Journal of King Saud University – Computer and Information Sciences 31 (2019) 426–435

Contents lists available at ScienceDirect

Journal of King Saud University – Computer and Information Sciences

journal homepage: www.sciencedirect.com

Attribute-centric anonymization scheme for improving user privacy and utility of publishing e-health data

Abdul Majeed

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

ARTICLE INFO

Article history:
Received 10 November 2017
Revised 4 March 2018
Accepted 26 March 2018
Available online 31 March 2018

Keywords:
Privacy
Quality of treatment
Healthcare
Regulations
Collaboration
Background-knowledge

ABSTRACT

The adoption of advanced technologies in the healthcare sector has brought about many improvements in the industry, including better communication between healthcare providers, improved quality of treatment, and reduced cost. For the most part, these improvements have come about due to collaboration between healthcare providers and the sharing of healthcare data. However, this introduces various security and privacy concerns pertaining to the data in question. Preserving the privacy of the patients while simultaneously sharing data that would facilitate medical research is absolutely essential, for it is not just an ethical requirement but is also dictated by the regulations. In this paper, we propose a new anonymization scheme of data privacy for e-health records which differs from existing approaches in its ability to prevent from identity disclosure even faced with adversaries having pertinent background knowledge. The proposed scheme is based on transforming data into fixed intervals and then replacing original values with averages. As a result, the proposed scheme offers improved data privacy and utility in privacy preserving data publishing. The simulation results show the effectiveness of the scheme and verify the aforementioned claims.

1. Introduction

Research on the healthcare's data is extremely critical for developing innovative medicines and improving the accuracy of diagnoses. To this end, many institutes and hospitals collaborate by sharing health records for effective data analysis between researcher and pharmaceutical companies. Electronic health records (EHRs) are shared to aid medical studies that play a key role in medical advancements (Hill and Powell, 2009). Other benefits include a decrease in treatment/consultancy costs, and easy accessibility. It allows doctors to access records and prescribe medicines from anywhere, enabling telemedicine and remote healthcare. Studies have shown that release of medical data is essential to bring innovation in the e-health sector and bringing innovative ways of treatments. E-health also reduces physical resources and administration costs (Hsu et al., 1999). Some major advantages offered by e-health services are summarized in Fig. 1.

Apart from several advantages, data-owners (e.g., hospitals) are hesitant to share the patient's data because of privacy issues. Privacy preserving data publishing (Chen et al., 2009) is the field which provides tool and methods for privacy protected data-sharing. While sharing patient data between different parties may lead to cutting edge research, it also introduces a major concern for users data privacy (Sweeney, XXXX). Data holders only publish data if they trust the cloud provider. Privacy is an ethical as well as the regulatory requirement in e-health. Privacy breaches can cause identity disclosure, sensitive information loss, identity theft, and unauthorized modification of the data. According to a privacy study (Sweeney, 2002a), a person-specific data released for research purposes in 2002 was analyzed by an adversary with minimal background knowledge. The attacker managed to figure out the individuals' identities by simply linking records from the published data. Avoiding such breaches is challenging for researchers and scientists who are exploring and researching on user's data privacy. When the data contains personally identifiable information (PIIs), attacks can cause individual's identity and private information disclosures.

Another privacy-related study (Grandison, 2007) conducted in 2007 by Ponemon Institute/MSNBC involved a poll where people were asked to answer the question: *who do you trust more to protect your privacy, government or private corporations?* In response,

Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

E-mail address: Abdulmajid09398@kau.kr

https://doi.org/10.1016/j.jksuci.2018.03.014
1319-1578/

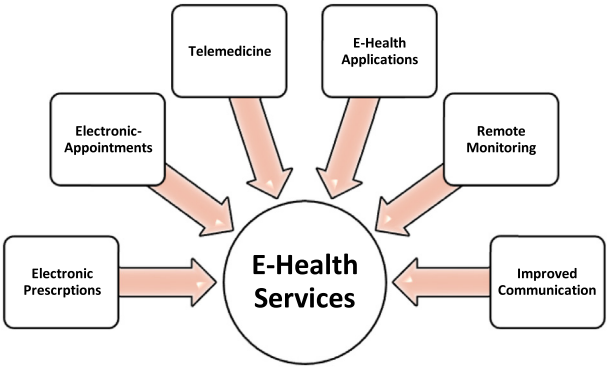

Fig. 1. E-health services overview.

88 people out of 100, picked the third option, "neither government nor private". This study concludes that privacy is an important concern. This study also highlights the need for novel privacy preserving techniques to protect user data in private clouds (Chen and Zhao, 2012). To avoid these issues, efficient measures should be taken while publishing e-health data in order to keep individual's information secure and private. A complete sanitization model with the collection of data from individuals to data dissemination to a 3rd party is described in Fig. 2. It is likely that attacker can obtain the data contents from an external source and can apply sophisticated techniques to link records (Ganta et al., 2008). Therefore, a revision in the model is necessary to protect data from misuse and explicit identity disclosures.

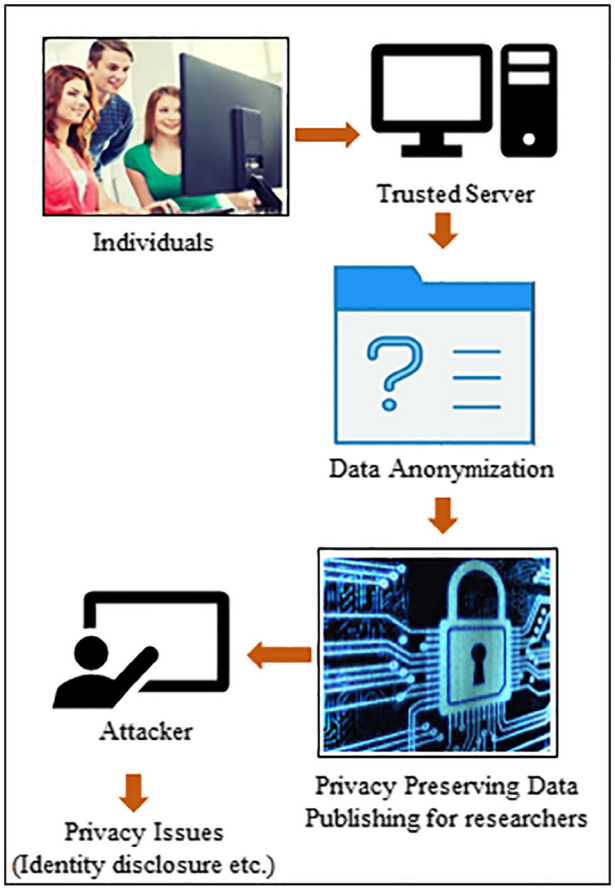

Fig. 2. Data collection, sharing and problem overview.

Numerous methods have been proposed to preserve the privacy of EHRs including k-anonymity (Sweeney, 2002a); *l*-diversity (Machanavajjhala et al., 2007), (α,k)-anonymity (Jian-min et al., 2008) and t closeness (Li et al., 2007) etc. However, attacker can still infer the sensitive information with the help of sophisticated techniques (Wernke et al., 2014). In this study, we studied the widely-used approach of data privacy used for *k*-anonymity and several other methods originating from *k*-anonymity named generalization. Our contribution in this work is twofold. First, we describe the generalization approach of classifying EHRs that offers a secure and easy mechanism for generalizing numerical attributes. Second, we propose a new approach based on fixed intervals for generalizing the numerical attributes of e-health records. The proposed scheme is solely based on data values. Hence, we hope that this work will provide essential knowledge to data publishers, researches and key-players in the privacy area to design and build more secure solutions for preserving the privacy of individuals while publishing data.

The rest of the paper is organized as follows: Section 2 presents few major requirements of the privacy in electronic health records. The description of privacy preserving techniques and comprehensive overview of generalization approach is provided in Section 3. Meanwhile, Section 4 explains the proposed approach with example, and Section 5 describes the key findings and the superiority of the proposed scheme in terms of privacy and utility, and the last section presents the conclusion where future directions in the same domain are also explained in detail.

2. Requirements of privacy in e-health sector

Recent trends in the healthcare sector are centering more and more on accessing, producing and sending information at any time and from anywhere. This encourages moving the e-health data to the cloud to ensure availability (Craig and Ludloff, 2011). Even though cloud computing offers numerous services, it also poses certain challenges in terms of privacy and security. As per recent studies, it is necessary for cloud service providers to fulfill privacy requirements in order to gain the trust of the intended users (Lederer et al., 2003). The fulfillment of these requirements guarantees a secure and privacy-preserving data publication and thwarts attackers from getting desired data. Detailed discussion about each requirement is explained below.

2.1. Assurance

Trust based relation between different entities is important to ensure that data is valid and free from errors. Trust management is subject to the enforcement of policies and permissions. Assurance in EHR provides confidence to the users and ensures that they are accessing and using accurate information from a trusted third-party cloud provider. Assurance increases information utility and results in better accessibility of the information.

2.2. Audit

The effective audit ensures that each and every activity on an e-health system is monitored and logged for carrying out inspections and investigation later. This will help to identify attacks and to understand adversarial capabilities. The audit is very important for conducting post attack analysis, this helps in understanding different activities happening between communicating parties (Shah et al., 2008). To ensure that data access only occurs when the requesting entities meet certain requirements such as: verify credentials, get a token, authenticate twice or physical presence etc.

2.3. Anonymity

Exchanging data between different parties with controlled modification is important to ensure users privacy (Chow and Mokbel, 2009; Zhou et al., 2008; LeFevre et al., 2008; Klosgen, 1995; Cormode and Srivastava, 2009; Byun et al., 2007; Bettini et al., 2009; Muntés-Mulero and Nin, 2009; Das et al., 2010; Masoumzadeh and Joshi, 2012; Majeed et al., 2017), such that the data cannot be linked back to an individual. It is necessary that data obtained from multiple sources is not sufficient for unauthorized users to reveal any information pertaining to a particular entity.

2.4. Non-repudiation

To ensure that each activity is monitored and logged with entity details so that no entity can deny after performing an activity. In e-health care system activities of each entity should be logged and stored for analysis.

2.5. Authorization

The users of her systems are patients, hospital staff (nurses, doctors, and pharmacy and laboratory staff), insurance companies and cloud service providers (Byun et al., 2005). This distributed nature and multiple client infrastructures are vulnerable to unauthorized access and internal attacks. Role-based access control can be used to limit the attacks originating from the organizations.

2.6. Spoofing

An adversary can get authorized access through forged credentials and can gain access to medical data as an authorized user. In e-health, every entity must have valid credentials to login to the system and these credentials must be verified before accessing to any computing resource.

2.7. Malicious insiders

Doctors who are the authorized consumers and producers of the e-health records can share this data with unauthorized pharmacies and laboratories causing information disclosure. In an e-Health system, data protection against such malicious insiders is required to ensure that authorized personnel do not take unfair advantage of their data access.

2.8. Consistency

Locations cause consistency and leakage issues, which in turn hurt the utility of the data and risk information disclosure. So, in e-health records, data consistency is important so that interested parties always get up-to-date data.

2.9. Nondisclosure of personal information

Disclosure of the record of medicine prescribed by a doctor may result in privacy breach. Expense information regarding insurance companies can be tampered with and cause financial loss. Sensitive information disclosure is necessary to be protected over the clouds.

2.10. Usage control

Usage control mediates who can access which data and how data will be used and distributed later. In e-health, usage control is important, which ensures that data will not be subject to inhibition, delay, modification or signaling. Usage control ensures the right usage of the data and controls data modifications.

2.11. Patient's consent

The patients must have the right to allow or disallow the dissemination of their information (Lunshof et al., 2008). In the cloud environment, patient's consent protection is necessary in order to restrict information to the desired entity. This requirement is crucial when someone is suffering from dangerous diseases.

2.12. Relevance

Only the relevant entities (i.e. doctors, patients, pharmaceutical and insurance companies) should have access to the patient's data. Moreover, fine-grained access control can be applied to ensure that personnel only have the information that is relevant to their tasks.

2.13. Archiving of the medical data

Archiving is not necessary for the EHR, as excessive data storage increases the hardware demand (Barrows and Clayton, 1996). A cloud provider should have complete information about the patient during his/her lifetime. Therefore, data is compressed and saved in repositories to meet this challenge.

3. Related work

In the recent years, cloud computing has managed to gain a significant amount of attention. E-Health is also benefiting from the services provided by the cloud (Rolim et al., 2010). However, the cloud itself is plagued with numerous security and privacy issues. Numerous privacy preserving techniques exist in literature for overcoming the privacy issues. Different techniques (Dinev and Hart, 2004; Young and Quan-Haase, 2013; Sheehan and Hoy, 1999; Sheehan, 2002; Bellman et al., 2004; Hann et al., 2007) like a generalization, slicing, packetization, suppression, pseudonymizing and cryptographic techniques are used to protect data from identity/attribute disclosure in the cloud, as. Privacy is very crucial over the internet. Therefore, it is essential to survey state of the art of privacy preserving data publication techniques. In this work, we focus on a well-known approach of data anonymization named "generalization" and provide its comparison with the new proposed approach is provided.

3.1. Privacy-preserving techniques

Even though cloud computing provides numerous services to the users, it is vulnerable to many attacks the distributed nature of e-health poses certain challenges like integrity, confidentiality, privacy breaches, identity disclosure, unavailability of the user data and alteration of the contents during transmission, etc. (Wernke et al., 2014). It is important for the data holder to have a combination of techniques to protect data efficiently in a third-party environment. Apart from the above mentioned non-cryptographic techniques for preserving privacy, some cryptography-based solutions such as authentication, access controls, password management and biometric schemes are also used (Craig and Ludloff, 2011). Implementation of such techniques leads to better results, but explicit monitoring and maintenance of such techniques are very challenging. Techniques like a generalization, suppression, randomization, pseudonymization, cryptographic techniques (Hsu et al., 1999) and anatomy are used to support privacy preserving data publication. However, each technique has its own merits and demerits. A comprehensive overview of the pri-

vacy techniques in term of privacy fulfillment is given in Table 1. A comprehensive overview in term of privacy requirements has been provided in the table. This paper specifically aims to compare generalization with the proposed approach of fixed intervals. All existing techniques are listed in Fig. 3.

Many of the privacy objectives are prone to explicit identity disclosures when user's data persist over cloud. However, these approaches are collectively helpful in achieving privacy objectives. Cryptographic techniques (Pinkas, 2002; Ristenpart et al., 2008; Diffie and Hellman, 1976; Clifton et al., 2002a; Goethals et al., 2004; Bogdanov et al., 2008; Bellare et al., 2000; Pedersen et al., 2007; Feistel et al., 1975; Fischer-Hübner, 2001) are more secure than non-cryptographic techniques and provide better data security in cloud computing environments. A related review of the literature on e-health records storage and retrieval from cloud based on various roles of stakeholders was offered by Bahga and Madisetti (2013), Zhang and Liu (2010), Proceedings of the 8th International Conference on Collaborative Computing (2012). They draw attention to secure e-health record sharing via clouds rather than building and maintaining dedicated data centers.

We refer the interested reader to Fernández-Alemán et al. (2013) for comprehensive literature concerning the security and privacy of electronic health record systems. A comprehensive study about the collaborative and secure sharing of healthcare data in multi-clouds was given by authors (Fabian et al., 2015). The proposed architecture helps in data sharing considering the privacy needs in semi-trusted cloud computing environments. A large scale framework (Benharref and Serhani, 2014), which relies on service-oriented architecture (SOA) and the Cloud, allows a seamless integration of different technologies, applications, and services. The proposed framework allows the easy access to the data by physicians, paramedics, or any other authorized entity. Recently, internet of things (IoT) for healthcare have gained popularity (Hossain and Muhammad, 2016). However, this requires the identification and modelling of privacy threats when collecting and releasing data (Deng et al., 2011). Data mining reveals knowledge

Table 1
Privacy techniques and requirements fulfilment analysis.

Sr. #	Technique	Supported privacy requirements	Description
1.	Generalization	Nondisclosure of identity and sensitive information.	It is widely-used approach to ensure privacy using predefined hierarchies or values.
2.	Suppression	Spoofing, patient consent.	It is another form of the generalization which completely hide original values with '*'.
3.	Pseudonymization	Consistency, anonymity	It replaces original values with pseudonyms.
4.	Bucktization	Nondisclosure of information, relevance	It is used to keep individual data in separate buckets, ensuring bucket protection.
5.	Slicing	Audit, nondisclosure of membership	It splits original values into slices.
6.	Randomization	Anonymity, increase attacker search space	Randomization increases the attacker's efforts by adding noise in the original data.
7.	Cryptographic Approaches	Assurance, authenticity, audit, authorization, confidentiality, integrity, non-repudiation	They are mainly used to secure data in the clouds.

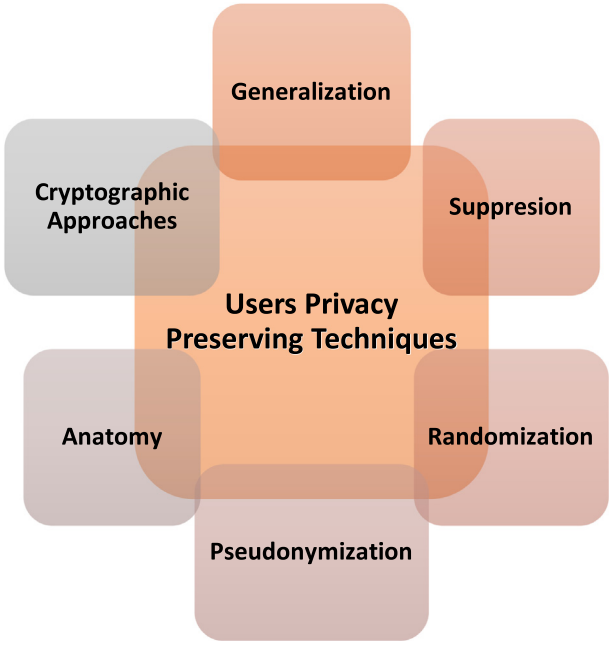

Fig. 3. Taxonomy of the privacy preserving techniques.

patterns that apply to many people in the data (Vercellis, 2011), and the existing privacy models often ignore these receptive patterns which may hurt users privacy. An independent study in 2008 (Friedman et al., 2008) discussed k-anonymity in from data mining perspective and seemed to be reliable solution for users privacy. Anonymity is achieved by extending the general k-anonymity model with the data mining model. A comprehensive literature concerning the privacy approaches and concepts in general and from cloud perspective can be obtained from the sources (Fung, 2011; Rass and Slamanig, 2013).

3.2. Data generalization technique

Generalization is a most widely-used technique for anonymization which replaces quasi-identifiers (attributes that potentially identify individual i.e., Age, zip code, gender, etc.) values with other less specific values which are consistent with the original data. Generalization protects an individual's identity by replacing quasi-identifiers with less specific records. Privacy models (Machanavajjhala et al., 2007; Sweeney, 2002b, 2000c; Wang et al., 2004; Bayardo and Agrawal, 2005; Aggarwal and Philip, 2008; Chen et al., 1996; Li et al., 2012; Clifton et al., 2002b; Aggarwal, 2005) like *k*-anonymity use generalization and suppression to anonymize data, but this technique does not provide protection against attribute disclosure. However, data holders also consider utility along with privacy to ensure that the data published by the data holder is useful for researchers despite anonymization. A popular model of *k*-anonymity uses generalization to create an anonymous version of the data. *K*-anonymity also provides protection against linking and attributes disclosure. This model applies a generalization to the quasi-identifiers, as discussed in subsection c, in order to replace explicit information with more general values. However, this technique does not guarantee the protection of disclosure based on background knowledge and linking in specific cases.

A simple example with at least eleven records is given in Tables 2 and 3, where the former shows original data and the latter shows the transformed data after generalization is applied. During generalization process, original values are replaced with the semanti-

Table 2
Original data.

Age	Sex	Zip code	Disease
21	M	53706	Anemia
26	M	53706	Flu
32	F	53710	Cancer
36	F	53715	Torn ACL
48	M	52108	Flu
56	F	52100	Whiplash

Table 3
Generalized data ($k = 2$).

Age	Sex	Zip code	Disease
20–30	M	53705–53710	Anemia
20–30	M	53705–53710	Flu
30–40	F	53710–53715	Cancer
30–40	F	53710–53715	Torn ACL
45–60	M	52100–52108	Flu
45–60	F	52100–52108	Whiplash

cally consistent values. A city can be generalized to country and country can be generalized to the continent to which a particular country belongs. For numerical data generalization is performed to replace original values with the range of values or showing one digit (age: 30–40 or 2*) only.

Generalization protects from privacy breaches but loses considerable information of the micro data and as a result data becomes useless. However, generalization is not effective at all when the adversary has strong background knowledge. From the data provided in Table 3, if an adversary knows that Bob is 21 years old and his zip code is 53706, then he can infer that Bob is either suffering from Anemia or flu. Studies based on generalized tables also conclude that executing queries on them lead to false results and implementation of techniques such as data mining and correlation is not effective.

4. Proposed anonymization scheme

This section focuses on clarifying the overall approach we used for data anonymization. A trusted publisher, say a hospital, collects data about the patients. This data contains information about different individuals. After a specific duration, trusted publishers release this information to pharmaceutical companies that conduct research on the data for the development of medicines for specific diseases. Their key findings include how these diseases correlate with age and gender. In an e-health data, these identifiers are usually present: name, age, zip code, address, disease, contact number, identity number, salary information, and designation along with disease and treatment details. These attributes are classified into specific categories (Table 4).

A comprehensive overview of the attributes collected by a publisher, along with their classification, is provided in Table 4. This general classification of identifiers clarifies, mapping, normalizing and identifying relations in the data. These identifiers are further classified as,

Table 4
Identifier's classification.

Identifier	Numerical	Categorical
Name	×	✓
Address	×	✓
Age	✓	×
Zip-code	✓	×
Disease	×	✓
Salary Information	✓	×
Diagnosis History	✓	✓
Designation	×	✓

I. *Personal Identifier(s)* (denoted by ID) are uniquely identifying attributes e.g., Social Security Number, Name etc. these attributes are removed in anonymization process as they identify users directly.
II. *Quasi-Identifier(s)* (denoted by QIs) e.g., age, zip code, designation and gender are the set of attributes that are publicly available and can be used to reveal individual's identities.
III. *Sensitive Attribute(s)* (denoted by SA) contain sensitive information about individuals that must be protected from the adversary. In Table 4, disease, treatment history etc. can be termed as sensitive attributes.

While publishing data, IDs are removed completely, QIs are either generalized, suppressed or bucketed depending upon the scenario. Sensitive's attributes are also published with micro data after some modifications and mostly as it is for researchers. Trusted publishers always try to ensure privacy-preserving distribution of e-health data so that the identity of the people who are the subjects of the data is not disclosed. In pre-processing stage in case zip code portions are separated by '–', we first convert into quantity by removing it and consider it as number during mean calculations.

Keeping in mind the importance of data privacy as an unmet need for e-health, we propose a variant of generalization which protects from identity disclosure and helps in generalizing QIs in such a way that privacy is preserved. It helps in meeting most of the privacy-related requirements expressed in Section 1. Additionally, we propose another approach for classifying categorical data as well. This complete model for anonymization assists in generalizing numerical and categorical data with ease. We propose formulas for identifying intervals and data distributions in each group. This state of the art solution helps e-health providers to publish their data with confidence. Fixed intervals approach protects the data from adversaries with background knowledge. It assists data providers in the current assessment of different security trends. It plays a vital role in the preservation of data privacy.

Our proposed approach has wide applications in anonymizing data with ease. It assists data publishers in creating an anonymized version of the data with less complexity as compared to other approaches for data privacy. This is a relatively better technique with respect to generalization in terms of data privacy.

Suppose that there is an original data collected by hospital XYZ. This micro-data will be published later with different companies for research purposes. Before publishing, an anonymized version of the data will be generated so that the privacy of the individuals is preserved. The complete data anonymization process is given with the help of the example below. First, we will anonymize numerical attributes and anonymize then categorical attributes. The core objective of the any privacy technique is to preserve maximum information for researchers while limiting privacy issues of background-knowledge and linking attacks. In background-knowledge attack the attacker possess some preliminary information about someone by knowing his/her QIs. They try to pinpoint individuals in the data and try to identify their information.

In order to anonymize age attributes, the sorted values are distributed into N number of buckets or bins of fixed size and then each value of the micro-data is replaced with bin averages. Divide the range into N intervals of equal size to create a uniform grid. If A and B are the lowest and highest values of the age attribute in the actual data which we want to anonymize, the width of intervals will be calculated using Eq. (1),

$$Interval\ Width(w) = \frac{B - A}{N} \quad (1)$$

where B is the highest value of the age attribute, A is the lowest value and N is the total number of values in the bins. In our original micro-data, the lower value of age attribute is 21 and the highest value is 56, and we want to create three anonymized data groups of equal sizes. The group's width is calculated as:

$$w = \frac{56 - 21}{3} = 12$$

After calculations, we get the following three groups based on the original micro-data:

i. 21–33
ii. 33–45
iii. 45–57

Later, the actual values which belong to these intervals can be replaced by mean or median of the values. Even percentage has more protection than averages, in our case, we replace actual values with the mean of the values in a particular interval.

In the first interval, there are three values so mean of the values is calculated as under,

$$Mean = \frac{\varkappa_1 + \varkappa_{2+\ldots.}\varkappa_n}{n} \quad (2)$$

where $\varkappa_1$ *and* $\varkappa_2$ are the actual values and n are the number of total values in that group or bin (e.g., Group 1).

$$Mean(G_1) = \frac{21 + 26 + 32}{3} = 26$$

The mean calculation process is same as mentioned in above equation for rest of the two groups.

In order to anonymize zip code values given in Table 5 as original micro-data, the lower value of zip code attribute is 52100 and the highest value is 53715 and we want to create three anonymized data groups of equal sizes same as gender. Using Eq. (1), we get the following results,

$$w = \frac{53715 - 52100}{3} = 539$$

After calculations, we got the following groups based on the original micro-data:

i. 52100–52639
ii. 52639–53178
iii. 53178–53717

Later, the actual values which belong to these intervals can be replaced by mean or median of the values as shown in Tables 6 and 7. For simplicity, we keep the original data values as it is. This anonymization is quite suitable when data is not too big and data contents are known.

For the anonymization of categorical data, we suggest id-based matching. In the original micro-data, we have two categorical attributes, namely sex and disease. Which are simply anonymized based on ids. Each id values are published in anonymized data. For example, anonymization of sex can be zero for male and 1 for female. The anonymization of sensitive categorical value is provided in Table 8.

This approach is suitable for making appropriate intervals for anonymizing data. This scheme protects from background knowledge attack. The proposed approach has a few similarities with binning approach used for modeling statistics data. However, this technique does not hurt data utility. This approach eliminates the manual tree making process of generalizing records as well. Complete working of the proposed scheme is depicted in Fig. 4 with major components.

Most of the existing generalization-based algorithm hurts the anonymous data utility too much. For example, if a medical student wants to model the diseases from which 25-years of age people usually suffer. Generalization will either give him/her either partial values of age attribute or either overfitted or underfitted intervals. For example, 10–25 or 20–35. It will really become difficult for him/her to model or carry out some disease analysis on these intervals. From the privacy protection point of view if an attacker knows someone age (e.g., 27) of gender male, and having zip code (e.g., 53710) based on the background-knowledge the attacker can infer someone sensitive information. The utility motivated generalization-based algorithms will present full information to attacker to identify someone as Table 9. Meanwhile, the proposed approach has better protection in such circumstances. The proposed approach overcomes the difficulty in creating generalization hierarchies and setting up the generalization degree. It provides better protection from well-known attacks regarding privacy protection and retains better semantics of the original data from utility point of view. It resolves the generalization underfitting and overfitting issues of quasi-identifiers values and effective data analysis is possible.

To anonymize any person-specific data containing users quasi identifiers such as, age, gender, zip code etc. and sensitive attributes such as, salary or disease information the following six principal concepts are introduced: (1) the pre-processing of the original e-health records; (2) highest similarity user ranking based on QIs values; (3) the formation of equivalence classes (C_i) using privacy parameter k; (4) attributes values range analysis, and outliers removal from data (5); classification of the attributes into two

Table 5
Original data.

Age	Sex	Zip code	Disease
21	M	53706	Anemia
26	M	53706	Flu
32	F	53710	Cancer
36	F	53715	Torn ACL
48	M	52108	Flu
56	F	52100	Whiplash

Table 6
Anonymization of age attribute.

Intervals	Values
21–33	26
33–45	36
45–57	52

Table 7
Anonymization of zip code attribute.

Intervals	Values
52100–52639	52104
52639–53178	0
53178–53717	53709

Table 8
Nominal attributes data anonymization.

Id	Value	Id	Value
0	Anemia	2	Torn ACL
1	Flu	3	Whiplash
4	Cancer		

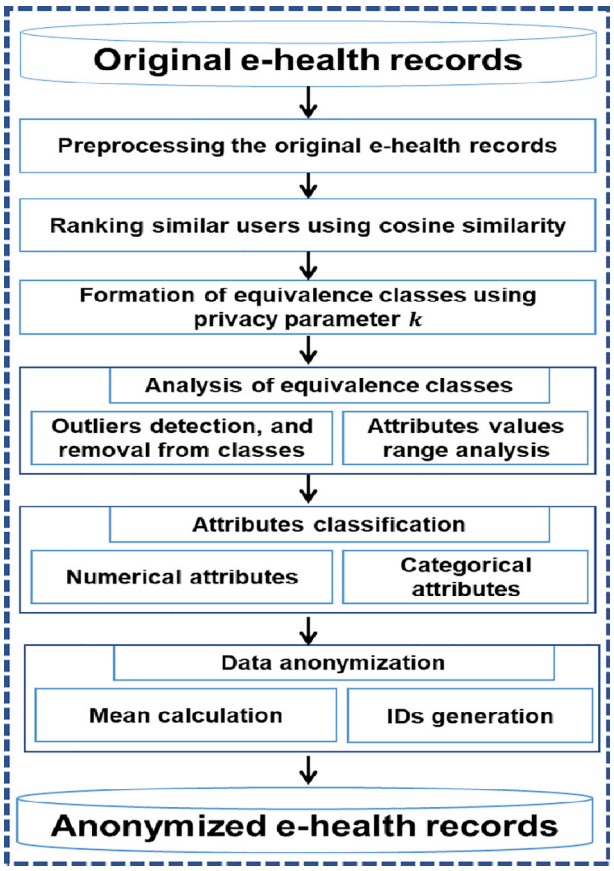

Fig. 4. Fixed interval approach working flowchart.

Table 9
Attributes taxonomy-based data anonymization.

Age	Sex	Zip code	Disease
20–30	M	53705–53710	Anemia
20–30	M	53705–53710	Flu
30–40	F	53710–53715	Cancer
30–40	F	53710–53715	Torn ACL
45–60	M	52100–52108	Flu
45–60	F	52100–52108	Whiplash

classes; and (6) data anonymization. This approach is chosen to enhance user privacy in any dataset and to reduce privacy breaches caused by the background knowledge and linking attacks. Apart from the users' privacy protection, the anonymous data utility in terms of building different classifiers, and preserving attributes values close to original values as much as possible are the main objectives of the proposed scheme. The proposed scheme working along with the procedural steps is shown in Fig. 4.

Brief details of the principal components with equations and procedures are as follows.

4.1. Pre-processing of the e-health records

Before anonymizing the e-health data, pre-processing is carried out to overcome the "garbage in, garbage out" issues. Data pre-processing is very important step since data gathering methods are often loosely controlled which yields inaccurate results. The problem such as, out of range values (i.e., income: −100), impossible data combinations (e.g., gender: male, pregnant: yes) and missing values etc. can produce misleading results. Apart from the issues mentioned above the presence of irrelevant and redundant values in the data increase the complexity of the anonymization process. Data pre-processing includes data cleaning, normalization, feature extraction, selection and transformations (e.g., categorical to numerical).

4.2. Highly similar users ranking

Based on QI values, similar users are ranked, this is done by means of cosine similarity given as:

$$Sim(P_1, Q_1) = \frac{\sum_{n=1}^{N} P1_{(n)} \times Q1_{(n)}}{\sqrt{\sum_{n=1}^{N} (P1_{(n)})^2} \times \sqrt{\sum_{n=1}^{N} (Q1_{(n)})^2}} \quad (3)$$

where P_1 and Q_1 are two different users having QIs, $P1_1$, $Q1_1$, $P1_2$, $Q1_2, \ldots, P1_n$, $Q1_n$. The resultant matrix contains highly similar users based on their QIs values.

4.3. Formation of equivalence classes

After ranking the similar users, the user matrix U is partitioned into different equivalence classes $(C_1, C_2, C_3, \ldots, C_N)$ based on the privacy parameter (k), where each class consists of at least k individuals. The value of k is selected by the data owner (e.g., hospitals), and it can be any whole number. However, it must be chosen carefully considering the data distribution, and the objectives of the data publishing. If highly similar users are N, the number of equivalence classes (C_i) can be obtained using following equation.

$$C_i = \frac{N}{k} \quad (4)$$

4.4. Analysis of equivalence classes

After forming the equivalence classes, we analyze the values of each attribute in equivalence classes by plotting the data. The outliers are removed by visual inspection of the data if any. Apart from outlier's removal, each attribute's values range analysis is carried to provide precise results in privacy protection. In some cases, it is also possible that attribute in one equivalence class can contain only a single value (e.g., all users have age value 30 years) or very small number of values (e.g., five users having age values 30,31,30,32, and 30 years respectively). The anonymization of such equivalence class can leak much more information about actual data in such cases. Therefore, if the range values are lower than defined threshold T then the actual values of the attributes are increased by constant factor to preserve users' privacy. The values of T in our experiments were set to five. However, it can be adjusted according to the objective, and the protection level which data owner want to ensure for EHRs.

4.5. Attributes classification

In EHRs, the records can be either numeric, characters or combined from. In proposed approach we deal two types of the data while producing the anonymized values. Therefore, we classify the data only into two types. A question which arises in the mind is, are these only two types of the data in EHRs? So, to clarify this, data may exist in several forms other than these two mentioned in our work, but we dealt with two types of the data only. The explicit identifiers such as, email address and postal address exists in the combination of both attributes. Meanwhile, according to privacy preserving data publishing (PPDP) concept direct identifiers (e.g., email address, postal address, phone number, name, and place etc) are removed before data anonymization process begins.

4.6. Data anonymization

Data anonymization is performed for replacing the original attributes values with generalized values to anonymize the data. The anonymization of the attributes in each equivalence class is based on the type and values of the specific attributes. The data anonymization considering the real attribute values facilitates superior data anonymity, thereby protecting identities and preventing confidential information disclosures. At the same time, the data utility is also preserved for effective analysis and building classifiers of several types. The proposed scheme helps in retaining the semantics of original values up to great extent possible thereby improving the utility of the data.

5. Results and discussion

This section presents the output of the concept discussed. Fixed interval approach has comparatively better results than closely related approaches. The proposed approach has wide adoption and ensures accurate results in terms of user's privacy and utility of anonymous data. Most importantly, this approach has numerous advantages as compared to all generalization approach like it protects from identity and membership disclosure attacks because it is one of the unmet requirement by of the most schemes. Alternatively, this approach is suitable for a web application where user interaction is high and query responses are generated simultaneously. It also increases attacker's search space for learning about the contents of the data up to great extent. Therefore, the proposed approach yields promising results in both user's privacy and anonymous data utility during the e-health data publishing.

5.1. Improvements in user privacy

The proposed approach is not only good in terms of space and utility but also protects from explicit identity disclosure. The proposed approach helps in avoiding background knowledge attacks, identity disclosure, and membership disclosure as well. The detailed comparison of both techniques is provided in Tables 9 and 10 respectively. For producing the anonymized values of each attribute, the records given in Table 5 are used.

If adversaries have some background knowledge about a person in terms of age, zip code or gender, they can easily identify the disease of that person. Even though the best technique of data privacy (e.g., generalization) is useless when adversary possesses some background knowledge. In most of the cases, when data publisher releases their data on a periodic basis, generalization fails badly to protect from identity and membership disclosure in the microdata. However, there are certain algorithms which can make better use of these techniques jointly.

5.2. Improvements in anonymous data utility

A question which arises in the mind is that loss of accuracy may be an issue with the proposed approach, which need to be clarified. So, to solve this ambiguity, we performed extensive simulations from utility point of view on the adult's datasets (Blake and Merz, 1998) available at the UCI machine learning repository using the two classifiers named random forest (Breiman, 2001) and support vector machines (Osuna et al., 1997). The original dataset contains 48,842 records, comprises of six numerical and eight categorical/non-numerical attributes and is 5.4 MB in size. Four attributes are used as quasi identifiers, and one attribute is used as the target class. The two-thirds division of actual data in the presence of missing values gives 32,561 instances as training data and 16,281 instances as testing data. We eliminated the records with unknown values before conducting experiments and resulting data set contains 45,222 tuples. The two-thirds division of refined data contains 30,162 instances as training data and 15,060 instances as testing data. We compared the obtained results with one of the existing and state of the art method named IACk (Li et al., 2011). The proposed algorithm performs consistently better as compared to the IACk algorithm. The accuracy results for the different values of K from $5 \sim 25$ are listed in Fig. 5 using random forest (RF).

Similarly, the accuracy results for the different values of K from $5 \sim 25$ are listed in Fig. 6 using support vector machines (SVM). The proposed scheme results are promising with respect to utility and privacy of anonymous data. The proposed scheme can work well with all types of the data.

Some readers may wonder that what are the applications of the anonymized attributes values that are produced by the proposed scheme. So, to solve this ambiguity, we highlight some of the potential uses of the anonymous values which are also clarified by the numerous studies. From different researches on EHRs, its proven that each attribute in the EHRs contribute differently for utility and privacy of users. Apart from the utility and privacy only, each attribute has different predictive power. Some attribute has different prediction ability than other. For example, the gender attribute has more predictive power than the attribute ID since male students are more likely to study computer science than female students and this trend has not changed for a few years. Similarly, the medical researchers are always interested in modelling the causes of disease using age attribute in most cases. Therefore, its desirable to present them the anonymized values that deviate less from original values to perform effective analysis. The proposed scheme retains the semantics of the original values well by means of averages, therefore it assists in accurate diseases modelling. Additionally, the zip code attribute represents the local region more clearly than country or state and if some dangerous

Table 10
Proposed approach data anonymization.

Age	Sex	Zip code	Disease
26	0	52104	0,1
26	0	52104	0,1
36	1	53715	2,4
36	1	53715	2,4
52	0,1	53709	1,3
52	0,1	53709	1,3

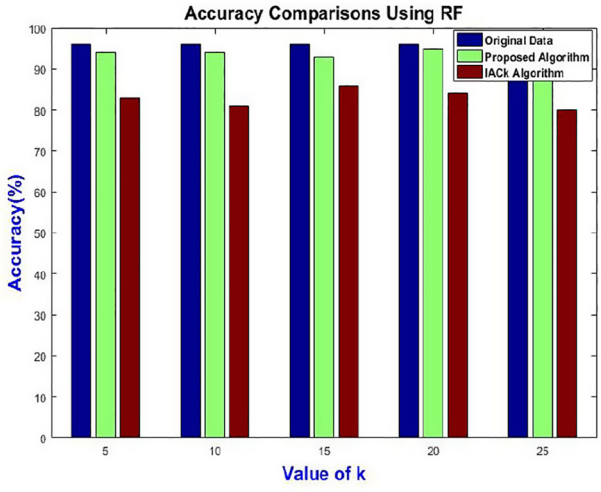

Fig. 5. Accuracies: proposed algorithm versus IACk algorithm.

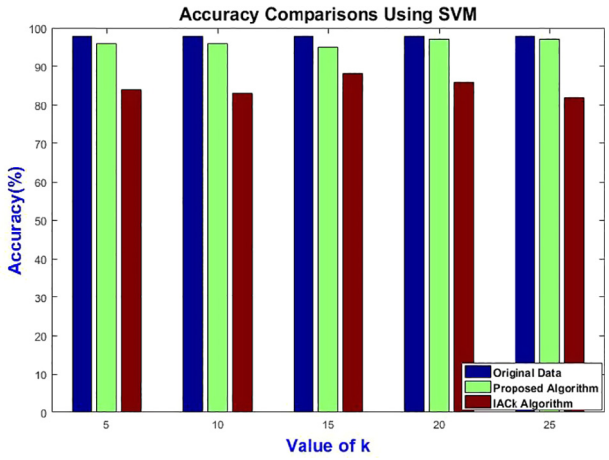

Fig. 6. Accuracies: proposed algorithm versus IACk algorithm.

disease (e.g., cancer) is very common in that locality. It will help the pharmaceutical companies, and hospitals to discover more sophisticated treatments and give special preferences to those regions based on extracted information. Apart from the common QIs such as, age, gender and zip code etc. some other related attributes such as, height, weights and body mass index reading of individuals can also cause privacy breaches when transferred through devices. The proposed scheme ensures the protection of all numerical attributes. Meanwhile, the combined numeric attributes values knowledge such as, people of certain territory having zip code 412-791 with the age of 80 or above always suffer from Alzheimer disease. Such informative analysis extracted from the data provided by proposed scheme while protecting the individual's privacy is helpful for improving the quality of treatment, better understanding of the disease trends, and reduced cost.

6. Conclusions and future work

The privacy of electronic health records in the cloud is a genuine issue that requires special consideration from the research community. Researchers have proposed and implemented different algorithms for protecting user's privacy. In this paper, we have discussed a state of the art and existing generalization technique for data anonymization and its limitations. In this paper, we proposed a fixed interval approach for data privacy of e-health data containing users' quasi-identifiers and sensitive attributes. This approach can be applied to other similar systems as well for preserving user's privacy. We have compared the usefulness of both generalization and fixed interval approach in terms of privacy and utility. The main idea behind the proposed technique is that the quasi-identifiers present in the electronic health records should be properly classified in fixed intervals, and then original values are replaced with the averages of original values. For the anonymization of categorical attributes, id-based anonymization is proposed. We found this approach very promising which efficiently resolve the privacy issues stemming from the adversary background knowledge and preserves better utility of anonymous data.

Few open issues which need further exploration are briefly mentioned below:

- Selection of appropriate trusted infrastructure, service provider and algorithms are still insufficient to fulfil user privacy needs in pervasive environments.
- Due to the diverse nature of the cloud, secure provenance is the key issue which needs to be explored. Generally, this secure provenance includes 1- actions (insert, delete, view, update etc.) that users take, 2- entities personal information security, 3- the location of the action (i.e., geographical location), and 4- the reason for the action. Although this environment is protected by different approaches, but still provenance has revealed sensitive information to the malicious users by several ways (e.g., sniffing & spoofing etc.).
- Cryptography-based algorithms work considerably slow. The searching and manipulating of the record in the large data set is a time-taking process. So, there is a significant need to implement efficient, scalable and usable data search strategies to improve the speed of these techniques.
- Extending our proposed scheme to multiple sensitive attributes, and effective mining of the interesting patters from the anonymous data are very interesting topics for future research.

References

Aggarwal, C., Philip, S., 2008. A general survey of privacy-preserving data mining models and algorithms. Privacy-preserving data Min.
Aggarwal, C., 2005. On k-anonymity and the curse of dimensionality. In: 31st Int. Conf. Very large data.
Bahga, A., Madisetti, V.K., 2013. A cloud-based approach for interoperable Electronic Health Records (EHRs). IEEE J. Biomed. Heal. Informatics 17 (5), 894–906.
Barrows, R.C., Clayton, P.D., 1996. Privacy, confidentiality, and electronic medical records. J. Am. Med. Informatics Assoc. 3 (2), 139–148.
Bayardo, R., Agrawal, R., 2005. Data privacy through optimal k-anonymization. Data Eng. 2005. ICDE 2005.
Bellare, M., Kilian, J., Rogaway, P., 2000. The security of the cipher block chaining message authentication code. J. Comput. Syst.
Bellman, S., Johnson, E.J., Kobrin, S.J., Lohse, G.L., 2004. International differences in information privacy concerns: a global survey of consumers. Inf. Soc. 20 (5), 313–324.
Benharref, A., Serhani, M.A., 2014. Novel cloud and SOA-based framework for e-health monitoring using wireless biosensors. IEEE J. Biomed. Heal. Informatics 18 (1), 46–55.
Bettini, C., Jajodia, S., Samarati, P., Wang, S., 2009. Privacy in location-based applications: research issues and emerging trends.
Blake, C., Merz, C., 1998. UCI repository of machine learning databases, University of California, Dept. Information and Computer Science, Irvine, CA, USA.
Bogdanov, D., Laur, S., Willemson, J., 2008. Sharemind: a framework for fast privacy-preserving computations. Eur. Symp. Res.
Breiman, L., 2001. Random forests. Mach Learn 45, 5–32.
Byun, J.-W., Bertino, E., Li, N., 2005. Purpose based access control of complex data for privacy protection. In: Proceedings of the Tenth ACM Symposium on Access Control Models and Technologies – SACMAT '05, pp. 102.
Byun, J., Kamra, A., Bertino, E., Li, N., 2007. Efficient k-anonymization using clustering techniques. Int. Conf. Database.
Chen, D., Zhao, H., 2012. Data security and privacy protection issues in cloud computing. In: 2012 International Conference on Computer Science and Electronics Engineering, 2012, pp. 647–651.
Chen, M., Han, J., Yu, P., 1996. Data mining: an overview from a database perspective. Trans. Knowl. Data.
Chen, B.-C., Kifer, D., LeFevre, K., Machanavajjhala, A., 2009. Privacy-Preserving Data Publishing. Found. Trends® Databases 2 (1–2), 1–167.
Chow, C., Mokbel, M., 2009. Privacy in location-based services: a system architecture perspective. Sigspatial Spec.
Clifton, C., Kantarcioglu, M., Vaidya, J., Lin, X., 2002. Tools for privacy preserving distributed data mining. ACM Sigkdd Explor.
Clifton, C., Kantarcioglu, M., Vaidya, J., 2002b. Defining privacy for data mining. Work. Next Gener Data.
Cormode, G., Srivastava, D., 2009. Anonymized data: generation, models, usage. Proc. 2009 ACM SIGMOD.
Craig, T., Ludloff, M.E., 2011. Privacy and big data. O'Reilly.
Das, S., Eğecioğlu, Ö., El Abbadi, A., 2010. Anonymizing weighted social network graphs. Data Eng. (ICDE).
Deng, M., Petkovic, M., Nalin, M., Baroni, I., 2011. A home healthcare system in the cloud–addressing security and privacy challenges. In: 2011 IEEE 4th International Conference on Cloud Computing, pp. 549–556.
Diffie, W., Hellman, M., 1976. Multiuser cryptographic techniques. In: Proc. June 7–10, 1976, Natl., 1976.
Dinev, T., Hart, P., 2004. Internet privacy concerns and their antecedents – measurement validity and a regression model. Behav. Inf. Technol. 23 (6), 413–422.
Fabian, B., Ermakova, T., Junghanns, P., 2015. Collaborative and secure sharing of healthcare data in multi-clouds. Inf. Syst. 48, 132–150.
Feistel, H., Notz, W., Smith, J., 1975. Some cryptographic techniques for machine-to-machine data communications. Proc IEEE.
Fernández-Alemán, J.L., Señor, I.C., Lozoya, P.Á.O., Toval, A., 2013. Security and privacy in electronic health records: a systematic literature review. J. Biomed. Inform. 46 (3), 541–562.

Fischer-Hübner, S., 2001. IT-security and privacy: design and use of privacy-enhancing security mechanisms.
Friedman, A., Wolff, R., Schuster, A., 2008. Providing k-anonymity in data mining. VLDB J. 17 (4), 789–804.
Fung, B.C.M., 2011. Introduction to Privacy-preserving Data Publishing : Concepts and Techniques. Chapman & Hall/CRC.
Ganta, S.R., Kasiviswanathan, S.P., Smith, A., 2008. Composition attacks and auxiliary information in data privacy. In: Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining – KDD 08, 2008, pp. 265.
Goethals, B., Laur, S., Lipmaa, H., Mielikäinen, T., 2004. On private scalar product computation for privacy-preserving data mining. Secur. Cryptol.
Grandison, T., 2007. Privacy in eHealth.
Hann, I.-H., Hui, K.-L., Lee, S.-Y., Png, I., 2007. Overcoming online information privacy concerns: an information-processing theory approach. J. Manage. Inf. Syst. 24 (2), 13–42.
Hill, J.W., Powell, P., 2009. The national healthcare crisis: Is eHealth a key solution? Bus. Horiz. 52 (3), 265–277.
Hossain, M.S., Muhammad, G., 2016. Cloud-assisted Industrial Internet of Things (IIoT) – Enabled framework for health monitoring. Comput. Networks 101, 192–202.
Hsu, J., H. DL, A. DE, B. RJ, B. JM, R. MD, Ortiz, E., 1999. Use of e-Health Services between 1999 and 2002: a growing digital divide. J. Am. Med. Informatics Assoc. 12 (2), 164–171.
Jian-min, H., Hui-qun, Y., Juan, Y., Ting-ting, C., 2008. A Complete (alpha,k)-Anonymity Model for Sensitive Values Individuation Preservation. In: 2008 International Symposium on Electronic Commerce and Security, 2008, pp. 318–323.
Klosgen, W., 1995. Anonymization techniques for knowledge discovery in databases. In: Proc. 1st Int. Conf. Knowl. Discov. Data Min.
Lederer, S., Mankoff, J., Dey, A.K., 2003. Who wants to know what when? Privacy preference determinants in ubiquitous computing. In: CHI '03 extended abstracts on Human factors in computing systems - CHI '03, pp. 724.
LeFevre, K., DeWitt, D., Ramakrishnan, R., 2008. Workload-aware anonymization techniques for large-scale datasets. ACM Trans.
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, 2007, pp. 106–115.
Li, T., Li, N., Zhang, J., Molloy, I., 2012. Slicing: a new approach for privacy preserving data publishing. Knowl. Data.
Li, J., Liu, J., Baig, M., Wong, R.C.-W., 2011. Information based data anonymization for classification utility. Data Knowl. Eng. 70 (12), 1030–1045.
Lunshof, J.E., Chadwick, R., Vorhaus, D.B., Church, G.M., 2008. From genetic privacy to open consent. Nat. Rev. Genet. 9 (5), 406–411.
Machanavajjhala, A., Kifer, D., Gehrke, J., 2007. l-diversity: Privacy beyond k-anonymity. Discov. from Data.
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.
Masoumzadeh, A., Joshi, J., 2012. Preserving structural properties in edge-perturbing anonymization techniques for social networks. IEEE Trans. Dependable.
Muntés-Mulero, V., Nin, J., 2009. Privacy and anonymization for very large datasets. Proc. 18th ACM Conf.
Osuna, E., Freund, R., Girosi, F., 1997. An improved training algorithm for support vector machines. In: Neural Networks for Signal Processing VII. Proceedings of the 1997 IEEE Signal Processing Society Workshop, pp. 276–285.
Pedersen, T., Saygın, Y., Savaş, E., 2007. Secret charing vs. encryption-based techniques for privacy preserving data mining.
Pinkas, B., 2002. Cryptographic techniques for privacy-preserving data mining. ACM Sigkdd Explor. Newsl.
Proceedings of the 8th International Conference on Collaborative Computing: Networking, Applications and Worksharing : CollaborateCom 2012, October 14–17, Pittsburgh, United States. Institute for Computer Sciences, Social Informatics, and Telecommunications Engineering, 2012.
Rass, S., Slamanig, D., 2013. Cryptography for Security and Privacy in Cloud Computing. Artech House.
Ristenpart, T., Maganis, G., Krishnamurthy, A., 2008. Privacy-preserving location tracking of lost or stolen devices: cryptographic techniques and replacing trusted third parties with DHTs. Usenix Secur.
Rolim, C.O., Koch, F.L., Westphall, C.B., Werner, J., Fracalossi, A., Salvador, G.S., 2010. A cloud computing solution for patient's data collection in health care institutions. In: 2010 Second International Conference on eHealth, Telemedicine, and Social Medicine, pp. 95–99.
Shah, M.A., Swaminathan, R., Baker, M., 2008. Privacy-preserving audit and extraction of digital contents privacy-preserving audit and extraction of digital contents *.
Sheehan, K.B., 2002. Toward a typology of internet users and online privacy concerns. Inf. Soc. 18 (1), 21–32.
Sheehan, K.B., Hoy, M.G., 1999. Flaming, complaining, abstaining: how online users respond to privacy concerns. J. Advert. 28 (3), 37–51.
Sweeney, L., 2000. Simple demographics often identify people uniquely. Health (San Francisco) 671, 1–34.
Sweeney, L., 2002b. Achieving k-anonymity privacy protection using generalization and suppression. Int. J Uncertainty, Fuzziness.
Sweeney, L., 2002c. k-anonymity: a model for protecting privacy. Int. J Uncertainty, Fuzziness.
Sweeney, L., 2002a. k-anonymity: a model for protecting privacy. Int. J. Uncertainty, Fuzziness Knowledge-Based Syst. 10 (5), 557–570.
Vercellis, C., 2011. Business Intelligence: Data Mining and Optimization for Decision Making. Wiley.
Wang, K., Yu, P., Chakraborty, S., 2004. Bottom-up generalization: a data mining solution to privacy protection. Data Mining, 2004. ICDM'04.
Wernke, M., Skvortsov, P., Dürr, F., Rothermel, K., 2014. A classification of location privacy attacks and approaches. Pers. Ubiquitous Comput. 18 (1), 163–175.
Young, A.L., Quan-Haase, A., 2013. Privacy protection strategies on facebook. Inf. Commun. Soc. 16 (4), 479–500.
Zhang, R., Liu, L., 2010. Security models and requirements for healthcare application clouds. In: 2010 IEEE 3rd International Conference on Cloud Computing, pp. 268–275.
Zhou, B., Pei, J., Luk, W., 2008. A brief survey on anonymization techniques for privacy preserving publishing of social network data. ACM Sigkdd Explor. Newsl.