THE JOURNAL OF DIGITAL FORENSICS, SECURITY AND LAW

Journal of Digital Forensics, Security and Law

Volume 12 | Number 3 Article 5

9-2017

Analysis of Security in Big Data Related to Healthcare

Isabel de la Torre
University of Valladolid, isator@tel.uva.es

Begoña García-Zapirain
University of Deusto

Miguel López-Coronado
University of Valladolid

Recommended Citation

de la Torre, Isabel; García-Zapirain, Begoña; and López-Coronado, Miguel (2017) "Analysis of Security in Big Data Related to Healthcare," *Journal of Digital Forensics, Security and Law*: Vol. 12 : No. 3 , Article 5.
DOI: https://doi.org/10.15394/jdfsl.2017.1448
Available at: https://commons.erau.edu/jdfsl/vol12/iss3/5

ANALYSIS OF SECURITY IN BIG DATA RELATED TO HEALTHCARE

Isabel de la Torre-Díez[1], Begoña Garcia-Zapirain[2], Miguel López-Coronado[1]
[1] Department of Signal Theory and Communications, and
Telematics Engineering. University of Valladolid
Paseo de Belén, 15, 47011 - Valladolid, Spain
Phone: +34 983423000 Ext. 3703; Fax: +34 983423667
Emails: isator@tel.uva.es, miglop@tel.uva.es
[2] University of Deusto
Avenida de las Universidades 24, 48007 Bilbao, Spain
Phone: +34 944139000
Email: mbgarciazapi@deusto.es

ABSTRACT

Big data facilitates the processing and management of huge amounts of data. In health, the main information source is the electronic health record with others being the Internet and social media. Health-related data refers to storage in big data based on and shared via electronic means. Why are criminal organisations interested in this data? These organisations can blackmail people with information related to their health condition or sell the information to marketing companies, etc. This article analyses healthcare-related big data security and proposes different solutions. There are different techniques available to help preserve privacy such as data modification techniques, cryptographic methods and protocols for data sharing, query auditing methods and others that are analysed in this research work. Although there remains much to do in the field of big data security, research in this area is moving forward, both from a scientific and commercial point of view.

Keywords: Big data, health, information, privacy, security

1. INTRODUCTION

Large volumes of data are processed using big data in order to obtain information and be able to create knowledge about them. In the field of healthcare, the main source of information is the Electronic Health Record (EHR), while other sources are the Internet and social media. The notion of new hypotheses and assessment of the real effectiveness of healthcare will be permitted with big data. Populations and healthcare services will be able to be analysed using this technology, while patients will gain access to information about healthcare results and assessments made by other patients about their care (Moseley, 2014; Shaikh, 2014). Examples of big data in healthcare include a calculation of the incidence of epidemics via online searches, and identification of patterns in response to the treatment of illnesses by reusing data obtained from electronic health records or by estimating prognostic factors involving exhaustive monitoring (De la Torre-Díez, et al., 2015).

One of the advances made as a result of the evolution of technology and big data is the investment required to sequence the genome, which has decreased from initially millions of dollars to some thousand dollars per genome – one of the main fields in which big data is gaining weight is that of genomics, which is an area that is constantly evolving and making new discoveries. It is in this field in which an attempt is being made to discover and analyse the individual genome of each person and, as we mentioned at the beginning, ensure personalised healthcare. The study of DNA is also included in this field together with the analysis of why bacteria are resistant to antibiotics. Another field is that of hospitals, where analysing the analyses carried out on patients using big data may lead to an earlier diagnosis, meaning the dosage of medication the patient needs to receive could thus be reduced.

The serious problem with security in healthcare-related big data is explained in this paper and different techniques are proposed for trying to deal with it.

2. PRIVACY IN BIG DATA: A MAJOR PROBLEM

The serious problem with healthcare-related big data security is analysed in this section and a possible solution is taken into consideration.

Healthcare data is at present stored in large data bases and shared via a range of electronic means. Yet, why can such data arouse the interest of organised, extremely dangerous criminal organisations? These organisations can use the data, among other things, to blackmail individuals via information about their illnesses, sell information to marketing firms for product advertising campaigns based on such information or obtain fraudulent prescriptions by creating false identities. Among examples of this type of healthcare data theft are the following: in 2004, data from the Home Depot medicine distributor was stolen which affected 53 million of its customers, while in 2015, healthcare information pertaining to 80 million customers was stolen from the American insurance company Anthem.

Healthcare data privacy is a generally recognised right in nearly all countries in the world. According to the Cloud Security Alliance (CSA), privacy issues are magnified by the three Vs associated with big data: velocity, volume and variety (Cloud Security Alliance, 2014; The White House, 2014). In the European Union, privacy of such data is covered by Directive 95/46/EC (European Union, 2016; Martínez-Pérez, 2015). In Spain, this directive is transferred to the Law governing Data Protection (LOPD), which establishes the highest level of protection for healthcare data. In the USA, healthcare data privacy is in accordance with the Health Insurance Portability and Accountability Act (HIPPA), which has established what is known as the Privacy Rule. This rule regulates the use by healthcare organisations of patients' confidential information. All institutions that handle patients' confidential data must adhere to the Privacy Rule, and the Office of Civil Rights (OCR) is responsible for ensuring its compliance. In February 2012, the White House released a report titled "Consumer Data Privacy in a Networked World: A Framework for Protecting Privacy and Promoting Innovation in the Global Digital Economy." Moreover, in March 2014, the Federal Trade Commission, together with different agency officials from the European Union and Asia-Pacific Economic Cooperation (APEC) economies, announced joint EU and APEC endorsement of a document that maps out the requirements of the European and APEC privacy frameworks (European Commission, 2012; European Commission, 2014).

2.1 Literature Review

When reviewing the literature available on health data privacy, we find an interesting paper by Lobato de Faria and Valente Cordeiro

(2014) that describes the rights to health data privacy and confidentiality in their classical contours. They analyse the new European data protection draft regulation, the purpose of which is to bring reinforced tools to this domain. Moreover, they conclude that the legal aura that still embraces the right to medical data confidentiality in Health Law and Bioethics seems to no longer have any practical sense and that further transformations of privacy and confidentiality in fields of medicine are necessary (Lobato et al., 2014).

Cho, et al. (2016) provide us with architecture based on a double layer for the working environment with big data. These two layers are the pre-filtering layer and the post-filtering layer. The first-mentioned works on the big data information gathering stage, the purpose of which is to seek and eliminate sensitive personal information from the data gathered, which it does in order to make the information anonymous and thus make it more difficult to identify the person in particular.

Liu, et al. (2015) define a series of steps in which the validity of the data stored is verified externally. External verification is as important as the privacy and security provided by the server and, as this is an external agent, certain steps need to be established to maintain both data privacy and integrity. Fabiano, et al. (2015), from the University of Wyoming, have been developing a variant of the MapReduce paradigm to be applied in security and which complies with HIPAA (the U.S. standard for medical record privacy), as well as using the OpenSSL encryption package. To ensure maximum scalability, they implemented a hybrid of the OpenMP-MPL programming paradigm, by means of which they enabled each processing core to be assigned a number of files and then for each core to subdivide these files, depending on the number of threads being used.

Hsu, et al. (2014) show us how to develop a protocol for the secure transfer of data, and also propose a protocol for the transfer of a group key. To do so, they create a variant of the Diffie Hellman algorithm which is designed for one-to-one communication rather than between several individuals. The motivation behind this protocol is to preserve the key refresh, its confidentiality and its authentication.

Jing (2014) comments on the growing use of the cloud as a storage space. To improve security, they propose a double-encryption data system, which consists of an initial encryption using the AES encryption algorithm and therefore a symmetric algorithm.

Hingwe, et al. (2014) explain a cloud-based data base model using architecture with additional two-layers which are used depending on whether the data is deemed sensitive or otherwise. Encryption is added to the data together with the layers, and Double encryption is used if the information is deemed sensitive. A key is needed for this type of encryption, and the authors use a symmetric key provided by the data base server.

Thilakanathan, el at. (2014) provide us with a security model to be used in monitoring patients via remote devices such as mobile phones and bracelets, etc. This model makes use of double encryption, symmetric encryption and encryption using the ElGamal algorithm, whereby mobile devices generate the patient's data and this data is encrypted using a symmetric key. The second encryption, which is asymmetric, is used to improve security, and its function will be to encrypt the public key being used. The disadvantage of using a symmetric key, however, is that it loses the identity of the data. Subashini and Kavitha (2011) explain a security model that does not prevent the data base from being hacked, but rather, ensures the data obtained is of no value. They cite the example of a user's login and password in which two unrelated, separate data are of no value. Their model involves splitting the data stored into a Public Data Segment (PDS) and a Sensitive Data Segment (SDS).

3. SOLUTIONS FOR PRIVACY OF BIG DATA IN HEALTH

A possible solution to the privacy problem could be a mechanism that ensures the privacy of medical information by providing individuals with a platform that allows them to monitor how their data has been used, by whom, and for what purposes.

Within this context, it proves necessary for projects that involve large volumes of clinical data to include the decision – taken in view of the information provided – by the patients themselves, in order to ensure its privacy.

Moreover, healthcare institutions that handle health-related data must meet a series of requirements in order ensure the privacy of data as follows (Yan, et al., 2016):

- Access to confidential information.
- Electronic means for secure storage.
- Availability of back-up copies of all information.
- Use of secure, encrypted information.
- System for managing security-related incidents.
- Use of physical media for confidential information.

Figure 1 shows the main challenges concerning data privacy for big data in health. These challenges should be applied to any project involving big data.

Privacy-preserving techniques fall into the following categories: data modification techniques, cryptographic methods and protocols for data sharing, query auditing methods, statistical techniques for disclosure and inference control, and randomization and perturbation-based techniques (Aggarwal, et al., 2008).

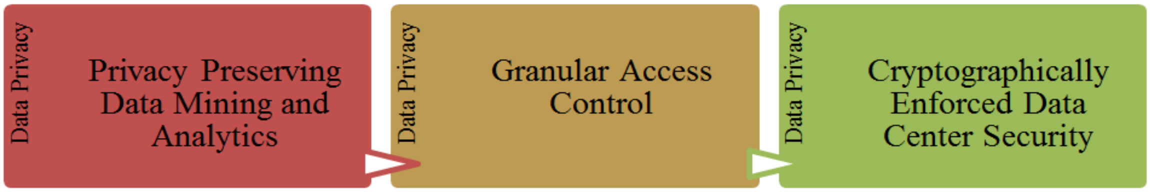
Data Privacy
Privacy Preserving Data Mining and Analytics
Data Privacy
Granular Access Control
Data Privacy
Cryptographically Enforced Data Center Security

Figure 1. Challenges concerning data privacy for big data in health

Granular access control is a mechanism for constraining the interaction between authenticated users and protected resources. Control is implemented via an authorisation service, which includes an authorisation decision function. The output of this decision function is determined by evaluating the access request with regard to an authorisation policy (Atallah, et al., 2009; Crampton, et al., 2006). Health protected data is encrypted, and authorised users are given the appropriate keys. Lazy re-encryption techniques and temporary hierarchical access control policies can also be used, and these techniques require that multiple keys be associated with a node in the hierarchy (Backes, et al., 2006; Brinkmann, et al., (2009).

Before describing a theoretical improvement to security, we should like to mention three possible threats according to their origin, namely: a) the internal agent, who belongs to the health system and is authorised to obtain the information but uses it for non-ethical purposes; b) the intermediary agent, who we identify as a member of the group in possession of the information storage system; and c) the external agent, who may be anyone outside the health and/or storage system. The improvement we propose will focus on improving security against intermediary and external agents.

The information we will wish to protect consists of a set of data expressed in the form of text or images. The first theoretical proposal involves using the double-layer scheme suggested by Subashini and Kavitha (2011) by means of which we split the text data in such a way that it is of no value on its own. By doing this we obtain a certain anonymity regarding the data, which is important in the area of health. As an alternative to this scheme, that put forward by Cho, et al. (2016) can be used, which involves splitting the data into sensitive and non-sensitive but for the same purpose, i.e. to make the information anonymous. This step will be the first of all of them. The next step will be to use the proposal made by Jing (2014), which describes the use of double encryption in order to protect data. Now that the data has been split, encryption will help to ensure that intermediary and external agents will not obtain information in flat text format. Although we will be using data splitting, this may reach a point in which the agents referred to could obtain all the information and would just need to rearrange it. Therefore, we use such encryption to ensure that it is not so easy to obtain the data in flat text format. Additionally, and by way of a complement, the algorithm developed by Zhou, et al. (2015) can be used due to the fact that it often tends to be preferable not to split the data into fragments, as medical images are deemed to be as valuable as written reports and hence also need to be protected. The algorithm proposed is a good system to ensure that only authorised persons may be able to correctly decrypt the images, although unfortunately and has been explained previously, care is required with these types of image as it is not possible to use the algorithm for all of them.

4. CONCLUSION

Big data is a new feature in all sectors although it represents a major advance in terms of healthcare. The amount of data that can be created daily needs to be both stored and processed, and hence cloud architecture is gaining prominence. Data generated in healthcare is of vital importance as it may enable us to ensure personalised healthcare by prescribing treatment to specific individuals that may in turn attain its maximum level of efficiency in combating illness. Not only will it help us to find personalised treatment but will also reduce healthcare costs – we do not only include treatment within this category, but also possibly a more reliable diagnosis and a more accurate and inexpensive analysis of tests. It will also help us to determine risk factors in the tests conducted or in epidemics such as the flu, speculate about the evolution of the virus and it's spreading through countries affected by it. All this data needs to be protected, as misuse may prove to be a catastrophe. If criminal organisations are able to get hold of important clinical data pertaining to patients, they are then able to blackmail them and wield great power. Chemists who have clinical data at their disposal may alter the results so as to counterfeit the positive aspect of the product and increase its adverse effects, and for similar reasons, data about how an epidemic spreads can also be altered, meaning that the prognosis will be worse than it otherwise would be and thus sowing panic among the population. These assumptions lead us to the conclusion that security systems are indeed necessary – something that either prevents or simply renders the idea of stealing such data unappealing. Security is a direct consequence of lack of trust and takes the form of mechanisms of the type we have summarised previously. We have reached the conclusion that there is no perfect security system, as the methods used are adapted to the sector to which they belong and to requirements at any given time. Technology has been taking huge steps forward over the years, which may help to create algorithms that cannot currently be used owing to the computational load they require. However, this technology is the same for hackers, meaning that they need increasingly less time to discover the keys. As security will never be perfect, all we can do is design methods to try and make things more difficult. Many of these methods, as we noted in the results associated with security, have been developed for specific environments and for specific types of file such as images or text with specific features. As years go by, security is something that will continue to gain prominence in any business to the extent that more investment will be made to improve it.

REFERENCES

Aggarwal, Charu C., Yu, Philip S. (2008) Privacy-Preserving Data Mining: Models and Algorithms. Springer.

Atallah MJ, Blanton M, Fazio N, Frikken KB. (2009). Dynamic and efficient key management for access hierarchies. *ACM Transactions on Information and System Security, 12(3), 1–43.*

Backes M, Cachin C, Oprea M. (2006). Secure key-updating for lazy revocation. In Proceedings of 11th European Symposium on Research in Computer Security, 327–346.

Brinkmann B, Bowera M, Stengel K, Worrell G, Steada M. (2009). Large-scale electrophysiology: Acquisition, compression, encryption, and storage of big data, *Journal of Neuroscience Methods, 180, 185–192.*

Cho D, Kim S, Yeo S. (2016). Double Privacy Layer Architecture for Big Data Framework. *International Journal of Software Engineering and Its Applications, 10(2), 271-278.*

Cloud Security Alliance. (2014). Retrieved on July 8 from https://cloudsecurityalliance.org

Crampton J, Martin K, Wild P. (2006). On key assignment for hierarchical access control. In Proceedings of 19th Computer Security Foundations Workshop, 98–111.

European Commission. (2012). Commission Proposes a Comprehensive Reform of the Data Protection Rules. Retrieved on July 10 from http://ec.europa.eu/justice/newsroom/data-protection/news/120125_en.htm.

European Commission. (2014). Article 29 Data Protection Working Party, Opinion 02/2014 on a referential for requirements for Binding Corporate Rules. Retrieved on July 10 from http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp212_en.pdf

European Union. (2016). Directive 95/46/EC of the European Parliament and of 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. Off J Eur Union, L 281 (1995, November), 0031–0050. Retrieved on July 8 from http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31995L0046:en:HTML

De la Torre-Díez I, Lopez-Coronado M, Garcia-Zapirain Soto B, Mendez-Zorrilla A. (2015). Secure Cloud-Based Solutions for Different eHealth Services in Spanish Rural Health Centers. *Journal of Medical Internet Research, 17(7), e157.*

Fabiano E, Seo M, Wu X, Douglas C. (2015). OpenDBDDAS Toolkit: Secure MapReduce and Hadoop-like Systems. *Procedia Computer Science, 51, 1675–1684.*

Hingwe K, Bhanu M. (2014). Two Layered Protection for Sensitive Data in Cloud. International Conference on Advances in Computing, Communications and Informatics (ICACCI), 1265-1272.

Hsu C, Zeng B, Zhang M. (2014). A novel group key transfer for big data security. *Applied Mathematics and Computation, 249, 436–443.*

Jing P. (2014). A New Model of Data Protection on Cloud Storage. *Journal of Networks, 9(3), 666-671.*

Liu C, Yang C, Zhang X, Chen J. (2015). External integrity verification for outsourced big data in cloud and IoT: A big picture. *Future Generation Computer Systems, 49, 58–67.*

Lobato de Faria P, Valente Cordeiro J. (2014). Health data privacy and confidentiality rights: Crisis or redemption?. *Revista Portuguesa de Saúde Pública, 32(2), 123-133.*

Martínez-Pérez, Borja. et al. (2015). Privacy and Security in Mobile Health Apps: A Review and Recommendations. *Journal of Medical Systems, 39(1), 181.*

Moseley, Edward T. et al. (2014). Journal of Medical Internet Research, 16(11), e259.

Shaikh, Abdul R. et al. (2014). Collaborative Biomedicine in the Age of Big Data: The Case of Cancer. Journal of Medical Internet Research, 16(4), e101.

Subashini S, Kavitha V. (2011) A Metadata Based Storage Model for Securing Data in Cloud Environment. International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery, 429-434.

The White House. (2014). President's Council of Advisors on Science & Technology, Big Data and Privacy: A Technological Perspective, May 1, 2014.

Thilakanathan D, Zhao Y, Chen S, Nepal S, Calvo R, Pardo A. (2014). Protecting and Analysing Health Care Data on Cloud. Second International Conference on Advanced Cloud and Big Data, 143-149.

Yan Z, Ding W, Niemic V, Vasilakos A. (2016). Two Schemes of Privacy-Preserving Trust Evaluation. Future Generation Computer Systems, 62, 175-189.

Zhou G, Zhang D, Liu Y, Yuan Y, Liu Q; A novel image encryption algorithm based on chaos and Line map. (2015). Neurocomputing, 169, 150–157.