Attribute susceptibility and entropy based data anonymization to improve users community privacy and utility in publishing data

Abdul Majeed¹ · Sungchang Lee¹

Published online: 12 March 2020
© The Author(s) 2020

Abstract

User attributes affect community (i.e., a group of people with some common properties/attributes) privacy in users' data publishing because some attributes may expose multiple users' identities and their associated sensitive information during published data analysis. User attributes such as gender, age, and race, may allow an adversary to form users' communities based on their values, and launch sensitive information inference attack subsequently. As a result, explicit disclosure of private information of a specific users' community can occur from the privacy preserved published data. Each item of user attributes impacts users' community privacy differently, and some types of attributes are highly susceptible. More susceptible types of attributes enable multiple users' unique identifications and sensitive information inferences more easily, and their presence in published data increases users' community privacy risks. Most of the existing privacy models ignore the impact of susceptible attributes on user's community privacy and they mainly focus on preserving the individual privacy in the released data. This paper presents a novel data anonymization algorithm that significantly improves users' community privacy without sacrificing the guarantees on anonymous data utility in publishing data. The proposed algorithm quantifies the susceptibility of each attribute present in user's dataset to effectively preserve users' community privacy. Data generalization is performed adaptively by considering both user attributes' susceptibility and entropy simultaneously. The proposed algorithm controls over-generalization of the data to enhance anonymous data utility for the legitimate information consumers. Due to the widespread applications of social networks (SNs), we focused on the SN users' community privacy preserved and utility enhanced anonymous data publishing. The simulation results obtained from extensive experiments, and comparisons with the existing algorithms show the effectiveness of the proposed algorithm and verify the aforementioned claims.

Keywords User attributes · Anonymization · Susceptibility · Entropy · Community privacy · Utility · Social network

1 Introduction

Due to the popularization of the social networks (SNs), all SN service providers collect and store user's data much like hospitals and banks. This data often contains information about the user's demographics, finance, SN activities and preferences, hobbies, community affiliations, medical status, and relationships with other online users. Aside from the required information necessary to create an account, most users readily post other valuable data including popular songs, favourite

talk shows, political and religious views, and interesting movies which third parties and analysts collect. Accordingly, this goldmine of data has attracted the attention of interested third parties who are interested in such data as part of their business strategy [1]. Such data is invaluable for understanding the social trends, users' behaviours, opinions, contents recommendation, and targeted advertisement [2].

On the one hand, releasing user's personal data is beneficial for extracting accurate, timely, detailed, and multifaceted insights about the users with advanced data mining tools and pattern analysis applications. On the other hand, the release of such data poses a potential privacy threat including identity disclosure, sensitive attribute disclosure, and membership disclosure [3]. Sweeney, in her study, determined that 87% people in the United States can be uniquely identified based only on five-digit zip code, gender, and date of birth [4]. Besides the individual's privacy problems, an adversary can infer the sensitive information based on the users quasi identifiers' values that can jeopardize the privacy of a specific community

✉ Sungchang Lee
sclee@kau.ac.kr

Abdul Majeed
abdulmajeid09398@kau.kr

¹ School of Information and Electronics Engineering, Korea Aerospace University, Deogyang-gu, Goyang-si, Gyeonggi-do 412-791, Korea

(i.e., group of users). Therefore, it is paramount that any sharing and mining of SN data must protect the privacy of users' community.

Privacy preserving data publishing (PPDP) provides a set of tools, models, and methods to safeguard against the privacy threats that emerge from the mining and sharing of this data. The original data about the users contains four types of attributes: explicit identifiers (i.e., name, SSN), quasi-identifiers (i.e., age, race, zip code, and gender), sensitive attributes (i.e., salary, disease, and political views), and non-sensitive attributes (i.e., height) [5]. In PPDP, the explicit identifiers and non-sensitive attributes are removed before the collected data is shared with third parties, while quasi-identifiers (QIs) are generalized or suppressed, and sensitive attributes (SA) are retained as it is for the analytical purposes. In general, those who have collected and mined this data are reluctant to share this data in any way, shape, or form because attackers often possess strong background knowledge or have access to the excessive amount of auxiliary information necessary to breach safeguards used to protect user's privacy [6].

The anonymization approaches used for preserving user's privacy in released data are classified into two major categories, perturbation-based approaches and permutation-based approaches [7]. The former approaches use generalization and suppression techniques [8] to modify the QI's original values to protect a user's privacy and this category of techniques is also called relational anonymization. The later approaches apply the permutation to a graph structure (i.e., by either adding random nodes and edges or both) [9] to preserve user's privacy and this category of techniques is called structural anonymization. k -anonymity [10], ℓ -diversity [11], t -closeness [12], slicing [13], anatomy [14], and their improved versions are well known relational anonymization techniques. Random walk [15], differential privacy [16], cluster based techniques [17], k -anonymity based techniques [18, 19], edge editing [20], k -degree anonymization [21], k -neighbourhood anonymization [22], k -isomorphism anonymization [23], k -automorphism anonymization [24], edge differential privacy [25], node differential privacy [26], vertex degree distribution and attribute value distribution [27, 28], and uncertainty semantics [29] are the most widely used structural anonymization techniques.

Most of the existing PPDP algorithms do not provide thorough insight into users' community privacy protection, particularly regarding the susceptibility of QIs and the entropy of SA-based adaptive data generalization. Existing privacy models mainly focus on the individual privacy protection and they do not quantify the QIs impact on individual as well as community privacy. In addition, each user in a SN can be associated with a community identity [30]. The community identity of a user can be highly sensitive to the public due to the personal information (i.e., political activity group, friend group association, and online disease support group etc.) it

contains. Therefore, it is mandatory to consider the QI's susceptibility and SA's entropy simultaneously in data anonymization to protect the users' community privacy breaches. Recently, advances in machine learning tools and data availability have made multiple users identities and their SA revelation more easier [31]. An even greater source of concern is the need to protect the SA prediction from the published data [32], and algorithm-based disclosures [33]. The capability of machine learning tools is increasing rapidly, and they enable adversaries to alter algorithm parameters to compromise user's privacy on a large scale. Therefore, considering the continually increasing capability of machine learning tools, and limitations of the existing work, the need to develop methods that use the capabilities of these tools to protect users' community privacy has become more pressing than ever.

The remainder of this paper is structured as follows; Section 2 explains the background and related work regarding well-known PPDP algorithms. Section 3 presents the proposed anonymization algorithm and explains its principal steps. Section 4 discusses the experiments and simulation results. Finally, conclusions and future directions are presented in Section 5.

2 Background and related work

The protection of a user's privacy in SN data publishing is an active area of research, resulting in numerous proposed solutions to solve this social problem. With the advancements in data mining tools, the scale and the scope of privacy breaches is expanding from the unique identification of an individual or SA disclosure to the behavioral advertising, online stalking, and users group identity theft [34]. Due to the information surge and maturity of the data mining tools, such privacy breaches in PPDP do occur more frequently. Protecting the privacy of the individual user and his/her community, and enhancing the anonymous data utility is a longstanding challenge in SN users' data publishing [35]. According to the Pham et al. [36] there are three major privacy areas in SN that are shown in Fig. 1.

From Fig. 1, it can be observed that user's privacy in SN can be breached through their activities (i.e., comments and posts), stored data in SN service providers database, and released anonymized data with the researchers/data analysts/ 3rd party applications. In this work, our focus is on the third area of privacy breaches. The k -anonymity [10] concept and its ramifications are a very popular privacy models within the PPDP research community. The k -anonymity [10] protects users' privacy by placing at least k users in an equivalence class with same QI's values. Hence, the probability of re-identifying someone from anonymized data becomes $\frac{1}{k}$. Due

Fig. 1 Overview of the three privacy areas in social networks

to the conceptual simplicity, this model has been widely used in PPDP of SN users [37–40]. The k -anonymity privacy model overview is shown in Fig. 2.

The anonymized data given in Fig. 2b hides each user within a crowd of two users (i.e., $k = 2$). The probability of each user being identified in class A is $\frac{1}{2} = 0.5$. Meanwhile, the equivalence classes (ECs) B and C have no diversity in SA values. Therefore, the probability of re-identification based on the background knowledge or in the presence of auxiliary information becomes 1, which violates the standard k -anonymity privacy protection threshold ($\frac{1}{k}$). The subsequent researches (e.g., ℓ -diversity, t -closeness, and (α, k) -anonymity) have considered this problem, but provide limited protection in preventing the release of SA disclosures. Generally, there are two settings of PPDP: non-interactive and interactive. In the former setting, the data owner, a trusted party, publishes the complete dataset in a sanitized form after applying some modifications to the original data and removing directly identifiable information (e.g., name). However, in the interactive

setting the data owner does not publish the whole dataset in an anonymized form like the former setting. Instead, data owner provides an interface to the users through which they may pose different queries about the related data and get (possibly noisy) answers. In this paper, we focus on non-interactive setting of PPDP, and we extend the k -anonymity concept to accomplish the stated assertions.

Differential privacy (DP) [41] is a well-known state of the art method for protecting user's privacy in interactive setting (i.e., in this setting data owner provides an interface for receiving queries and answer queries by respecting user's privacy). Xie et al. [42] proposed a differentiated k -anonymity ℓ -diversity social network anonymity algorithm. The main goals of the presented algorithm are to protect user's privacy and enhance the anonymous data utility. Some research recommends the DP concept be used in recommendation in SN applications [43], high-dimensional PPDP [44], frequent sequence pattern mining without degrading user's privacy [45], and privacy preserving collaborative filtering [46]. Researchers have extended the traditional anonymization concepts for anonymizing SN users' graph G by either guaranteeing k -degree, k -edges, vertex and edge modification, clustering, and ℓ -diverse sensitive node label retention [17, 47–49]. Despite the success of these privacy techniques, in most cases, either individual user private information is inferred through the community's detection, group membership, and friendship information, or users' community privacy breached if a group of users largely share the same SA value.

A growing body of literature has examined the problems of privacy breaches based on the user attributes. Yin et al. [50] defined a new type of attack named, attribute couplet attack in SN data publishing, and they proposed a k -couplet anonymity concept to avoid this attack. Zhang et al. [51] proposed a user's de-anonymization method from SN published data based on the user attribute similarity measure. The proposed method finds the significance values of attributes and utilize these values in re-identifying users uniquely through seed

Quasi identifiers		Sensitive attribute	
ID	Age	Country	Political views
1	35	Greenland	Liberal
2	35	Canada	Conservative
3	38	Belize	Liberal
4	40	Belize	Liberal
5	37	Canada	Conservative
6	37	Canada	Conservative

(a) Original data about the users

Quasi Identifiers		Sensitive Attribute	
ID	Age	Country	Political views
1	35-40	America	Liberal
2	35-40	America	Conservative
3	35-40	America	Liberal
4	35-40	America	Liberal
5	35-40	America	Conservative
6	35-40	America	Conservative

(b) 2-anonymous data about the users

Fig. 2 2-anonymity applied to the user's data with six records

node identification and propagation. Another study suggested that user's privacy cannot be effectively protected by assuming that QIs and SAs are different. In some cases, the QIs can behave as SA and can jeopardize the privacy of the user [52]. Social identity linkage [53], users profile cloning [54], group affiliation link disclosure [55], and community identities disclosures [56], among others, are possible threats on SN users released data. Recently, many sophisticated tools in the form of open source libraries, frameworks, and prototypes have been developed for data anonymity (DA), privacy preserving data mining (PPDM), and privacy preserving utility mining (PPUM). Lin et al. [57] presented a unified privacy preserving and security framework (PPSF) for the DA, PPDP, and PPUM. The developed framework provides thirteen algorithms' implementation. It provides a user-friendly interface for the algorithm's respective parameters selection. PPSF is a practical tool in offering data security and providing results in a easily understood text file format. Zhang et al. [58] presented a (k, p) -anonymity framework to anonymize a dataset by considering the privacy needs of each user for multiple pieces of his/her sensitive information. The proposed (k, p) anonymity framework solves the sensitive information disclosure problems of the k -anonymity and ℓ -diversity models. Several solutions have been proposed to sanitize transaction databases containing sensitive items [59–62]. PPDM allows the knowledge extraction from the dataset, while preserving a user's privacy, and it has become an important research topic [63–65]. Jimmy et al. [66] proposed a method for hiding the user's sensitive information based on density clustering in sanitized datasets. This proposed method achieves desirable results especially in a sparse dataset compared to the single objective algorithms. PPUM has also been extensively studied in the literature [67–69].

Few studies have explored closely related methods used for PPDP with superior privacy protection, the partial k -anonymity (PK) algorithm [70], chaos and perturbation based anonymization (CPA) algorithm [71], and the related weighted full domain anonymization (WFDA) algorithm [72]. The PK algorithm [70] combines the k -anonymity and randomization concepts to anonymize the data. This algorithm effectively preserves the privacy of a user and utility in PPDP. However, the proposed algorithm fails to quantify the QI's susceptibility which enables the adversary to derive sensitive attribute inferences and thus, compromise the privacy of users' community. The CPA algorithm [71] analyses the frequency of unique attribute values of QIs, determine crucial unique values based on the frequency analysis, and perform perturbation only for the crucial values using the chaotic function. The proposed algorithm yields superior performance in terms of utility and privacy. The proposed approach, however, does not consider the SA's diversity, which makes private information disclosure and precise rules extraction easier. The WFDA algorithm [72] determines attribute weights based

on information gain ratio, and uses these weights to determine the generalization order and level, and anonymizes data based on global generalization and suppression. However, the proposed algorithm does not consider the vulnerability of the QIs from a privacy point of view for an individual or a group of individuals (i.e., community) and generates an anonymous data which increases the privacy risks. Therefore, sanitized data produced by the existing algorithms have less privacy protection for the user. Accordingly, QI's susceptibility and SA's entropy have not been jointly used in SN users' community privacy preserved and utility enhanced anonymous data publishing. The improvements and advantages of the proposed algorithm compared to the existing algorithms are summarized as: (1) it limits multiple users' unique identifications caused by the highly susceptible QIs by identifying susceptible QIs from the user's dataset, and by paying considerable attention to these susceptible QIs during their anonymization; (2) it limits SA disclosure about a specific users' community; (3) it performs data generalization in an adaptive manner, whereas most of existing research performs data generalization in a uniform or fixed manner; (4) it enhances anonymous data usefulness by controlling unnecessary generalization; (5) it limits QIs couplet attack to safeguard multiple users' privacy in data publishing; (6) it can be tailored with the data owner's privacy and legitimate information consumers' utility expectations.

The contributions of this research in the field of PPDP can be summarized as follows: (i) it proposes a novel data anonymization algorithm based on QI's susceptibility and SA's entropy that has the potential to significantly reduce multiple users' unique identifications and sensitive information inferences without sacrificing guarantees on anonymous data utility in SN users' data publishing; (ii) it quantifies the susceptibility of each QI using random forest to reduce the multiple users' unique identifications caused by the highly susceptible QIs; (iii) it uses the information entropy concept to determine the uncertainty about SA categories' values present in equivalence classes, and considerable attention is paid to low or no uncertainty classes to overcome the explicit disclosure of private information about a set of individuals; (iv) it performs data generalization adaptively (i.e., generalizes susceptible QIs to higher level of QI's taxonomy in the equivalence classes having high privacy leakage score about users' community privacy and vice versa.) to effectively resolve privacy and utility trade-off; (v) it can be used for the anonymization of any dataset, balanced or imbalanced.

3 The proposed anonymization algorithm

The QI susceptibility- and SA entropy-based SN users data anonymization algorithm is necessary to account for the users' community privacy issues stemming from the highly

susceptible QIs and low entropy equivalence classes (ECs). This algorithm not only protects privacy of the users' community privacy, it enhances anonymous data utility for legitimate information consumers by controlling over-generalization of less susceptible QIs in EC having high entropy values (i.e., more uncertainty for the attacker). Several kinds of threats to the privacy of individuals associated with publishing databases of customers, patients, subscribers, consumers, and SN users exist. The consequences of these potential privacy threats from published data analysis including identity and private information disclosures of multiple users through community clustering, discrimination about the specific ethnic people if a group of users largely share the same religion/race values, loan declining to multiple people if the analysis of customers' previous ratings is not satisfactory, and political interference in an election from the spreading of false information or selectively withholding information etc. demand a theoretical and practical solution to thwart these threats. The novelty of our proposed anonymization algorithm lies in the ability to exploit the intrinsic characteristics of user's attributes that may reveal private information to the adversaries during published data analysis to effectively protect the privacy of multiple users, something that previously proposed anonymization algorithms based on similar principles could not do. This section presents the conceptual overview of the proposed anonymization algorithm and outlines its procedural steps. Figure 3 shows the conceptual overview of our proposed anonymization algorithm.

To anonymize any SN user's dataset D containing N users with multiple QIs and single SA, we introduced six principal concepts: (1) users attributes' classification; (2) the concept of QI's susceptibility; (3) ranking similar users, and formation of the ECs based on privacy parameter k ; (4) calculating the entropy of SA values in ECs; (5) determining crucial unique values only for the highly susceptible QIs; and (6) adaptive data generalization considering both the susceptibility of the QIs and entropy of the SA simultaneously. All concepts are complimentary, and result in the generation of anonymous data D' . The proposed algorithm involves fewer parameter's settings with the exception of step two where the RF model building and re-building requires alternate parameters to quantify the susceptibility weights of the QIs. In addition, the proposed algorithm does not enforce any hard constraints regarding SA's values in ECs during anonymization. Brief details of each concept along with equations and procedures are presented below.

3.1 Users attributes classification and users' data pre-processing

The original user's dataset D contains four types of attributes, explicit identifiers (EI), quasi identifiers (QIs), sensitive attribute (SA), and non-sensitive attributes (NSA). After getting the D as an input, the proposed algorithm classifies users' attributes into four types, and removes EI and NSA as a standard PDP practice. After the removal of EIs and NSA, D

Fig. 3 Conceptual overview of the proposed user's data anonymization algorithm

contains only two types of attributes, QIs and SA, represented as $D\{Q, S\}$. We use set $Q = \{q_1, q_2, \dots, q_p\}$ to denote the QIs present in D , where q_p is one type of QI such as race or sex. Set S represents the SA which can be of a single type (i.e., disease) or of multiple types (i.e., disease and salary) depending upon the scenario. In this paper, we consider the former case in which D contains a single SA with n distinct values, $S = \{s_1, s_2, \dots, s_n\}$. Hence, each user u_i has two types of the attributes in D , Q_{u_i} and S_{u_i} , respectively.

Each user in D referred as a tuple, denoted with t . For example, the Table given in Fig. 2a has six tuples, and each tuple has unique id. The original dataset about users can contain outliers, missing values for some QIs, and incomplete tuples. Therefore, the proposed algorithm pre-processes the user's data before feeding it into the anonymization process. It removes the outliers present in the data which are not feasible for the analysis. Furthermore, it eliminates records with unknown (i.e., missing) values from the data. In some cases, the data processing model needs information in a specified format. Therefore, the proposed algorithm formats the data and performs enrichment if required for data processing. With the help of data pre-processing, a cleaned dataset can be obtained which contains complete information about each user.

3.2 Quantifying the susceptibility weights of the quasi identifiers present in users' dataset

A number of studies have examined the QI's effect on user privacy [50, 51, 53, 56, 71]. Some studies reported the attacks that can be launched by targeting the QIs having more unique values and quantifying the significance of QIs. However, most of the existing studies do not consider the susceptibility of the QIs in terms of SA inferences to be made based on QI's values that may explicitly disclose private information about a set of users rather than an individual user from the released data. Hence, such susceptible QIs are risky for the users' community privacy disclosure rather than identity

or attribute disclosure of a single user. Susceptible QI refers to a user attribute having many similar values and it allows attacker to make user's community based on these values, which is subsequently used to infer the SA of the users associated with that community. In this paper, we identify the susceptible QIs from D , and quantify their weights using random forest (RF) [73] to reduce the unique identifications of multiple users. The steps used to quantify the susceptibility weights of QIs are depicted in Fig. 4. RF [73] is a well-known machine learning algorithm, and it yields superior accuracy values among comparable algorithms. It builds an ensemble of classification and regression trees (CARTs) from samples of the training data, and provides accuracy values by averaging all trees errors or votes.

It provides many valuable statistics about the data used in the underlying process. It has two main parameters: the number of trees (*ntree*) and the number of variables (i.e., QIs) required at the tree node split (*mtry*)—so it is very easy to adjust them. The rationale for using the RF method is its ability to consider the impact of each QI individually as well as in multivariate interactions with other QIs. To provide a deeper insight into the underlying QI's susceptibility weights computation process, we categorize the whole mechanism into five steps—the users data input, RF parameter (s) settings, building of the CARTs, accuracy values analysis, and data alteration and RF parameters tuning for the model re-building. The user's data D , where $D = \{N, Q, S\}$ is provided as an input to the RF. The set Q and S are used as predictors and target class, respectively. We divide the D into two partitions, two-thirds as training (e.g., the data on which the algorithm was trained) and one-third as testing (i.e., the data used for validation and testing purposes) data.

The RF parameters (*ntree* and *mtry*) were chosen considering the data size. RF takes the training data, *ntree* and *mtry*, formula (predictors and target class labels) and other required functions as an input, and starts building the CARTs t_b , with b

Fig. 4 Steps of determining the susceptibility weights of QIs in user's dataset

$\in \{1, \dots, ntree\}$. Later, prediction is carried out using the testing data and out-of-bag (oob) error $\tau^{(o)}$ is recorded. If the $\tau^{(o)}$ value is high, then RF parameter's values are modified to lower the $\tau^{(o)}$ value and to bring the accuracy Acc into an acceptable range. These values are then used as reference values in subsequent steps.

After obtaining the $\tau^{(o)}$ reference value, the values of each QIs are permuted column wise. By doing so, the QI x_{q_j} (a.k.a predictor variable) association with the target class (i.e., SA) Y is broken. Now, when the permuted x_{q_j} is used together with the non-permuted QIs to predict the Y , the prediction accuracy decreases significantly if x_{q_j} was strongly associated with the Y , and vice versa. Accordingly, the new out-of-bag (oob) $\tau^{(a)}$ error increases. Thus, the difference in $\tau^{(a)}$ (after permutation of x_{q_j} values) and $\tau^{(o)}$ (before permutation of x_{q_j} values), averaged over all trees $ntree$ can be used to quantify the susceptibility weight w_{q_j} of the j th QI. This implies that all those QIs which have less unique values while each unique value frequency is very high will have very little effect on the prediction accuracy. Hence, such attributes are highly susceptible for the community privacy. Let $\tau^{(b)}$ is the oob error for a tree b . Then the QI importance QII of the q_j in tree b can be computed using Eq. 1.

$$QII^b(q_j) = \frac{\sum_{i \in \tau^{(b)}} I(y_i = \hat{y}_i^{(b)})}{|\tau^{(b)}|} - \frac{\sum_{i \in \tau^{(b)}} I(y_i = \hat{y}_{i, \pi_j}^{(b)})}{|\tau^{(b)}|} \quad (1)$$

where $\hat{y}_i^{(b)}$ is the predicted SA value for i th observation before permutation and $\hat{y}_{i, \pi_j}^{(b)}$ is the predicted SA value for i th observation after q_j values permutation. The value V of the $QII(q_j)$ in b can be of following two types.

$$V(QII(q_j)) = \begin{cases} QII^b(q_j), & \text{if } q_j \in b. \\ 0, & \text{otherwise.} \end{cases} \quad (2)$$

The $V(QII(q_j))$ is zero when QI q_j is not in tree b or having all similar values in a column. The QII for each QI is then calculated as the mean (x_{q_j}) importance from all trees using Eq. 3.

$$x_{q_j} = \frac{\sum_{b=1}^{ntree} QII^b(q_j)}{ntree} \quad (3)$$

where x_{q_j} gives the mean score from all trees. The standard deviation s_{q_j} and susceptibility weight w_{q_j} can be computed using Eqs. 4 and 5, respectively.

$$s_{q_j} = \sqrt{\frac{1}{ntree-1} \sum_{b=1}^{ntree} (QII^b(q_j) - x_{q_j})^2} \quad (4)$$

$$w_{q_j} = \frac{x_{q_j}}{s_{q_j}} \quad (5)$$

Equation 5 gives the susceptibility weight w for the j th QI.

Through this process, we can compute the weights of all p QIs and can categorize the QIs into high, medium, and less susceptible QIs. Careful attention is paid to highly susceptible QIs because such QIs can enable unique identifications of multiple users more easily from ECs having low entropy values. We normalize the weight values to maintain the relative weight values' range between 0 and 100, and obtain weight set W of QIs.

3.3 Ranking similar users and formation of the equivalence classes based on privacy parameter k

To effectively preserve the privacy of users' community and minimize the loss of information in the anonymization process, each user in the EC must have nearly identical attributes. In this paper, we rank similar users based on QI values using the cosine similarity (Sim) measure. It is a very simple and reliable measure that ranges between $[0, 1]$. Sim value 1 means that two users are exactly similar, and 0 means that they are dissimilar. The similarity Sim value between two different users U_1 and V_1 can be calculated using Eq. 6.

$$Sim(U_1, V_1) = \frac{\sum_{i=1}^Q U_{1(i)} \times V_{1(i)}}{\sqrt{(\sum_{i=1}^Q U_{1(i)})^2} \times \sqrt{(\sum_{i=1}^Q V_{1(i)})^2}} \quad (6)$$

where i are the QIs of user U_1 and V_1 , and Q is the total of the QIs. With the help of Eq. 6, the similarity between all N users can be computed, and the resultant matrix M containing highly similar users is obtained. Later, the M is partitioned into set C of different EC, where $C = \{C_1, C_2, C_3, \dots, C_N\}$ based on the privacy parameter k . The value of k is chosen by the data owner, and it can be any whole number (i.e., $k > 1$). The number of ECs nc for N highly similar users can be determined using Eq. 7.

$$nc = \frac{N}{k} \quad (7)$$

The complete pseudo-code used to rank highly similar users based on QIs values, and formation of ECs using k value is presented in Algorithm 1.

Algorithm 1: Similar users' ranking and formation of the equivalence classes.

Input : (1) Users Dataset D containing N number of users. Each users u_i has Q QIs and one SA S , where $Q = \{q_1, q_2, q_3, \dots, q_p\}$ and $S = \{s_i\}$.
 (2) Privacy parameter k .

Output : Set C of equivalence classes

Procedure:

```

1  $N = |D|$ 
2 if  $(N \leq k)$  then
3   Return
4 else
5   Initialize, set  $C = \emptyset$ 
6   for  $i = 1$  to  $N$  do
7     for  $j = 1$  to  $N$  do
8       Compute  $Sim(i, j)$  value between users  $i$  and  $j$  based on
        QI values using Eq. 6
9       Add the computed  $Sim(i, j)$  values in Matrix  $R$ .
10      Repeat steps 8 and 9 for all subsequent values of  $i$  and  $j$ .
11    End for
12  End for
13  Sort  $R$  in descending order based on  $Sim$  values to find
    highly similar user matrix  $M$ .
14  Find the value of  $nc$  using Eq. 7
15  Generate equivalence classes,  $\{C_1, C_2, C_3, \dots, C_{nc}\}$  from
     $M$  with at least  $k$  user in each  $C_i$ .
16   $C = C \cup \{C_1, C_2, C_3, \dots, C_{nc}\}$ 
17  End if
18 end
19 return  $C$ 

```

In algorithm 1, users' dataset D and privacy parameter k are provided as an input. The set C of ECs are obtained as an output. The similarity values between users are computed and results are stored in a user matrix M (Lines 6-13). Lines 14-16 determine the number of ECs and generate the users ECs with at least k users in each class. Finally, set C of ECs is returned (Line 19). These ECs will be used for further processing before the generalization of QI's values.

3.4 Calculating entropy of the SA's values in equivalence classes

After the ECs have been generated, the users of each class share a different proportion of the SA's values. All users can share either similar or different SA's category values in the EC. To prevent the explicit disclosure of private information about a set of individuals, the SA values in ECs must be shared on average. Meanwhile, in most cases, the SA values distribution is not uniform and existing methods such as ℓ -diversity and p -sensitive k -anonymity etc. introduce constraints for the SA values in ECs. Meanwhile, making every EC ℓ -diverse or p -sensitive is very challenging especially in an imbalanced dataset. Furthermore, it will cause significant information losses by introducing constraints regarding SA values in ECs. In this paper, we compute the entropy of SA values in each EC to quantify the level of uncertainty for explicit disclosure and careful attention is paid to those ECs having low entropy values. The following procedure is used to calculate the entropy E of an i th EC C_i .

1. Identify the distinct SA category values, $S = \{s_1, s_2, \dots, s_n\}$ present in C_i .

2. Calculate the total occurrences (i.e., frequency) of each category value s in (C_i) , $F(S, C_i) = \{f_{s_1}, f_{s_2}, f_{s_3}, \dots, f_{s_n}\}$.
3. Calculate the proportion p of each SA's category value in an EC. The proportion (p_i) of i th SA value given its frequency f_{s_1} can be computed using Eq. 8.

$$p_i = \frac{f_{s_1}}{k} \quad (8)$$

4. The entropy E of an EC C_i can be computed using Eq. 9.

$$E(C_i) = - \sum_{i=1}^S p_i \log_2 p_i \quad (9)$$

The value of E ranges between 0 and 1, $E[0, 1]$. An E of 0 indicates that there is no uncertainty in the EC for the attackers regarding SA values because all users share the same SA value. In contrast, an E value of 1 means that there is enough uncertainty for the attackers, and SA values are distributed on average among users. With an E value of 0, the SA disclosure is easier, and the sensitive user information is easily seen. With an E value of 1, this is ideal for protecting the privacy of the users. In our work, we take into account the E values during data anonymization process to effectively protect the privacy of the users' community and anonymous data utility, respectively.

3.5 Determining the crucial unique values for highly susceptible QIs in equivalence classes

With the help of the RF method, we can identify the highly susceptible QIs and can determine their susceptibility weights. Meanwhile, these weights represent the vulnerability of the QIs in relation with whole dataset. However, a very highly susceptible QI can be less vulnerable to the privacy concerns related to a users' community if it contains large number of unique values in an EC. Therefore, in order to minimize the unwanted data disclosures and to control unnecessary QIs generalization, the crucial unique values that can allow an adversary/analyst to jeopardize the privacy of the users' community are determined by analysing the frequency of those unique values. The crucial unique values are the particular QI values with high frequency, and which dominate the prediction results on SA, which can significantly impact the users' community privacy from the published data. According to the authors in [71], the number of crucial unique values r for each QI in EC can be calculated using Eq. 10.

$$r = \text{round}(\log_2 x_{q_i}) \quad (10)$$

where x_{q_i} represents the number of unique values for i th QI present in an EC. If the number of unique values are lower for

a highly susceptible QI, then the QI can reveal more information about the individuals identity. We employed a five-step process to determine the crucial unique values for each highly susceptible QI in the ECs. The steps are: identifying the unique values for each highly susceptible QI, computing the frequency of each unique value, sorting the unique values based on the frequency in ascending order, identifying the crucial unique values having high frequency value, and computing the proportion β^c of most crucial unique value c . With the help of β^c , we can determine the susceptibility score SS of the QI in EC. The SS value of an i th QI in C_i can be computed using Eq. 11.

$$SS(q_i, C_i) = w_{q_i} \times \beta_{q_i}^c \quad (11)$$

where w_{q_i} represents the susceptibility weight of the i th QI, and $\beta_{q_i}^c$ denotes the proportion of the most crucial unique value c of a particular QI. The $SS(q_i, C_i)$ gives the susceptibility score of the i th QI in C_i . The main reason to take the maximum value of proportion in the SS computation process is to handle the worst cases effectively. We can compute the SS of each highly susceptible QI using Eq. 11. The SS of the i th EC can be computed by adding all susceptible QI's SS using Eq. 12.

$$SS(C_i) = \sum_{i=1}^Q SS(q_i, C_i) = \sum_{i=1}^Q (w_{q_i} \times \beta_{q_i}^c) \quad (12)$$

where w and β are the susceptibility weights and proportion of the most crucial unique value c belonging i th QI, and Q is the total of susceptible QIs. The $SS(C_i)$ denotes the score of the C_i containing at least k users. We use both $SS(C_i)$ and $E(C_i)$ values of each EC to determine the best generalization level from the QI's generalization taxonomy in the generalization process.

3.6 Adaptive data generalization considering both the susceptibility of the QIs and entropy of the SA

Adaptive generalization selects the generalization level for each QI based on its susceptibility weight, crucial unique values' proportion, and the corresponding class's entropy value. Data generalization is used to replace the original QI values with less specific but semantically consistent values, and it is done by means of each QI's taxonomy T . The T of QI can be constructed by analysing the domain values of the QIs or it can be application specific. The levels of the T can be classified into three categories, lower, intermediate, and higher levels. Each level has a unique and specific impact on privacy and utility. A

pictorial overview of the value T of a QI zip code with relevant details is shown in Fig. 5.

The privacy leakage risk is higher at lower levels of T and the utility is maximum. In contrast, there is no privacy risk at higher levels of T but the utility is very low. Hence, there exists a trade-off between privacy and utility which can be exploited by designing an adaptive data generalization algorithm that integrates the susceptibility weight of each QI and entropy of SA to address the privacy issues, while enhancing the anonymous data utility. The key principle of adaptive generalization is to retain the semantics of the original data as much as possible by utilizing the entropy and susceptibility statistics jointly to effectively preserve both users' community privacy and anonymous data utility. For example, the higher levels of generalization are preferred when there exists a risk that would uniquely identify a multiple of users and their SA's disclosure.

We consider both SS and E values, and compare them with the relevant thresholds (T_s and T_e) to decide the level of generalization for QI values in each EC. We used the T_s value of 0.75 and T_e value of 0.65 as threshold to determine the best generalization level during data sanitization. However, both these values can be adjusted according to the protection level the data owner wants, the target data analyzers, and the objectives for publishing the data. For example, if a bank wants to release its customer's data to a data mining firm for classification analysis on loan return rating, but the bank does not want the firm to infer the rating 'good' no more than 75% based on bankruptcy state 'Discharged' using the attributes of job and country. So, in this case, the T_s and T_e values will be adjusted as per the data owner's privacy requirement and any attributes' combination that violates the privacy requirement must be protected.

The complete pseudo-code used to perform adaptive generalization is listed in Algorithm 2 where set C of ECs, QIs' generalization taxonomies, and QI's susceptibility weights set W are provided as an input. The anonymized dataset D' is obtained as an output. Lines 2-5 implement the SS and E values computation of each EC C_i (where every C_i contains at least k users, and each user U_i has Q QIs and single SA.) and comparison with the relevant thresholds (T_s and T_e), respectively. Lines 6 – 11 perform the lower levels generalization for the classes having low SS and high E values, and Lines 13 – 18 implement the higher/intermediate-level generalization for low entropy ECs. Lines 6 – 11 and 13 – 18 are the two blocks of *if-else* that separates the ECs based on SS and E statistics to meet the stated privacy requirements and to control over-generalization of users' QIs. Finally, anonymous data D' is returned by combining both classes of anonymity (Lines 21 and 22).

Algorithm 2: Adaptive data generalization algorithm to anonymize user's data.

Input : (1) Set C of ECs where $C = \{C_1, C_2, \dots, C_n\}$ and each EC contains at least k users.
 (2) QI's generalization taxonomies
 $T_{q_1}, T_{q_2}, \dots, T_{q_p}$
 (3) QI's Susceptibility weights set W where
 $W = \{w_{q_1}, w_{q_2}, \dots, w_{q_p}\}$

Output : Anonymous dataset D'

Procedure:

```

1 for each equivalence class  $C_i \in C$ , where  $i = 1, 2, 3, \dots, n$  do
2   Compute  $SS$  value of the  $C_i$  using Eq. 11 and 12.
3   Find  $E$  value of the  $C_i$  using Eq. 8 and 9.
4   Compare obtained  $SS(C_i)$  and  $E(C_i)$  values with  $T_s$  and  $T_e$ .
5   if  $(SS(C_i) < T_s) \wedge ((1 - E(C_i)) < T_e)$  then
6     for each QI  $j \in C_i$ , where  $j = 1, 2, 3, \dots, p$  do
7       Acquire generalization taxonomy  $T_j$  of  $q_j$ .
8       Perform adaptive generalization ( $d'$ ), Preferable at lower levels of  $T_j$ .
9       Repeat:  $\forall$  QIs,  $q_1, q_2, q_3, \dots, q_p \in C_i$ 
10      return  $C'_i$ 
11    End for
12   else
13     for each QI  $j \in C_i$ , where  $j = 1, 2, 3, \dots, p$  do
14       Acquire generalization taxonomy  $T_j$  of  $q_j$ .
15       Perform adaptive generalization ( $d''$ ), Preferable at higher or intermediate levels of  $T_j$ .
16       Repeat:  $\forall$  QIs,  $q_1, q_2, q_3, \dots, q_p \in C_i$ .
17      return  $C''_i$ 
18    End for
19   end
20 End for
21  $D' = \text{combine } (C'_1 + C'_2 + C'_3 + \dots + C'_n \text{ and } C''_1 + C''_2 + C''_3 + \dots + C''_n)$ 
22 return  $D'$ 

```

4 Simulation results and discussion

This section presents the simulation results and key findings from the proposed concept. The improvements of the proposed algorithm were compared using two criteria; the improvements in privacy protection of users' community, and anonymous data utility with the existing closely related algorithms. To benchmark the proposed algorithm, we compared the proposed algorithm results with the PK -anonymity

algorithm [70], CPA algorithm [71], and the WFDA algorithm [72]. The simulation results were generated and compared on a PC running Windows 10, with a CPU of 2.6 GHz and 8.00 GB of RAM. The simulations were carried out with the help of two software MATLAB (version 9.4.0.81 (R2018a)) and R-tool (version 3.6.1). In simulations of the proposed PPDP algorithm, we consider a relational D having multiple QIs and a single SA about users. We present the detailed overview of the five datasets that were used in the experiments in Table 1. The first three datasets are publicly available [74, 75], and the last two were created synthetically by respecting the attributes values distribution and percentage in real SN [76]. We assigned two SA (political views and online disease community affiliation) with distinct values to synthetically created datasets.

To verify the proposed algorithm effectiveness, we created two replicas (i.e., R_1, R_2) of adult's dataset [74] considering salary and occupation as SA, respectively. Adults dataset has become a benchmark dataset for analysing the feasibility of k -anonymity based algorithms for PPDP. Table 2 shows the susceptibility weight values of different QIs present in users' datasets listed in Table 1. The symbol '-' indicates that respective QI is not a part of that dataset. These measures were calculated with RF through a series of experiments. We selected the best combinations of the RF parameters while computing the QI's susceptibility weights for each dataset. The following values of each parameters were chosen in order to determine the susceptibility weights of the QIs from Adults (R_1) dataset: Number of trees ($ntree$) = 495, QIs used to split the classification tree node ($mtry$) = 4, RF model type = classification, keep. forest = true, variable importance = true, data = users - data, predictors = (age, gender, race, country) and target = salary.

In adult datasets (R_1, R_2), most records have the U.S.A. listed as the country value and the remaining records list many other countries that appear only when k is small. Therefore, the susceptibility weight of the country is very high. In addition, all those QIs whose values are not concentrated in a certain value have lower weights, respectively. Furthermore, we verified these weights by analyzing each QI values distribution in the original datasets.

Fig. 5 Overview of the value generalization taxonomy of the QI zip code

Table 1 Description about the datasets used in simulations

Datasets	No. of users (N)	No. of QIs	QIs names (Distinct values, Taxonomy tree levels)	SA name (Distinct values)
Adults (R_1)	32,561	4	Age (74,7), Gender (2,2), Race (5,3), Country (41,4)	Salary (2)
Adults (R_2)	32,561	4	Age (74,7), Gender (2,2), Race (5,3), Country (41,4)	Occupation (14)
Bkseq-data	16,160	3	Age (30, 5), Gender (2,2), Weight (30, 3)	Result of Medical Exam (19)
Synthetic Dataset 1	5196	6	Age (60,4), Gender (2,2), Race (5,3), Zipcode (10, 4), Job (7, 3), DoB year (86,5)	Political views (3)
Synthetic Dataset 2	10,000	5	Age (30,3), Gender (2,2), Zipcode (2255, 5), Work Class (8,3), Martial Status (7,3)	Online Disease community Affiliation (5)

4.1 Improvements in users' community privacy protection

We first show the improvements in users' community privacy protection provided by the proposed algorithm. The overall response to this metric is good, and the anonymous data produced by the algorithm is more resilient toward SA inference (SAI) breach as compared to existing algorithms. Multiple users' identity revelation and their associated SAI are reduced via the adaptive generalization, which considers the entropy of the ECs and susceptibility weight of each QI simultaneously. We select the table with six records shown in Fig. 2a to highlight the difference in anonymization between proposed and existing works. In this small subset of data originally extracted from the adult's dataset with political views as SA, the country was determined as highly susceptible for users' community privacy compared to the age because most of the tuples belong to the same location and could leak the SA of a set of users explicitly. Country and age have relative susceptibility weight values of 78.01 and 1.79, respectively. These findings confirm that $q_{country}$ needs considerable attention during anonymization due to very high susceptibility weight. Three ECs— C_1 , C_2 , and C_3 are considered, as mentioned in Fig. 2b. Determining EC entropy is equally important because it limits explicit disclosure of SAs about a set of individuals. C_1 has an entropy of 1, which is higher than the threshold, but C_2 and C_3 have 0 entropy.

This indicates that country can assist attackers in deriving SAs, which is helpful information when anonymizing data. This study performs higher levels generalization rather than

suppression to yield superior data utility during published data analysis. D' produced by the proposed algorithm using the original data from Fig. 2a is presented in Fig. 6, and three-scenarios were chosen to show the improvements from community's privacy protection point of view. In this paper, we employed the PS rules-based SAIs, and probabilistic disclosures of the identities and SAs of an unknown users' community to quantify the privacy of anonymization after applying the proposed algorithm.

The first two scenarios show the percentage of SAI that can be revealed using PS rules to invade privacy of users in an online or offline community from the anonymous data D' . The third scenario is about the prediction of an unknown community's SA based on the information gained from the D' . The PS rule antecedents are QI's values and consequent is SA value, respectively. It models the association between QIs and SA as, $((q_{1_v}, q_{2_v}, q_{3_v}) \rightarrow s_1)$. The analysts/attackers can construct such PS rules by analyzing the QI's domain values or based on the auxiliary information obtained from other sources to link multiple users' identities. The very high percentage of SAI value of the existing algorithms put the privacy of many specific communities at risk because many such PS rules can easily be derived by the data mining firms or attackers. The ECs created by using large values of k with low entropy values are much more vulnerable to SAIs extraction. A possible third scenario is assuming that users in a community σ are likely to share some common SAs. The SA values prediction of a user or multiple users of a community creates unanticipated privacy breaches.

Table 2 Susceptibility weight values of the QIs present in the five users' datasets

Sr. no.	Datasets	Quasi identifiers									
		Age	Gender	Race	Country	Weight	Zip code	Job	DoB year	Work class	Martial status
1	Adults (R_1)	1.75	1.2	19.04	78.01	—	—	—	—	—	—
2	Adults (R_2)	2.69	2.77	26.35	68.19	—	—	—	—	—	—
3	Bkseq-data	42.1	37.8	—	—	20.1	—	—	—	—	—
4	Synthetic Dataset 1	2.98	32.89	28.91	—	—	30.08	3.43	1.71	—	—
5	Synthetic Dataset 2	63.94	0.97	—	—	—	34.05	—	—	0.86	0.18

(a) Anonymous data produced by WFDA algorithm ($k = 2$)			
ECs	Age	Country	Sensitive Attribute Political views
C_1	35-37	North America	Liberal
	35-37	North America	Conservative
C_2	38-40	Central America	Liberal
	38-40	Central America	Liberal
C_3	35-37	North America	Conservative
	35-37	North America	Conservative
Scenario 1	PS-Rule: {35, Canada} → Conservative SAI: (Pol. Views are Conservative)= 75%		
Scenario 2	PS-Rule: {39, Belize} → Liberal SAI: (Pol. Views are Liberal)= 100%		
Scenario 3	σ : $N=100$, Age=40, Country= Honduras $P(\text{People having Pol. Views Liberal})= 1$		

(b) Anonymous data produced by Proposed Algorithm ($k = 2$)			
ECs	Age	Country	Sensitive Attribute Political views
C_1	35-40	North America	Liberal
	35-40	North America	Conservative
C_2	35-40	America	Liberal
	35-40	America	Liberal
C_3	35-40	America	Conservative
	35-40	America	Conservative
Scenario 1	PS-Rule: {35, Canada} → Conservative SAI: (Pol. Views are Conservative)= 50%		
Scenario 2	PS-Rule: {39, Belize} → Liberal SAI: (Pol. Views are Liberal)= 50%		
Scenario 3	σ : $N=100$, Age=40, Country= Honduras $P(\text{People having Pol. Views Liberal})= 0.5$		

Fig. 6 Comparison of privacy protection between the WFDA algorithm and the proposed algorithm

If the SA's values about multiple users are revealed to have a very high probability P from D' , then some unknown users' SAs could also be accurately predicted. Consequently, implicit disclosure of SA of a specific users' community can occur based on the knowledge gained from the D' . Therefore, the D' must be resilient to both these attacks to protect the privacy of users' in an online/offline community. The D' produced by our algorithm shows a 31% improvements in terms of the multiple users' unique identifications and SAIs protection, and the probability of SA's prediction about an unknown σ is also close to the threshold (e.g., $1/k$) as of standard k -anonymity privacy model.

We applied various PS rules on each dataset's anonymous versions that were produced with different k values to perform quantitative analysis on the privacy protection. We selected highly susceptible QI's distinct values of high frequency as antecedents, and SA category value that is shared by most users in dataset as consequents of PS rules. We used k -anonymity [10], ℓ -diversity [11], and t -closeness [12] privacy models as the baseline in the experiments. The values of ℓ and t were chosen by considering the distinct values of SA, and SA values' distribution in each D , respectively. Table 3 presents the privacy protection results obtained by performing PS rules based analysis on each version of anonymized dataset. As shown in Table 3, the k -anonymity privacy model [10] has higher SAIs values compared to the other methods because it does not consider the diversity and distributions of SA's values in ECs. In contrast, ℓ -diversity model [11] considers the diversity of SA's values to resolve the k -anonymity limitations, but it does not consider the distribution of SA's values in ECs. Hence, it explicitly discloses private information of users from the ECs having skewed distribution of SA's values. The t -closeness model [12] resolves the limitations of k -anonymity and ℓ -diversity models by considering the SA's values distribution, but it is prone to higher SAIs when a specific SA's value exists with high frequency in a dataset. In addition, the t -closeness model is prone to identity disclosures

and it may even provide worse privacy protection than ℓ -diversity in some cases [77, 78]. The proposed algorithm, on average, has lower SAIs values compared to the existing baseline privacy models in most cases. Hence, for the users' community privacy protection, the proposed algorithm proves to be the best algorithm among those compared.

4.2 Improvements in Anonymous Data Utility

During the anonymization of any dataset, information losses (IL) are inevitable. When data is sanitized, some specific values of QIs are always lost. On the other hand, data analysis applications want that anonymization algorithm must retain the data as original as possible to enhance anonymous data utility. However, this can only be achieved when the data owner is fully aware of the original data statistics (i.e., QIs having less susceptibility and ECs with sufficient entropy). The proposed algorithm provides such valuable statistics to the data owner prior to data sanitization. By making use of these statistics, over-generalizations can be significantly controlled to reduce IL, and the anonymous data preserve better utility compared to the existing PPDP algorithms. To evaluate anonymous utility, two criteria are used—information loss and classification/regression model accuracy. IL were calculated using two metrics—distortion measures (DM) and coverage of generalized values. Fung et al. [79] explained the comprehensive details about both these IL metrics. Both metrics' values are computed by assessing the taxonomy levels on which QIs' values are generalized during QI values generalization process. The value of DM can be calculated using Eq. 13.

$$DM = \sum_{n=1}^N \sum_{q=1}^Q \frac{l_i}{l_t} \times w_q \quad (13)$$

where l_i is the level in the generalization taxonomy for which QI's value is generalized, and l_t represents the total number of

Table 3 Comparison of the SAIs values of the proposed algorithm with the existing privacy models

Datasets (Distinct values of SA)	PPDP models/algorithm	Privacy parameter values (i.e., k values)								
		2	5	10	15	20	25	30	35	40
Adults (R_1) Dataset (2)	k -anonymity Model	0.91	0.80	0.84	0.93	0.82	0.88	0.85	0.92	0.94
	ℓ -diversity Model	0.83	0.73	0.76	0.78	0.71	0.81	0.78	0.87	0.88
	t -closeness Model	0.77	0.60	0.63	0.74	0.69	0.80	0.77	0.80	0.81
	Proposed Algorithm	0.55	0.41	0.52	0.66	0.63	0.71	0.68	0.74	0.76
Adults (R_2) Dataset (14)	k -anonymity Model	0.81	0.69	0.63	0.77	0.81	0.84	0.75	0.85	0.87
	ℓ -diversity Model	0.76	0.60	0.53	0.71	0.73	0.78	0.72	0.77	0.81
	t -closeness Model	0.66	0.46	0.43	0.59	0.60	0.72	0.68	0.76	0.74
	Proposed Algorithm	0.45	0.33	0.39	0.51	0.53	0.65	0.59	0.69	0.66
Bkseq Dataset (19)	k -anonymity Model	0.70	0.66	0.73	0.75	0.78	0.80	0.72	0.75	0.78
	ℓ -diversity Model	0.66	0.60	0.63	0.68	0.71	0.73	0.69	0.71	0.73
	t -closeness Model	0.50	0.34	0.39	0.53	0.57	0.65	0.63	0.67	0.69
	Proposed Algorithm	0.50	0.26	0.31	0.44	0.48	0.54	0.51	0.61	0.63
Synthetic Dataset 1 (3)	k -anonymity Model	0.87	0.73	0.76	0.84	0.80	0.85	0.83	0.89	0.91
	ℓ -diversity Model	0.85	0.61	0.73	0.76	0.73	0.77	0.76	0.83	0.85
	t -closeness Model	0.83	0.53	0.57	0.70	0.65	0.76	0.75	0.78	0.79
	Proposed Algorithm	0.52	0.38	0.44	0.57	0.56	0.68	0.65	0.72	0.73
Synthetic Dataset 2 (5)	k -anonymity Model	0.75	0.69	0.71	0.74	0.76	0.81	0.82	0.87	0.89
	ℓ -diversity Model	0.60	0.57	0.63	0.66	0.69	0.79	0.74	0.81	0.84
	t -closeness Model	0.55	0.47	0.54	0.61	0.63	0.72	0.73	0.75	0.76
	Proposed Algorithm	0.51	0.35	0.39	0.53	0.55	0.62	0.64	0.67	0.69

levels in a taxonomy of QI. w_q is the susceptibility weight of a QI.

All distortion values from each QI and tuple are summed for all users in the dataset. If the value of a particular QI is retained as an original in D' , the DM will be zero. The IL caused by generalizing a specific value v_s of a particular QI q_j to a general value v_g is computed as: $IL(v_g) = \frac{|v_g|-1}{|D_{q_j}|}$ where $|v_g|$ is the number of the domain values which are descendants of v_g in T_{q_j} and $|D_{q_j}|$ denotes the number of domain values of the j th QI. The value of v_g can be equal to v_s in some ECs. Hence, the IL value of $IL(v_g)$ can be of following two types depending upon the generalization of v_s in an EC.

$$IL(v_g) = \begin{cases} IL(v_g), & \text{if } v_g \neq v_s. \\ 0, & \text{otherwise.} \end{cases} \quad (14)$$

$IL(v_g) = 0$ only if no modification was applied to the v_s during data generalization. The IL caused by the complete tuple t values generalization can be calculated using Eq. 15.

$$IL(t^*) = \sum_{q \in Q} (IL(t^*, q_1) \times w_{q_1}) \quad (15)$$

where $IL(t^*, q_1)$ is the IL caused by the coverage of the generalized values of a single QI, Q represents total of the QIs, and w is the susceptibility weight of a specific QI. One more

summation is used to estimate the total value of this metric for whole D' .

Table 4 shows the IL comparison of the proposed algorithm with the existing algorithms using eight different values of k and five datasets (listed in Table 2). These results are the averaged results obtained from the ECs. The rise in k value causes an increase in IL for both the metrics due to increase in the records in each EC, and their corresponding generalized values. The proposed algorithm IL values are lower compared to existing algorithms in most cases because the proposed algorithm slightly distorts the original data. The rationale regarding better utility is derived from the fact that entropy and susceptibility are considered simultaneously during data sanitization. The proposed algorithm generalizes the values only for the highly susceptible QIs in ECs having less E values. All these values were computed by analysing the v_g of all QIs present in D' . From Table 4, it can be observed that the IL of the proposed algorithm is lower compared to the existing algorithms for the most values of k . The proposed algorithm IL values are higher for a smaller k because most ECs have less uncertainty for the analysts (i.e., low E value) and degree of generalization is relatively higher to effectively protect users' community privacy. However, with increasing k values, the proposed algorithm IL value in ECs is lower compared to the existing algorithms. The proposed algorithm, on average, lowers the IL by 22.95% compared to the existing closely related algorithms. The average IL

Table 4 Information losses: proposed algorithm results comparison with three existing PPDP algorithms

PPDP algorithms	Avg. IL values computed from the DM metric								Avg. IL values computed from the coverage of generalized values' metric							
	5	10	15	20	25	30	35	40	5	10	15	20	25	30	35	40
PK Algorithm	0.029	0.031	0.037	0.039	0.041	0.045	0.063	0.091	7.31	13.41	18.13	23.24	31.23	39.34	48.43	60.25
WFDA Algorithm	0.019	0.026	0.044	0.061	0.064	0.069	0.082	0.15	2.25	7.47	19.32	27.98	31.8	43.75	51.5	65.75
CPA Algorithm	0.018	0.028	0.036	0.048	0.052	0.061	0.078	0.098	2.05	5.75	14.95	26.25	33.75	40.35	51.5	67.24
Proposed Algorithm	0.021	0.027	0.032	0.035	0.037	0.039	0.042	0.047	2.75	4.27	6.19	10.69	18.5	24.25	30.45	34.75
PK Algorithm	0.02	0.025	0.032	0.037	0.04	0.048	0.066	0.088	5.31	10.31	14.23	21.14	30.03	36.54	45.13	56.35
WFDA Algorithm	0.011	0.020	0.039	0.059	0.063	0.071	0.085	0.12	0.25	4.45	15.92	25.08	30.18	40.95	49.75	61.15
CPA Algorithm	0.009	0.022	0.031	0.046	0.051	0.064	0.081	0.095	0.51	2.95	10.85	23.15	32.95	38.5	48.15	63.54
Proposed Algorithm	0.012	0.019	0.027	0.033	0.036	0.043	0.044	0.046	0.45	1.27	4.19	8.99	17.75	21.35	27.5	31.35
PK Algorithm	0.011	0.023	0.027	0.033	0.047	0.058	0.071	0.089	1.31	4.69	8.33	13.54	18.34	22.59	25.53	29.45
WFDA Algorithm	0.008	0.026	0.031	0.039	0.048	0.059	0.072	0.91	0.36	3.45	9.22	16.98	18.18	21.54	24.95	30.15
CPA Algorithm	0.011	0.021	0.025	0.031	0.049	0.058	0.074	0.092	0.43	3.05	8.85	15.25	19.95	23.52	24.15	31.34
Proposed Algorithm	0.007	0.015	0.019	0.024	0.037	0.045	0.058	0.079	0.37	2.27	6.39	7.95	11.75	15.95	21.58	27.45
PK Algorithm	1.39	3.95	5.5	6.98	12.05	16.85	23.25	28.89	7.61	16.51	19.13	24.75	34.73	38.45	49.43	62.25
WFDA Algorithm	1.42	5.12	5.75	7.35	13.1	19.25	27.15	34.75	7.95	13.97	20.52	28.98	39.85	45.75	53.75	68.95
CPA Algorithm	1.38	4.16	5.69	7.25	12.25	17.12	24.81	31.57	7.85	12.75	19.95	25.51	35.57	41.55	52.57	70.84
Proposed Algorithm	1.41	3.58	4.98	6.07	11.87	15.51	21.55	26.57	7.75	10.21	16.82	21.54	29.75	36.75	45.95	53.75
PK Algorithm	0.014	0.023	0.035	0.039	0.041	0.044	0.049	0.058	3.11	5.31	10.13	15.24	24.33	29.44	33.53	40.57
WFDA Algorithm	0.016	0.025	0.039	0.051	0.059	0.067	0.072	0.092	5.75	9.46	14.42	19.48	27.88	32.75	37.75	44.76
CPA Algorithm	0.016	0.027	0.037	0.045	0.051	0.059	0.069	0.078	3.65	6.75	10.49	16.95	25.56	31.35	36.65	41.21
Proposed Algorithm	0.012	0.019	0.024	0.026	0.029	0.034	0.038	0.045	2.25	3.21	5.95	9.69	17.55	23.75	29.55	35.52

values calculated from all records of each dataset are shown in the Fig. 7a, b. From the results it can be observed that the proposed algorithm has lower ILs compared to the existing algorithms in all five datasets, and both IL metrics, respectively. Generally, a dataset having more number of tuples lead to the high ILs. This confirms to the theoretical analysis that a dataset

containing a high number of records is more prone to changes in the data semantics and therefore utility decreases. Meanwhile, the proposed algorithm incurs less ILs compared to the existing algorithms on all five datasets because it performs lower level generalization of less susceptible QIs in ECs having high entropy values.

(a) IL computed from the DM metric.

(b) IL computed from the coverage of generalized values.

Fig. 7 Comparison of IL of the proposed algorithm with the existing algorithms on three real-world and two synthetic SN users' datasets

Table 5 Information losses: proposed algorithm results comparison with the existing privacy models

Datasets	k values	Avg. IL values computed from the DM metric			
		k -anonymity	ℓ -diversity	t -closeness	Proposed algorithm
Adults R_1	2	0.011	0.024	0.028	0.016
	5	0.019	0.025	0.032	0.021
	10	0.029	0.031	0.041	0.026
	20	0.047	0.059	0.064	0.035
	30	0.061	0.071	0.081	0.040
	40	0.091	0.125	0.161	0.048
Synthetic 1	2	0.64	0.94	1.07	0.76
	5	1.38	1.74	2.07	1.41
	10	4.26	5.18	6.39	3.58
	20	8.28	9.76	11.56	6.07
	30	18.33	20.41	23.51	15.57
	40	31.91	34.93	41.29	26.57

Furthermore, experiments were performed to compare the IL of the proposed algorithm with three existing privacy models k -anonymity, ℓ -diversity, and t -closeness. We used both real-world and synthetic datasets to compare the proposed algorithm performance. Results are shown in Tables 5 and 6. From the results, it can be seen that as k value is increased, the IL values of both metrics also increase. The proposed algorithm yields less IL compared to the existing privacy models in most cases by controlling unnecessary generalization of less susceptible QI in ECs of high entropy. In contrast, the existing privacy models do not apply the concept of susceptibility and entropy to control unnecessary generalization, and thus produce a relatively higher IL's values. Moreover, when the k values are small, there is an inadequate heterogeneity in the SA's values in most of the ECs, and proposed algorithm applies higher level generalization to protect users' privacy.

Therefore, its IL's values are marginally higher than k -anonymity privacy model. But the proposed algorithm still achieves better utility results compared to the existing privacy models in most cases.

Accuracy has been widely used to evaluate the anonymous data quality for data mining tasks. Generally, the accuracy values close to the original data are preferred for the informative data analysis. For achieving better accuracy domain consistency is important while generalizing QI values. Accuracy values can be computed using machine learning methods (i.e., decision tree, RF, and SVM etc.) with the help of Eq. (16).

$$Accuracy(Acc.) = \frac{T_p + T_n}{N} \quad (16)$$

where T_p is the number of true positives, T_n is the number of true negatives, and N is the total number of users in the D .

Table 6 Information losses: proposed algorithm results comparison with the existing privacy models

Datasets	k values	Avg. IL values computed from the coverage of generalized values			
		k -anonymity	ℓ -diversity	t -closeness	Proposed algorithm
Adults R_1	2	0.87	1.49	2.91	1.11
	5	2.81	3.85	5.31	2.75
	10	6.45	9.42	12.43	4.27
	20	16.25	21.98	25.34	10.69
	30	37.39	49.70	57.77	24.25
	40	52.94	63.75	72.79	34.75
Synthetic 1	2	2.98	3.92	4.49	3.19
	5	7.93	9.25	10.61	7.75
	10	12.95	15.97	18.61	10.21
	20	29.91	36.98	44.95	21.54
	30	53.95	60.05	75.15	36.75
	40	71.94	85.25	89.25	53.75

Figure 8a–c show the accuracy values of the proposed algorithm, and its comparison with the existing methods on three real-world datasets. Figure 9a–c shows the accuracy values obtained from two synthetic datasets and the averaged accuracy obtained from all datasets during the simulations, respectively. It can be observed from Figs. 8 and 9 that a rise in k value causes a nincrease in accuracy for each dataset due to less modifications in the QI's original values. The rationale regarding the better accuracy of the proposed algorithm compared to the existing methods is derived from the fact that domain consistency is maintained in the data anonymization process. The proposed algorithm yields superior accuracy values by generating the consistent levels of the generalization and original data statistics. All these values were obtained using the decision trees with R- programming (*R-tool*) scripts. Through extensive simulations and comparison with three existing algorithms, on average, our algorithm improves accuracy by 6.5% on both replicas (R_1 , R_2) of the adult's dataset, 8.05% on the Bkseq-dataset, and 6.25% on the synthetic datasets. Furthermore, the proposed algorithm accuracy values are only marginally lower than the original dataset. To further demonstrate the effectiveness of the proposed algorithm, accuracy values have been measured and compared with the result of three existing privacy models. Results are shown in Table 7. The proposed anonymization algorithm has produced higher accuracy values from the anonymous data than other privacy models. The reason is that proposed algorithm applies the concept of similarities in the ECs formation, and maintains the QI's values domain consistency during the data generalization process to improve accuracy.

The simulation results obtained from five different datasets emphasize the validity of the proposed algorithm with respect to achieving better users' community privacy protection and improved anonymous data utility. It effectively resolves the privacy and utility trade-off in PPDP. The proposed algorithm is an offline approach for user's data anonymization. The time complexity of the proposed algorithm depends on the number of users N , the number of QIs, the distinct values of SA, and the privacy parameter k . The complexity of algorithm 1 lies in step 6 to step 12. The 'for' loops at Step 6 and Step 7 iterate $O(n)$ times. There are $O(\lambda)$ instructions to be executed in each iteration inside the inner 'for' loop. Meanwhile, λ has a constant upper bound. Thus the overall complexity of Algorithm 1 is $O(n^2)$. Algorithm 2 complexity lies in Step 1 to Step 20. The 'for' loop at Step 1 iterates $O(n)$ times, and in each iteration, $O(\chi)$ instructions are executed. Hence, the overall complexity of algorithm 2 is $O(n)$ since χ has a constant upper bound. Meanwhile, the execution speed of the proposed mechanism is significantly reduced when we provide pre-computed ECs containing similar users, and SS and E scores of each EC to the data generalization algorithm (i.e., algorithm 2). In addition, due to lessing of the parameter's setting and by not forcing any hard constraints as stated in

Fig. 8 Comparison of accuracy of the proposed algorithm with the existing algorithms on real-world datasets

Sec.3, the proposed mechanism requires less time and space complexity. The overall time complexity of the proposed mechanism is $O(n^2)$ with pre-computed values of susceptibility weights of QIs, ECs containing similar users, and SS and E scores of ECs.

The proposed algorithm performs well with respect to protecting the privacy of users' community and anonymous data utility for two reasons: (1) the susceptibility of QIs is

(a) Synthetic Dataset 1

(b) Synthetic Dataset 2

(c) All datasets (overall accuracy analysis)

Fig. 9 Comparison of accuracy of the proposed algorithm with the existing algorithms on two synthetic datasets and overall comparison

introduced, which helps to treat QIs according to their impact on users' community privacy and it helps to effectively protect users' community privacy; and (2) the adaptive data generalization, which considers both the susceptibility of the QIs and entropy of the SAs simultaneously, improves the anonymous data utility by controlling over-generalization of QIs in ECs having high E value. It can handle both types (i.e.,

Table 7 Accuracies: proposed algorithm versus three privacy models

Datasets	PPDP methods	Avg. accuracy (%)
Adults R_1	k -anonymity Model	85.36
	ℓ -diversity Model	84.70
	t -closeness Model	83.17
	Proposed Algorithm	87.78
Adults R_2	k -anonymity Model	81.34
	ℓ -diversity Model	80.29
	t -closeness Model	79.13
	Proposed Algorithm	83.28
Bkseq	k -anonymity Model	88.06
	ℓ -diversity Model	86.61
	t -closeness Model	85.37
	Proposed Algorithm	91.14
Synthetic 1	k -anonymity Model	86.08
	ℓ -diversity Model	83.83
	t -closeness Model	82.99
	Proposed Algorithm	88.63
Synthetic 2	k -anonymity Model	83.13
	ℓ -diversity Model	81.56
	t -closeness Model	80.33
	Proposed Algorithm	86.34

numerical and categorical) of QIs present in users' dataset. Furthermore, it is applicable for anonymizing highly imbalanced (i.e., the datasets in which SA distribution is not uniform) datasets.

5 Conclusions and future work

In this paper, we have presented a user attributes' susceptibility and entropy based anonymization algorithm for preserving the privacy of users' data given or sold to analysts and researchers. We propose a mechanism for quantifying the susceptibility of each item of QIs using random forest to reduce the unique identifications of multiple users (i.e., users' community) caused by the highly susceptible QIs. We adapt the information entropy concept for calculating the uncertainty about SA' values in equivalence classes (ECs) to overcome the users' community SA disclosure caused by the low uncertainty ECs. In addition, the proposed adaptive generalization algorithm anonymizes users' data considering both the susceptibility of QIs as well as the entropy of SA simultaneously to effectively resolve the users' community privacy and anonymous data utility trade-off. It resolves the users' community privacy issues stemming from the highly susceptible QIs and low entropy ECs, and improves anonymous data utility by controlling over generalization of the less susceptible QIs. We conducted extensive experiments on different real-world

and synthetic social network users' datasets to demonstrate the effectiveness of the proposed anonymization algorithm. The anonymous data produced by the proposed algorithm is more resilient towards multiple users' unique identifications and their associated SA' inferences, and it yields higher anonymous data utility for performing analyses and building classification models compared to the existing algorithms. Furthermore, it is potentially applicable for fulfilling the two competing requirements of users' community privacy preservation and anonymous data utility enhancement of many data holders such as hospitals, banks, insurance companies, and analytics firms during their customers/subscribers data publishing. In future work, we plan to extend the proposed algorithm for multiple SAs scenarios. Furthermore, we also intend to investigate the other types of user privacy threats that emerge from publishing the dynamically evolving datasets.

Acknowledgements This work was supported by the Basic Science Research Programs through the National Research Foundation of Korea (NRF) funded by the Ministry of Education under Grant NRF-2017R1A2B1010817.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Wieringa J, Kannan PK, Ma X, Reutterer T, Risselada H, Skiera B (2019) Data analytics in a privacy-concerned world. *J Bus Res*. <https://doi.org/10.1016/j.jbusres.2019.05.005>
- Adhikari K, Panda RK (2018) Users' information privacy concerns and privacy protection behaviors in social networks. *J Glob Mark* 31(2):96–110
- Gkoulalas-Divanis A, Loukides G, Sun J (2014) Publishing data from electronic health records while preserving privacy: a survey of algorithms. *J Biomed Inform* 50:4–19
- Sweeney L (2000) Simple demographics often identify people uniquely. *Health (San Francisco)* 671:1–34
- Victor N, Lopez D, Abawajy JH (2016) Privacy models for big data: a survey. *Int J Big Data Intell* 3(1):61–75
- Al-Rubaie M, Chang JM (2019) Privacy-preserving machine learning: threats and solutions. *IEEE Secur Priv* 17(2):49–58
- Watanabe C, Amagasa T, Liu L (2011) Privacy risks and countermeasures in publishing and mining social network data. In: 7th international conference on collaborative computing: networking, applications and worksharing (CollaborateCom). IEEE. p 55–66
- Bayardo RJ, Agrawal R (2005) Data privacy through optimal k-anonymization. In: 21st International conference on data engineering (ICDE'05). IEEE. p 217–228
- Sun Y, Yuan Y, Wang G, Cheng Y (2016) Splitting anonymization: a novel privacy-preserving approach of social network. *Knowl Inf Syst* 47(3):595–623
- Sweeney L (2002) k-anonymity: a model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems* 10(05):557–570
- Machanavajjhala A, Gehrke J, Kifer D, Venkatasubramanian M (2006) l-diversity: privacy beyond k-anonymity. In: 22nd international conference on data engineering (ICDE'06). IEEE. p 24–24
- Li N, Li T, Venkatasubramanian S (2007) t-closeness: privacy beyond k-anonymity and l-diversity. In: 2007 IEEE 23rd international conference on data engineering. IEEE. p 106–115
- Li T, Li N, Zhang J, Molloy I (2010) Slicing: a new approach for privacy preserving data publishing. *IEEE Trans Knowl Data Eng* 24(3):561–574
- Xiao X, Tao Y (2006) Anatomy: simple and effective privacy preservation. In: Proceedings of the 32nd international conference on Very large data bases. VLDB Endowment. p 139–150
- Liu Y, Ji S, Mittal P (2016) Smartwalk: Enhancing social network security via adaptive random walks. In: Proceedings of the 2016 ACM SIGSAC conference on computer and communications security. ACM. p 492–503
- Liu P, Xu Y, Jiang Q, Tang Y, Guo Y, Le W et al (2019) Local differential privacy for social network publishing. *Neurocomputing*
- Bhagat S, Cormode G, Krishnamurthy B, Srivastava D (2009) Class-based graph anonymization for social network data. In *PVLDB* 2(1):766–777
- Liu K, Terzi E (2008) Towards identity anonymization on graphs. In: Proceedings of the 2008 ACM SIGMOD international conference on management of data. ACM. p 93–106
- Yuan M, Chen L, Yu PS (2010) Personalized privacy protection in social networks. In *PVLDB* 4(2):141–150
- Ying X, Wu X (2008) Randomizing social networks: a spectrum preserving approach. In: Proceedings of the 2008 SIAM international conference on data mining. SIAM. p 739–750
- Casas-Roma J, Herrera-Joancomarti J, Torra V (2013) An algorithm for k-degree anonymity on large networks. In: Proceedings of the 2013 IEEE/ACM international conference on advances in social networks analysis and mining. ACM. p 671–675
- Zhou B, Pei J (2008) Preserving privacy in social networks against neighborhood attacks. In: *ICDE*. vol 8. Citeseer, p 506–515
- Cheng J, Fu AWc, Liu J (2010) K-isomorphism: privacy preserving network publication against structural attacks. In: Proceedings of the 2010 ACM SIGMOD International Conference on Management of data. ACM. p 459–470
- Zou L, Chen L, Özsu MT (2009) K-automorphism: a general framework for privacy preserving network publication. In *VLDB* 2(1):946–957
- Hay M, Li C, Miklau G, Jensen D (2009) Accurate estimation of the degree distribution of private networks. In: 2009 Ninth IEEE International Conference on Data Mining. IEEE. p 169–178
- Day WY, Li N, Lyu M (2016) Publishing graph degree distribution with node differential privacy. In: Proceedings of the 2016 International Conference on Management of Data. ACM. p 123–138
- Kifer D, Machanavajjhala A (2011) No free lunch in data privacy. In: Proceedings of the 2011 ACM SIGMOD International Conference on Management of data. ACM. p 193–204
- Wang Y, Wu X (2013) Preserving differential privacy in degree-correlation based graph generation. *Trans Data Privacy* 6(2):127
- Nguyen HH, Imine A, Rusinowitch M (2015) Anonymizing social graphs via uncertainty semantics. In: Proceedings of the 10th ACM symposium on information, computer and communications security. ACM. p 495–506
- Leskovec J, Lang KJ, Dasgupta A, Mahoney MW (2008) Statistical properties of community structure in large social and information

- networks. In: Proceedings of the 17th international conference on World Wide Web. ACM, p 695–704
31. Siraj MM, Rahmat NA, Din MM (2019) A survey on privacy preserving data mining approaches and techniques. In: Proceedings of the 2019 8th international conference on software and computer applications. ACM, p 65–69
 32. Gong NZ, Liu B (2018) Attribute inference attacks in online social networks. *ACM Transactions on Privacy and Security* 21(1):3
 33. Wong RCW, Fu AWC, Wang K, Pei J (2007) Minimality attack in privacy preserving data publishing. In: Proceedings of the 33rd international conference on Very large data bases. VLDB Endowment, p 543–554
 34. Siddula M, Li Y, Cheng X, Tian Z, Cai Z (2019) Anonymization in online social networks based on enhanced Equi-Cardinal clustering. *IEEE Transactions on Computational Social Systems* 6(4):809–820
 35. He Z, Cai Z, Yu J (2017) Latent-data privacy preserving with customized data utility for social network data. *IEEE Trans Veh Technol* 67(1):665–673
 36. Pham VVH, Yu S, Sood K, Cui L (2017) Privacy issues in social networks and analysis: a comprehensive survey. *IET Networks* 7(2):74–84
 37. Yuan M, Chen L, Philip SY, Yu T (2011) Protecting sensitive labels in social network data anonymization. *IEEE Trans Knowl Data Eng* 25(3):633–647
 38. Kiabod M, Dehkordi MN, Barekatin B (2019) TSRAM: a time-saving k-degree anonymization method in social network. *Expert Syst Appl* 125:378–396
 39. Ros-Martin M, Salas J, Casas-Roma J (2019) Scalable non-deterministic clustering-based k-anonymization for rich networks. *Int J Inf Secur* 18(2):219–238
 40. Majeed A, Ullah F, Lee S (2017) Vulnerability-and diversity-aware anonymization of personally identifiable information for improving user privacy and utility of publishing data. *Sensors* 17(5):1059
 41. Dwork C (2006) Differential privacy. In: Proceedings of the International Colloquium on Automata, Languages and Programming (ICALP) (2):1–12
 42. Xie Y, Zheng M (2016) A differentiated anonymity algorithm for social network privacy preservation. *Algorithms* 9(4):85
 43. Li G, Cai Z, Yin G, He Z, Siddula M (2018) Differentially Private recommendation system based on community detection in social network applications. *Secur Commun Netw* 2018
 44. Wang N, Gu Y, Xu J, Li F, Yu G (2019) Differentially private high-dimensional data publication via grouping and truncating techniques. *Front Comput Sci* 13(2):382–395
 45. Zhou F, Lin X (2018) Frequent sequence pattern mining with differential privacy. In: International conference on intelligent computing. Springer, p 454–466
 46. Yang J, Li X, Sun Z, Zhang J (2019) A differential privacy framework for collaborative filtering. *Math Probl Eng* 2019
 47. Mohapatra D, Patra MR (2019) Anonymization of attributed social graph using anatomy based clustering. *Multimedia Tools Appl* 1–32
 48. Namdarzadegan M, Khafaei T (2019) Privacy preserving in social networks using combining Cuckoo optimization algorithm and graph clustering for anonymization. *Asian Journal of Research in Computer Science* 1–12
 49. Casas-Roma J (2019) An evaluation of vertex and edge modification techniques for privacy-preserving on graphs. *Journal of Ambient Intelligence and Humanized Computing* 1–17
 50. Yin D, Shen Y, Liu C (2017) Attribute couplet attacks and privacy preservation in social networks. *IEEE Access* 5:25295–25305
 51. Zhang C, Jiang H, Wang Y, Hu Q, Yu J, Cheng X (2019) User identity De-anonymization based on attributes. In: International conference on wireless algorithms, systems, and applications. Springer, p 458–469
 52. Sei Y, Okumura H, Takenouchi T, Ohsuga A (2017) Anonymization of sensitive quasi-identifiers for l-diversity and t-closeness. *IEEE Transactions on Dependable and Secure Computing* 16(4):580–593. <https://doi.org/10.1109/TDSC.2017.2698472>
 53. Li X, Yang Y, Chen Y, Niu X (2018) A privacy measurement framework for multiple online social networks against social identity linkage. *Appl Sci* 8(10):1790
 54. Kontaxis G, Polakis I, Ioannidis S, Markatos EP (2011) Detecting social network profile cloning. In: 2011 IEEE international conference on pervasive computing and communications workshops (PERCOM Workshops). IEEE, p 295–300
 55. Zheleva E, Getoor L (2011) Privacy in social networks: a survey. In: Social network data analytics. Springer, p 277–306
 56. Tai CH, Philip SY, Yang DN, Chen MS (2013) Structural diversity for resisting community identification in published social networks. *IEEE Trans Knowl Data Eng* 26(1):235–252
 57. Lin JCW, Fournier-Viger P, Wu L, Gan W, Djenouri Y, Zhang J (2018) PPSF: An open-source privacy-preserving and security mining framework. In: 2018 IEEE international conference on data mining workshops (ICDMW). IEEE, p 1459–1463
 58. Zhang B, Lin JCW, Liu Q, Fournier-Viger P, Djenouri Y (2019) A (k, p)-anonymity framework to sanitize transactional database with personalized sensitivity. *J Internet Technol* 20(3):801–808
 59. Lin C, Liu Q, Fournier-Viger P, Hong TP (2016) PTA: an efficient system for anonymizing transaction databases. *IEEE Access* 4: 6467–6479
 60. Wang SL, Tsai YC, Kao HY, Hong TP (2014) On anonymizing transactions with sensitive items. *Appl Intell* 41(4):1043–1058
 61. Lin JCW, Wu TY, Fournier-Viger P, Lin G, Zhan J, Voznak M (2016) Fast algorithms for hiding sensitive high-utility itemsets in privacy-preserving utility mining. *Eng Appl Artif Intell* 55:269–284
 62. Lin JCW, Zhang Y, Zhang B, Fournier-Viger P, Djenouri Y (2019) Hiding sensitive itemsets with multiple objective optimization. *Soft Computing* 1–19
 63. Zhang L, Wang W, Zhang Y (2019) Privacy preserving association rule mining: taxonomy, techniques, and metrics. *IEEE Access* 7: 45032–45047
 64. Mendes R, Vilela JP (2017) Privacy-preserving data mining: methods, metrics, and applications. *IEEE Access* 5:10562–10582
 65. Inuiguchi M, Ichida H, Torra V (2019) Data anonymization with imprecise rules and its performance evaluations. *Journal of Ambient Intelligence and Humanized Computing* 1–13
 66. Wu JMT, Lin CW, Fournier-Viger P, Djenouri Y, Chen CH, Li Z (2019) The density-based clustering method for privacy-preserving data mining. *Math Biosci Eng*
 67. Gan W, Chun-Wei J, Chao HC, Wang SL, Philip SY (2018) Privacy preserving utility mining: a survey. In: 2018 IEEE international conference on big data (Big Data). IEEE, p 2617–2626
 68. Li S, Mu N, Le J, Liao X (2019) A novel algorithm for privacy preserving utility mining based on integer linear programming. *Eng Appl Artif Intell* 81:300–312
 69. Jisna J, Salim A (2018) Privacy preserving data utility mining using perturbation. In: International conference on distributed computing and internet technology. Springer, p 112–120
 70. Liu P, Bai Y, Wang L, Li X (2017) Partial k-anonymity for privacy-preserving social network data publishing. *Int J Softw Eng Knowl Eng* 27(01):71–90
 71. Eyupoglu C, Aydin M, Zaim A, Sertbas A (2018) An efficient big data anonymization algorithm based on chaos and perturbation techniques. *Entropy* 20(5):373
 72. Han J, Yu J, Lu J, Peng H, Wu J (2017) An anonymization method to improve data utility for classification. In: International symposium on cyberspace safety and security. Springer, p 57–71
 73. Breiman L (2001) Random forests. *Mach Learn* 45(1):5–32

74. Blake CL, Merz CJ (1998) UCI Repository of Machine Learning Databases, Dept. Inf. Comput. Sci., Univ. California, Irvine, CA. <http://www.ics.uci.edu/~mlearn/MLRepository.html>
75. Amiri F, Yazdani N, Shakery A, Chinaei AH (2016) Hierarchical anonymization algorithms against background knowledge attack in data releasing. *Knowl-Based Syst* 101:71–89
76. Nettleton DF (2016) A synthetic data generator for online social network graphs. *Soc Netw Anal Min* 6(1):44
77. Sondeck LP, Laurent M, FREY V (2017) The semantic discrimination rate metric for privacy measurements which questions the benefit of T-closeness over L-diversity. In: *SECRYPT 2017: 14th international conference on security and cryptography*. vol 6. Madrid, Spain: Scitepress, p 285 – 294. Available from: <https://hal.archives-ouvertes.fr/hal-01576996>
78. BinJubier M, Ahmed AA, Ismail MAB, Sadiq AS, Khan MK (2019) Comprehensive survey on big data privacy protection. *IEEE Access* 8:20067–20079
79. Fung BC, Wang K, Fu AWC, Philip SY (2010) *Introduction to privacy-preserving data publishing: concepts and techniques*. Chapman and Hall/CRC

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Sungchang Lee received the B.S. degree from Kyungpook National University, in 1983, the M.S. degree in electrical engineering from the Korea Advanced Institute of Science and Technology (KAIST), in 1985, and the Ph.D. degree in electrical engineering from Texas A&M University, in 1991. From 1985 to 1987, he was with KAIST, as a Researcher, where he worked on Image Processing and Pattern Recognition Projects. From 1992 to 1993, he was a Senior Researcher with the Electronics and Telecommunications Research Institute (ETRI), South Korea. From 2004 to 2009, he was the Director of Government Project on Intelligent Smart Home Security & Automation Service Technology. In 2009, he was the Vice President of the Institute of Electronics and Information Engineers (IEIE), South Korea, and also the Director of the Telecommunications Society, South Korea. Since 1993, he has been a Faculty with Korea Aerospace University, Goyang, South Korea, where he is currently a Professor with the School of Electronics, Telecommunication & Computer Engineering.

Abdul Majeed did his B.Sc. (2013) in Information technology from University Institute of Information Technology (UIIT), PMAS-UAAR, Rawalpindi, Pakistan and his MS (2016) in Information Security from COMSATS University, Islamabad, Pakistan. He worked as Security Analyst at Trillium Information Security Systems (TISS), Rawalpindi, Pakistan from 2015 to 2016. He is currently pursuing his Ph.D. degree from Korea Aerospace University, Goyang,

South Korea. His research interests include data privacy, information security, robotics, data mining, and machine learning.