortion $D(G)$ is minimized.*

Note that Problem 2 always has feasible solutions. In the worst case, all edges not present in each connected component of the input graph can be added. In this way, the graph becomes the union of cliques and all nodes in each connected component have the same degree; thus, any privacy requirement is satisfied (due to Proposition 1). Because of Corollary 1, Problem 1 always has a feasible solution as well.

If a given survey rating data set T satisfies the anonymity requirement, we can publish the data directly. On the other hand, if T is not (k, ϵ)-anonymous, we need to do some modifications in order to make it anonymous. Due to the hardness of computing Problem 1, in this paper, we investigate the solutions of Problem 2. We provide the heuristic methods to compute (k, ϵ)-anonymous solution, which starts

from each connected component. More specifically, we consider three scenarios that may happen during the computation. Firstly, if each connected component is already k-decomposable, then we do nothing since it has satisfied the privacy requirements. Secondly, if some connected components are k-decomposable while others are not. We reinvestigate the their Hamming groups to see whether two different connected components belonging to the same Hamming group can be merged together. Third, if none of the above situations happen, we consider to borrow nodes from connected components that belonging to different Hamming groups. In Sect. 4.4, we discuss the possible graphical modification operations, and in Sect. 4.5, we apply the graphical modifications to the survey rating data sets by the metrics defined in Sect. 4.2.

4.4 Graphical modification

Given the survey rating data set T with its graphical representation G, the number of connected components in G can be determined by the flag matrix of T. If two transactions are in different Hamming groups in the flag matrix, there must be no edge between these two nodes in G. For instance, the flag matrix of Table 2 is shown in Eq. (2), obviously there are two connected components in G (shown in Fig. 4). However, the converse is not true, since it may happen that two transactions are in the same Hamming group in the flag matrix, but their distance is greater than the given ϵ. For instance, although there are still two groups in the flag matrix of Table 3, there would be three connected components in its graphical representation (see Fig. 7a).

The number of Hamming groups decided by the flag matrix is not sufficient to determine the number of connected components of G, but it is enough to determine the minimum number of connected graphs of G. The graph anonymisation process starts from the connected component of the graphical representation. We test the (k, ϵ) requirements for each connected component of G, and have the following three cases:

Case 1 (*Trivial case*) If all the connected components of G are k-decomposable, then we publish the survey rating data without any changes.

Table 3 Sample survey rating data (II)

ID	Non-sensitive			Sensitive
	Issue 1	Issue 2	Issue 3	Issue 4
t_1	3	6	*null*	6
t_2	2	5	*null*	1
t_3	4	7	*null*	4
t_4	5	6	*null*	1
t_5	1	*null*	5	1
t_6	2	*null*	6	5
t_7	6	*null*	6	3
t_8	5	*null*	5	2

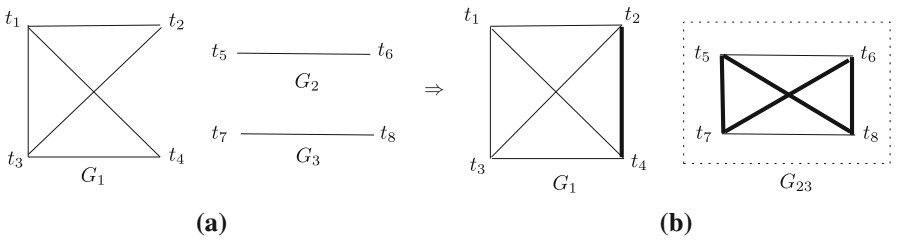


Fig. 7 Merging and modification process for subcase 2.1

Case 2 (*Merging case*) There exists at least one connected component containing at least two nodes that is not k-decomposable. If some of the connected components do not satisfy the requirement, it may happen that some of them belong to the same Hamming group in the flag matrix. For example, with $k = 3$ and $\epsilon = 2$, the two connected components G_2 and G_3 do not satisfy this requirement, but they belong to the same Hamming group in the flag matrix of Table 3 whose graphical representation is shown in Fig. 7a. In this situation, we merge them first, and then do modifications in order to make them meet the requirement. Figure 7b illustrates how the merging process and modification works.

At the initial stage, there are three connected components G_1, G_2 and G_3. If the privacy requirement is $k = 3$ and $\epsilon = 2$, we verify this requirement for each component, and it turns out that none of the components satisfy the requirement. We further know that records t_5, t_6, t_7, t_8 are in the same Hamming group of the flag matrix of Table 3, so we merge them into one connected components G_{23} by adding four edges among them. To make G_1 meet the requirement, it is enough to add one edge between t_2 and t_4. The added edges are shown in bold Fig. 7b. After the merging and modification process, Fig. 7b is 4-decomposable, and according to Corollary 1, the survey rating data set shown in Table 3 satisfies the privacy requirement. Now, we could make the graph k-decomposable by edge addition operations.

Case 3 (*Borrowing case*) There exists at least one connected component that is not k-decomposable and in the case that we could not make Gk-decomposable through merging and modification process, we need to borrow some nodes from other connected components without affecting other connected components. In order to produce no effect to other groups, we find the maximal k-clique.

Take Table 2 (graphical representation in Fig. 4a) as an example with $k = 3, \epsilon = 2$. We need to borrow at least one point from G_1 for G_2 in order to satisfy the given k. In order not to affect the structure of G_1, we find the maximal 3-clique $G_{1,3,4}$ of G_1, and the left point t_2 is the one we borrow from G_1. Then, we add edges between t_2, t_5 and t_2, t_6 to make it 3-decomposable. The process is shown in Fig. 8.

Case 3.1 If the k-clique is unique in the connected graph, then we borrow the point from the left ones. However, there might not be a unique k-clique. For example, either t_1, t_2, t_3 or t_1, t_3, t_4 form a 3-clique of G_1. In either case, the left point is t_4 or t_2. In order to determine which one we should choose, we need to define the objective of our problem and measure the information loss. We discuss appropriate metrics in the next section. Generally speaking, our objective is to find a solution with minimum distortion.

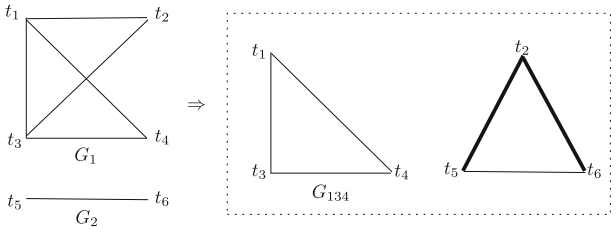


Fig. 8 Borrowing nodes from other connected graph

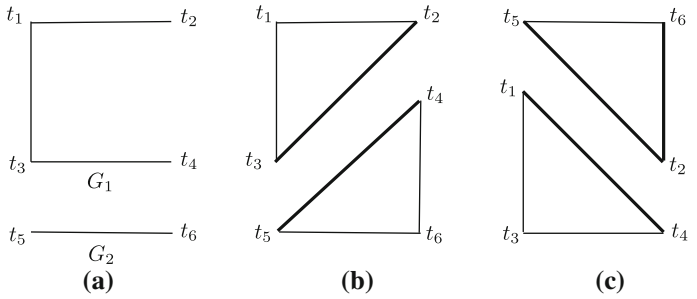


Fig. 9 Combining two 2-cliques

Case 3.2 It might happen that there is no k-clique in some connected components. For example, the graphical representation of some sample data is shown in Fig. 9 with the privacy requirement $k = 3, \epsilon = 2$. In Fig. 9a, there are two connected components G_1 and G_2. With the requirement of $k = 3$, there is no 3-clique in G_1. Instead, we find a 2-clique. Generally, if there is no k-clique, we find a $(k-1)$-clique, and since 2-clique always exists, this recursive process will end.

If we find the 2-cliques, the next question is how to combine them into a 3-clique. In the example above, there are three possible 2-cliques consisting of $\{t_1, t_2\}$, $\{t_1, t_3\}$ and $\{t_3, t_4\}$. If we choose $\{t_1, t_2\}$ and $\{t_1, t_3\}$ to merge together, there will be information loss in adding the edge between t_2 and t_3 (Fig. 9b). If we choose $\{t_1, t_3\}$ and $\{t_3, t_4\}$ to merge together, there will be information loss in adding the edge between t_1 and t_4 (Fig. 9c). The decision of choosing which kind of operation is depended on the distortion incurred by the edge addition operation. Distortion metrics are introduced in the next section.

4.5 Data modification

In the previous section, we discussed how to modify the graph to make it k-decomposable. In this section, we reflect such changes in the corresponding survey rating data set.

Recall we have the survey rating data set $T = (t_1, t_2, \ldots, t_n)$, $t_i = (x_{i1}, x_{i2}, \ldots, x_{im})$, where x_{ij} is the rating of survey participant i on issue j $(1 \leq i \leq n,\ 1 \leq j \leq m)$. $x_j = (x_{1j}, x_{2j}, \ldots, x_{nj})$ denotes the vector of ratings on issue j by all the survey participants $(1 \leq j \leq m)$. Given the privacy requirement ϵ, k, we construct the

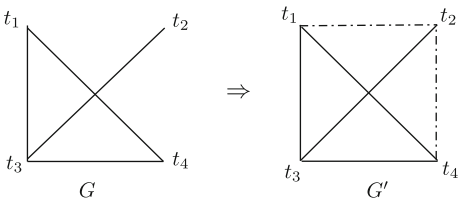


Fig. 10 The modification of graphical representation G for Case 2.1.1

graphical representation G of the data set T, and publish $T' = (t'_1, t'_2, \ldots, t'_n)$, $t'_i = (x'_{i1}, x'_{i2}, \ldots, x'_{im})$ $(1 \le i \le n$ and $1 \le j \le m)$.

Case 1 If G is already a k-clique with given ϵ, then output T', the same as T.

Case 2 (*Edge addition*) If G is not yet a k-clique, add necessary edges to make G a k-clique. We publish T' as follows: Firstly, we compute the centroid $t_c = (t_{c1}, t_{c2}, \ldots, t_{cm})$, where $t_{ci} = \frac{x_{1i}+x_{2i}+\ldots+x_{ni}}{n}$, $1 \le i \le n$. There are several cases that may happen to t_c:

Case 2.1 (*Integer strategy*) If t_{ci} is an integer, $\forall i = 1, 2, \ldots, m$, we sort the ratings of the jth issue of T ascending order. Without loss of generality, we assume the ratings on the jth issue of T, $x_j = (x_{1j}, x_{2j}, \ldots, x_{nj})$ is sorted ascended $(1 \le j \le m)$.

Case 2.1.1 If $\epsilon \ge 1$ and n is even, the first $\frac{n}{2}$ ratings $x_{1j}, x_{2j}, \ldots, x_{\frac{n}{2}j}$ are modified to $t_{cj} - 1, t_{cj} - 1, \ldots, t_{cj} - 1$, and the remaining $\frac{n}{2}$ ratings $x_{(\frac{n}{2}+1)j}, \ldots, x_{nj}$ are modified to $t_{cj} + 1, t_{cj} + 1, \ldots, t_{cj} + 1$. For example, if $T = (t_1, t_2, t_3, t_4)$, where $t_1 = (5, 6, 0)$, $t_2 = (2, 5, 0)$, $t_3 = (4, 7, 0)$ and $t_4 = (5, 6, 0)$ and $\epsilon = 2, k = 4$, the centroid is $t_c = (4, 6, 0)$, then after the modification $T' = (t'_1, t'_2, t'_3, t'_4)$, where $t'_1 = (5, 5, 0)$, $t'_2 = (3, 5, 0)$, $t'_3 = (3, 7, 0)$ and $t'_4 = (5, 7, 0)$. See matrix (3) for a more visualized transformation. The numbers in bold indicate that they are modified. The modification of the graphical representation G to the 4-clique G' is shown in Fig. 10.

$$T = \begin{pmatrix} 5 & 6 & 0 \\ 2 & 5 & 0 \\ 4 & 7 & 0 \\ 5 & 6 & 0 \end{pmatrix} \Rightarrow \begin{pmatrix} 5 & \mathbf{5} & 0 \\ \mathbf{3} & 5 & 0 \\ \mathbf{3} & 7 & 0 \\ 5 & \mathbf{7} & 0 \end{pmatrix} = T' \tag{3}$$

Case 2.1.2 If $\epsilon > 1$ and n is odd, the first $\frac{n-1}{2}$ ratings $x_{1j}, x_{2j}, \ldots, x_{\frac{n-1}{2}j}$ are modified to $t_{cj} - 1, t_{cj} - 1, \ldots, t_{cj} - 1$, the $\frac{n}{2}$th is modified to t_{cj}, and the remaining $\frac{n+1}{2}$ ratings $x_{\frac{n}{2}j}, x_{(\frac{n+1}{2})j}, \ldots, x_{nj}$ are modified to $t_{cj} + 1, t_{cj} + 1, \ldots, t_{cj} + 1$. For example, if $T = (t_1, t_2, t_3, t_4, t_5)$, where $t_1 = (5, 6, 0)$, $t_2 = (2, 5, 0)$, $t_3 = (4, 7, 0)$, $t_4 = (5, 6, 0)$ and $t_5 = (4, 6, 0)$ and $\epsilon = 2, k = 5$, the centroid is $t_c = (4, 6, 0)$, then after the modification $T' = (t'_1, t'_2, t'_3, t'_4, t'5)$, where $t'_1 = (5, 5, 0)$, $t'_2 = (3, 5, 0)$, $t'_3 = (4, 7, 0)$, $t'_4 = (5, 6, 0)$, and $t'_5 = (3, 7, 0)$. See the matrix (4) for a more visualized transformation. The numbers in bold indicate that they are modified. The modification of the graphical representation G to the 5-clique G' is shown in Fig. 11.

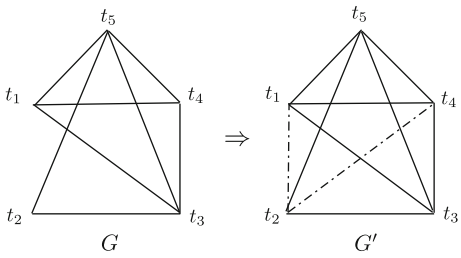


Fig. 11 The modification of graphical representation G for Case 2.1.2

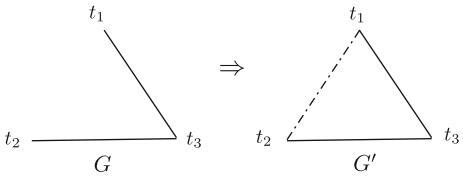


Fig. 12 The modification of graphical representation G for Case 2.2.1

$$T = \begin{pmatrix} 5 & 6 & 0 \\ 2 & 5 & 0 \\ 4 & 7 & 0 \\ 5 & 6 & 0 \\ 4 & 6 & 0 \end{pmatrix} \Rightarrow \begin{pmatrix} 3 & 6 & 0 \\ \mathbf{3} & 5 & 0 \\ 4 & 7 & 0 \\ 5 & 6 & 0 \\ \mathbf{3} & \mathbf{7} & 0 \end{pmatrix} = T' \tag{4}$$

Case 2.1.3 If $\epsilon = 1$ and n is odd, the ratings $x_j = (x_{1j}, x_{2j}, \ldots, x_{nj})$ are all changed to the $t_{cj}, t_{cj}, \ldots, t_{cj}$, $1 \leq j \leq m$.

Case 2.2 (*Fraction strategy*) If t_{ci} is a fraction, $\forall i = 1, 2, \ldots, m$, then since $t_{ci} = \frac{x_{1i}+x_{2i}+\ldots+x_{ni}}{n}$, $1 \leq i \leq n$, write it in another form $t_{ci} = \lfloor t_{ci} \rfloor + \frac{r}{n}$, where $\lfloor t_{ci} \rfloor$ is the largest integer that is smaller than t_{ci} and r is an integer with $0 < \frac{r}{n} < 1$.

Case 2.2.1 If $r \leq \epsilon$, the ratings $x_{1j}, x_{2j}, \ldots, x_{nj}$ are modified to $\lfloor t_{ci} + r \rfloor, \lfloor t_{ci} \rfloor, \ldots, \lfloor t_{ci} \rfloor$. Actually, r can be added to any one ratings. For simplicity, we add it to the first rating. For example, if $T = (t_1, t_2, t_3)$, where $t_1 = (5, 6)$, $t_2 = (2, 5)$ and $t_3 = (4, 6)$ with $\epsilon = 2, k = 3$. The centroid is $t_c = (\frac{11}{3}, \frac{17}{3})$. For $t_{c1} = \lfloor t_{c1} \rfloor + \frac{r}{n} = 3 + \frac{2}{3}$ and $t_{c2} = \lfloor t_{c2} \rfloor + \frac{r}{n} = 5 + \frac{2}{3}$. After the modification $T' = (t'_1, t'_2, t'_3)$, where $t'_1 = (5, 7)$, $t'_2 = (3, 5)$ and $t'_3 = (3, 5)$. See the matrix (5) for a more visualized transformation. The numbers in bold indicate that they are modified. The modification of the graphical representation G to the 3-clique G' is shown in Fig. 12.

$$T = \begin{pmatrix} 5 & 6 \\ 2 & 5 \\ 4 & 6 \end{pmatrix} \Rightarrow \begin{pmatrix} 5 & \mathbf{7} \\ \mathbf{3} & 5 \\ \mathbf{3} & \mathbf{5} \end{pmatrix} = T' \tag{5}$$

Case 2.2.2 If $r > \epsilon$, then r can be written in the form $r = p \times \epsilon + s$, where the integers $p \geq 1$ and $0 \leq s < \epsilon$. The ratings $x_{1j}, x_{2j}, \ldots, x_{nj}$ are modified to $\lfloor t_{ci} \rfloor + 1, \lfloor t_{ci} \rfloor + 1, \cdots, \lfloor t_{ci} \rfloor + s$. p is added to the first $p \times \epsilon$ ratings, and s is added to the last rating. For example, if $T = (t_1, t_2, t_3, t_4)$, where $t_1 = (5, 6)$, $t_2 = (2, 5)$, $t_3 = (4, 7)$ and $t_3 = (4, 5)$ with $\epsilon = 2, k = 4$. The centroid is $t_c = (\frac{13}{4}, \frac{23}{4})$. For $t_{c1} = \lfloor t_{c1} \rfloor + \frac{r}{n} = 3 + \frac{3}{4}$ and $t_{c2} = \lfloor t_{c2} \rfloor + \frac{r}{n} = 5 + \frac{3}{4}$. Since $r = 3 > \epsilon = 2$,

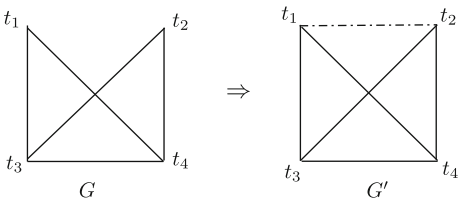


Fig. 13 The modification of graphical representation G for Case 2.2.2

we write $r = p \times \epsilon + \frac{s}{\epsilon} = 1 + \frac{1}{2}$. After the modification $T' = (t'_1, t'_2, t'_3, t'_4)$, where $t'_1 = (4, 6)$, $t'_2 = (4, 6)$, $t'_3 = (3, 5)$ and $t'_4 = (4, 6)$. See the matrix (6) for a more visualized transformation, and the numbers in bold indicate that they are modified. The modification of the graphical representation G to the 4-clique G' is shown in Fig. 13.

$$T = \begin{pmatrix} 5 & 6 \\ 2 & 5 \\ 4 & 7 \\ 4 & 5 \end{pmatrix} \Rightarrow \begin{pmatrix} \mathbf{4} & 6 \\ \mathbf{4} & \mathbf{6} \\ \mathbf{3} & \mathbf{5} \\ 4 & \mathbf{6} \end{pmatrix} = T' \tag{6}$$

Case 2.3 (*Mixed strategy*) If t_{ci} is an integer, for some $i = 1, 2, \ldots, m$ and for others t_{ci} is a fraction, then apply the integer strategy to the ratings whose t_{ci} is an integer, and apply fraction strategy to the ratings whose t_{ci} is a fraction.

The following theorem prove that the cases are complete and the modified data set indeed satisfies (k, ϵ)-anonymity requirement.

Theorem 3 *(Correctness and completeness) Given a survey rating data set T, ϵ and k, the modified data set T' satisfies (k, ϵ)-anonymity after applying modification cases.*

Proof Suppose a survey rating data set $T = (t_1, t_2, \ldots, t_n)$, $t_i = (x_{i1}, x_{i2}, \ldots, x_{im})$, where x_{ij} is the rating of survey participant i on the issue j ($1 \le i \le n$, $1 \le j \le m$). $x_j = (x_{1j}, x_{2j}, \ldots, x_{nj})$ denotes the vector of ratings on issue j by all the survey participants ($1 \le j \le m$). In order to discuss the modification of the data, without loss of generality, we assume that T forms one (k, ϵ)-anonymous group after the modification. Given the privacy requirement ϵ, k, we construct the graphical representation G of the data set T. We publish $T' = (t'_1, t'_2, \ldots, t'_n)$, and $t'_i = (x'_{i1}, x'_{i2}, \ldots, x'_{im})$, $1 \le i \le n$ and $1 \le j \le m$. We verify the statement case by case.

Case 2.1.1 For the jth issue, the first $\frac{n}{2}$ ratings $x_{1j}, x_{2j}, \ldots, x_{\frac{n}{2}j}$ are modified to $t_{cj} - 1, t_{cj} - 1, \ldots, t_{cj} - 1$, and the remaining $\frac{n}{2}$ ratings $x_{\frac{n}{2}j}, x_{(\frac{n}{2}+1)j}, \ldots, x_{nj}$ are modified to $t_{cj} + 1, t_{cj} + 1, \ldots, t_{cj} + 1$. It is easily verified that the distance between any two ratings is bounded by 2, which is no more than ϵ.

Case 2.1.2 For the jth issue, the first $\frac{n-1}{2}$ ratings $x_{1j}, x_{2j}, \ldots, x_{\frac{n-1}{2}j}$ are modified to $t_{cj} - 1, t_{cj} - 1, \ldots, t_{cj} - 1$, and the $\frac{n}{2}$th is modified to t_{cj}, and the remaining $\frac{n+1}{2}$ ratings $x_{\frac{n}{2}j}, x_{(\frac{n+1}{2})j}, \ldots, x_{nj}$ are modified to $t_{cj} + 1, t_{cj} + 1, \ldots, t_{cj} + 1$. It is easy to verify that the distance between any two ratings is bounded by either 1 or 2, which is no more than ϵ as well.

Case 2.1.3 This is the most trivial case where all the ratings are the same for issue j. Of course, the ϵ requirement is satisfied since the distance between any two ratings is 0.

Case 2.2.1 For issue j, the ratings $x_{1j}, x_{2j}, \ldots, x_{nj}$ are modified to $\lfloor t_{ci} + r \rfloor$, $\lfloor t_{ci} \rfloor, \ldots, \lfloor t_{ci} \rfloor$. The distance between two ratings is bounded by r, which is no more than ϵ under this case.

Case 2.2.2 For issue j, the ratings $x_{1j}, x_{2j}, \ldots, x_{nj}$ are modified to $\lfloor t_{ci} \rfloor + 1, \lfloor t_{ci} \rfloor + 1, \ldots, \lfloor t_{ci} \rfloor + s$. The distance between two ratings is bounded either by 1 or $s - 1$, which is no more than ϵ under this case. □

In practice, applying one single data modification methods is not adequate. Usually a combination of several strategies is needed to meet the (k, ϵ) requirements. In order to test the efficiency and effectiveness of our proposed approaches, we have conducted extensive experiments which are described and discussed in the next section.

5 Proof-of-concept experiments

In this section, we experimentally evaluate the effectiveness and efficiency of the proposed survey rating data publication method. Our objectives are three-fold. Firstly, we verify that publishing the survey rating data satisfying (k, ϵ)-anonymity via our proposed approaches is fast and scalable. Secondly, we show that the anonymous survey rating data sets produced permit accurate data analysis. Finally, we perform the statistical analysis on both original and anonymized data sets.

5.1 Data sets

Our experimentation uses two real-world databases, MovieLens[2] and Netflix.[3] The MovieLens data set was made available by the GroupLens Research Project at the University of Minnesota. The data set contains 100,000 ratings (5-star scale), 943 users and 1682 movies. Each user has rated at least 20 movies. The Netflix data set was released by Netflix for competition. This movie rating data set contains over 100,480,507 ratings from 480,189 randomly-chosen Netflix customers over 17,000 movie titles. The Netflic data were collected between October, 1998 and December, 2005 and reflected the distribution of all ratings received during this period. The ratings are on a scale from 1 to 5 stars. In both data sets, a user is considered as a survey participant while a movie is regarded as an issue to respond. Many entries are empty since each participant only rated a small number of movies. We consider all the movies as non-sensitive attributes, and add one sensitive issue "income level" to each data set, in which the ratings scales from 1 to 5. We randomly generate a rating from 1 to 5 and assign it to each record. The correspondence of the income level and income

[2] http://www.grouplens.org/taxonomy/term/14.

[3] http://www.netflixprize.com/.

Table 4 Correspondence of income level and income interval

Income level (1–5)	Income interval
1	$0–$6,000
2	$6,001–$35,000
3	$35,001–$80,000
4	$80,001–$180,000
5	>$180,001

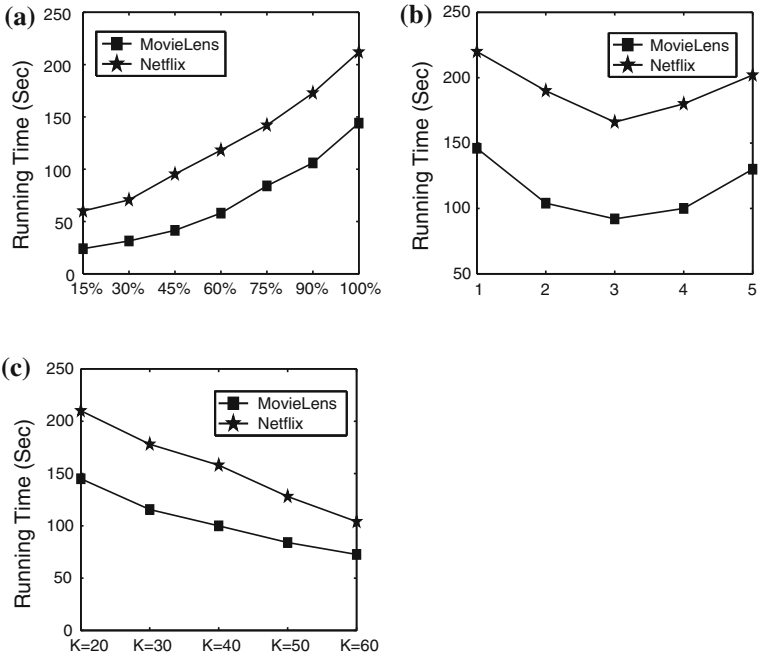


Fig. 14 Running time on MovieLens and Netflix data sets vs. **a** Data percentage varies; **b** ϵ varies; **c** k varies

interval is shown in Table 4, where the classification of income interval is referred as the Australia Tax Rates 2008–2009.[4]

5.2 Efficiency

Data used for Fig. 14a is generated by re-sampling the Movielens and Netflix data sets while varying the percentage of from 15 to 100%. For both data sets, we evaluated the running running time for the (k, ϵ)-anonymity model with the default setting $k = 20, \epsilon = 1$. For both testing data sets, the execution time for (k, ϵ)-anonymity

[4] http://www.ato.gov.au/individuals/content.asp?doc=/content/12333.htm.

is increased by enlarging the percentage of both data sets. This is because as the percentage of data increases, the computation cost increases too. The result is expected since the overhead is increased with the more dimensions.

Next, we evaluated the effect of the parameters k, ϵ on the cost of computing. The data sets used for this experiment are the whole MovieLens and Netflix databases and we evaluate by varying the value of ϵ and k. With $k = 20$, Fig. 14b shows the computational cost as a function of ϵ, in determining (k, ϵ)-anonymity for both data sets. Interestingly, in both data sets, as ϵ increases, the cost initially becomes lower but then increases monotonically. This phenomenon is due to a pair of contradicting factors that push up and down the running time, respectively. At the initial stage, when ϵ is small, fewer edges are contained in the graphical representation of the data set, and therefore, more computation efforts are put into edge addition and data modification operations. This explains the initial descent of overall cost. However, as ϵ grows, there are more possible (k, ϵ)-anonymous solutions and searching for the one with least distortion requires a larger overhead, and this causes the eventual cost increase. Setting $\epsilon = 2$, Fig. 14c displays the results of execution time by varying k from 20 to 60 for both data sets. The cost drops as k grows. This is expected because fewer search efforts for possible (k, ϵ)-anonymous solutions are needed for a greater k, allowing our algorithm to terminate earlier.

5.3 Data utility

Having verified the efficiency of our technique, we proceed to test its effectiveness. We measure data utility as the error in answering average rating queries ion the anonymized survey rating data it produces by running 100 random queries of the rating of a movie. Each query has the form:

SELECT COUNT(*) FROM MovieLens/NetFlix
WHERE $pred(A_1^n)$ AND $\cdots\cdots$ AND $pred(A_w^n)$ AND $pred(A^s)$.

Specifically, a query involves w random non-sensitive attributes $A_1^n, \ldots, A_w^n$ (in the underlying Movielens/Netlflix), and the sensitive attribute A^s, where w is a parameter called query dimensionality. For instance, if the the survey rating data is Table 1a and $w = 2$, then $\{A_1^n, A_2^n\}$ is a random 2-sized subset of non-sensitive issues {issue 1, issue 2, issue 3}. For any issue A, the predicate $pred(A)$ has the form of $(A = x_1$ OR $A = x_2$ OR $\cdots$ OR $A = x_b)$, where $x_i (1 \leq i \leq b)$ is a random value in the domain of A. The value of b depends on the expected query selectivity s: $b = \lceil |A| \cdot s^{1/(w+1)} \rceil$, where $|A|$ is the domain size of A. A higher s leads to more selection conditions in $pred(A)$. We derive the estimated answer of a query using the approach explained in LeFevre et al. (2006a). The accuracy of an estimate is evaluated as its relative error. Let *act* and *est* be the actual and estimated results respectively. The relative error then equals $|act - est|/act$.

We first study the influence of ϵ (i.e., the length of a proximate neighborhood) on data utility. Towards this, we set k to 10. With $(10, \epsilon)$-anonymity, Fig. 15a plots the average error on both data sets as a function of ϵ. (k, ϵ)-anonymity produces useful anonymized data with average error below 15%. The anonymisation strategies incur higher error as ϵ decreases. This is expected, since a smaller ϵ demands

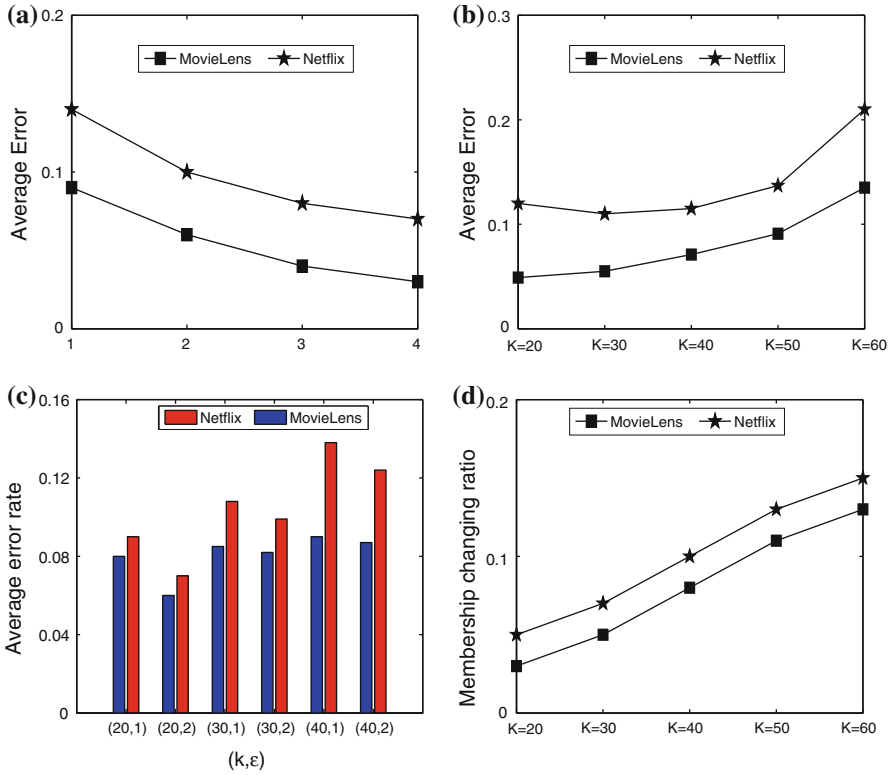


Fig. 15 Performance comparison on MovieLens and Netflix data sets: **a** Query accuracy vs. ϵ; **b** query accuracy vs. k; **c** query accuracy vs. (k, ϵ); **d** clusters changes vs. k

stricter privacy preservation, which reduces data utility. Next, we examined the utility of (k, ϵ)-anonymous solutions with different k when $\epsilon = 2$. Figure 15b presents the average error of 100 random queries of the average rating as a function of k. The error grows with k because a larger k demands tighter anonymity control. Nevertheless, even for the greatest k, the data still preserves fairly good utility by our technique, incurring an error of no more than 20% for Movielens and 25% for Netflix. Figure 15c plots the query accuracy by changing the parameter pair (k, ϵ). We vary the k from 20 to 40 with ϵ changing from 1 to 2. From the graph we can see, when ϵ (k) is fixed, the accuracy is increasing with k (ϵ), which is consistent with the results obtained from Fig. 15a and b.

Since our objective is to anonymize large survey rating data, we adopt another criterion to evaluate data utility called membership changing ratio. This the proportion of data points changing cluster memberships from clusters on the original data set to clusters on the anonymized data set when a clustering algorithm [e.g., k-means algorithm (Kanungo et al. 2002)] runs on both data sets. We first anonymize the original dataset by our anonymisation method, and then we run a k-means algorithm over both the original and anonymous data sets, keeping the same initial seeds and identical k.

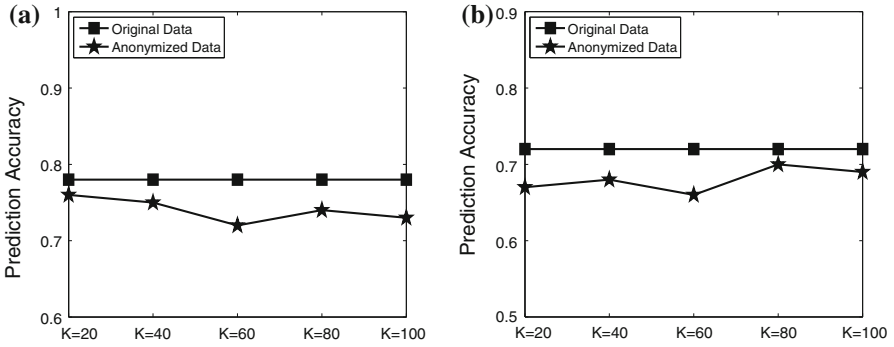


Fig. 16 Prediction accuracy. **a** Movielens; **b** Netflix

We use the proportion of data points changing cluster memberships as another measure of utility. Generally, the lower the membership changing ratio is, the higher the data utility is preserved. Figure 15d plots clustering membership changing ratio versus k. The membership changing ratio increases with increasing k. When $k = 60$, the less than 15% for both data sets. This shows that our data This shows that our data modification approach preserves the grouping quality of anonymized data very well.

Figure 16a and b evaluate the classification and prediction accuracy of the graph anonymization algorithm. Our evaluation methodology is that we first divide data into training and testing sets, and we apply the graph anonymization algorithm to the training and testing sets to obtain the anonymized training and testing sets, and finally the classification or regression model is trained by the anonymized training set and tested by anonymized testing set. The Weka implementation (Witten and Frank 2005) of simple Naive Bayes classifier was used for the classification and prediction. Using the Movielens data, Fig. 16a compares the predictive accuracy of classifier trained on Movielens data produced by the graph anonymization algorithm. In these experiments, we generated 50 independent training and testing sets, each containing 2000 records, and we fixed $\epsilon = 1$. The results are averaged across these 50 trials. For comparison, we also include the accuracies of classifier trained on the (not anonymized) original data. From the graph, we can see that the 75%, very close to the original accuracy, which preserves better original accuracy, which preserves better utility for data mining purposes. In Fig. 16b, similar results are obtained by using the Netflix rating data.

5.4 Statistical properties

We further performed the statistical analysis on the original and anonymous data sets. In this series of evaluations, we compare some key statistical properties, centroid and standard deviation with the original and anonymized data, since these statistics are extremely useful in the data mining environment for anonymous data sets. For the centroid comparison, we first calculated the average vector of the ratings that are not *null* of each attribute, then compared the inner product of this vector with the result of the same operation on the anonymous data set. The results were evaluated by varying

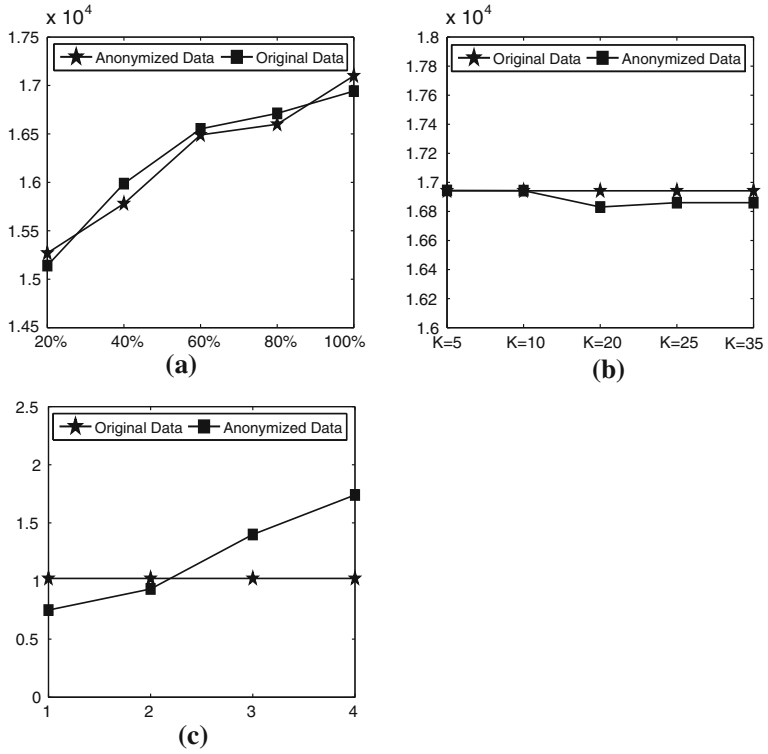
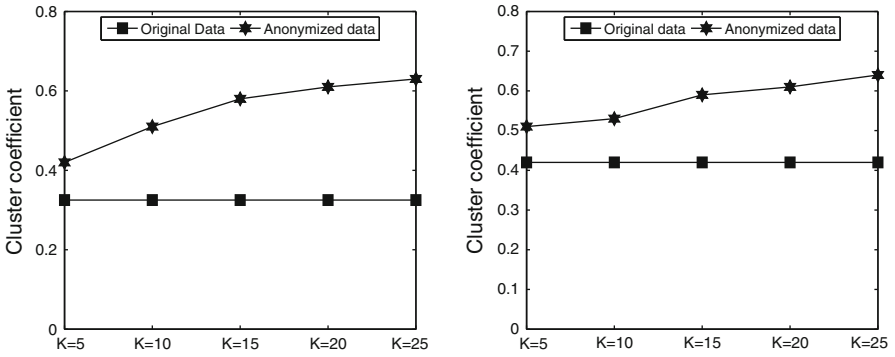


Fig. 17 Statistical properties analysis (Movielens Data set): **a** centroid vs. data percentage; **b** centroid vs. k; **c** standard deviation vs. ϵ

the percentage of the data and the privacy requirement k. For the standard deviation, we computed the average standard deviation among all the attributes for the original and anonymous data sets. The experiments were conducted by varying ϵ.

We first compared the centroid before and after anonymisation while varying the percentage of the data set. We set $k = 20, \epsilon = 2$ and let the percentage of the data vary from 20 to 100%. The result is shown in Fig. 17a. We can see that although the centroid between original and anonymous data sets are different, they do not differ much which makes the data useful for the data mining purposes, and the results suggest that our modification strategies preserve the centroid of the data. We then fixed the data set with $\epsilon = 2$ and varied the privacy requirement k from 5 to 35. The result is shown in Fig. 17b. No matter what kind of operations are used, the centroids before and after the operation are similar to each other. Figure 17c compares average standard deviations before and after data anonymisation. The average standard deviation remains constant for the original data, since parameter ϵ has no effect on it. For the anonymous data set, the standard deviation is bounded by some specific value for a given ϵ. It is not difficult to prove the upper bound of the standard deviation for issue s is $\frac{s_{\max}-s_{\min}}{2}$, where $s_{\max}$ and $s_{\min}$ are the maximum and minimum ratings of s. With



(a) Movielens **(b)** Netflix

Fig. 18 Average cluster coefficient comparison

the parameter ϵ, the standard deviation is bounded by $\frac{\epsilon}{2}$. Similar results were obtained on Netflix data sets as well.

Besides the centroid and standard deviation, we also test with the statistical measure the average cluster coefficient. Clustering coefficient is a property of a node in a graph or a network. Roughly speaking it tells how well connected the neighborhood of the node is. If the neighborhood is fully connected, the clustering coefficient is 1 and a value close to 0 means that there are hardly any connections in the neighborhood. Average clustering coefficient of the graph is the average of the clustering coefficient over all nodes in the graph.

We construct the graphical representations of the original and anonymized survey rating data, and evaluate the average cluster coefficient by setting $\epsilon = 2$ and varying the value of k from 5 to 25. Figure 18a and b show the comparison results on Movielens and Netflix data sets. On the one hand, As we can see, in both data sets, the average of cluster coefficient is increased in the anonymized data. This is because in order to satisfy the (k, ϵ)-anonymity, we need to add more edges into the graph, which leads to the increase of the cluster coefficient. On the other hand, before anonymization, since the data sets are very sparse, the value of cluster coefficient is relatively low, and after the anonymization, although the value of cluster coefficient has been increased, it does not reach to 1, the trivial case when the graph becomes fully connected. This reflects the fact that our anonymziation method does not add too many edges to reduce its usefulness.

6 Conclusion and future work

In this paper, we alleviate a privacy threat to a large survey rating data set with an anonymisation principle called (k, ϵ)-anonymity. We apply a graphical representation to formulate the problem and provide a comprehensive analysis of the graphical modification strategies. Extensive experiments confirm that our technique produces anonymized data sets that are highly useful and preserve key statistical properties.

This work also initiates several directions for future investigations on our research agenda. Firstly, the (k, ϵ)-anonymity model introduced in the paper is targeted at identify protection in a large survey rating data set, it is also important to address the issue of how to prevent attribute disclosures. The privacy principle similar to l-diversity might be considered. Secondly, in this paper, we consider only edge addition operations for graph modification, and it is interesting to investigate the case when the privacy requirement is achieved by deleting some transactions from the survey rating data. Thirdly, since we have proven that finding an optimal solution is NP-hard, it might be possible to find an approximate solution with a better approximate ratio. Finally, it is also interesting to employ dimensionality-reduction techniques for more effective anonymisation.

Acknowledgment This research is supported by Australian Research Council (ARC) grant DP0774450, DP0663414 and DP110103142.

References

Aggarwal C (2005) On k-anonymity and the curse of dimensionality. In: VLDB, pp 901–909
Atzori M, Bonchi F, Giannotti F, Pedreschi D (2005a) Blocking anonymity threats raised by frequent itemset mining. In: ICDM, pp 561–564
Atzori M, Bonchi F, Giannotti F, Pedreschi D (2005b) k-anonymous patterns. In: PKDD, pp 10–21
Atzori M, Bonchi F, Giannotti F, Pedreschi D (2008) Anonymity preserving pattern discovery. VLDB J 17(4):703–727
Bayardo RJ, Agrawal R (2005) Data privacy through optimal k-anonymisation. In: ICDE, pp 217–228
Frankowski D, Cosley D, Sen S, Terveen LG, Riedl J (2006) You are what you say: privacy risks of public mentions. In: SIGIR, pp 565–572
Fung BC, Wang K, Yu PS (2005) Top-down specialization for information and privacy preservation. In: ICDE, pp 205–216
Garey MR, Johnson DS (1979) Computers and intractability: a guide to the theory of $\mathcal{NP}$-completeness. Freeman, San Francisco
Ghinita G, Tao Y, Kalnis P (2008) On the anonymisation of sparse high-dimensional data. In: Proceedings of international conference on data engineering (ICDE), April, pp 715–724
Hafner K (2006) And if you liked the movie, a Netflix contest may reward you handsomely. New York Times, Oct 2
Hamming RW (1980) Coding and information theory. Prentice Hall, Englewood Cliffs
Hansell S (2006) AOL removes search data on vast group of web users. New York Times, Aug 8
He Y, Naughton J (2009) Anonymization of set-valued data via top-down, local generalization. In: VLDB 2009: proceedings of the thirtieth international conference on very large data bases. VLDB endowment
Iyengar V (2002) Transforming data to satisfy privacy constraints. In: SIGKDD, pp 279–288
Kanungo T, Mount DM, Netanyahu NS, Piatko CD, Silverman R, Wu AY (2002) An efficient k-means clustering algorithm: analysis and implementation. IEEE Trans Pattern Anal Mach Intell 24:881–892
Kifer D, Gehrke J (2006) Injecting utility into anonymized datasets. In: SIGMOD conference, pp 217–228
LeFevre K, DeWitt D, Ramakrishnan R (2006a) Mondrian multidimensional k-anonymity. In: ICDE, pp 25–25
LeFevre K, DeWitt DJ, Ramakrishnan R (2006b) Workload-aware anonymisation. In: KDD, pp 277–286
Li T, Li N (2009) On the tradeoff between privacy and utility in data publishing. In: Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining (SIGKDD), pp 517–526
Li N, Li T, Venkatasubramanian S (2007) t-Closeness: privacy beyond k-anonymity and l-diversity. In: ICDE, pp 106–115
Li T, Li N, Zhang J (2009) Modeling and integrating background knowledge in data anonymization. In: ICDE, pp 6–17
Liu K, Terzi E (2008) Towards identity anonymization on graphs. In: SIGMOD

Machanavajjhala A, Gehrke J, Kifer D, Venkitasubramaniam M (2006) *l*-Diversity: privacy beyond *k*-anonymity. In: ICDE, p 24

Meyerson A, Williams R (2004) On the complexity of optimal *k*-anonymity. In: Proceedings of the 23rd ACM-SIGMOD-SIGACT-SIGART symposium on the principles of database systems, Paris, France, pp 223–228

Narayanan A, Shmatikov V (2008) Robust de-anonymisation of large sparse datasets. In: IEEE security and privacy, pp 111–125

Samarati P (2001) Protecting respondents' identities in microdata release. IEEE Trans Knowl Data Eng 13(6):1010–1027

Samarati P, Sweeney L (1998a) Generalizing data to provide anonymity when disclosing information (abstract). In: PODS, p 188

Samarati P, Sweeney L (1998b) Protecting privacy when disclosing information: *k*-anonymity and its enforcement through generalization and suppression. Technical report SRI-CSL-98-04, SRI Computer Science Laboratory

Sweeney L (1997) Weaving technology and policy together to maintain confidentiality. J Law Med Ethics 25(2–3): 98–110

Sweeney L (2002) *k*-Anonymity: a model for protecting privacy. Int J Uncertain Fuzziness Knowl Syst 10(5):557–570

Verykios VS, Elmagarmid AK, Bertino E, Dasseni E, Saygin Y (2004) Association rule hiding. IEEE Trans Knowl Data Eng 16(4):434–447

Wang K, Fung BCM (2006) Anonymizing sequential releases. In: ACM SIGKDD, pp 414–423

Wang K, Yu PS, Chakraborty S (2004) Bottom-up generalization: a data mining solution to privacy protection. In: The fourth IEEE international conference on data mining (ICDM 2004), pp 249–256

Witten I, Frank E (2005) Data mining: practical machine learning tools and techniques. 2. Morgan Kaufmann, San Francisco

Wong R, Li J, Fu A, Wang K (2006) (α, k)-anonymity: an enhanced *k*-anonymity model for privacy preserving data publishing. In: KDD, pp 754–759

Xu Y, Wang K, Fu Ada W-C, Yu PS (2008) Anonymizing transaction databases for publication. In: KDD, pp 767–775

Zhang Q, Koudas N, Srivastava D, Yu T (2007) Aggregate query answering on anonymized tables. In: ICDE, pp 116–125

Zhou B, Pei J, Luk WS (2008) A brief survey on anonymization techniques for privacy preserving publishing of social network data. ACM SIGKDD Expl 10(2):12–22