DOUBLE-PHASE MICROAGGREGATION FOR DE-IDENTIFICATION OF BIOMEDICAL DATA

G.RAMYA (ME), A.SELVARAJ (ME), N.ANANDH (ME)
Department of Computer Science and Engineering,
Muthayammal Engineering College, INDIA.
ramya14be@gmail.com, engineerselvaraj@gmail.com, anandhme1983@gmail.com

Abstract- The concept of Double-phase microaggregation is used as an improvement to classical microaggregation. This concept is applied in the context of mobile health in distributed scenarios without fully trusted third parties. The distributed architecture consists of patients and several intermediate entities whose data are released to third parties for secondary use. The distributed architecture allows private gathering, storing, sharing of biomedical data. The mobile device like cell phones is used to monitor the patient health. The double-phase microaggregation properly fits the need of privacy preservation of biomedical data. The location and time based approach is also considered to enhance the protection of privacy.

***Keywords*- Privacy protection, mobile health, microaggregation.**

I. INTRODUCTION

Mobile communications are experiencing a tremendous development that leads to new ways of providing healthcare services, the so-called mobile health. The biomedical data storage and analyzes the patient details for various processes like research. M-health is used for storing and analyzing the biomedical data. To enhance the security of biomedical data, the data should be encrypted. The number of encryption technique has been introduced. But attackers break the key and get the original data. The biomedical data transferred from one location to another location. The numbers of third party entities are present during the transfer of information. M-health services in three main aspects: (a) The ubiquity of mobile devices allows services to be accessed everywhere, anytime. As a result, data could be collected more easily regardless of the location of patients. (b) m-Health is patient-oriented. Patients play a key role in an m-health service, because in most cases they are responsible for the remote control of the service and (c) m-Health is personalized. Patients receive customized services that fit their specific needs. M-Health and, in general, e-health significantly contribute to the efficiency and immediacy of healthcare services and, as a result, to the treatment quality received by patients. Last but not least, the generalization of m-health can drastically increase the amount of biomedical data that can be collected, stored and analyzed.

The mobile device is attached to proper gadgets is used to monitor variables such as heartbeat rate or blood pressure of patients constantly. The new ability of gathering huge amounts of personal, highly sensitive data opens the door to use diverse information coming from multiple sources. According to privacy legislation, patients give their consent to allow researchers to study their data for secondary use. Unfortunately, it has been observed that requiring explicit consent for the participation in different forms of health research can negatively affect the process and out- comes of the very research. In fact, people who consent tend to be different from those who decline. Hence, the results of the studies can be biased. With the aim to alleviate this problem, many research ethics boards will waive the consent requirement if the data used are deemed to be de-identified. As a result, there is an increasing need for methods that allow the gathering and de-identification of biomedical data for secondary use.

A. Statistical Disclosure Content

Protecting individual privacy is paramount for many institutions, namely statistical agencies, healthcare centers, Internet companies, manufacturers, etc. Many efforts have been devoted to develop techniques that assure a given degree of privacy to the people, whose data are collected and shared. Statistical disclosure control (SDC) was the first to consider the problem; initially on tabular data, and later on microdata (i.e., data from individuals).

The protection of individual privacy by means of protecting their microdata. So, it aims at avoiding the re-identification of individuals through their released microdata. To achieve this goal microdata sets have to be properly modified prior to their publication. The degree of modification varies between two extremes: (i) encrypting the microdata and, (ii) leaving the microdata intact.

The utility of the data is almost nonexistent because the encrypted micro- data can be hardly studied or analyzed by others than the owner of the decryption key. The microdata are totally useful. However, the privacy of the individuals is endangered because sensitive private data can be seen without limitation.

SDC methods aim at distorting the original microdata sets to protect individuals privacy and avoid their re-identification while maintaining some of the statistical properties of the data and minimizing the information loss as much as possible.

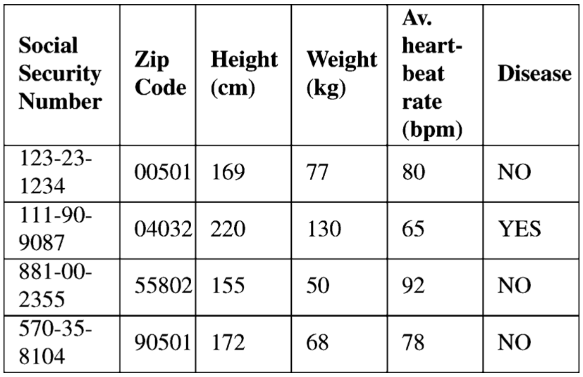
B. Basic Definitions and Concepts

Some basic definitions related to the field of statistical disclosure control that are used through the paper:

Microdata In opposition to macrodata, that refers to large aggregates of information generally represented in tables, microdata refers to individual data such as the social security number (SSN), age, ethnicity, height, income, etc. that are represented with records. In our example shown in Table I, each row represents a microdata record.

TABLE I

Example Of Microdata Set With Four Records And Six Attributes



Social Security Number	Zip Code	Height (cm)	Weight (kg)	Av. heart-beat rate (bpm)	Disease
123-23-1234	00501	169	77	80	NO
111-90-9087	04032	220	130	65	YES
881-00-2355	55802	155	50	92	NO
570-35-8104	90501	172	68	78	NO

Microdata set. A microdata set is the union of microdata records sharing the same attributes. Thus, a microdata set is understood as a twodimensional matrix in which rows represent individual data and columns represent specific attributes. Table I is an example of a microdata set with six attributes belonging to four individuals.

Identifiers. Those attributes in a microdata set that point out to a unique individual are called identifiers. In our ex- ample the social security number (SSN) is an identifier. (This attributes are deleted before releasing microdata for public use).

Quasi-identifiers. Those attributes containing information about an individual that, when taken individually, do not identify him/her. Note that, the combination of quasi-iden- tifiers might lead to the identification of a unique indi- vidual. Examples of this kind of attribute in Table I could be zip code, weight and height (e.g., a very tall person in a small village could be easily identified).

Confidential outcome attributes. Those attributes that have sensitive information like religion, salary, health con- dition, etc. In our example the existence of a given disease is a confidential outcome attribute.

C. The Microaggregation Problem

To protect a microdata set it is essential to remove all identifiers before releasing the data. Thus, attributes like the SSN or the full name should be deleted. However, removing identifiers is not enough protection because the combination of quasi-identifiers might lead to the re-identification of individuals. For example, consider the case of two attributes (Occupation and Town). If these attributes were kept untouched in the micro- data set, it would be very simple to identify the doctor of a little village because, in general, there is a single doctor in a small village and everybody knows him/her.

To solve this problem, a variety of techniques have been proposed, namely noise addition, rank swapping, generalization and deletion, microaggregation, etc

Microaggregation is a technique that protects the privacy of individuals by aggregating similar records and producing microaggregated data sets satisfying the property of k-anonymity. Many classifications can be applied to microaggregation techniques, however, for the sake of brevity we classify them in two classes: (i) fixed-size microaggregation and, (ii) variable-size microaggregation. The former builds clusters of a fixed number of elements k, while the latter generates clusters with variable number of records in the range [k,2k-1]. A classical, well-known fixed size microaggregation is the maximum distance to average vector (MDAV). Regarding variable size methods, amongst other alternative we find the variable maximum distance to average vector (V-MDAV).

D. Basics of Public key Cryptography

Public key cryptography, also known as asymmetric cryptography, refers to cryptographic systems that require a pair of different keys to operate: one key (the public key) is used to encrypt messages and, the other key (the private key) is used to decrypt them. The public key can be released so as to allow everyone to know it, while the private key is kept secret and is only known by its owner. In this way, private communications are possible because everyone can use the known public keys to encrypt messages and send them to the owner of the corresponding private key, who will be the only one able to decrypt.

II. PROPOSED ARCHITECTURE

The proposed architecture consists of four main actors namely mobile devices (patients), health centers (HC), research centers and a centralized storage and aggregation server (SAS). These actors/entities interact so as to guarantee the private collection and sharing of data.

Patients have mobile devices with communication capabilities able to collect data and encrypt them by using a public key cryptosystem. Each healthcare center have pair of keys {K_P, K_{PR}} of public key cryptosystem. Each patient is assigned to a healthcare center that is responsible for the patient and his/her data. Patients know the public key (K_P) of the healthcare center to which they are assigned, and they share the same cryptosystem.

Fig. 2 graphically depicts the different actions carried out by each actor of the system to gather and share data privately. First mobile devices collect biomedical data from patients (See step1 of Fig. 2).

We refer to these data as,

$$D = (d_{u1}, d_{u2},\ldots\ldots\ldots,d_{ui},\ldots\ldots.d_{un})$$

where d_{ui}, represents the data of user/patient $_{ui}$.

After collecting the data, the mobile device of the patient encrypts the gathered data with the public key of the healthcare center (HC_i) to which he/she is assigned (See step 2 of Fig. 2) and generates a message (m) with the following :

- User ID
- Healthcare Center ID
- Encrypted data

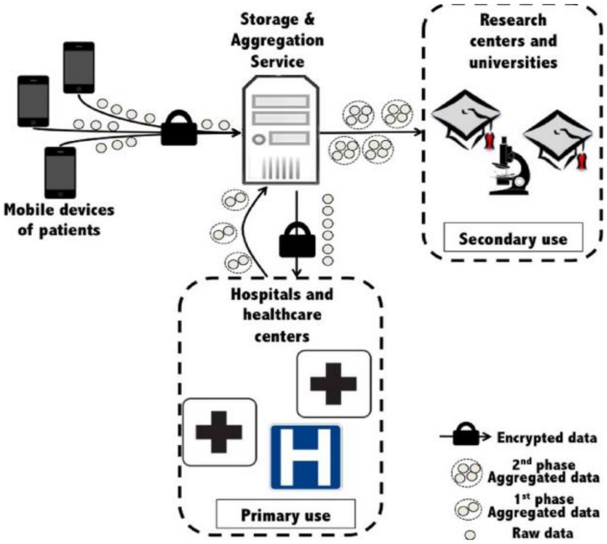
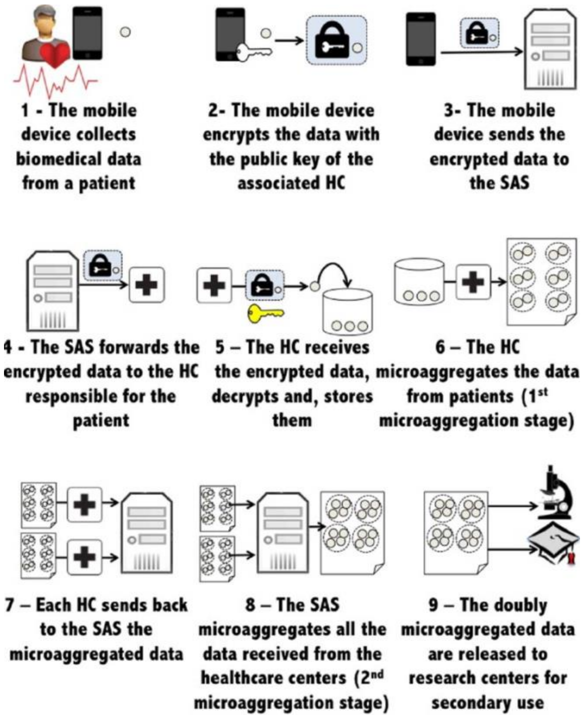


Fig. 1. Graphical scheme of the architecture and the double-phase microaggregation model. It can be observed that patients send data by means of mobile devices to a storage and aggregation service (SAS). Healthcare centers (HC) and research centers (RC) can access those data


1 - The mobile device collects biomedical data from a patient

2- The mobile device encrypts the data with the public key of the associated HC

3- The mobile device sends the encrypted data to the SAS

4 - The SAS forwards the encrypted data to the HC responsible for the patient

5 – The HC receives the encrypted data, decrypts and, stores them

6 – The HC microaggregates the data from patients (1st microaggregation stage)

7 – Each HC sends back to the SAS the microaggregated data

8 – The SAS microaggregates all the data received from the healthcare centers (2nd microaggregation stage)

9 – The doubly microaggregated data are released to research centers for secondary use

Fig. 2. Graphical description of the protocol and the flow of data within the proposed architecture.

After encrypting the data, the mobile device sends m to the storage and aggregation server (SAS) (see step 3 of Fig. 2). Once the SAS receives the patients' data, it checks the healthcare center ID of each message and forwards it to right HC_i (see step 4 of Fig. 2). After receiving and decrypting the data from the patients (see step 5 of Fig. 2), each HC microaggregates all the data (by using a microaggregation algorithm such as MDAV or V-MDAV) with a given security parameter HC (k) (see step 6 of Fig. 2) and sends the resulting microaggregated data set back to the SAS (see step 7 of Fig. 2). When the SAS receives the microaggregated data sets from all healthcare centers, it merges them all and microaggregates them again by using again a microaggregation algorithm such as MDAV or V-MDAV, with a given security parameter SAS(k) (see step 8 of Fig. 2).

III. RESULTS

Analysis of Correlations Preservation

The analysis of correlations between all the attributes in the data set reveals that they are all preserved and the use of double- phase microaggregation does not affect the correlations present in the original data set more than classical microaggregation.

IV. CONCLUSION

The concept of double-phase microaggregation is to limit the information accessible by intermediate entities such as the SAS. The MDAV and V-MDAV perform similarly over biomedical data set. The effect of double-phase microaggregation over the IL is negligible with respect to the classical microaggregation and that it preserves the correlations of the original data set. Then it concludes that the distributed double-phase microaggregation proposed can be applied in distributed environment to protect the privacy of individuals with the same effects of classical microaggregation. The encryption technique protects the data from the third party. Although this concept is applied over biomedical data, and method proposed might be smoothly applied to other fields like economics, tourism, energy, and so on. It also includes the analysis of the influence of time in the series of data collected using this model. It is possible that the use of time information might help attackers to get extra knowledge and increase the probability of re-identification. Also, the use and protection of location data should be considered.

REFERENCES

[1] Agusti Solanas and Antoni Mart´ınez-Ballest´, "V-MDAV: A Variable Group Size with multivariate microaggregation" Rovira i Virgili University. Av.Pa¨ısos Catalans 26. 43007 Tarragona. Catalonia

[2] Chris Skinner, "Statistical Disclosure Control for Survey Data" Univ. of Southampton

[3] J. Domingo-Ferrer et al., "Statistical Disclosure Risk & Statistical Disclosure Control"

[4] J. Domingo-Ferrer, F. Sebé, and J. Castellà, "On the security of noise addition for privacy in statistical databases," *Lecture Notes in Computer Sci.*, vol. 3050, pp. 149–161, 2004

[5] Josep Domingo-Ferrer,Antoni Martínez-Ballesté, Josep Maria Mateo-Sanz,Francesc Sebé, "Multivariate microaggregation with variable group size"

[6] Latanya Sweeney, k-Anonymity: "A Model for protecting privacy", School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA

[7] M. Naehrig, K. Lauter, and V. Vaikuntanathan, "Can homomorphic encryption be practical?," in Proc. 3rd ACM Workshop on Cloud Computing Security Workshop (CCSW'11), New York, NY, USA, pp113–124

[8] Matthias Templ, “Data Access and Personal Privacy: Appropriate Methods of Disclosure Control”, Department of Methodology, Statistics Austria, Guglgasse 13, 1110 Vienna, Austria

[9] Oleg Chertov and Anastasiya Pilipyuk, “Statistical Disclosure Control Methods for Microdata”, National Technical University “Kyiv Polytechnic Institute”, Kyiv, Ukraine

[10] Pierangela Samarati, “Protecting Respondents’ Identities in Microdata Release”

[11] R.L. Rivest, A. Shamir, and L. Adleman,” A Method for Obtaining Digital Signatures and Public-Key Cryptosystems”

[12] T. ElGamal, “A public-key cryptosystem and a signature scheme based on discrete logarithms,” IEEE Trans. Inf. Theory, vol. 31, no. 4, pp. 469–472, Jul. 1985