Minimal Information Loss for Privacy-Preserved eHealth Applications

Ya-Ling Chen[1], Bo-Chao Cheng[2], Hsueh-lin Chen[1], Chia-I Lin[1], Guo-Tan Liao[2], Bo-Yu Hou[2] and Shih-Chun Hsu[2]

[1]Service Systems Technology Center
Industrial Technology Research Institute, Taiwan
[2]Department of Communications Engineering
National Chung Cheng University, Taiwan
{sophiachen@itri.org.tw, bcheng@ccu.edu.tw, Sherry_Chen@itri.org.tw, Simon_Lin@itri.org.tw, {loboyoh | y2k1122335 | hisa918203}@gmail.com}

Abstract

With the rise of privacy protection awareness and legal norms, we should preserve individual health data confidentiality through de-identification operations while providing "as needed" health information for the doctor's diagnosis and treatment, the heath research study and other health management applications. Traditional privacy risk management systems mainly focus on reducing the re-identification risk but fail to consider the information loss. In addition, when faced with a high-risk situation, they cannot efficiently locate the source of the problem. This paper proposes the Hiatus Tailor (HT) system, which maintains low re-identification risk while providing more authenticated information to database users and identifying high-risk data in the database for better system management. The experimental results prove that compared to traditional risk management methods, the HT system achieves much lower information loss with the same risk of re-identification.

Keywords: *Privacy, k-anonymity, De-identification, Risk Assessment*

1. Introduction

Electronic medical records and cloud storage have been introduced in hospitals in recent years. Digital records provide convenience, but such a system also introduces the new challenge of storing personal information securely. Based on personal information, a specific person can be identified directly or indirectly. Medical institutions save large amounts of personal information in databases whose contents can be divided into three categories: Direct Identifiers (DID), Quasi-identifiers (QID), and Sensitive Information (SI). Information that allows direct identification, such as the Social Security Number, is called DID. Details such as date of birth, level of education, and postcode, which can be combined to identify a person, are QID. Information that is private and confidential, such as medical conditions, is categorized as SI. When eHealth practitioners want to access medical records, the hospital can de-identify the database to protect patient privacy. However, when multiple users need to access the database, they each have unique requirements. The hospital must release several de-identified databases which are difficult to manage and differ from the original database. In other words, the de-identified database will be altered and the degree of alteration is represented by the information loss (IL). The database provider prefers high IL to lower the possibility of re-identification of the information, but researchers prefer databases with low IL. Therefore, the challenge is to strike a balance between them.

De-identification is the primary method of protecting private information, where the original database is modified to prevent direct identification of a person through their records even if multiple databases are combined. Some common de-identification techniques are data reduction, data modification, data suppression, perturbation and pseudonymisation 0. The k-anonymity model is commonly used to assess the performance of a de-identification technique. The higher the k value is, the lower is the risk of re-identification [2]. Some previous studies have focused on reducing the risk of re-identification. However, limited research effort has been spent on safeguarding privacy while minimizing data distortion. El Emam et al. [3] proposed a set of programs that balance the risk and the extent of data distortion. But such a system is unable to identify the data that is responsible for the higher risk effectively. In this study, we propose the Hiatus Tailor (HT) system. By using the Execution Chain Graph (ECG) to progressively de-identify data, people's privacy can be protected. The name Hiatus Tailor refers to the fact that the proposed system is capable of identifying the missing element within the system and fixing it. It balances re-identification risk and data distortion using progressive risk assessment and mitigation. Among the scenarios where the re-identification risk requirement is satisfied, the proposed method chooses the one that minimizes the distortion level. The main contributions of this paper are: (1) In contrast to other de-identification methods that de-identify the entire database once, resulting in high IL, the HT system not only meets the privacy protection requirements, but also categorizes data into QID blocks using ECG. The risk is assessed progressively for each block. Based the re-identification risk estimated by this assessment, an optimal de-identification method is selected. As de-identification is not required at every stage, the HT system reduces IL. (2) Tradition risk assessment methods can only indicate whether the risk is high or low. However, for most databases, the source of the risk cannot be identified. Therefore, the process of identifying the source of the increased risk is time consuming. The HT system uses QID and progressively assesses risk for a database. ECG allows an examination of the entire system and assists medical institutions in evaluating whether the target system satisfies privacy safeguard requirements. If the system is found to have a high level of risk, it is easier to identify and handle the QID data block that is responsible for the high risk level.

2. HT System Architecture and Operation Method

The HT system architecture consists of two major components: ECG Composer and the Privacy Tailor. ECG Composer compiles the information obtained from users' requirements and generates the Execution Chain Graph, which is sent to the Privacy Tailor for further processing and risk assessment. The operation of the ECG Composer is based on information from the following elements:

- Database Schema: Defines the properties of the database, such as the type of the tables in the database and the attributes of the table.
- Application Context: Includes components related to SQL statements.
- Privacy Policy: Delineates the privacy policy associated with the user or company, such as the threshold k (k-anonymity) for the QID.

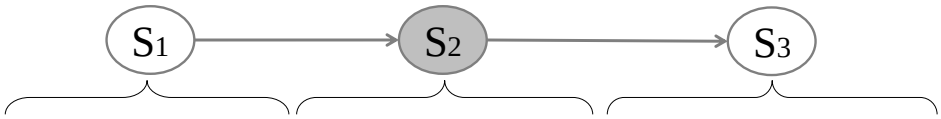
A Privacy Tailor is analogous to a privacy management department. It can be described as two stages: (1) Risk Assessment: it executes the risk assessment procedure and estimates the re-identification risk of the current assessment phase; (2) De-identification: on completing the risk assessment, if the re-identification risk is higher than the threshold, Privacy Tailor identifies the QID that has relatively high risk and needs to be de-identified. After the ECG composer defines the threshold for the ECG, the Privacy Tailor will calculate the re-

identification risk and extent of data alteration at the level of the node and record it in the node data. If the risk value is higher than the threshold, the Privacy Tailor will first evaluate and analyze each node to estimate re-identification risk and chose the most appropriate data for identification. If not, it proceeds to the next stage for analysis. When the re-identification value at each stage is below the threshold, the Privacy Tailor completes execution.

2.1. Execution Chain Graph (ECG) and Example

Database access task execution is modeled and structured in various stages aimed at clients in several stages of database retrievals. As described earlier, the ECG Composer compiles the user requirements, consisting of the database schema, application context and privacy policy, and then generates the ECG to represent "stored procedure" that accesses database system, and the directed edge denotes execution sequence each node. Each stage consists of several atomic "stored procedure" node which has a set of associated attributes (as shown in **Figure**). These properties can be further classified as Local and Aggregate. The Local value is the result of evaluating the QID combination of the current node. Aggregate value is the result of adding the evaluation of all QID combinations of all previous nodes. The attribute information that the Privacy Tailor requires to execute risk assessment includes:

- Information Loss: the magnitude of the difference between the original database and the database after de-identification.
- Re-identification Risk: the QID information after de-identification and the possibility of identifying the original information after various combinations and comparisons.
- Table Access: the table name where information is stored and accessed.
- QID: quasi-identifier, The QID itself can only indirectly identify a specific person, by combining different QIDs, it may be possible to directly identify the person.
- Condition: the relevant part of the SQL statement.



Attributes	Local	Aggregate
Information Loss	0	0
Re-Identification Risk	0.03	0.03
Table Access	E_Table	E_Table
QID	age	age
Condition	age ≥30	age ≥30

Node information in S_1

Attributes	Local	Aggregate
Information Loss	0	30%
Re-Identification Risk	0.23	0.26
Table Access	E_Table	E_Table
QID	region	age x region
Condition	age ≥30	age ≥30

Node information in S_2

Attributes	Local	Aggregate
Information Loss	0	30%
Re-Identification Risk	0.1	0.36
Table Access	E_Table	E_Table
QID	sex	age x region x sex
Condition	age ≥30	age ≥30

Node information in S_3

Figure 1. Execution Chain Graph (ECG)

Each node in the ECG represents one stored procedure. We use S_n to represent the n^{th} level of ECG. In Figure 1, the ECG can be divided into three levels, node in terms of nodes S_1, S_2, and S_3. Using S_1 as an example, there is no re-identification value initially. Next, the Privacy Tailor performs an evaluation and fills in the current node information. In node S_1, all QIDs belong to E_table, the age attribute. It satisfies the Conditions (comparison predicate) restricting the rows returned by the query (e.g., age ≥30), as the re-identification risk is 0.03. Thus, de-identification is no required and data distortion is zero.

Assume that a user requires access to information stored in the electronic hospital records database. The information in the database may include patients' age, address, and gender. Based on the user's requirements, Privacy Tailor performs risk assessment. The detailed processes are described as follows. At node S_1, the Privacy Tailor begins evaluation using the QID combination of the chosen table, which is the re-identification risk of the patients' age. Assuming that the threshold of the privacy policy equals to 2, the re-identification value

calculated is 0.03, which is less than the threshold value 0.5. Thus, the Privacy Tailor decides that age is low risk and de-identification is needless, and the IL value is 0. After evaluating S_1, node S_2 is evaluated, which involves calculating the re-identification risk of the combination of age and address (age×address). The result obtained is 0.73, which exceeds the threshold. Therefore, the Privacy Tailor must proceed with de-identification at this level. There are three possible de-identification ways (age, region and age×region), each associated with re-identification risk and information loss are 0.55 and 50%, 0.23 and 30%, 0.36 and 70% respectively. After calculating the results for the three different de-identification approaches, the Privacy Tailor will choose to perform de-identification on "region" because it has a relatively low re-identification risk (0.23) and the lowest data distortion level (30%). After finishing this step, the Local re-identification risk will change from 0.73 to 0.23. The Aggregate risk value will add S_1 to S_2, which is 0.03 plus 0.23 equals to 0.26; Local IL equals 30%, and Aggregate IL equals the sum of IL and that for S_1, which is 0% plus 30%, equals 30%. After finishing the assessment of S_2, it will calculate the re-identification risk of the (age×region×sex) combination at S_3, and the result obtained is 0.1, which is lower than the threshold value. Therefore, the Privacy Tailor will stop de-identification at this level. The risk at this level is 0.26 + 0.1 = 0.36.

This example demonstrates that the Privacy Tailor decides whether to perform de-identification based on the risk level, and locate the optimal QID information combination from different conditions; de-identification is not performed on all QID information. This multi-level method only needs to deal with local information combinations most of the time and therefore can effectively reduce IL value. In addition, it can also identify the high-risk data in a database and help improve privacy safeguards.

3. Simulation and Results

The environment developed in C language is used to simulate the workflow of the HT system. Under the considerations of the re-identification risk threshold between k=2 and k=15, the target attributes are age, address and income. All attributes are sourced from the Microdata (demodata.asl) and Macrodata (demodata.rda) of μ-Argus. Based on assumptions above, the ECG composer outputs an ECG with three levels of attributes, age, address and income. The three levels are provided to the Privacy Tailor for further processing. In each stage, the Privacy Tailor assesses whether the re-identification risk is higher than the threshold. If the risk is within an acceptable range, the information will be passed to the next stage without de-identifying the attribute. The risk of each de-identification combination of the attributes needs to be assessed. There are seven possible de-identification combinations: address, age, income, address×age, age×income, address×income, address×age×income. When the risk values of all stages are lower than the threshold, we perform data de-identification with only some of the attributes, which result in low distortion. The following paragraphs present the results plotted from the experiments.

The HT system uses the same de-identification techniques as μ-Argus. With the same re-identification risk threshold (k), we compared the distortion levels between de-identifying with the optimal combination of HT and de-identifying with the entire dataset of μ-Argus. The distortion level is represented by Modification Rate (MR) which represents the IL based on the amount of data being modified. Equation (1) is to calculate the ratio between the numbers of modified attributes and the total attribute numbers.

$$MR = \frac{N_A}{N_T} \tag{1}$$

Where N_A is the modified attribute numbers of a dataset;
and N_T is the total attribute numbers of a dataset

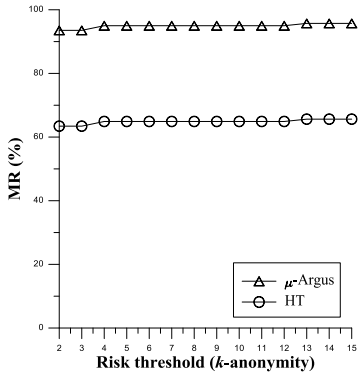


Figure 2. Data Distortion on Modification Rate (MR)

Figure 2 demonstrates the MR of both the HT system and the μ-Argus system. The x-axis represents the re-identification risk k, and the y-axis represents the MR of the de-identified dataset. As shown in the figure, the amount of data that needs to be modified is 65% and 95% for the HT system and μ-Argus system, respectively. Thus, the distortion level of the HT system is 30% lower than the μ-Argus system, and the HT system is superior.

4. Conclusion

Traditional methods, which perform de-identification on the entire database, can reduce the re-identification risk and protect private information, but they cannot provide authentic information to researchers. Based on experimental results, this paper proposes the HT system, which maintains a low re-identification risk in the required area, but is still able to effectively reduce the level of information loss and satisfy the needs of medical and research groups. HT system enables administrators to completely customize a privacy-preserved database system for eHealth applications and ensure that all service requests are managed in a consistent and reliable manner.

References

[1] A. Appari and M. E. Johnson, "Information Security and Privacy in Healthcare: Current State of Research", Proceedings of International Journal of Internet and Enterprise Management Issue. Vol. 6, pp. 279-314, **(2010)**.

[2] P. Samarati and L. Sweeney, "Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression", Proceedings of the IEEE Symposium on Research in Security and Privacy, **(1998)** Oakland, CA.

[3] K. E. Emam, F. K. Dankar, R. Vaillancourt, T. Roffey and M. Lysyk, "Evaluating the Risk of Re-identification of Patients from Hospital Prescription Records", Proceedings of The Canadian Journal of Hospital Pharmacy, Vol. 62, no. 4, pp. 307–319, **(2009)**.

Authors

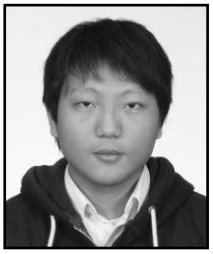
Ya-Ling Chen is a project manager of Industrial Technology Research Institute in Taiwan. Chen received a B.S. degree in Information Management from Yuan-Ze University in 2000. Chen works in Medical Informatics department, and her main job is project management and system analysis.

Bo-Chao Cheng is an Associate Professor of Department of Communications Engineering at National Chung-Cheng University. Cheng received a PhD degree in CIS from New Jersey Institute of Technology in 1996. After graduations, he also worked for Transtech Network (2000–2002), Bellcore (1998–2000) and Racal DataCom (1996–1998) respectively. His broad interests include network security, network management and real-time embedded system design.

Hsueh-Lin Chen is an associate engineer of Industrial Technology Research Institute in Taiwan. Chen received a B.A. degree (Information Management) from Hsuan Chuang University in 2001. Her ITRI's main work is the system design and development.

Chia-I Lin is an Associate Engineer of Service Systems Technology Center at Industrial Technology Research Institute in Taiwan. Lin received a Master degree in College of Engineering from National Tsing-Hua University in 2004.

Guo-Tan Liao received a B.S. degree in Information Engineering and Computer Science from Feng Chia University in 2003, and M.S. degree in Communications Engineering at National Chung-Cheng University in 2008. Currently, he is a Ph.D. candidate in National Chung-Cheng University. His research interests include network security, WSN and MANET.



Bo-Yu Hou received a B.S. degree in Electrical Engineering from Chung Yuan Christian University, Taiwan, in 2011. He is a master student in National Chung-Cheng University, Chia-Yi, Taiwan. His research interests include Network Security and Privacy.

Shih-Chun Hsu received a B.S. degree in Communication Engineering from Yuan Ze University, Taiwan, in 2011. And she is a master student in National Chung-Cheng University, Chia-Yi, Taiwan. Her research interests include Network Security and WSN.