2014 Second International Conference on Advanced Cloud and Big Data

Protecting and Analysing Health Care Data on Cloud

Danan Thilakanathan*†, Yu Zhao*, Shiping Chen*†, Surya Nepal†, Rafael A. Calvo* and Abelardo Pardo*
* School of Electrical and Information Engineering, the University of Sydney, Australia
† Digital Productivity & Services, Commonwealth Science Industry Research Organization (CSIRO), Australia
Corresponding Email: Danan.Thilakanathan@sydney.edu.au

***Abstract*—Various health care devices owned by either hospitals or individuals are producing huge amount of health care data. The big health data may contain valuable knowledge and new business opportunities. Obviously, cloud is a good candidate to collect, store and analyse such big health care data. However, health care data is very sensitive for its owners, and thus should be well protected on cloud. This paper presents our solution to protecting and analyzing health care data stored on cloud. First, we develop novel technologies to protect data privacy and enable secure data sharing on cloud. Secondly, we show the methods and tools to conduct big health care data analysis. Finally, both the security technology and the data analysis methods are evaluated to show the usefulness and efficiency of our solution.**

I. Introduction

There is a growing trend for doctors to remotely monitor and diagnose patients through the Cloud. This is mainly due to greater convenience for patients as they dont have to visit a clinic regularly and also for reduced healthcare costs for both the government and patients [11]. In Australia alone, the government spends in excess of 130 billion dollars on healthcare [33]. However, privacy and security issues associated with the Cloud make patients apprehensive about using the Cloud to store their personal health information.

One of the main trust and privacy issues arise from Cloud insider attacks [3]. It is well known about malicious insiders who steal data, since they have direct access to the owners data. Malicious Cloud service providers may steal data in order to sell to third parties in order to gain profit [7] [8]. Such privacy attacks affect the trust of the data owner and make them sceptical of using the Cloud for sensitive data storage. This is one of the main reasons why patients have a lack of trust for using the Cloud for storage and sharing of highly critical medical information. There has been multiple studies around privacy and trust in health systems in research [1][2][3][4][5][6]. Another major issue with private sharing of health information, and hence the major focus of this paper, is key management. A trivial solution to private data sharing in social networks involves a patient first encrypting health information on their own machine and distributing the encryption keys to each user, doctor, etc, he wishes to share his health information with. The authorised user can then download the encrypted health information from the social network and decrypt using the supplied encryption key. However, when the patient wishes to revoke a user, he must re-encrypt the data with a new encryption key and redistribute the new key to all the remaining users, hence, is computationally inefficient and places a huge burden on the patient, especially if he wishes to share his health information with plenty of users.

Large amounts of data are being generated by the healthcare system everyday [19], which is a valuable resource for better informed decision-making and treatment. Patient healthcare record, clinical data, diagnostic machine generated data are all information highly valued by doctors, and on top of that, social media information of patients is also a useful data source for doctors to help people with mental illnesses. However, paper-based health record tracking is inefficient and error-prone, and it is impossible to apply data mining and machine learning techniques on paper-based data. Thus, big data analytics application on e-health record is an obvious ideal solution to this problem [20], which is able to help doctors to come to more insightful diagnoses with lower costs [21].

In this paper, we explore the above vision by presenting some enabling technologies and methods.

- First, we present our security technology to protect health data privacy and enable secure health data sharing on the Cloud.
- We then conduct a simple health care data analysis with existing data analysis techniques and real public health data stored on the Cloud.
- We demonstrate the feasibility and usefulness of using the Cloud to collect and store health data for secure sharing and data analysis.

Our paper is organised as follows. In Section II, we describe and detail our system and protocol for secure health data storage and sharing in the Cloud. In Section III, we discuss health care data analysis techniques related to the Cloud followed by examples using real health data. In Section IV, we discuss related work and finally we conclude the paper in Section V.

II. Secure Health Data Storage and Sharing on Cloud

A. Architecture

The data model of our system is illustrated in Figure 1. The data producers of our system include smartphones, health sensors and trackers as well as social media. In this paper, we use the smartphone as our main data producer. Data is stored securely in a Cloud server and various data consumers are able to access the data in order to provide diagnosis and treatment to patients. We primarily focus on big data analysis for this paper to demonstrate our ideas. An authorized doctor for instance can then log-in and access the protected patients health data from the Cloud servers. If the doctor has the necessary permissions, he will be able to decrypt the patients data on his own machine and carry out diagnosis as well as analysis.

978-1-4799-8085-7/14 $31.00 © 2014 IEEE
DOI 10.1109/.24

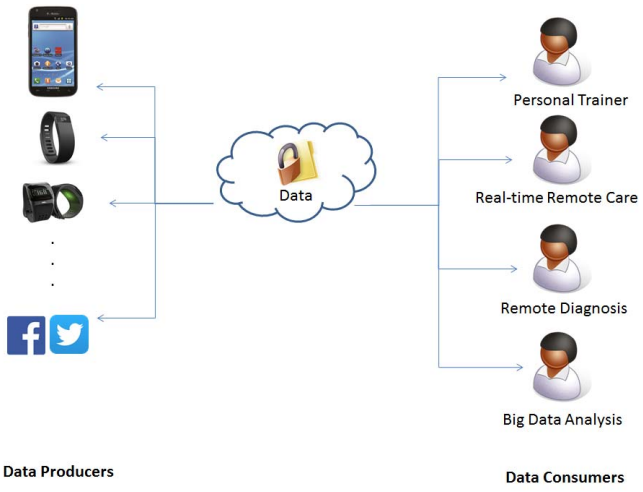

Fig. 1. Data Model

B. Technologies

1) ElGamal Encryption: ElGamal encryption, invented by T. ElGamal [29] is a public-key cryptography system. We take advantage of ElGamal Encryption in our work since the algorithm is both simple and efficient and can provide simple consumer revocation with low cost and overhead. There are three main steps of the ElGamal encryption algorithm:

- Initialisation: Given a prime p, a primitive root c of p, compute $b = c^x mod\ p$, where x is a randomly selected secret key. The public key is thus $\{p, b, c\}$ and private key is x.
- Encryption: Generate random value r and encrypt data m as follows:

$$\begin{aligned} E(m) &= m \cdot b^r\ mod\ p \\ &= m \cdot c^{\text{rx}}\ mod\ p \end{aligned} \tag{1}$$

  Also note: $g = c^r mod\ p$
- Decryption: This decrypts m with secret key x as follows:

$$\begin{aligned} D_x(E(m)) &= g^{\text{-x}} \cdot E(m)\ mod\ p \\ &= (c^r)^{\text{-x}} \cdot m \cdot c^{\text{rx}}\ mod\ p \\ &= c^{\text{-rx}} \cdot m \cdot c^{\text{rx}}\ mod\ p \\ &= m\ mod\ p \end{aligned} \tag{2}$$

2) Symmetric Cryptography: We use symmetric encryption cryptography algorithm in our work to protect the health data from being accessed by untrusted Cloud servers. Note, that in our work, we do not specify which symmetric algorithm is used. In theory, any symmetric encryption algorithm can be used based on the level of sensitivity of the health data.

Pros: While asymmetric encryption may be more secure compared to symmetric encryption due to the greater degree of difficulty in guessing the key, symmetric encryption is far better suited to data sharing in the Cloud. This is due to symmetric encryption using the one key to encrypt and decrypt data whereas asymmetric encryption uses two keys. When sharing data with very large groups, sometimes in excess of thousands, key management is more efficient when there is only key to protect. In our work, we make use of both symmetric encryption and asymmetric encryption. The symmetric key is used to protect the data and the ElGamal-based asymmetric key is used to protect the symmetric key which also has the added benefit of bundling the user's identity with the data.

Cons: One of the drawbacks of using symmetric encryption for data sharing, is that the data owner's identity is not bundled with the data. This makes it difficult for the data owner to claim ownership of the data. Since our solution also uses asymmetric encryption via ElGamal algorithm, the data owner's identity is also bundled along with the symmetric key.

C. Secure Sharing Protocol

In this section, we describe a scenario for our solution and provide the protocol for our work. We also provide a brief security evaluation of our protocol.

1) Mental Health Scenario: There is a now a much stronger need for people to complain about work stress, sleeplessness and depression as well as receive feedback in a convenient and secure manner. There are a growing number of people who feel the need to vent and complain about problems and issues that may occur in the household or at work and how much it is affecting their lives and their mental state. Doctors need to communicate with patients in order to determine symptoms and appropriately provide diagnosis. Visiting a clinic regularly to report about work stress and depression symptoms can be costly for both the patient and the doctor. For patients, a lot of time and effort is spent visiting the clinic in order to receive sometimes minor feedback. This is particularly true for rural patients or the elderly. For doctors, they may need to tend to more serious patients at the time. The government spends millions of dollars for healthcare every year, and are looking for ways to drastically reduce healthcare costs [11]. In our solution, we conveniently leverage the use of smartphones, as they are highly accessible. Patients can report and receive help wherever they are, such as at home or at work, as long as they have access to a smartphone. This can help cut healthcare costs as patients do not have to regularly travel in order to receive advice and feedback for mental health related problems. In fact, studies have also shown that the use of smartphones that provide support for mental health problems have shown significant reductions in depression, stress and substance use [10]. We provide a new way of protecting data without revealing the full encryption key to both user and the Cloud provider. Our high scalable solution provides the ability to share data with many users, such as doctors and nurses, while allowing the simple revocation of a user without the need to re-encrypt the data every time user revocation occurs. In this paper, we focus on creating a secure system that will enable patients to share mental health information with doctors and mental health specialists from the comfort of their own home.

2) Protocol: We now describe our protocol in detail. Throughout this paper, we assume the Cloud Service Provider (CSP) to be honest-but-curious, in the sense that the CSP will carry out the steps of the protocol as expected but is willing to find out any information about the patient as much as possible.

We also assume the smartphone app to be trusted and that it will not inadvertently or intentionally send information to the CSP without the patients knowledge.

Data Storage: The patient first runs the prototype app and inputs a text string, a number value and uploads an image on their smartphone. When the patient presses the Send button, the app will then generate an arbitrary symmetric key and then encrypt the text, number and image. The symmetric key will then be encrypted using the ElGamal public key. The ElGamal private key will then be partitioned into two parts. The encrypted symmetric key will then be partially decrypted using the first half of the key partition. The encrypted data contents and encrypted symmetric key will then be sent to the CSP for storage.

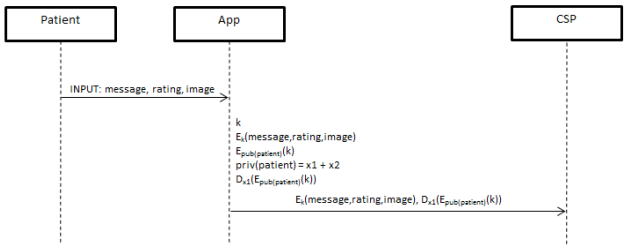

Fig. 2. Secure Data Storage

Data Sharing: When the patient decides to share the data with a doctor, they press the Share button on the app and enter the doctors social network username. The app will then partition the patients private key into two random parts. The first partition will be sent to the social network and the other will be sent to the doctor. By doing this, the untrusted CSP has no knowledge of the full private key since the other partition is stored on the doctors local machine.

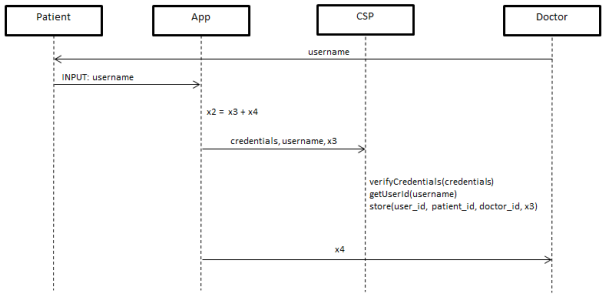

Fig. 3. Secure Data Sharing

Data Access: When the doctor wishes to access the patients data, they simply call the social network to retrieve the data. The CSP partially decrypts the symmetric key using the partial key supplied by the patient, and sends the encrypted data contents and partially decrypted symmetric key to the doctor. The doctor uses the partial key supplied by the patient to fully decrypt the symmetric key and finally decrypt the data contents.

Access Revocation: When the patient decides to revoke a specific data from access to his e-Health data, the patient

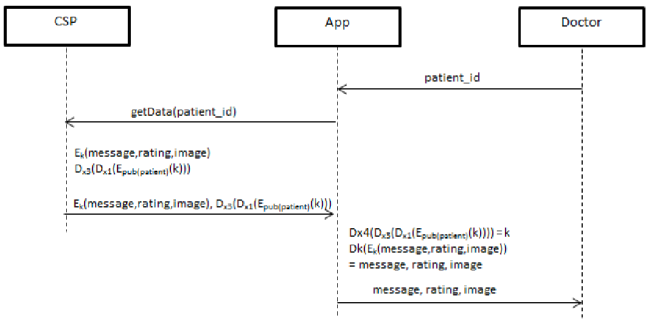

Fig. 4. Secure Data Access

simply calls the CSP to delete the doctors partial key entry. If the doctor attempts to download the data from the CSP, he will only see the ciphertext since the symmetric key will never be decrypted.

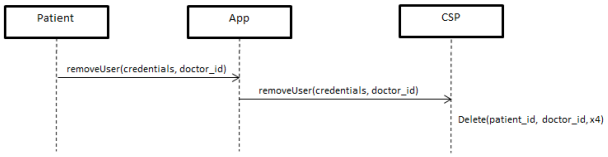

Fig. 5. Access Revocation

D. Security Analysis

We now provide a brief security analysis of our work.

- *Insider Attacks* - Our protocol is also secure under insider attacks since there is never a stage in our protocol where the data is decrypted in the untrusted Cloud. The data remains encrypted at all times on the untrusted Cloud servers as well as on untrusted public communication channels.
- *User Revocation* - In our protocol, user revocation can be achieved efficiently without the need to re-encrypt the data each time. The doctors key partition is simply removed from the Cloud storage. This way, if the revoked doctor attempts to access the health data, he will not be able to retrieve the full plaintext without the remaining key partition.
- *Update Secrecy* - Since health data is constantly changing, the patient may wish to update their health data. This is made possible in our protocol, as long as the updated version is encrypted with the same symmetric key that was used to encrypt the original health data, the patient may update their health data any number of times as they wish. Hence making our solution feasible to be deployed in a real-world scenario.
- *Mobile Stealing* - In the event someone steals the patients smartphone, they will not be able to access the personal health information as they would need to know the patients credentials such as email id and password in order to access the smartphone app.

Hence, a patient does not need to be tied down to only one smartphone device and can keep changing their device as often as theyd like without any loss of health information.

III. Smarter Health Care Data Analysis

Large amount of data is being generated by the healthcare system everyday [19], which is a valuable resource for better informed decision-making and treatment. Patient healthcare record, clinical data, diagnostic machine generated data are all information highly valued by doctors, and apart from that, social media information of patients is also a useful data source for doctors to help people with mental illnesses. However, paper-based health record tracking is inefficient and error-prone, and it is impossible to apply data mining and machine techniques on paper-based data. Thus, big data analytics application on e-health record is an obvious ideal solution to this problem [20], which is able to help doctors to come to more insightful diagnoses with lower costs [21].

Almost all data mining techniques can be applied in big data analytics in healthcare. To demonstrate the application of data mining techniques on healthcare data, two common data mining algorithms are selected as examples. Both algorithms will be applied on the Diagnosed Diabetes Incidence dataset [32].

A. Methods for Mining Healthcare Data

Linear Regression Linear regression [22] is one of the most basic data mining techniques that are still being used extensively. It tries to model the relationship between a dependent variable y and one or more explanatory variables X. The special case, in which X only contains one variable, is called simple linear regression. The most common usage of linear regression is prediction. First, a predictive model is built based on existing observed data containing both y and X. Then, when additional X are observed, the model can be used to predict the value of y.

$$y_i = a_1 x_{i1} + ... + a_p x_{ip} = x_i^T a + c_i \ (i = 1, ..., n) \quad (3)$$

Linear regression is based on the assumption that the relationship between a dependent variable y and a set of explanatory variables X is linear.

The T denotes transpose and the ε_i is the error term or noise, which is all other factors that affect the dependent variable y.

K-Means Clustering K-means clustering [23] is a cluster analysis technique in data mining, which aims at partitioning n observed items into k clusters in which each observed item belongs to the cluster with the closest mean.

$$\arg\min \sum_{a=i} \sum_{x_j \in S_i} ||x_j - u_i||^2 \quad (4)$$

The closeness is defined as the within-cluster sum of squares (WCSS):

where x1, x2, ..., xn is a set of observations, S = S1, S2, ..., Sk, and k (k ¡= n) is the number of sets, and i is the mean of observations in Si.

First a set with k initial mean should be given to the k-mean clustering algorithm. The most common method to initialize the k mean is the Forgy algorithm [24]. In the Forgy method, k observations are randomly selected from the whole dataset as the initial means.

Then the algorithm alternates between two steps: - Assignment step: each observation to the cluster whose mean yields the least within-cluster sum of squares (WCSS). - Update step: the new means are calculated as the centroids of the observations in the new clusters

B. Software Tools for Mining Healthcare Data

The biggest difference between big data analysis and more ordinary data analysis is that the data is so large that is not able to fit into the main memory of a single computer. Therefore algorithms aiming at data streaming, dimensionality reduction, data compression and data processing are available.

Weka [25], for example, is a general purpose data mining tool implemented by the University of Waikato (New Zealand). It is an open source software under the GNU General Public License (GPL). Weka is implemented in the Java programming language containing an GUI allowing easier interaction and an API which can be used in server-side data mining tasks.

Both algorithms introduced in the previous section have been implemented in Weka. The next section shows the examples of the two algorithms.

C. Examples

Dataset

The Diagnosed Diabetes Incidence dataset is used as the example dataset, which contains diabetic incidence information of all states of America from 2004 to 2011. For each year, 7 types of data is available, upper confidence limit, lower confidence limit, age-adjusted rate per 1000, age-adjusted lower confidence limit, age-adjusted upper confidence limit, rate per 1000, and number of new cases. Some of the data types are missing from the year 2011, thus data of the year 2011 is not included. The total number of instance is 3224, after removing instances containing missing value, the remaining number of instances is 3138.

Apply Linear Regression on ”number of new cases of diabetes (2004 2010)”

The ”number of new cases” of the year 2010 is used as the dependent variable, and all other values are used as explanatory variables. 10-fold cross-validation is used to validate the linear model. The resulting linear regression model is:

$$\begin{aligned} year2010 &= -0.1707 \times year2004 \\ &+ 0.1809 \times year2005 \\ &+ 0.0491 \times year2006 \\ &+ 0.0481 \times year2007 \\ &+ 0.8615 \times year2009 \\ &- 5.9506 \end{aligned} \quad (5)$$

Attribute	Full Data (3138)	Cluster# 0 (86)	Cluster# 1 (16)	Cluster# 2 (284)	Cluster# 3 (2)	Cluster# 4 (2750)
2004	594.5376	5306.7326	13597.5625	1749.0035	39112	224.2829
2005	640.4758	5651.4884	14795.9375	1885.2218	44090	241.2607
2006	648.6399	5701.8488	14917.6875	1925.5352	42898	244.9971
2007	643.7495	5697.8837	14920.1875	1911.4331	42757.5	241.0847
2008	637.2097	5642.5233	14943.0625	1899.9296	41425.5	237.3771
2009	628.9146	5543.2791	14474.6875	1868.8803	40342.5	237.7345
2010	612.9876	5347.186	14409.5625	1828.6092	40274.5	

TABLE I. CLUSTER CENTROIDS OF NUMBER OF NEW CASES FROM 2004 TO 2010

The correlation coefficient is 0.9979, the mean absolute error is 35.569, and the root mean squared error is 114.4877.

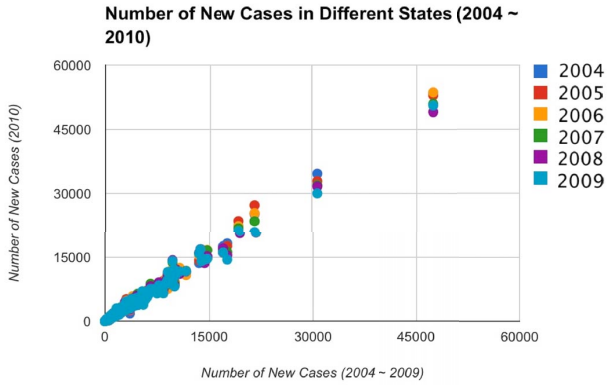

Fig. 6. Correlation between Number of New Cases in Different States from 2004 to 2009 with Number of New Cases in Different States of 2010

Figure 6 shows the correlation between number of new cases in 2010 and that in years from 2004 to 2009. The high correlation coefficient suggest that by using linear regression on existing "number of new cases" is able to predict the number of new cases of diabetes of the coming year. Also, from the plots in Figure 1, we can see that most of the states have a relative low number of new diabetic cases (under 15000) and there are states constantly having larger numbers of new diabetic cases, two of which even have more than 30000 new cases.

Apply K-Means Clustering on "number of new cases of diabetes (2004 2010)"

The K-means clustering algorithm is applied to the same dataset. The EuclideanDistance is chosen as the distance function and the number of clusters is 5. The clustered instances are:

This result, shown in the above table, suggests that there are 2 states, which are in cluster number 3, that have significantly larger number of new diabetes cases from 2007 to 2010, which is consistent with the result of linear regression. . Government and hospitals should pay special attention to those states. Special policies can be employed and doctors can even develop special treatment mechanisms to lower the large number of new diabetic cases.

Applying Linear Regression on 2010 data with 7 data types

The "number of new cases" of the year 2010 is used as the dependent variable, and all other values are used as explanatory variables. 10-fold cross-validation is used to validate the linear model. The resulting linear regression model is:

$$\begin{aligned} no.of newcases \quad = \quad & -152.3749 \times F \\ & + \quad 411.2871 \times f \\ & - \quad 463.266 \times \alpha \\ & + \quad 914.1189 \times \beta \\ & - \quad 316.8561 \times \gamma \\ & + \quad 1654.429 \end{aligned} \tag{6}$$

where F = upper confidence limit, f = lower confidence limit, α = age adjusted rate per 1000, β = age adjusted lower confidence limit, γ = rate per 1000

The correlation coefficient is 0.4511, the mean absolute error is 642.291, and the root mean squared error is 1577.0953.

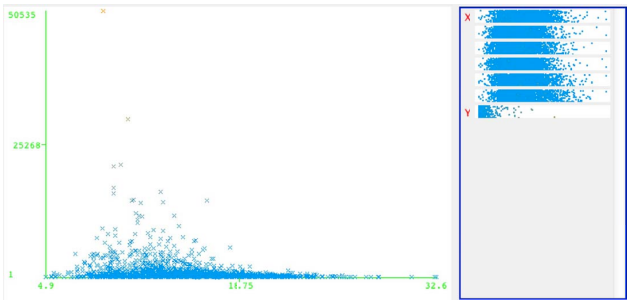

Fig. 7. Correlation between Number of New Cases in Different States in 2010 with Upper Confidence Limit of 2010

Figure 7 shows the distribution of number of new cases is quite different from the other properties. The relative low correlation coefficient suggests that the dataset is not able to be well defined by a linear model, which means more sophisticated models or data mining mechanisms should be used to look for hidden patterns,. and if possible , more fine-grained data should be collected.

All the above demonstrated data mining algorithms are the most basic ones; however, they are also the most effective ones for most of the time. When applied in real world systems, those algorithms are seldom re-implemented, but available in several frameworks that are designed from big data analysis, such as Mahout.

IV. Related Work

A. Secure Technologies for Data Protection

Proxy re-encryption and Attribute-Based Encryption [13] are two current techniques aimed at secure and private data sharing in the Cloud [15]. In our previous work [14], we focused on secure sharing of ECG data using a sensor, smartphone and the Cloud. The sensor would connect to the smartphone via Bluetooth, and stream encrypted ECG data to the Cloud. A form of proxy re-encryption is used where keys are partitioned and shared with other doctors. Revoking a user would simply involve removing the corresponding doctors key partition in the Cloud. We build upon this work and incorporate this to social networks. Tran et al. [12] utilises the idea of a proxy re-encryption scheme where the data owners private key is divided into two parts where one is stored in the data owners machine and the other on the proxy. We also use this concept in our work and applied it to data sharing with many users instead of just one user. Silva BM et al. [16] presents a data encryption solution for mobile health apps and conducts a performance evaluation comparing both symmetric and asymmetric encryption algorithms. Our work takes advantage of both symmetric and asymmetric cryptographic algorithms in order to achieve both strong security and high performance e-Health data using smartphones. The THEWS architecture proposed by Ruotsalainen et al. [17] developed a privacy management architecture to help the data owner create and manage the network as well as maintain information privacy. Routsalanien pointed out there is an asymmetric relationship between health information systems and their users because users rarely have the power "to force a system to put personal rules into effect". Our paper contributes a novel security architecture that can help balance this power difference. Even when data is encrypted, it may still be possible for a malicious Cloud provider to deduce information from the encrypted data. Zhang et al. [18] proposes a novel solution that adds noise obfuscation based on a time-series pattern to client data stored in the Cloud. This can help protect the privacy of the owners data since it prevents malicious service providers from deducing information from the encrypted data. Little of this work however, has focused on private data sharing between patients and doctors using untrusted Cloud-based servers. We presented a security model and protocol that would allow users to have a much more fine grained control of their health data.

B. Data Analysis Methods and Software Tools

1) There are already initiatives in big data analytics in healthcare data [26, 27], big IT companies like Google, Microsoft, Apple, and other web service providers have already devote their effort in healthcare data management and analysis. Different platforms have been built by those companies for storing, sharing, and analysing health related data. For example, Google Health, which is a personal health record service that allowing centralisation of different types of personal health information to one profile. However, this project was cancelled in 2011 due to low adoption rate. HealthVault, a web-based platform built by Microsoft, is designed to store and maintain personal health information. Apart from manually importing data into the platform, HealthVault allows collecting data through different types of tracking devices, such as heart rate monitors and blood pressure monitors.

2) Different data mining techniques have been proved effective when applied on health-related data. For example, in [28], researcher used patients medical record including age, sex, blood pressure and blood sugar to derive the likelihood of patients getting a heart disease. Methods that improves the Naive Bayes algorithm and the Naive Bayes algorithm itself were used in the prediction of heart diseases, and as the result showed, for predicting heart diseases the Naive Bayes algorithm achieves better result. [29] described a surveillance system that uses data mining techniques for infection control. Association rules were used in the system to generate monthly patterns of patient care data which were reviewed by experts in infection control. American Healthways used patient information to predict the likelihood of short-term health problems and provided intervention to achieve better short-term and long-term results [30]. In [31] more sophisticated data mining algorithms were used to improve healthcare operations and reducing fraud, waste, and abuse. The analysis for the case studies described in [31] was using the Hadoop/Hive data platform and open source software such as Mahout, R, and Python networkx. Specifically, topic modelling was performed on the a year of claim data and 20 hidden topics were revealed, which can be used in identifying the costly areas which need to be addressed and in comparing providers to identify fraudulent or wasteful providers. Affiliationis between providers were studied as social networks, which can be a mechanism to identify organised fraud. And temporal analysis methods are also helpful for timely detection of transient billing practices that are anomalous.

V. Conclusion

In this paper, we presented our vision of enabling secure health data sharing and analysis through technologies and methods. First, we presented our security technologies to protect health data privacy and enable secure health data sharing on cloud. Secondly, we conducted a simple health care data analysis with existing data analysis techniques and real public health data stored on the cloud. We also evaluated our security technologies and discussed the related work on data security and data analysis. This paper demonstrates the feasibility and usefulness of using cloud to collect and store the valuable health data for secure sharing and data analysis.

References

[1] J. J.P.C. Rodrigues, I. de la Torre,G. Fernndez, M. Lpez-Coronado (2013): Analysis of the Security and Privacy Requirements of Cloud-Based Electronic Health Records Systems. J Med Internet Res 2013;15(8):e186

[2] N. Regola ,N.V. Chawla NV (2013): Storing and Using Health Data in a Virtual Private Cloud. J Med Internet Res 2013;15(3):e63

[3] A. Cavoukian (2008): Privacy in the Clouds. Identity in the Information Society: Vol. 1, Issue 1: 89 - 108.

[4] F. Sabahi (2011): Cloud computing security threats and responses. Communication Software and Networks, 2011 IEEE 3rd International Conference: 245-249.

[5] J. Yao, S. Chen, S. Nepal, D. Levy, J. Zic (2010): TrustStore: Making Amazon S3 Trustworthy with Services Composition,. Cluster, Cloud and Grid Computing, 2010 10th IEEE/ACM International Conference: 600-605.

[6] T. ElGamal (1985): A public key cryptosystem and a signature scheme based on discrete logarithms. Advances cryptology, 1985: 469 - 472

[7] A. Patrizio (2007). Salesforce.com Scrambles To Halt Phishing Attacks. InternetNews.com. Published: November 7, 2007. http://www.internetnews.com/ent-news/article.php/3709836/ Salesforce-com+Scrambles+To+Halt+Phishing+Attacks.htm

[8] C. Arthur (29/04/2011). PlayStation Network: hackers claim to have 2.2m credit cards. The Guardian Technology Blog. http://www.guardian.co.uk/technology/blog/2011/apr/29/playstation-network-hackers-credit-cards

[9] F. Rocha, S. Abreu, M. Correia (2011): The Final Frontier: Confidentiality and Privacy In the Cloud: 44-50.

[10] T. Donker, K. Petrie, J. Proudfoot, J. Clarke, M.R. Birch, H. Christensen (2013): Smartphones for Smarter Delivery of Mental Health Programs: A Systematic Review. J Med Internet Res 2013;15(11):e247

[11] P. Karvelas: Australias mental health system must become more efficient. The Australian. March 11, 2014. Source: http://www.theaustralian.com.au/national-affairs/policy/australias-mental-health-system-must-become-more-efficient/story-fn59nokw-1226850819260#

[12] D.H. Tran, N. Hai-Long, Z. Wei, N.W. Keong (2011): Towards security in sharing data on cloud-based social networks. 8th International Conference on Information, Communications and Signal Processing(ICICS) 2011: 1 - 5.

[13] V. Goyal, O. Pandey, A. Sahai, B. Waters (2006): Attribute-based encryption for fine-grained access control of encrypted data. 13th ACM Conference on Computer and Communications Security (CCS 06): 89 - 98.

[14] D. Thilakanathan, S. Chen, S. Nepal, R. A. Calvo (2013): Secure data sharing in the Cloud. Book chapter in Security, Privacy, and Trust in Cloud Systems, by Springer (2013): 45 - 72

[15] D. Thilakanathan, S. Chen, S. Nepal, R. A. Calvo (2013): Secure and Controlled Sharing of Data in Distributed Computing. 2nd IEEE International Conference on Big Data Science and Engineering (2013): 825 - 832

[16] B.M. Silva, J.J. Rodrigues, F. Canelo, I.C. Lopes, L. Zhou (2013): A Data Encryption Solution for Mobile Health Apps in Cooperation Environments. J Med Internet Res 2013; 15(4):e66

[17] P.S. Ruotsalainen, B. Blobel, A. Seppala, P. Nykanen (2013): Trust Information-Based Privacy Architecture for Ubiquitous Health. JMIR Mhealth Uhealth 2013;1(2):e23

[18] G. Zhang, X. Liu, Y. Yang J. Chen (2014): A Time-Series Pattern Based Noise Generation Strategy for Privacy Protection in Cloud Computing. Cluster, Cloud and Grid Computing (CCGrid). 2012 12th IEEE/ACM International Symposium: 458,465

[19] W. Raghupathi (2010): Data Mining in Health Care. In Healthcare Informatics: Improving Efficiency and Productivity. Edited by Kudyba S. Taylor & Francis; 2010:211223.

[20] Explorys: Unlocking the Power of Big Data to Improve Healthcare for Everyone. https://www.explorys.com/docs/data-sheets/explorys-overview.pdf.

[21] Knowledgent: Big Data and Healthcare Payers; 2013. http://knowledgent. com/mediapage/insights/whitepaper/482.

[22] K. P. Murphy (2012): Machine learning: a probabilistic perspective. MIT press.

[23] S. P. Lloyd (1957): Least square quantization in PCM. Bell Telephone Laboratories Paper. Published in journal much later: S. P. Lloyd (1982): Least squares quantization in PCM. IEEE Transactions on Information Theory 28 (2): 129 - 137.

[24] G. Hamerly, C. Elkan (2002): Alternatives to the k-means algorithm that find better clusterings. Proceedings of the eleventh international conference on Information and knowledge management (CIKM): 600 - 607

[25] I. H. Witten, E. Frank (2005): Data Mining: Practical Machine Learning Tools and Techniques. Second edition, Elsevier: San Francisco, ISBN 0-12-088407-0.

[26] R. Steinbrook (2008): Personally controlled online health data-the next big thing in medical care?. New England Journal of Medicine, 358(16), 1653.

[27] B. Kayyali, D. Knott, S. Van Kuiken (2013): The big-data revolution in US health care: Accelerating value and innovation. Mc Kinsey & Company.

[28] K. Srinivas, B. K. Rani, A. Govrdhan (2010): Applications of data mining techniques in healthcare and prediction of heart attacks. International Journal on Computer Science and Engineering (IJCSE): 250-255.

[29] S.E. Brosette, A.P. Spragre, W.T. Jones, S.A. Moser (2000): A data mining system for infection control surveillance. Methods Inf Med 2000: 303-310.

[30] M. Ridinger (2002): American Healthways uses SAS to improve patient care. DM Review:139.

[31] V. Chandola, S. R. Sukumar, J. C. Schryver (2013): Knowledge discovery from massive healthcare claims data. In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining:1312-1320

[32] Centers for Disease Control and Prevention. (2014) County_EXCELstateINCIDENCE [excel]. Retrieved from http://www.cdc.gov/diabetes/atlas/countydata/DMINCID/INCIDENCE_ALL_STATES.xls

[33] "Healthcare costs rise to $130bn, or $5800 per Australian: report". http://www.news.com.au/lifestyle/health/health-spending-reaches-130b-report/story-fneuz9ev-1226481443042