Double-phase Microaggregation for Protecting Bio-medical Data in Mobile Health

S. Saranya, M. Madan mohan, X. Rexeena

Abstract— **Double phase microaggregation is used for protecting biomedical data.It provide the privacy for patient without fully trusted party. Microaggregation perform using V-MDAV algorithm over patient data's then send the patient data to third party.so no one can reveal the patient data's. Distributed architecture that allows the private gathering, storage, and sharing of biomedical data. Before microaggregation perform encryption and decryption operation using elliptic curve cryptography. Double phase microaggregation reduce the information loss, disclosure risk and preserve the correlation using correlation analysis.**

Keywords—**Privacy protection, distributed environments, microaggregation,
Mobile health.**

I. INTRODUCTION

Due to the recent advances in information and communication technologies the gathering, storage and sharing of data are becoming simpler and faster than ever. In this regard, one of the most significant break throughs is the use of mobile devices (e.g., cell phones) to monitor patients. Mobile communications are experiencing a tremendous development that leads to new ways of providing healthcare services, the so-called mobile health (m health). m-Health affects the way we understand healthcare services in three main aspects: (a) m-Health simplifies the access to classical and new services. 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 data that can be collected, stored and analyzed. Thanks to commonly used mobile devices such as cell phones attached to the proper gadgets, it is possible to monitor variables such as the heartbeat rate or the blood pressure of patients easy gathering of data does not directly translate into the easy access to and use of those data by researchers. As a result, there is an increasing need for methods that allow the gathering and de-identification of biomedical data for secondary use.

II. STATISTICAL DISCLOSURE CONTROL

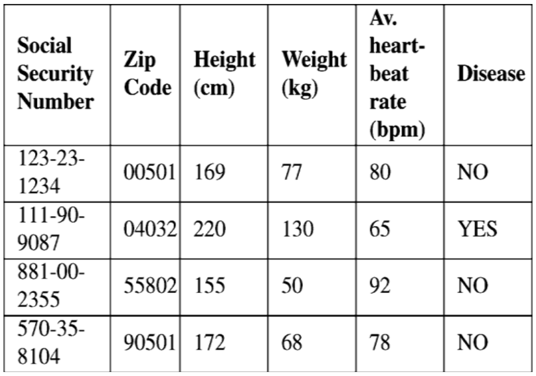
Protecting individual privacy is paramount for many institutions, namely statistical agencies, healthcare centers, Internet companies, manufacturers, etc.Among all these disciplines, statistical disclosure control (SDC) [17], [18],[9] was the first to consider the problem; initially on tabular data, and later on microdata. Protection of individual privacy by means of protecting their microdata. By doing so, we aim 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.

III. BASIC DEFINITIONS AND CONCEPTS

• **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 Table1, each row represents a microdata record.

Table1: 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 two-dimensional matrix in which rows represent individual data and columns represent specific attributes. Table 1 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 example the social security number (SSN) is an identifier.

• **Quasi-identifiers**.
Those attributes containing information about an individual that, when taken individually, do not identify him/her. Note that, the combination of quasi-identifiers might lead to the identification of a unique individual. 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 condition, etc. In our example the existence of a given disease is a confidential outcome attribute.

Proposed technique

Noise addition[1],[3]
Rank swapping[2],[7]
Generalization and Deletion

IV. MICROAGGREGATION

Microaggregation is a technique that protects the privacy of individuals by aggregating similar microdata records and producing microaggregated data sets satisfying the property of K-anonymity [15], [15].

Types of microaggregation technique

1)Fixed size Microaggregation[4]

Fixed number of elements and use MDAV Algorithm.

2)Variable size Microaggregation[14]

Variable number of elements and use VMDAV algorithm.

V. 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 and read the messages. Given a message and a pair of keys, the sender encrypts by using ,and the receiver uses to decrypt the message and obtain the message.Well known public key cryptosystem is ECC[11],[12].

VI. METHOD

The use of information and communication technologies and, more specifically, mobile devices to monitor patients has significantly increased the chances of researchers to gather unprecedented amounts of timely data from multiple sources. However, in the context of m-health, if the data are not de-identified, patients have to consent that their data are released for secondary use. It has been shown that consent requirement can introduce bias in the analyses and their results [5]. Thus, it would be desirable to have architectures and methods allowing the automatic de-identification of distributed biomedical data to avert the need for patients consent and the associated bias. The following properties should be fulfilled by the architecture:

• **The architecture is not fully trusted**. The raw biomedical data sent by patients through the architecture should not be seen by any entity other than an authorized healthcare center.

• **The resulting microdata sets should guarantee the privacy of patients** by assuring K-anonymity, where the security parameter could be determined according to the specific needs of each microdata set. The disclosure risk should be kept low.

• **The released data set should be useful**. The information loss caused by the distributed microaggregation procedure should be low, and correlations between attributes should be maintained

VII. PROPOSED ARCHITECTURE

The proposed architecture considers four main actors, namely mobile devices (patients), healthcare centers (HC), research centers and a centralized storage and aggregation server (SAS) (Fig. 1 for a graphical representation of the architecture). These actors/entities interact so as to guarantee the private collection and sharing of data.

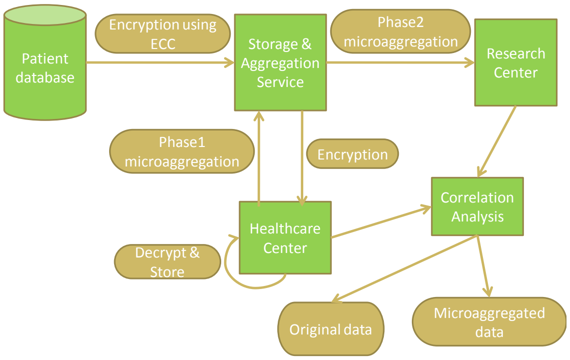


Fig 1:Architecture of double phase microaggregation

Patients have mobile devices with communication capabilities able to collect data and encrypt them by using a public key cryptosystem (e.g., ECC[11],[12]).

- The data collected from patients have the same attributes for all patients participating in a given study/trial. Note that this is quite common because clinical studies tend to be highly parametrized and strict with the treated variables.
- Each healthcare center (doctors) has a pair of keys of a public key cryptosystem(eg., ECC[11],[12]).
- Each patient is assigned to a healthcare center that is responsible for the patient and his/her data. Although the interaction of patients with healthcare centers is kept to the minimum to maximize the benefits of using ICT, in most studies, regular supervision is required.
- Patients know the public key of the healthcare center to which they are assigned, and they share the same cryptosystem

.

VIII. RESULTS AND DISCUSSION

The proposed architecture guarantees the private exchange of data between patients and healthcare centers through the SAS thanks to the use of public key cryptography. In this regard, even if the SAS is not a trusted party the architecture remains private. However, to protect the privacy of the patients with respect to research centers, public key cryptography is not suitable and we propose the use of our double-phase microaggregation procedure. With the aim to assess the usefulness of the proposed double phase microaggregation we have performed the following experiments:

- Compare the performance of a fixed-size microaggregation algorithm (MDAV)[4],a variable size microaggregation algorithm (V-MDAV)[14], and a hybrid microaggregation algorithm based on genetic algorithms and MDAV (GA)[15] over a biomedical data set.
- Compare the performance in terms of information loss and disclosure risk of single-phase microaggregation and double-phase microaggregation based on MDAV and V-MDAV.
- Compare the ability of preserving correlations of single phase and double-phase microaggregation

Table 2-Correlations analysis between the quasi-identifiers and attributes for the proposed methods

Original data set	**Age**	**Weight**	**Height**
Age	**1**	**-0.11437**	**-0.17760**
Weight	**-0.11437**	**1**	**0.47084**
Height	**-0.17760**	**0.47084**	**1**
V-MDAV-single(K=4)	**Age**	**Weight**	**Height**
Age	**1**	**-0.11518**	**-0.17957**
Weight	**-0.11518**	**1**	**0.47592**
Height	**-0.1795**	**0.47592**	**1**

IX. CONCLUSION

Proposed architecture allows the private gathering and sharing of biomedical data in the context of m-health. Concept of double-phase microaggregation to limit the information accessible by intermediate entities and V-MDAV algorithm perform microaggregation over biomedical data set. Before microaggregation perform encryption using ECC. Double-phase microaggregation reduce the information loss and that it preserves the correlations of the original data set. Then, we can conclude that the double-phase microaggregation proposed can be applied to protect individual patient privacy.Although we have applied our proposal over biomedical data, the architecture and method proposed might be smoothly applied to other fields like economics, tourism, energy, and so on. Further research might include the analysis of the influence of time in the series of data collected using our 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 in the future.

REFERENCES

[1] "Brand R.Microdata protection through noise addition," *Lecture Notes in Computer Sci.*, vol. 2316, pp. 97–116, 2002.
[2] T. Dalenius and S. P. Reiss, "Data swapping: A technique for disclosure control," *J. Statistical Planning and Inference*, vol. 6, no. 1, pp. 73–85, 1982.
[3] J. Domingo-Ferrer, F. Sebé "On the security of noise addition for privacy in statistical databases," *Lecture Notes in Computer Sci.*, vol. 3050, pp. 149–161, 2004.
[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] K. Emam *et al.*, "Globally optimal k-anonymity for de-identification of health data," *J. Amer. Med. Inform. Assoc.*, vol. 16, no. 5, pp. 670–682, 2009.
[6] 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.
[7] B. Greenberg, Rank Swapping for Masking Ordinal Microdata Tech. report. U.S. Bureau of the Census, 1987, unpublished.
[8] 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, pp. 113–124.
[9] G. J. Matthews and O. Harel, "Data confidentiality: A review of methods for statistical disclosure limitation and methods for assessing privacy," *Statist. Surveys*, vol. 5, pp. 1–29, 2011.
[10] D. Pagliuca and G. Seri, Some Results of Individual Ranking Method on the System of Enterprise Accounts Annual Survey, Esprit SDC Project, 1998, Deliverable MI3/D2.
[11] Stallings, William. Cryptography and Network Security: Principles and Practice, Second Edition. New Jersey: Prentice Hall Inc., 1999.
[12] Crutchley, Duncan Alexander. "Cryptography And Elliptic Curves." May 1999. URL: http://www.dacrutchley.plus.com/files/ecc_project.zip (2 Jan. 2004)
[13] P. Samarati, "Protecting respondents' identities in microdata release," *IEEE Trans. Knowl. Data Eng.*, vol. 13, no. 6, pp. 1010–1027, Nov./Dec. 2001.
[14] A. Solanas and A. Martínez-Ballesté, "V-MDAV: Variable group size multivariate microaggregation," in *Proc. COMPSTAT*, 2006, pp.917–925.
[15] A. Solanas, A. Martínez-Ballesté, and Ú. González-Nicolas, "A variable-MDAV-based partitioning strategy to continuous multivariate microaggregation with genetic algorithms," in *Proc. Int. Joint Conf. Neural Networks (IJCNN)*, 2010, pp. 1–7.
[16] L. Sweeney, "k-anonymity: A model for protecting privacy," *Int. J.Uncertainty, Fuzziness and Knowledge-Based Syst.*, vol. 10, no. 5, pp.557–570, 2002.
[17] L. Willenborg and T. DeWaal, "Statistical Disclosure Control in Practice," in *Lecture Notes in Statistics*. New York, NY, USA: Springer-Verlag, 1996, vol. 111.
[18] L.Willenborg and T. DeWaal, "Elements of Statistical Disclosure Control,"
in *Lecture Notes in Statistics*. New York, NY, USA: Springer-Verlag, 2001, vol. 155.

S.Saranya is a PG with computer science department at Anna university,Chennai.She received the BE degree from Roever Engineering College, Perambalur in 2012 and pursuing the ME degree from Ranganathan Engineering College, Coimbatore in 2014.Her area of interest include Network security, Data mining.

M.Madan mohan is an assistant professor at Ranganathan Engineering College, Coimbatore. He received the B.Tech degree from Adhiyamaan College of Engineering, Hosur (2007-2010) and the ME degree from Anna University of Technology,Coimbatore (2010-2012).

X.Rexeena pursuing her master of engineering in computer science from Ranganathan engineering college,coimbatore and she done bachelor of engineering in computer science from ssk college of engineering and technology,coimbatore.Her research interest are datamining,network security,cloud computing.