VOL. 9, NO. 7, JULY 2014 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences


www.arpnjournals.com

ANONYMIZATION CLINICAL DATA: PRIVACY CASE STUDY

Yahaya Abd. Rahim[1], Mohd Azlishah thman[2] AhmadNaim Che Pee[1] and Mohd Fairuz Iskandar Othman[1]
[1]Faculty of Information and Communication Technology, Malaysia
[2]Faculty of Computer Electronic and Electrical Engineering, Universiti Teknikal Malaysia Melaka (UTeM), Durian Tunggal, Melaka, Malaysia
E-Mail: mohdfairuz@utem.edu.my

ABSTRACT

Privacy includes the right of individuals and organizations to determine for themselves when, how and to what extent information about them is communicated to others. The growing need of managing large amounts of data in hospital or clinical raises important legal and ethical challenges. This paper introduces and show the testing implementation of the privacy-protection problems, and highlights the relevance of trusted third parties and of privacy-enhancing techniques (PETs) in the context of data collection, e.g., for research. Practical approache on the pseudonymization model for batch data collection are presented. The actual application of the described techniques today proves the possible benefits for medicine that innovative privacy-enhancing techniques can provide. Technical PET solutions can unlock valuable data sources, otherwise not available.

Keywords: pseudomization, privacy-enhancing, clinical data, public released data.

INTRODUCTION

Organizations like hospital, clinic or pharmacy have vast amounts of personal data that are collected, stored and processed during doctor-patient consultation. They are very keen to release the clinical data for research purposes. However, the clinical data typically has much sensitive nature (e.g. medical data, disease and name), and although generally used for the benefit of the community, it can be easily abused by malicious people.

Currently incidents are frequently reported in the public media, without concern about a proper treatment of sensitive data. However, people tend to become more apprehensive when their personal healthcare-related data are at stake, mainly because they can easily imagine motives for abuse and assess its impact. Other an obvious case in point is that at some point in their life practically everyone is confronted with loan and insurance applications. Recent incidents such as the one in which an outsourced transcribers' threatened to disclose all medical records she had been processing for a US hospital [1] clearly illustrate that the threat to privacy is genuine. Public authorities are also sharply aware of these repercussions, and they are putting considerable effort into privacy protection legislation [2, 3]. Nowadays, we can't deny that privacy protection directly impacts personal well-being as well as society as a whole. Indeed, some go as far as to believe that failure to protect privacy might lead to our ruin [4]. Privacy is in fact recognised as a fundamental human right.

In Malaysia; until now there are none special bodies that are pay careful attention to the requirement of obtaining the informed consent from subjects. Because of that, most of the hospital or clinics are very cautions on assessing their information because they know the impact of the information is very complex; thus a real danger that informed consent is rather an ill-informed consent. Research ethics and security guidelines demand research units to divert more and more resources and time to privacy and identity protection, but burdensome requirements governing the transmission of medical information could unnecessarily discourage research. Well-intentioned privacy laws should not clash with the legitimate use of information when clearly to the public's benefit.

Protecting human rights for example like privacy while maximizing research productivity is one of the coming challenges. A first step towards this goal is the research and implementation of technical solutions to the privacy problem. Privacy-enhancing techniques (PETs) should be provided with to unlock invaluable data sources for the benefit of society without endangering individual privacy.

This paper focuses on the possible use of privacy enhancing techniques in the context of research and statistics for health care.

PRIVACY-ENHANCING TECHNIQUES

There are many situations in which privacy can be an issue; until now many research covers many different areas, including:

- Anonymous communication (for example anonymous remailers, anonymous surfing, etc.),
- Anonymous transactions,
- Anonymous publication and storage,
- Anonymous credentials,
- Anonymity in files and databases

By focusing at medical applications, in which privacy issues are raised by the information content of the stored data, hence the paper is discussed in. Privacy-enhancing techniques for privacy protection within databases help us to protect the privacy of the database like person records or organisation records that maintain in the database where these privacy-enhancing techniques allow storing relevant and useful information without anyone can ever find out, who the information is actually

about. Lists are some of the examples of these techniques are (non exhaustive list):

- Hard de-identification by the owner of the data;
- Various types of anonymization and/or pseudonymization;
- Privacy risk assessment techniques;
- Controlled database alteration (modification, swapping or dilution of data);
- Data flow segmentation;

Today, privacy-enhancing technique has proven for privacy protection in marketing and research data collection in United State [5]. However, in this research project we tried to focus and enhance lays the implementation of pseudonymization techniques, and complementary PETs at one of the general hospital in Johore state, Malaysia.

PSEUDONYMIZATION TECHNIQUE

Pseudonymization refers to privacy-enhancing techniques and methods used to replace the true (nominative) identities of individuals or organizations in databases by pseudo-identities (pseudo-IDs) that cannot be linked directly to their corresponding nominative identities [6].

With this technique, the data that contains, identifiers and ‘‘payload data” (non-identifying data) are separated. The pseudonymization process translates the given identifiers into a pseudo-ID by using secure, dynamic and preferably irreversible cryptographic techniques (the identifier transformation process should not be performed with translation tables). For an observer, the resulting pseudo-Ids are thus represented by complete random selections of characters. This transformation can be implemented differently according to the project requirements. Pseudonymization can:

- always map a given identifier with the same pseudo-ID;
- map a given identifier with a different pseudo-ID;
- time-dependant (e.g. always varying or changing over specified time intervals);
- location-dependant (e.g. changing when the data comes from different places);
- content-dependant (e.g. changing according to the content);

Pseudonymization is used in data collection scenarios where large amounts of data from different sources are gathered for statistical processing and data mining (e.g. research studies). In contrast with horizontal types of data exchange (e.g. for direct care), vertical communication scenarios (e.g. in the context of disease management studies and other research) do not require identities but the pseudonymization can help find solutions. It is a powerful and flexible tool for privacy protection in databases, which is able to reconcile the two following conflicting requirements: the adequate protection of individuals and organizations with respect to their identity and privacy, and the possibility of linking data associated with the same data subject (through the pseudo-IDs) irrespective of the collection time and place.

Because of this flexibility, however, correct use of pseudonymization technology is not as straightforward as often suggested. Careless use of pseudonymization technology could lead to a false feeling of privacy protection. The danger mainly lies within the separation of identifiers and payload.

The important things that should be alert before we precede this process make sure that payload data does not contain any fields that could lead to indirect re-identification. For example, the re-identification is based on content, not on identifiers.

The key to good privacy protection through pseudonymization is thus careful privacy assessment. Privacy gauging or privacy risk assessment is measuring the risk that a subject in a ‘‘privacy protected” database can be re-identified without cooperation of that subject or against his or her will. This consists in measuring the likelihood that a data subject could be re-identified using the information that is available (hidden) in the database. The lower this re-identification risk, the better the privacy of the subject listed in that database is protected. Conducting a privacy analysis is a difficult task. At this point in time, no single measure for database privacy is fully satisfying and this matter is still a hot topic in scientific communities. However, extensive research, mainly conducted by statisticians (area of statistical databases, etc.) and computer scientists like data miners or security experts are making significant progress.

From our literature review, using privacy risk assessment techniques, pseudonymization performance can be guaranteed. Data collection models are used to estimate the risk level for re-identification by attackers (a priori risk assessment). How the data should be separated (identifiers versus payload), filtered (removal of information) and transformed (transforming payload information in order to make it less identifying) is subsequently determined on the basis of these results. This is meaning one of the uses of privacy risk assessment techniques is to determine correct configuration of PETs.

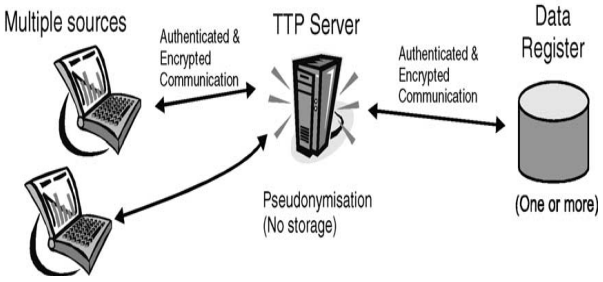

Figure-1. Communicating entities.

Many more aspects of the pseudonymization process are closely linked and key to ensuring optimum privacy protection, as for example, the location of the

identifier and payload processing, the number of steps in which the pseudonymization is performed.

PSEUDONYMIZATION IMPLEMENTATION

The pseudonymization as described above provides privacy protection for data collection for research and market studies. Two logical entities involved in handling the data are:

- The data suppliers or 'sources';
- The data collectors, one or several 'data registers' where the pseudonymized data are stored. Data suppliers typically have access to nominative data (e.g. treating doctors); the data collectors should only have access to anonymous data.

BATCH DATA COLLECTION

In this research, a possible scenario is the use of pseudonymization in batch data collection. The three interacting entities are shown in Figure-1. In contrast to traditional data collection, the sources (e.g. electronic medical record systems) do not necessarily interact directly with the database and vice versa. Communication is routed through a pseudonymization server (TTP server), where the pseudonymization and the processing of relevant data take place, as required.

Data is gathered and packed at the sources, typically in local databases. An example could be a local patient database which is managed at a clinic. The data is transmitted on a regular basis to the register through the TTP server where it is pseudonymized.

The data that can be extracted from the local databases is split into two variables, identities and (screened) payload data according to rules determined during the privacy risk assessment stage. Identifiers are pre-pseudonymized at the source, like a first transformation into pre-pseudo-IDs is performed. The payload data (assessment data) is filtered for indirect identifying data and transformed it to avoid re-identification of the anonymous data. Finally, the pre-pseudo-IDs are encrypted using a public-key scheme for decryption by the TTP server exclusively. The payload data are public-key encrypted to the register, so that only the register can read the data. Both are then transmitted to the TTP over secure links (authenticated and encrypted).

Full trustworthiness and integrity of the service is thus guaranteed not only by means of policy but also on a technical level. First, because the TTP never actually processes real identities (there is a pre-pseudonymization stage). Second, because although payload information passes through the TTP server, the latter can neither interpret nor modify the assessment data and to fully trusted this data is encrypted for decryption by the final destination (data register) only.

As a researcher, we believe and understood that although the pre-pseudonymized information leaving the source no longer contains any real identities, but this does not always guarantee absolute privacy because, as the prepseudonymization software is available at many sources, a smart intruder might find a way to map identities with their corresponding pseudoidentities for a 'dictionary attack' by entering known identities and creating a translation table. This technique may be like such an attack can be prevented by use of tamper-proof pseudonymization devices; these are however not yet deployed in real data collection scenarios.

From the previous research, we believe by performing a second transformation in a centrally controlled location for example in the TTP server, optimum security can be offered against such malicious attacks and etc. But as already mentioned there are more advantages to the use of an intermediary party. As the TTP server dynamically controls the pseudonymization process, additional privacy protecting functionality can be added like monitoring of incoming identities against such attacks, re-mappings of identifies, data flow segmentation, data source anonymization, etc.

After this second stage, we propose at the TTP in which the pre-pseudonymized identifiers are transformed into the final pseudo-Ids may be by using cryptographic algorithms, both the payload data and the pseudo-Ids are transferred to the register via secure communication.

At the register, the data can then be stored and processed without raising any privacy concerns.

CONCLUSIONS

Privacy includes the right of individuals and organisations to determine for themselves on when, how and to what extent information about themselves can be communicated to others. Several types of privacy-enhancing technologies exist that can be used for the correct treatment of sensitive data in medicine, but in this paper we focus that advanced pseudonymization techniques can provide optimal privacy protection of individuals. The research also shows that the privacy-enhancing techniques currently deployed for medical research, which proves that the use of pseudonymization and other innovative privacy enhancing techniques can unlock valuable data sources, otherwise legally not available.

REFERENCES

[1] D. Lazarus, A tough lesson on medical privacy: Pakistani transcriber threatens UCSF over back pay, San Francisco Chronicle Wednesday, October 22, 2003.

[2] Directive 95/46/EC of the European Parliament and the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data.

[3] Directive 2002/58/EC of the European Parliament and of the Council of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector (Directive on privacy and electronic communications).

[4] M. Caloyannides, Society cannot function without Privacy, IEEE Security and Privacy, vol. 1, No. 3, May-June 2003.

[5] F. De Meyer, B. Claerhout, G.J.E. De Moor, The PRIDEH project: taking up privacy protection services in e-health, in: Proceedings MIC 2002 ‘‘Health Continuum and Data Exchange”, IOS Press, 2002, pp. 171-177.

[6] G.J.E. De Moor, B. Claerhout, F. De Meyer, Privacy enhancing techniques: the key to secure communication and management of clinical and genomic data, Meth. Inf. Med. 42 (2003) 148-153.

[7] D.J. Solove, M. Rotenberg, Information Privacy Law, Aspen Publishers, New York, 2003.

[8] First draft of AURTAF, Anonymity User Requirements for Trusted Anonymisation Facilities. CEN/TC 251/WG III N 02- 018 (2002-07-17).