Sustainable Cities and Society 40 (2018) 326–334

Contents lists available at ScienceDirect

Sustainable Cities and Society

journal homepage: www.elsevier.com/locate/scs

Privacy preserving data by conceptualizing smart cities using *MIDR*-Angelization

Adeel Anjum[a], Tahir Ahmed[a], Abid Khan[a], Naveed Ahmad[a], Mansoor Ahmad[a], Muhammad Asif[b], Alavalapati Goutham Reddy[c,*], Tanzila Saba[d], Nayma Farooq[a]

[a] Department of Computer Science, Comsats Institute of Information Technology Islamabad, Pakistan
[b] Ernst & Young, Milan, Italy
[c] Department of Computer & Information Security, Sejong University, Seoul 05006, South Korea
[d] College of Computer and Information Sciences, Prince Sultan University, Riyadh, Saudi Arabia

ARTICLE INFO

Keywords:
Big data
IoT data management
Disclosure risk
HIPAA
Patient privacy
Re-identification risk
Smart city

ABSTRACT

Smart City and IoT improves the performance of health, transportation, energy and reduce the consumption of resources. Among the smart city services, Big Data analytics is one of the imperative technologies that have a vast perspective to reach sustainability, enhanced resilience, effective quality of life and quick management of resources. This paper focuses on the privacy of big data in the context of smart health to support smart cities. Furthermore, the trade-off between the data privacy and utility in big data analytics is the foremost concern for the stakeholders of a smart city. The majority of smart city application databases focus on preserving the privacy of individuals with different disease data. In this paper, we propose a trust-based hybrid data privacy approach named as "*MIDR*-Angelization" to assure privacy and utility in big data analytics when sharing same disease data of patients in IoT industry. Above all, this study suggests that privacy-preserving policies and practices to share disease and health information of patients having the same disease should consider detailed disease information to enhance data utility. An extensive experimental study performed on a real-world dataset to measure instance disclosure risk which shows that the proposed scheme outperforms its counterpart in terms of data utility and privacy.

1. Introduction

Numerous technological advancements of smart meters such as the sensors, Internet of Things (IoT), smart cards etc as a part of the smart city electronic devices, are used to share the different type of information. Hence these technologies and equipment facilitate us to be smarter and make diverse features of smart cities more reachable and pertinent. Smart City is a city, based on information combined with its operational infrastructure to help; efficient management of resources enhanced decision making and being more practical toward the important incidents (Harrison and Donnelly, 2018). In different components of the smart city, the information is shared, analyzed and processed. Examples can be seen in smart health organizations which greatly facilitate the medical research, as well as they, help to provide better healthcare facilities. This information is used for knowledge-based decision making and for creating new research opportunities. A Large amount of data is used for analysis, computation and for statistical inference to make decisions and gather information for social and human development. But this shared data can also reveal the privacy of individuals. According to a research that was done in the United States, the ratio of identified individuals from a publicly published data was extremely high as 87% were identified (Ljiljana, Islam, & Giggins, 2007). The challenge is to preserve the privacy of individuals in the data shared by the smart city without compromising on data utility. To overcome this issue,

various privacy models and algorithms have been proposed. A well-known model proposed by Sweeney (Sweeney, 2002) called *k*-anonymity is a good example of the statistical standard approach. This approach focuses on the quasi-identifiers that are defined by Safe Harbour rule and demographic attributes like age, gender and zip code. The values of QI attributes are available publicly in form of voter list or census data. From this publicly available data, the published data can be identified which cause the major privacy disclosure. To overcome this privacy disclosure k-anonymity privacy model generalizes the QI attribute values such that an individual is unable to distinguish by k-1 other individuals. This way is better to protect an individual's

* Corresponding author.
E-mail address: goutham.ace@gmail.com (A.G. Reddy).

https://doi.org/10.1016/j.scs.2018.04.014
Received 30 October 2017; Received in revised form 31 January 2018; Accepted 7 April 2018
Available online 22 April 2018

information but had a drawback of information loss.

A case study proves that big data is managing several privacy issues in smart health component. Health must be provided in an adapted manner as it is a part of the important infrastructure in a smart city. In an IoT framework, there are several ways to collect patient's data such as cloud distributed systems, Wireless Sensor Networks, smart cards etc. Due to the cloud distributed systems usage, the maintenance cost of smart health organizations has diminished. Hence the cloud services play a crucial role in smart health services. The main concern in both IoT and the Cloud is to keep up the privacy and trust of smart health infrastructure. As the health data consist of individual's private and sensitive information, which may lead to privacy risks.

Most of the time in the medical field, data is collected to study about a specific disease. In that case, it's obvious that all the patients in the data set having the same disease. This is known as "Same-Disease micro-data". Analysis of the same disease microdata is the most common problem in medical research. Our goal is to ensure the data privacy and utility during data publication. To have a clear picture of same disease microdata consider the following example. For instance, cancer registry (Kingsley, Schmeichel, & Rubin, 2007), diabetes cohort studies (Van Dam, Willett, Manson, & Hu, 2006) and registry of HIV patients (Rabeneck et al., 2001). When sharing same-disease data privacy of an individual can be compromised in case of, if the adversary knows that individual's information is part of microdata. This risk is very high as all patients having the same sensitive attribute. This privacy risk is known as "Instance disclosure risk". Another type of privacy risk that found in same-disease microdata for an individual is a chance to get a correct match of an individual's record in population data. This type of risk is known as "re-identification risk" (Golle, 2006). The Scheme proposed by Xiaoping (Liu et al., 2016) provides a way to calculate accurate instance disclosure risk and an algorithm to provide security against instance disclosure risk.

Xiaoping (Liu et al., 2016) proposed a formula to calculate actual instance disclosure risk and an algorithm which reduces the privacy risk (referred to as instance disclosure risk) by dividing the data into classes or groups having minimum instance disclosure risk and also generalize the QI values. In data privacy, this technique is referred to as "*MIDR*-Generalization" (Minimum Instance Disclosure Risk- Generalization). According to above example, they are not publishing the sensitive attribute information which causes a utility loss as without sensitive information researchers are unable to work further regarding research or data mining tasks etc. If we assume that given published data is about cancer patients then we can say that there are several types of cancer. Which type of cancer found in maximum individuals is a question mark in this situation? We have purposed a hybrid scheme called "*MIDR*-Angelization" to protect the identification of any individual's record in microdata. We reduced the instance disclosure risk by using random sampling (Chaudhuri and Mishra, 2006) and Angelization (Tao, Chen, Xiao, Zhou, & Zhang, 2009) technique and enhanced the data utility by publishing the sensitive attribute information and specified the sensitive attribute with given characteristic to differentiate the specific disease of each patient of same disease microdata. The proposed approach is named as MIDR-Angelization which will assure the privacy and utility when sharing same disease data of patients. We have considered two datasets i.e. Population Dataset and Same Disease microdata.

To sum up, our main contributions are as follows:

- We thoroughly investigated the problem of data privacy in the context of same disease data. Specifically, we highlighted the drawbacks in the literature that focuses on same disease data.
- We proposed a novel privacy model, named *MIDR-Angelization*, that effectively caters the attacks on individuals with same disease data.
- We prove that our proposed privacy model substantially reduces the *so-called instance disclosure risk* as compared to its counterparts.
- Furthermore, we proposed an effective algorithm that follows MIDR-Angelization and provides an efficient privacy vs. utility tradeoff.

The rest of the paper is structured as follows: In Section 2, related work concerning previous privacy models and algorithms is illustrated. Section 3 is about preliminaries. The proposed scheme is explained in Section 4. The experimental results and analysis are demonstrated in Section 5. The paper is concluded in Section 6 along with some future work directions in this area of research.

2. Related work

There are different privacy preservation techniques and privacy models like Generalization (Sweeney, 2002; Garfinkel, Gopal, & Thompson, 2007), Suppression (Sweeney, 2002), *l*-diversity (Gehrke, Kifer, & Venkitasubramaniam, 2006), and *m*-invariance (Xiao and Tao, 2007) that are used to preserve individual's privacy in case of different disease, but these schemes do not provide sufficient privacy guarantee in case of same disease data. There are some other approaches like noise base perturbation (Liew, Choi, & Liew., 1985; Li and Sarkar, 2013) and data swapping (Dalenius and Reiss, 1982; Li and Sarkar, 2011) that add noise into data or exchange the values of records to anonymize the data to preserve individual's privacy. Therefore, such approaches are needed to implement in a different way to achieve sufficient privacy level and maintain the data utility in case of same disease data.

There are two main categories of privacy models which provide privacy and utility based on attacks. In the first category, models that deal with attribute linkage attacks; in which attacker can link different published data tables attributes to get sensitive information of an individual. This type of attack is known as attribute linkage attack in which it is assumed that attacker has some background knowledge which helps the attacker to link specific individual's sensitive information [X. B. Li, 2011; Anjum and Raschia, 2017; Anjum et al., 2017; Anjum et al., 2018; Khan et al., 2017; Moqurrab et al., 2017; Nasir et al., 2017]. In the second category, the models reside that deal with probabilistic attacks. In these types of attacks, the attacker has some additional information with background knowledge. The attacker has some posterior and prior beliefs about data and attacker get some information from published data that helps the attacker to infer sensitive information. But our research will focus on attacks to preserve privacy in case of same disease data.

There are two major attack types in same-disease data that can be performed by an attacker to identify the accurate sensitive information of an individual. First one is instance disclosure attack which specifies the presence of an instance in the dataset which results in the privacy breach of individual's private information i.e. disease. The second one is the identity disclosure or re-identification attack. In this type of attack, an adversary can re-identify an individual in the published data by using correlation.

In 1997 Sweeney and Samarati found that a record can be identified from published data even if their unique identifiers are removed (Barth-Jones, 2012). To prove their belief, they identified the information of governor of Massachusetts from a medical data from which unique identifiers are removed. After that, they got proof that the approach of "removal of unique identifiers from the published data" is not a good countermeasure to preserve data from the adversary. Because other attributes in the dataset which is known as QI attributes may help the adversary to link different datasets to get sensitive information published by different publishers. There are also some data quality measures (Pipino, Lee, & Wang, 2002; Madnick, Lee, Wang, & Zhu, 2009) that were proposed to assess data and quality of information.

3. Preliminaries

Let D be the microdata that needs to be published. D contains h Quasi-identifier (QI) attributes {qi1, qi2 … qih} and a sensitive attribute *sa*. Following the assumption, each QI attribute i.e. qi_i ($1 \leq i \leq h$) can be either categorical or numerical but Bs should be categorical. For any tuple $t \in D$, we denote t[i] ($1 \leq i \leq h$) as the *QI* value of t and $t[h+$