Hindawi
Mathematical Problems in Engineering
Volume 2017, Article ID 8632183, 7 pages
https://doi.org/10.1155/2017/8632183

Research Article

A Privacy-Preserving Intelligent Medical Diagnosis System Based on Oblivious Keyword Search

Zhaowen Lin,[1,2,3] Xinglin Xiao,[1,2,3] Yi Sun,[2,3,4] Yudong Zhang,[5,6] and Yan Ma[1]

[1]*Network and Information Center, Institute of Network Technology, Beijing University of Posts and Communications, Beijing 100876, China*
[2]*Science and Technology on Information Transmission and Dissemination in Communication Networks Laboratory, Beijing 100876, China*
[3]*National Engineering Laboratory for Mobile Network Security (No. [2013] 2685), Beijing 100876, China*
[4]*Network and Information Center, Institute of Network Technology and Institute of Sensing Technology and Business, Beijing University of Posts and Communications, Beijing 100876, China*
[5]*School of Computer Science and Technology, Nanjing Normal University, Nanjing, Jiangsu 210023, China*
[6]*Department of Informatics, University of Leicester, Leicester LE1 7RH, UK*

Correspondence should be addressed to Xinglin Xiao; 252666437@qq.com and Yi Sun; sybupt@bupt.edu.cn

Received 28 June 2017; Accepted 30 August 2017; Published 8 October 2017

Academic Editor: Chanho Jung

One of the concerns people have is how to get the diagnosis online without privacy being jeopardized. In this paper, we propose a privacy-preserving intelligent medical diagnosis system (IMDS), which can efficiently solve the problem. In IMDS, users submit their health examination parameters to the server in a protected form; this submitting process is based on Paillier cryptosystem and will not reveal any information about their data. And then the server retrieves the most likely disease (or multiple diseases) from the database and returns it to the users. In the above search process, we use the oblivious keyword search (OKS) as a basic framework, which makes the server maintain the computational ability but cannot learn any personal information over the data of users. Besides, this paper also provides a preprocessing method for data stored in the server, to make our protocol more efficient.

1. Introduction

Due to the increasing health consciousness, the intelligent medical diagnosis system has received immense popularity around the world. Nowadays, patients can get personalized online medical services based on the original health data submitted by themselves, without the need for doctors. That is, patients can obtain a reasonable diagnosis anytime and anywhere, as long as they have health data and are willing to transfer them to the server. People are getting more and more accustomed to convenience that is made possible by the IMDS. Moreover, this kind of intelligent medical service is inexpensive, which means, via popularizing IMDS, we can reduce the expenditure of public medical services. All the above shows that IMDS has a bright future and will be an essential part in the future life [1–3].

However, the privacy disclosure issue has become a big obstacle for developing IMDS. Patients are afraid that their sensitive personal information may fall into wrong hands. In most cases, poor security mechanisms and weak safety awareness are main reasons for the information leakage. For example, in 2009, the AvMed Health Plans, a large nonprofit US health plans org, exposed the personal information of 200,000 subscribers and their dependents, as a result of the theft of two company laptops that contain sensitive information [4]. Types of divulged personal information include names, addresses, phone numbers, social security numbers, and protected health information. For this reason, designing a privacy-preserving IMDS, which can ensure the privacy of patients, is an urgent task. Besides, the security of the server also cannot be ignored. Currently, on this respect, various schemes have been proposed [5–7]. One example is [8]; it

designs a privacy-preserving recommendation system using homomorphic encryption. This system allows patients to rate physicians based on their satisfaction, so that other patients can choose a popular physician via these ratings. Another paper that also describes a physicians recommendation system based on hybrid matrix factorization is raised by [9]. This paper applies text sentiment analysis to analyze patient comments that increase accuracy in grading physicians.

In this paper, combining the additive homomorphic cryptosystem and the oblivious keyword search, we propose a privacy-preserving IMDS, and the security of data stored in server can also be ensured. Patients can get an effective and reasonable diagnosis in our system, after uploading their health examination data to the server. The diagnosis will tell the patient which parameters are not in the normal range, and which diseases he is likely to get. Moreover, our system contains a preprocessing phase to reduce the calculation amount of search.

The rest of the paper is organized as follows. Section 2 gives a brief overview of related works. In Section 3, we show the framework and notions. Section 4 provides the detailed description of our protocol. Section 5 analyzes the security of this system. In the end, Section 6 discusses the performance of this system and concludes the paper.

2. Preliminaries

2.1. Paillier Cryptosystem. In 1999, Paillier proposed a homomorphic public key cryptosystem called Paillier cryptosystem, which is a common encryption scheme used for data protection [10]. The homomorphic property of this cryptosystem means that people can get the sum of two plaintexts, according to decrypting the product of their corresponding ciphertexts. Besides, Paillier cryptosystem is a semantically secure cryptosystem, which means no information about the plaintext can be obtained from the according ciphertext. Therefore, calculations over ciphertexts in Paillier cryptosystem will not reveal any extra information. The Paillier cryptosystem is briefly described as follows.

Let m be a message, p and q are two large prime numbers, $n = pq$, g is a random integer, and r is a random number. The encryption of m is defined as

$$c = g^m \cdot r^m \bmod n^2. \tag{1}$$

The decryption of m is defined as

$$m = L\left(c^\lambda \bmod n^2\right) \cdot \mu \bmod n, \tag{2}$$

where $L(u) = (u-1)/n$, $\mu = (L(g^\lambda \bmod n^2))^{-1} \bmod n$, and $\lambda = lcm(p-1, q-1)$.

2.2. Oblivious Keyword Search (OKS). Oblivious keyword search (OKS) protocol was proposed by Wakaha Ogata and Kaoru Kurosawa in 2002; it is a secure keyword search scheme between two parties [10]. In an OKS protocol, there is a server S that maintains some secret data and a user U that is allowed to search for the data containing the keyword chosen by the user; this chosen keyword is secret to S. Next, we introduce an efficient k-out-of-n OKS protocol based on RSA blind signature (OKS_k^n protocol) as follows.

In the OKS_k^n protocol, S stores data $B_1, \ldots, B_n$. Define

$$B_i = (w_i, c_i), \tag{3}$$

where $\Delta = \{w_1, \ldots, w_n\}$ is the set of keywords, $w_i \in \Delta$, and c_i is the corresponding content.

2.2.1. Submit Phase. S generates a public key (N, e) and a secret key d of RSA and then for $i = 1, \ldots, n$ computes

$$\begin{aligned} K_i &= \left(H\left(w_i\right)\right)^d \bmod N, \\ E_i &= G\left(w_i \parallel K_i \parallel i\right) \oplus \left(0^l \parallel D\left(c_i\right)\right), \end{aligned} \tag{4}$$

where H is a hashing function and G is a pseudorandom generator. Then S send $E_1, \ldots, E_n$ to U.

2.2.2. Transfer Phase. At each transfer round j, firstly, U choose a keyword w_j^* and a random element r to compute

$$Y = r^e H\left(w_j^*\right) \bmod N, \tag{5}$$

and then U send Y to S. Secondly, S computes $K' = Y^d \bmod N$ and sends it to U. U computes $K = K'/r = H(w_j^*)^d \bmod N$. Finally, let $J = \emptyset$; J is the set that stores the search results. For $i = 1, \ldots, n$, U computes

$$\left(a_i \parallel b_i\right) = E_i \oplus G\left(w_j^* \parallel K \parallel i\right). \tag{6}$$

If $a_i = 0^l$, then U add (i, b_i) to J.

2.3. Useful Tools. We introduce here two generic subprotocols from literatures [8, 11–13], which will be employed in our protocol, and all these protocols can be easily implemented by using the Scalar Product Protocol; it is a standard protocol mentioned in [13]. Moreover, let $[a]$ denote an encrypted value of a. The subprotocols are described as follows:

(1) Bits(l, $[a]$): return $[a]_{lB}$, which is an encryption of l least significant bits of the plaintext of $[a]$; that is, $[a]_{lB} = \langle [a_0], \ldots, [a_{l-1}] \rangle$;

(2) Min($[a_1]_{lB}$, $[a_2]_{lB}$): return an encrypted bit $[b]$, where $b = 1$ iff $a_1 \le a_2$.

3. The Framework

This section we develop the framework of our privacy-preserving IMDS. At the beginning, we describe two entities interacting with each other: the user U and the server S.

In the field of U, each U measures n private health examination parameters of his own body in advance, such as blood glucose, vital capacity, and vitamin content. U uploads those health data to S for getting health service. We denote n parameters that U measured as $p_1, \ldots, p_n$ and denote the corresponding measured value of $p_1, \ldots, p_n$ as $r_{p1}, \ldots, r_{pn}$.

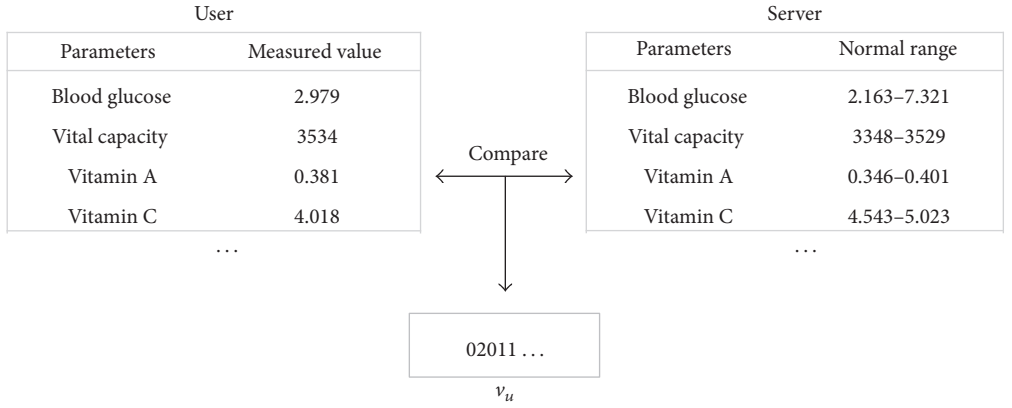

Figure 1: The structure of our intelligent medical diagnosis system.

In the aspect of S, it stores a wide range of diseases $d_1, \ldots, d_h$ and generally more health examination parameters $p_1, \ldots, p_m$ than U, $n \leq m$. Besides, we define v_{di} as the parameter vector of $d_i \in \{d_1, \ldots, d_h\}$, which represents the relationship between the disease d_i and all kinds of parameters $p_1, \ldots, p_m$ stored in S. Suppose the jth bit of v_{di} is 1, which means when disease d_i occurs, the measured value of parameter p_j will be smaller than the normal range. Similarly, 2 means p_j will be larger than the normal range, and 0 means p_j will in the normal range. Obviously, v_{di} has m bits. Thus, each $p_i \in \{p_1, \ldots, p_m\}$ has the upper limit u_{pi} and the lower limit l_{pi} to represent the normal range. In S, the above-mentioned d_i and v_{di} are represented by $D_1, \ldots, D_h$. Define

$$D_i = (d_i, v_{di}). \tag{7}$$

Next we introduce our framework. It contains three phases: submit phase, preprocessing phase, and search phase. The submit phase builds on the Paillier cryptosystem. U firstly uploads his health examination data $[r_{p1}], \ldots, [r_{pn}]$ to S in the form of encryption, using a secret key generated by the Paillier cryptosystem. At the meantime, U directly tells S which parameters he has uploaded, namely, $p_1, \ldots, p_n$. Then, S compares all $\{[r_{pi}]\}_{i\in[1,n]}$ with $[u_{pi}]$ and $[l_{pi}]$ to get the parameter vector $[v_u]$, which represents whether each health examination parameter of U is in the normal range. The generation procedure of v_u is depicted in Figure 1. In addition, the comparison is completed by functions Bits and Min mentioned in Section 2. In this comparison, S has no knowledge of the uploaded data but maintains the computational ability. Finally, S returns $[v_u]$ to U in encrypted form.

To facilitate the following search operation and make it faster, we propose a transform method in the preprocessing phase. S uses this method to split and reshape all v_{di} into the keywords set WP, and reorganizes $D_1, \ldots, D_h$ to the new data structures $B_1, \ldots, B_{h^*}$, which are pairs of keyword and disease. After that, to prepare the oblivious keyword search (OKS) used in the search phase, S calculates $E_1, \ldots, E_{h^*}$ and sends them to U. $E_1, \ldots, E_{h^*}$ can be understood as the encrypted form of $B_1, \ldots, B_{h^*}$.

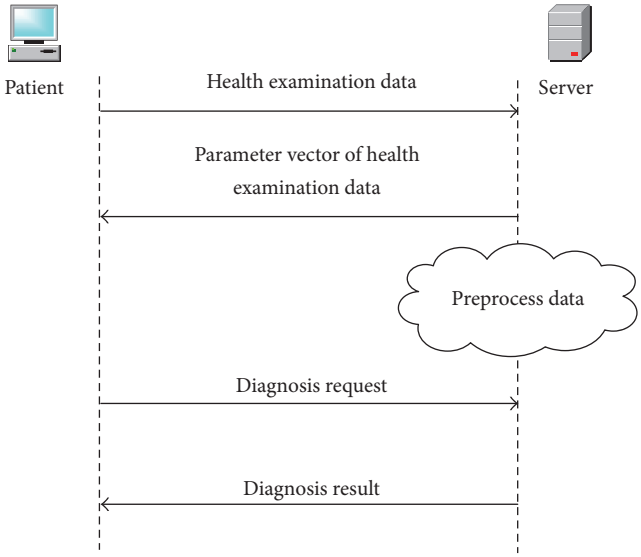

Figure 2: The structure of our privacy-preserving intelligent medical diagnosis system.

In the search phase, we utilize the oblivious keyword search (OKS) to realize privacy-preserving search. Firstly, U constructs a set of keywords on the basis of v_u and successively calculates Y for each keyword and sends them to S. We can also understand Y as the encrypted form of a keyword. Then, S computes K using Y and returns it to U. Next, U retrieves keywords constructed by v_u in the $E_1, \ldots, E_{h^*}$ via (6) and thus gets their corresponding diseases. Finally, according to the search results, S finds the disease with the highest frequency as the diagnosis for U. The whole structure of our IMDS is depicted in Figure 2.

4. Our Protocol

In this section, we detail the implementation of our protocol.

4.1. Submit Phase. The submit process is established on the Paillier cryptosystem; it is an additive homomorphic

Input: health examination data $r_{p1}, \ldots, r_{pn}$, length parameter l
Output: user parameter vector v_u
(1) U: executes $[r_{pi}]_{lB} \leftarrow \text{Bits}(l, [r_{pi}])$
(2) S: executes $[u_{pi}]_{lB} \leftarrow \text{Bits}(l, [u_{pi}])$ and $[l_{pi}]_{lB} \leftarrow \text{Bits}(l, [l_{pi}])$
(3) S: executes $[m_1] \leftarrow \text{Min}([r_{pi}]_{lB}, [l_{pi}]_{lB})$ and $[m_2] \leftarrow \text{Min}([u_{pi}]_{lB}, [r_{pi}]_{lB})$
(4) S: sends $[m_1]$ and $[m_2]$ to U
(5) U: decrypts $[m_1]$ and $[m_2]$, get m_1 and m_2
(6) **if** $m_1 == 1$ **then**
(7) executes the ith bit of $v_u \leftarrow 1$
(8) **else**
(9) **if** $m_2 == 1$ **then**
(10) executes the ith bit of $v_u \leftarrow 2$
(11) **else**
(12) executes the ith bit of $v_u \leftarrow 0$
(13) **end if**
(14) **end if**
(15) **return** v_u

Algorithm 1: Submission of health examination data.

cryptosystem. First, U generates a secret key k and uses the public key encryption to distribute k to S. Then, for all $r_{pi} \in \{r_{p1}, \ldots, r_{pn}\}$, U sends the $[r_{pi}]_{lB}$ to S using Bits. Similarly, S executes Bits to get $[u_{pi}]_{lB}$ and $[l_{pi}]_{lB}$. Next, Min can help S to compare $[r_{pi}]_{lB}$ with $[l_{pi}]_{lB}$ and $[u_{pi}]_{lB}$. The comparison results will be transferred to U after being encrypted. According to these results, U constructs the parameter vector v_u. We define the parameter vector v_u of the user U as follows:

$$\text{the } i\text{th bit of } v_u = \begin{cases} 0, & l_{pi} \le r_{pi} \le u_{pi} \\ 1, & r_{pi} < l_{pi} \\ 2, & r_{pi} > u_{pi}. \end{cases} \tag{8}$$

The algorithm is shown in Algorithm 1.

4.2. Preprocessing. To make following search process faster, data $D_1, \ldots, D_h$ stored in S need to be reorganized into a new structure. Let $W = \{w_{i1}, w_{12}\}_{i \in [1,m]}$ be the set of keywords. Define

$$\begin{aligned} w_{i1} &= i \parallel 1, \\ w_{i2} &= i \parallel 2, \end{aligned} \tag{9}$$

where i represents the mark number of p_i. That is, a keyword is obtained by concatenating the mark number of a parameter with its values in v_{di}. We should note that only those abnormal parameters will be selected to create keywords. Now we define the new data structure as

$$B_j = \left(w_j, d_j^*\right), \tag{10}$$

where $w_j \in W$ and d_j^* is one of the diseases whose parameter vector v contains w_j. The specific transformation method is described in Algorithm 2.

Next, S generates a public key (N, e) and a secret key d of RSA and then only publishes (N, e). With the secret key d and the hash value of each $w_j \in W$, S computes E_j and K_j. Finally, S outputs $\{E_j\}_{j \in [1,h^*]}$ to U, where $h^* = 2n$. In addition, let l be a security parameter, G be a pseudorandom generator, and H be a hash function. The computational process is also described in Algorithm 2.

4.3. Transfer Phase. In the search phase, U successively finds out abnormal parameters by judging whether their values in v_u equal 0 and constructs keywords for them. For example, if the ith bit of v_u equals 0, then p_i is a normal parameter for U; otherwise we say p_i is an abnormal parameter. At each transfer round j, after choosing an abnormal parameter, U creates a keyword w_j^* by concatenating the mark number of chosen parameter with its values in v_u. Then, U calculates Y and sends it to S, which is the encryption of w_j^*. S returns the search result K'. According to decrypting K', U retrieves keywords constructed by v_u in the $E_1, \ldots, E_{h^*}$ via (6) and thus gets their corresponding diseases. We maintain these diseases in a list as suspected diseases at each transfer round and denote this list as J. When all abnormal parameters are traversed, we compute the frequency of occurrence of diseases in J, and output a diseases list $D_{\max}$ that consists of diseases with the highest frequency as the final search result. The algorithm is described in Algorithm 3.

5. Safety Analysis

We present here the analysis of the security of our system.

Lemma 1. *The problem of computing nth residue classes is believed to be computationally difficult.*

Lemma 2. *RSA known target inversion problem (RSA-KTI) is hard.*

Firstly, we discuss the security of U. In the submit phase, $\{r_{pi}\}_{i \in [1,n]}$, the health examination data of the patient U are

Input: $D_1, \ldots, D_h$, security parameter l, pseudo-random generator G, hash function H
Output: $E_1, \ldots, E_{h^*}$
(1) **for** $i \in [1, h]$ **do**
(2) **for** $j \in [1, m]$ **do**
(3) **if** the jth bit of v_{di} equals to 1 **then**
(4) create a new $B = (w_{j1}, d_i)$ and save it
(5) **else**
(6) **if** the jth bit of v_{di} equals to 2 **then**
(7) create a new $B = (w_{j2}, d_i)$ and save it
(8) **end if**
(9) **end if**
(10) **end for**
(11) **end for**
(12) generates a public key (N, e) and a secret key d of RSA, then publishes (N, e)
(13) **for** $k \in [1, h^*]$ **do**
(14) compute $K_k \leftarrow (H(w_k))^d \bmod N$
(15) compute $E_k \leftarrow G(w_k \parallel K_k \parallel k) \oplus (0^l \parallel d_k^*)$
(16) **end for**
(17) **return** $E_1, \ldots, E_{h^*}$ to U

Algorithm 2: Preprocessing of server data.

Input: user parameter vector v_u, $B_1, \ldots, B_{h^*}$
Output: diseases list $D_{\max}$
(1) **for** $i \in [1, n]$ **do**
(2) U: $j \leftarrow 1$
(3) **if** the ith bit of v_u equals to 1 **then**
(4) execute $w_j^* \leftarrow i \parallel 1$
(5) **else**
(6) **if** the ith bit of v_u equals to 2 **then**
(7) execute $w_j^* \leftarrow i \parallel 2$
(8) **else**
(9) continue
(10) **end if**
(11) **end if**
(12) chooses a random element r
(13) compute $Y \leftarrow r^e H\ (w_j^* \bmod N)$
(14) send Y to S
(15) S: compute $K' \leftarrow Y^d \bmod N$ and send to U
(16) U: compute $K \leftarrow K'/r = H(w_j^*)^d \bmod N$
(17) execute $J \leftarrow \emptyset$
(18) **for** $k \in [i, h^*]$ **do**
(19) compute $(a_k \parallel b_k) = E_k \oplus G(w_j^* \parallel K \parallel k)$
(20) **if** $a_k = 0^l$ **then**
(21) U insert b_k into J
(22) break
(23) **end if**
(24) **end for**
(25) $j \leftarrow j + 1$
(26) **end for**
(27) U: compute frequency of occurrence of diseases in the J, output a diseases list $D_{\max}$ that consists of diseases with the highest frequency.
(28) **return** $D_{\max}$

Algorithm 3: Search for diagnosis results.

these data that the attacker T wants to steel. However, the only message that T may get is ciphertext $[r_{pi}]$, which is encrypted by the private key k of U. The Paillier cryptosystem produces this k and it is the basis of this phase. If T wants to get the plaintext r_{pi} without knowing private key, his task is, given a composite n and an integer z, deciding whether z is n-residue modulo n^2 or not. By Lemma 1, we know that to complete this task is hard. Therefore, T can not get the plaintext r_{pi}. In the search phase, U's important and private data is w_j^*, but S has no information on w_j^* because they are blinded in the RSA blind signature scheme.

Next, we prove the security of S, assuming attacker T is allowed to make at most t queries to S. At first, T behaves as if it were S. T generates and sends (N, e) to U. After that, T randomly chooses E_j and sends them to U. From Lemma 2, we could know it is hard for T to get the plaintext.

6. Conclusion

In this paper we propose a privacy-preserving intelligent medical diagnosis system and also discuss how privacy-preserving protocols can be used for protecting sensitive patient data in medical scenarios. This system applies two security protocols, Paillier cryptosystem and oblivious keyword search (OKS), to medical diagnosis, and it can be put into practice. Besides, our system also has following advantages:

(1) Previously mentioned information security requirements are achieved. That is, privacy of patient data and security of server get properly protected. Thus, server is blind to the personal information of patients; patients also know nothing about the data maintained in server.

(2) Our system reduces the calculation amount of search by adding the preprocessing phase. This phase can link each keyword w with the corresponding disease name d. Hence, we can directly get the d after w is retrieved instead of looking up for d in the database. In the search phase, system only needs to retrieve n submitted parameters instead of every parameter in the database, which also makes search phase more efficient.

(3) System is able to support multiple possible diseases instead of a single result for patients. When the submitted data are not enough to determine only one disease, system will show patients several results for reference.

Therefore, our privacy-preserving intelligent medical diagnosis system is capable of providing efficient and reasonable diagnosis for patients. We view this work as a start of our follow-up work. There are still a lot of work to be done. In the future, we will focus on the research for multiuser and multiserver IMDS. It is necessary for developing a system that can simultaneously and securely provide service for users.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work is supported by the National High Technology Research and Development Program of China (863 Program) (Grant no. 2013AA014702), the Fundamental Research Funds for the Central Universities (BUPT2016RC48, Grant 2014ZD03-03), and the National Natural Science Foundation of China (Grant no. 61601041).

References

[1] M. H. Tekieh and B. Raahemi, "Importance of data mining in healthcare: a survey," in *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM '15)*, pp. 1057–1062, Paris, France, August 2015.

[2] L. Xu, C. Jiang, J. Wang, J. Yuan, and Y. Ren, "Information security in big data: privacy and data mining," *IEEE Access*, vol. 2, pp. 1151–1178, 2014.

[3] V. S. Verykios, E. Bertino, I. N. Fovino, L. P. Provenza, Y. Saygin, and Y. Theodoridis, "State-of-the-art in privacy preserving data mining," *SIGMOD Record*, vol. 33, no. 1, pp. 50–57, 2004.

[4] A. Moscaritolo, "Sensitive laptops stolen from Fla. Health insurance provider," http://www.scmagazine.com//sensitive-laptops-stolen-from-fla-health-insurance-provider/article/163618/.

[5] M. Adjedj, J. Bringer, H. Chabanne, and B. Kindarji, "Biometric identification over encrypted data made feasible," in *International Conference on Information Systems Security*, Lecture Notes in Computer Science, pp. 86–100, 2009.

[6] S. Katzenbeisser and M. Petkovic, "Privacy-preserving recommendation systems for consumer healthcare services," in *Proceedings of the 3rd International Conference on Availability, Security, and Reliability (ARES '08)*, pp. 889–895, Barcelona, Spain, March 2008.

[7] I. Song and N. V. Marsh, "Anonymous indexing of health conditions for a similarity measure," *IEEE Transactions on Information Technology in Biomedicine*, vol. 16, no. 4, pp. 737–744, 2012.

[8] T. R. Hoens, M. Blanton, A. Steele, and N. V. Chawla, "Reliable medical recommendation systems with patient privacy," *ACM Transactions on Intelligent Systems and Technology*, vol. 4, no. 4, article 67, 2013.

[9] Y. Zhang, M. Chen, D. Huang, D. Wu, and Y. Li, "IDoctor: personalized and professionalized medical recommendations based on hybrid matrix factorization," *Future Generation Computer Systems*, vol. 66, pp. 30–35, 2017.

[10] P. Paillier, "Public-key cryptosystems based on composite degree residuosity classes," in *EUROCRYPT99*, 238, p. 223, Springer, Berlin, Germany, 1999.

[11] B. Schoenmakers and P. Tuyls, "Efficient binary conversion for paillier encrypted values," in *Proceedings of the Advances in Cryptology - EUROCRYPT, 2006, International Conference on the Theory and Applications of Cryptographic Techniques*, pp. 522–537, St. Petersburg, Russia, 2006.

[12] P. Bunn and R. Ostrovsky, "Secure two-party k-means clustering," in *Proceedings of the 14th ACM Conference on Computer*

and Communications Security (CCS '07), pp. 486–497, Alexandria, Va, USA, November 2007.

[13] T. R. Hoens, M. Blanton, and N. V. Chawla, "A private and reliable recommendation system for social networks," in *Proceedings of the 2nd IEEE International Conference on Social Computing (SocialCom '10)*, pp. 816–825, Minneapolis, Minn, USA, August 2010.