Privacy-Aware Blind Cloud Framework for Advanced Healthcare

Subhadeep Sarkar, *Student Member, IEEE*, Subarna Chatterjee, *Student Member, IEEE*, Sudip Misra, *Senior Member, IEEE*, and Rajesh Kudupudi

***Abstract*— This letter proposes a novel privacy-aware "*blind*" cloud infrastructure to be utilized for storage, processing, and organization of health data. Traditional healthcare systems rely on cloud computing servers for back-end storage and processing. However, cloud servers are heavily vulnerable to privacy threats and the problem is even more intense as physiological data carry sensitive information. To resolve the aforementioned issue, this letter proposes the blind cloud framework. The goal is to take advantage of the enormous computing and storage abilities of the cloud servers, and yet maintain data anonymity simultaneously. To preserve the privacy of the medical data, the cloud server is forcefully blinded, i.e., the identities of the patients are masked off and a pseudo-identity is generated, thereby, obtaining unidentified in-cloud data for storage and analysis. We also propose a parallel method to be executed within the non-cloud servers for efficient and lossless identity management and retrieval. Results indicate that the performance of the processes of pseudo-identity generation and identity retrieval is independent of the data volumes, and negligibly vary with the increase in the number of the clients of the system.**

***Index Terms*— Cloud computing, privacy, blindness, wireless body area networks (WBANs), Internet of Things (IoT).**

I. Introduction

THE use of Wireless Body Area Networks (WBANs) is emerging at a rapid rate due to their potential to offer the advanced healthcare services. A WBAN consists of many wearable wireless sensor nodes that collect various physiological data (e.g. heart rate, body temperature, ECG, and galvanic skin response) of a person and send it to a Local Data Processing Unit (LDPU), which is generally a personal digital assistant. Data from the LDPU are processed and aggregated and subsequently transmitted to doctors remotely, wherein the data are analyzed for medical diagnosis. In today's world of Internet of Things (IoT), cloud servers primarily serve the backbone of medical data repository. However, the storage of health data within cloud servers can be dangerous, as these data are very sensitive and the servers are tremendously vulnerable to privacy attacks.

Lin and Squicciarini [1] proposed a policy-based framework for the protection of a client's privacy within the cloud computing core. The proposed framework consists of three key sectors, viz., policy ranking, policy integration, and policy enforcement to enable data protection. A privacy-preserving public auditing scheme was proposed by Wang et al. [2], which uses a public key-based authentication through homomorphic encryption and random masking techniques to meet their requirements. However, none of these works preserve the anonymity of the data. In other words, the privacy of data – specially in the context of modern e-healthcare, faces the biggest challenge in the preservation of the anonymity of the patient's data. The problem looms larger when a third-party (such as the cloud service provider) is invoked into the architecture for the improvement of the quality of service (QoS). Existing data monitoring policies are deficient in preserving the privacy and retaining the anonymity of the data – specifically in the cloud computing environment [3], [4]. For few works which do not directly expose the anonymity of the subjects, it is still possible for the *honest-but-curious* cloud service provider (CSP) [5] to derive meaningful information from the anonymous data.

In order to address the above-mentioned problem, in this work, we propose a novel, privacy-aware cloud framework – the "blind" cloud framework. The goal is to make the cloud platform "blind" forcibly and yet take the advantage of the enormous computing and storage abilities of such servers. *Blindness* of a cloud server essentially refers to the inability to see or retrieve identities of patients. The cloud is just equipped with the intelligence of building innumerable rule-based clusters and aggregate the data meaningfully. Additionally, for the purpose of storage, the cloud servers follow a pattern for storing anonymous clusters of physiological data.

TABLE I
TABLE OF NOTATIONS

Symbols	Corresponding sets	Meaning
h_i	$H = \{h_1, h_2, \ldots, h_n\}$	Hospitals
$S_{i,1}$	–	Hospital server 1 of h_i
$S_{i,2}$	–	Hospital server 2 of h_i
$k_i(t)$		No of patients served by h_i at t
$P_{i,j}$	$P_i(t) = \{P_{i,j}\} 1 \leq i \leq k_i(t)$	j^{th} patient served by h_i
$\mathcal{W}_{i,j}$	$\mathcal{W}_i$	WBAN corresponding to $P_{i,j}$
$L_{i,j}$	$L_i(t) = \{L_{i,j}\}, 1 \leq j \leq k_i(t)$	LDPU of h_i serving $P_{i,j}$
$d_i(t)$	$D_i(t) = \{D_{i,j}\}, 1 \leq j \leq d_i(t)$	No of doctors in h_i at time t

II. System Model

The proposed work considers a multi-organizational scenario comprising of n number of hospitals, as shown in Fig. 1. Table I highlights some of the important symbols and their meaning that are used in this work. Additionally, it should be noted that as a particular doctor may choose to serve multiple patients, $d_i(t) \leq k_i(t)$.

The architecture within cloud involves a set of n distinct Virtual Machines (VMs) denoted by $\mathcal{V} = \{\mathcal{V}_1, \mathcal{V}_2, \ldots, \mathcal{V}_n\}$, such that VM $\mathcal{V}_i$ is allocated for hospital h_i. As indicated in Fig. 1, there are three distinct levels of data processing:

Manuscript received June 27, 2017; accepted August 8, 2017. Date of publication August 14, 2017; date of current version November 9, 2017. The associate editor coordinating the review of this letter and approving it for publication was M. Khabbaz. *(Corresponding author: Subarna Chatterjee.)*

S. Sarkar, S. Chatterjee, and S. Misra are with IIT Kharagpur, Kharagpur 721302, India (e-mail: subhadeepsarkarybs@yahoo.com; chatterjeesubarna@yahoo.com; smisra.editor@gmail.com).

R. Kudupudi is with the Indian Institute of Information Technology Design and Manufacturing, Chennai 600127, India.

Digital Object Identifier 10.1109/LCOMM.2017.2739141

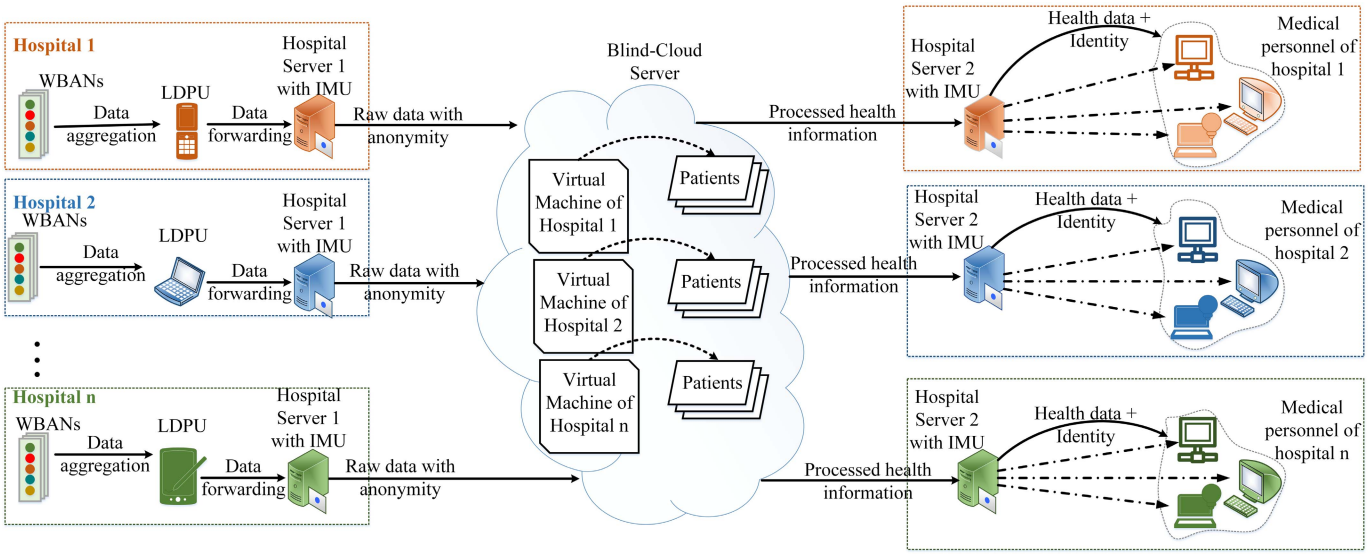


Fig. 1. Overall business architecture for a multi-organization multi-sensor deployment.

1. *Identity Masking*: In the first level, we propose a module for intelligent masking – Identity Management Unit (IMU). The IMU operates at two levels of processing and comprises of two components. The first component operates in Layer 1 and is responsible for masking identities. $\forall h_i \in H$, data packets from multiple WBANs are routed via the LDPUs to the $S_{i,1}$. Within the IMU, the data packets are made anonymous, i.e., the identity information from every data packet is masked and chopped off. The resultant anonymous packets are transmitted to $\mathcal{V}_i$ for further processing.

2. *In-cloud Processing*: Within the cloud servers, $\forall \mathcal{V}_i \in \mathcal{V}$, the data are meaningfully aggregated. Aggregation of the data happens over data clusters formed over a sequence of numbers that are algorithmically generated by the IMU in the previous level. Followed by the aggregation, the aggregated packets are routed to $S_{i,2}$ to be processed by the set of doctors.

3. *Identity Retrieval*: At $S_{i,2}$ lies the second component of IMU that rebuilds the identity of the patients. The anonymous data packets are combined and ratified with the new identity information that is generated at this level. This complete packet is thereby transmitted at the doctors' end.

III. Pre-Processing at the Hospital Servers

In this Section, we thoroughly present the processing of the algorithms required to be executed at the hospital servers to ensure blindness in cloud. The algorithms are executed to carry out two different functionalities – *data aggregation* and *identity management*, as discussed in the following subsections.

A. Algorithms for Data Packaging

A WBAN $\mathcal{W}_{i,j}$, of patient $P_{i,j}$, is assumed to be comprised of σ wireless sensors. The set of sensors is denoted by $s^{\mathcal{W}_{i,j}} = \{s_1, s_2, \ldots, s_\sigma\}$. Each sensor $s_i, 1 \leq i \leq \sigma$ is heterogeneous in terms of its sensing hardware and is responsible for sensing the physiological parameter p_i. $\forall s_i \in \mathcal{W}_{i,j}$, there is a unique identifier and the data sensed by the s_i at time t is denoted by $\rho_{s_i,t}$. The data aggregation algorithms of the blind cloud framework work on two different types of clusters – *sensor-level cluster*, and *patient-level cluster*.

Definition 1: A sensor-level cluster $C_t^{s_k}$ of a sensor s_k of WBAN $\mathcal{W}_{i,j}$ at time t comprises of the last τ data generated from sensor s_k, and is expressed as $C_t^{s_k} = \{\rho_{s_k,t-\tau+1}, \rho_{s_k,t-\tau+2}, \ldots, \rho_{s_k,t}\}$.

Therefore, at time t, $\forall s_k \in s^{\mathcal{W}_{i,j}}$, we have,

$$C_t^{\mathcal{W}_{i,j}} = \{C_t^{s_1}, C_t^{s_2}, \cdots, C_t^{s_\sigma}\} \tag{1}$$

The data of every sensor-level cluster of a WBAN $\mathcal{W}_{i,j}$ is aggregated using an aggregator function $f_1(\cdot)$. f_1 is a mapping that a cluster as input and outputs a single aggregated value. Thus, f_1 is expressed as $f_1 : C_t^{\mathcal{W}_{i,j}} \rightarrow C^s$ and defined as $f_1(C_t^{\mathcal{W}_{i,j}}) = f_1(C_t^{s_1}) \cup f(C_t^{s_2}) \cup \cdots \cup f(C_t^{s_\sigma})$, i.e., $f_1(C_t^{\mathcal{W}_{i,j}}) = \{\rho^*_{s_1,[t-\tau,t]}\} \cup \{\rho^*_{s_2,[t-\tau,t]}\} \cup \cdots \cup \{\rho^*_{s_\sigma,[t-\tau,t]}\}$ where $\forall s_k \in \mathcal{W}_{i,j}$, $f_1(C_t^{s_k}) = \{\rho^*_{s_k,[t-\tau,t]}\}$.

Definition 2: A patient-level cluster $C_t^{P_{i,j}}$ of a patient $P_{i,j}$ (or a WBAN $\mathcal{W}_{i,j}$) at time t, comprises of the aggregated data generated from all the sensor-level clusters $C_t^{s_k}$ of $\mathcal{W}_{i,j}$ and is expressed as $C_t^{P_{i,j}} = \{\rho^*_{s_1,[t-\tau,t]}, \rho^*_{s_2,[t-\tau,t]}, \cdots, \rho^*_{s_\sigma,[t-\tau,t]}\}$.

B. Algorithms for Identity Management

In the proposed system, every patient $P_{i,j}$ is assigned a unique identification number $\mathcal{D}_{i,j}$ for documentation, record verification and other purposes. Thus, $\forall P_{i,j} \in P_i(t)$ and $\forall h_i \in H$, the set of unique identification number $\mathcal{D}_i$ for all patients of h_i is denoted by $\mathcal{D}_i = \{\mathcal{D}_{i,1}, \mathcal{D}_{i,2}, \cdots, \mathcal{D}_{i,k_i(t)}\}$.

To enforce "blindness" in the cloud server, our goal is to mask the identity $\mathcal{D}_{i,j}$ of a patient. Thus, we propose a pseudo-identity $\Psi_{i,j}$ for every patient. Pseudo-identification of a patient $P_{i,j}$ at time t is the process of assigning a temporary identification to the sensor-level clusters of $\mathcal{W}_{i,j}$ at time t and the pseudo-identity of $P_{i,j}$ at time t is denoted by $\Psi_{i,j}(t)$. $\Psi_{i,j}(t)$ comprises of several component factors which includes the pseudo-identities for the individual sensors comprising $\mathcal{W}_{i,j}$. Assuming $\mathcal{I}(s_k)$ and $\mathcal{T}(s_k)$ to be the unique identification number and the type of sensing hardware,

respectively, of sensor s_k, the pseudo-identity of $P_{i,j}$ for sensor s_k is denoted by $\Psi_{i,j}^{s_k}(t)$ and is defined as,

$$\Psi_{i,j}^{s_k}(t) = f_c(a(t), \mathcal{T}(s_k), I(s_k)) \tag{2}$$

where f_c is the string concatenation operator for comma separated strings and $a(t)$ is a function that hashes into a single integer from a very large set of integers $\{0, 1, \cdots, r\}, r \in \mathbb{I}$. The magnitude of $a(t)$ varies after every time interval of τ. Collisions of hashing are resolved by linear probing and for the purpose of prototyping, we assume r to be too large to be exhausted. At time t, $\Psi_{i,j}(t)$ is formed as,

$$\Psi_{i,j}(t) = f_c\big(\Psi_{i,j}^{s_1}(t), \Psi_{i,j}^{s_2}(t), \cdots, \Psi_{i,j}^{s_\sigma}(t)\big) \tag{3}$$

It is to be noted that, for a particular patient $P_{i,j}$, this holds true: $\Psi_{i,j}(t_1) \neq \Psi_{i,j}(t_2)$ if $|t_1 - t_2| \geq \tau$, as $a(t_1) \neq a(t_2)$ if $|t_1 - t_2| \geq \tau$. A data packet from $\mathcal{W}_{i,j}$, generated at time t, denoted by $\Lambda_{i,j}(t)$, is expressed as:

$$\Lambda_{i,j}(t) = \Psi_{i,j}(t)\big|C_t^{P_{i,j}} \tag{4}$$

where the fields are separated by $|$. Therefore, at any time instant $\Lambda_{i,j}$ is fed into the cloud, thereby making it blind to the identity of all the patients. At both $S_{i,1}$ and $S_{i,2}$, back-end entries are maintained for all possible values of $\Psi_{i,j}(t)$ corresponding to a single $\mathcal{D}_{i,j}$. Thus, at $S_{i,2}$, we have a tuple of $< \Psi_{i,j}(t), \mathcal{D}_{i,j} >$ for every patient. Using Equation (3), the set $\{\Psi_{i,j}^{s_k}(t), 1 \leq k \leq \sigma\}$ is obtained by reverse mapping. Therefore, for every $\mathcal{D}_{i,j}$, we have, $\{\Psi_{i,j}^{s_k}(t)\}$ from which the values of $\mathcal{T}_{s_k}$ and I_{s_k} are extracted using Equation (2). Thus, the complete identity tuple of the patient is re-constructed as $< \mathcal{D}_{i,j}, \{\mathcal{T}_{s_k}, I_{s_k}, \rho_{s_k}\}_t >$ where $\{\mathcal{T}_{s_k}, I_{s_k}, \rho_{s_k}\}_t$ denote the the temporal description of the sensors (type, identity, and value, respectively) assigned to the patient.

Theorem 3: There exists a near impossibility of re-identification by the CSP of a patient's identity, $\mathcal{D}_{i,j}$ from the data packets $\Lambda_{i,j}(t)$ at any time t.

Proof: The content of $\mathcal{V}_i'$ at time t is as follows: $\Xi_{\mathcal{V}_i'}(t) = \{\Lambda_{i,1}(t), \Lambda_{i,2}(t), \cdots, \Lambda_{i,k_i(t)}(t)\}$. Thus, $\forall \Lambda_{i,j}(t) \in \Xi_{\mathcal{V}_i'}(t)$, $\exists \Psi_{i,j}(t)$. For any two sensors s_k and s_l, let us assume that $\Psi_{i,j}^{s_k}(t) = \Psi_{i,j}^{s_l}(t)$, which implies that $\mathcal{T}_{s_k} = \mathcal{T}_{s_l}$ and $I_{s_k} = I_{s_l}$. We note that, as the computation of $a(t)$ is independent of any patient or sensor, there may be multiple sensors with the same magnitude of $a(t)$. However, $I_{s_k} \neq I_{s_l}$, as I is unique for every sensor. Therefore, $\Psi_{i,j}^{s_k}(t)$ can never be identical to $\Psi_{i,j}^{s_l}(t)$. The same magnitude of $a(t)$ in $\Psi_{i,j_1}^{s_k}(t)$ and $\Psi_{i,j_2}^{s_l}(t)$ also does not guarantee that both comprise of the pseudo-identity generator of the same patient. Thus, for $\Psi_{i,j_1}^{s_k}(t) = f_c(a(t), \mathcal{T}(s_k), I(s_k))$ and $\Psi_{i,j_2}^{s_l}(t) = f_c(a(t), \mathcal{T}(s_l), I(s_l))$, $\mathcal{D}_{i,j_1} \not\Rightarrow \mathcal{D}_{i,j_2}$. It is only in $S_{i,1}$ and $S_{i,2}$ that the correct mappings of $\big(\mathcal{D}_{i,j}, \Psi_{i,j}(t)\big)$ are maintained. Thus, for the CSP which manages $\mathcal{V}$, it is *impossible* to trace back the identity $\mathcal{D}_{i,j}$ of patient $P_{i,j}$ from the set $\Xi_{\mathcal{V}_i'}(t), 1 \leq i \leq n$. This proves the near impossibility of re-identification by the CSP of a patient's identity. □

IV. Performance Evaluation: Clinical Trial

A clinical trial was performed at the B. C. Roy Technology Hospital, Indian Institute of Technology, Kharagpur, India over 100 patients within the age-group of $10 - 70$ years where we included people who were either suffering from chronic diseases or were receiving post-operative care.

TABLE II
Large Scale Performance Evaluation

No. of clients	Latency (ms)	
	Identity masking	Identity regeneration
5	7.704	0.625
10	12.342	1.565
15	17.045	2.14
20	21.735	2.837
25	26.433	3.452

A. Experimental Setup

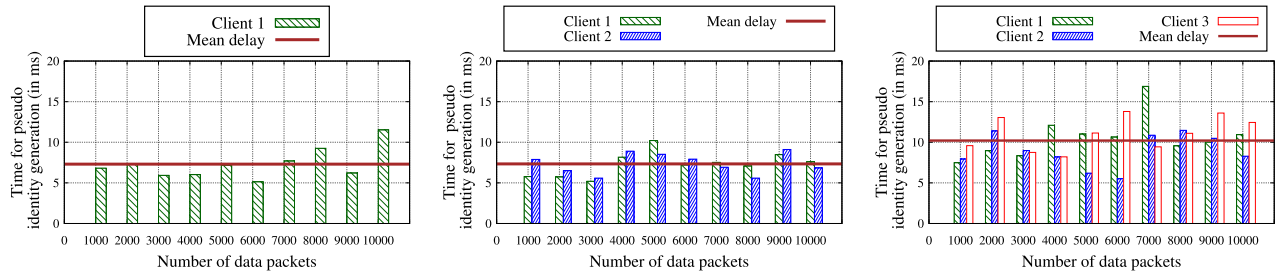
The LDPU is taken as machine with Intel Pentium processor clocked at 2.7 GHz with 2 computational cores each which can execute 2×10^3 MIPS and 2 GB DDR3 RAM. The IMUs in the hospital server 1 ($S_{i,1}$) and the hospital server 2 ($S_{i,2}$) run on Intel Core $i5-2400$U processor clocked at 3.1 GHz supported by 4 GB DDR3 RAM and 4 computational cores each capable of executing 83×10^3 MIPS. Each VM is configured to use 2 GB RAM with its clock rate varying from 2.7 GHz. Connections to the cloud server, at both end, take place over 10 Gbps dedicated links. The client machines for data retrieval at the doctor's end, however, may have any configuration as it does not take part in the data channelization or data processing processes.

B. Analysis of the Pseudo Identity Generation and Retrieval

In Fig. 2, the mean time for generation of pseudo-identity at the client server 1 is plotted against variable traffic flow and different number of clients. The simplest case with a single client is demonstrated in Fig. 2(a). The mean time $(\mu_{1,1})^*$ for generating the pseudo-identity is 7.31 ms with a standard deviation $(\sigma_{1,1})$ of 1.87 ms in this case. Fig. 2(b) depicts the variation in the pseudo-identity generation time under similar circumstances for a 2-client system. With data from both the clients arriving simultaneously at the client server 1, $\mu_{1,2} = 7.29$ ms and $\mu_{1,2} = 7.38$ ms and the standard deviation is obtained as $\sigma_{1,2} = 1.49$ ms and $\sigma_{1,2} = 1.28$ ms. Finally, in Fig. 2(c), the results corresponding to the experiments conducted on a 3-client system is illustrated. The mean and standard deviation for pseudo-identity generation are observed as: $\mu_{1,3} = 10.59, \mu_{2,3} = 8.93, \mu_{3,3} = 11.1, \sigma_{1,3} = 2.6$, $\sigma_{2,3} = 2.1, \sigma_{3,3} = 2.05$ in ms. However, the system is monitored to be tolerant to high network traffic pressure, as the pseudo-identity generation time is observed to be completely unaffected with the increase in the number of data packets processed per unit time by the server.

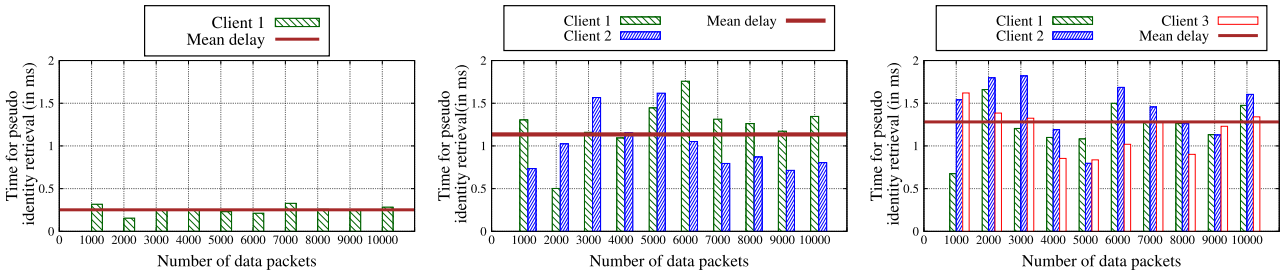
Inference: From the perspective of the overall system, the mean time for generation of the pseudo-identity is 7.31 ms, 7.33 ms, and 10.21 ms for a single-client, 2-client, and 3-client system, respectively. Comparing the three subplots, we remark that the increase in the time to generate the mask off the

*The symbols $\mu_{i,j}$ and $\sigma_{i,j}$ represent the mean and standard deviation, respectively, for the i^{th} client of the j-client ($i \leq j$) system.



(a) Single client (b) 2 clients (c) 3 clients

Fig. 2. Analysis of pseudo-identity generation time.



(a) Single client (b) 2 clients (c) 3 clients

Fig. 3. Analysis of identity retrieval time.

individual's identity vary negligibly as the number of clients in the system increases. Also, it is observed that for any (i, j) pair mentioned above, $\sigma_{i,j} < \mu_{i,j}$, which indicates that the time taken to generate the pseudo-identities is centered around the corresponding mean irrespective of the number of clients simultaneously operating in the system.

The time for identity retrieval is analyzed in Fig. 3. The identity retrieval time in a single-client system for variable network traffic is plotted in Fig. 3(a). The mean retrieval time for the i^{th} client in a j-client system is indicated by $\mu^r_{i,j}$ and the corresponding standard deviation is denoted by $\sigma^r_{i,j}$. The values of $\mu^r_{1,1}$ and $\sigma^r_{1,1}$ are computed to be 0.25 ms and 0.05 ms, respectively. In situations where two clients simultaneously requests for the aggregated data, as shown in Fig. 3(b), the values for the mean and standard deviation of the identity retrieval time for the first client are $\mu^r_{1,2} = 1.24, \sigma^r_{1,2} = 0.32$ in ms and for the second client are $\mu^r_{2,2} = 1.03, \sigma^r_{2,2} = 0.32$ in ms. Fig. 3(c) depicts the variation of the identity retrieval time for a 3-client system. The corresponding mean and standard deviation values are obtained as: $\mu^r_{1,3} = 1.24, mu^r_{2,3} = 1.43, mu^r_{3,3} = 1.18, \sigma^r_{1,3} = 0.28, \sigma^r_{2,3} = 0.23, \sigma^r_{3,3} = 0.26$ in ms.

Inference: Analyzing the three subplots, we observe that the mean time for identity retrieval is independent on the amount of data fetched from the server. Variation of the identity retrieval time is significantly low in a single-client system. This duration is observed to increase for multi-client systems. However, comparing Fig.s 3(b) and 3(c), it is noted that, as the number of clients simultaneously requesting for the data from the server increases, the increase in the mean in the identity retrieval time is negligible. We conclude that the time for identity retrieval is considerably lesser than that for pseudo-identity generation, which indicates that the developed system is capable of handling multiple client requests simultaneously without incurring any unwanted delay.

Table II depicts a large-scale analysis of the proposed work. The results obtained clearly indicate that the rate of increase in the latency incurred for identity masking and identity regeneration with the increase in the number of clients is in the order of tens of milliseconds, thereby confirming the suitability of the work in practical health-care application scenarios.

V. Conclusion

In this work, we propose a novel privacy-aware blind cloud framework with the help of of IMU which masks the identity of a patient and generates a pseudo-identity for every patient. The experiments performed for the evaluation of the proposed system reveals that the variation of mean time for pseudo-identity generation and identity retrieval remain constant with the increase in number of data packets and increases negligibly with the increase in number of clients.

References

[1] D. Lin and A. Squicciarini, "Data protection models for service provisioning in the cloud," in *Proc. 15th ACM Symp. Access Control Models Technol.*, 2010, pp. 183–192.

[2] C. Wang, Q. Wang, K. Ren, and W. Lou, "Privacy-preserving public auditing for data storage security in cloud computing," in *Proc. IEEE INFOCOM*, Mar. 2010, pp. 1–9.

[3] M. Meingast, T. Roosta, and S. Sastry, "Security and privacy issues with health care information technology," in *Proc. 28th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Aug. 2006, pp. 5453–5458.

[4] J. Xu, E.-C. Chang, and J. Zhou, "Weak leakage-resilient client-side deduplication of encrypted data in cloud storage," in *Proc. 8th ACM SIGSAC Symp. Inf., Comput. Commun. Secur.*, New York, NY, USA, 2013, pp. 195–206.

[5] Q. Chai and G. Gong, "Verifiable symmetric searchable encryption for semi-honest-but-curious cloud servers," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2012, pp. 917–922.