Collaborative eHealth Meets Security: Privacy-Enhancing Patient Profile Management

Rosa Sánchez-Guerrero, *Member, IEEE*, Florina Almenárez Mendoza, *Member, IEEE*,
Daniel Díaz-Sánchez, *Member, IEEE*, Patricia Arias Cabarcos, *Member, IEEE*,
and Andrés Marín López, *Member, IEEE*

***Abstract*—Collaborative healthcare environments offer potential benefits, including enhancing the healthcare quality delivered to patients and reducing costs. As a direct consequence, sharing of electronic health records (EHRs) among healthcare providers has experienced a noteworthy growth in the last years, since it enables physicians to remotely monitor patients' health and enables individuals to manage their own health data more easily. However, these scenarios face significant challenges regarding security and privacy of the extremely sensitive information contained in EHRs. Thus, a flexible, efficient, and standards-based solution is indispensable to guarantee selective identity information disclosure and preserve patient's privacy. We propose a privacy-aware profile management approach that empowers the patient role, enabling him to bring together various healthcare providers as well as user-generated claims into an unique credential. User profiles are represented through an adaptive Merkle Tree, for which we formalize the underlying mathematical model. Furthermore, performance of the proposed solution is empirically validated through simulation experiments.**

***Index Terms*—EHR, merkle tree, minimal disclosure, privacy, profile management.**

I. Introduction

MEDICAL records have moved from paper-based repositories to electronic records, which are a communication tool that supports clinical treatments, services coordination among health stakeholders, efficient care and legal protection. This change is allowing better health information sharing over the last years and it permits users to have control over their personal data more easily, for example, clinical documentation about diagnosis.

Health information sharing has become a vital part of modern healthcare delivery, in which patients are collaboratively treated by multiple healthcare institutions. In collaborative e-health environments, mobile internet devices, connected wirelessly to wearable, portable, and even embeddable sensors; provide efficient and effective ways to share medical information for several purposes [1], [2]: on the one hand, by enabling physicians to remotely monitor their patients' health and improve the quality of healthcare, and on the other hand, by enabling patients to spend less time in the hospital or make fewer visits to their doctor.

Nevertheless, these scenarios also raise important challenges in regards to security and privacy, being ethical priorities. Personal health data are generally very sensitive information and consequently must be protected appropriately from adversaries that try to capture the electronic medical behavior of a patient and construct "patient profiles" or reveal sensitive information related to patient's medical history, leading to the violation of the patient's privacy. Likewise, clinical professionals delivering care want fast access to relevant data, and to be sure that what they see on the screen is a faithful representation of what has been said about the patient. Emergency access to health records is sometimes needed by carers otherwise unrelated to the normal care of a patient; such accesses can only be consented in a general way, since the specific providers involved will not usually be known. Moreover, in these scenarios where the patient wears a body network in which lightweight, battery-operated wireless sensors monitor various health variables of interest, the requirements for strong cryptography must often be balanced against the requirements for energy efficiency.

However, current standards and specifications to share Electronic Health Records (EHRs) [3]–[5] are not ready to cope with some aspects of privacy. Specifically, it is necessary to develop techniques for the storage, maintenance, and fine-grained control of sensitive data that permit open sharing across different healthcare stakeholders, while data protection against unauthorized use and minimal disclosure according to patient's consent preferences is provided. To achieve these goals, we study and define a flexible privacy-aware approach for the management of patient's EHR profiles based on a generalized, adaptive and unbalanced Merkle structure. The solution enables to bring together various patient identity sources to be part of a single credential, while avoiding the creation of bogus patient's EHR profiles. Thus, a healthcare service accesses only the specific personal information it requires without being able to inspect any other details.

Manuscript received February 15, 2016; revised June 7, 2016 and October 6, 2016; accepted January 10, 2017. Date of publication August 9, 2017; date of current version November 3, 2017. This work was supported in part by the Spanish Ministry of Science and Innovation (MINECO) through the project Incident Monitoring in Smart Communities (INRISCO), TEC2014-54335-C4-2-R, and in part by eMadrid (S2013/ICE-2715). *(Corresponding author: Rosa Sánchez-Guerrero.)*

The authors are with the Department of Telematic Engineering, Carlos III University of Madrid, Leganés 28911, Spain (e-mail: rosa.rsg@gmail.com; florina@it.uc3m.es; dds@it.uc3m.es; ariasp@it.uc3m.es; amarin@it.uc3m.es).

Digital Object Identifier 10.1109/JBHI.2017.2655419

The rest of the article is organized as follows. Section II provides a brief background on main privacy principles, current eHealth specifications and Authenticated Dictionary structures, identifying the challenges to be faced. Section III illustrates a use-case motivating the work and highlights the advantages of our solution. Section IV explains our proposal to improve privacy in collaborative mobile eHealth scenarios, including the general architecture, a detailed explanation of the designed data structure for privacy-enhanced patient profile management, as well as the underlying mathematical model. Then, Section VI shows simulation results concerning the proposed structure. Section VII provides an overview about related work. Finally, Section VIII presents the conclusions and future work.

II. Background

A. Main Principles of Privacy

In this work we will address four fundamental privacy principles whose definition, according to [6], is provided below:

Anonymity: is defined as the state of being not identifiable within a set of subjects or entities. This property ensures that a user may use a resource or service without disclosing his identity. Encryption does not guarantee anonymity, since an observer can still analyze traffic, eavesdrop the sender or follow the message up to the receiver, establishing certain relationships; therefore, healthcare systems must provide additional mechanisms, such as opaque identifiers to prevent inferences.

Pseudonymity: is the use of pseudonyms as identifiers. An advantage of pseudonymity is that accountability for misbehavior can be enforced. Thus, this enables healthcare providers to link identifiers to real identities in order to make appropriate decisions when a user commits an attack.

Unlinkability: ensures that a user may consume multiple resources or services without letting other entities link these multiple resource or service accesses together. In particular, this allows users to interact with multiple organizations, each of them able to map a user to a given identity, using different identities. Healthcare systems should provide mechanisms to prevent collaborating organizations from linking a given user profile at one organization with the same user profile at another.

Unobservability: permits a user to access resources or services avoiding other entities, especially third parties, to observe that the resource or service is being used. Moreover, this property is closely related to anonymity, since in terms of item of interest (IOI), unobservability means anonymity of the subject(s) involved in the IOI even against the other subject(s) involved in that IOI.

B. Current e-Health Standards and Related Concepts

Nowadays, there are several EHR standards as HL7 [3], OpenEHR [7], and ISO EN 13606 [8] that are compliant with the HIPAA (*Health Insurance Portability and Accountability Act*) [9]. These are based on a dual model architecture, which defines two conceptual levels: reference model and archetype model. The reference model defines the set of entities that form the generic building blocks of the electronic healthcare record. The archetypes define clinical concepts in the form of structured and constrained combinations of the entities contained in the reference model, so clinical knowledge is defined at this level. Both OpenEHR and ISO EN 13606 use this modeling architecture, which has also influenced HL7 CDA. For this work, we have selected OpenEHR, because it offers an open and extensible framework, as well as archetypes for many clinical terms widely used in hospitals and summary EHR systems in multiple countries.

C. Privacy-Aware Profile Management Structures

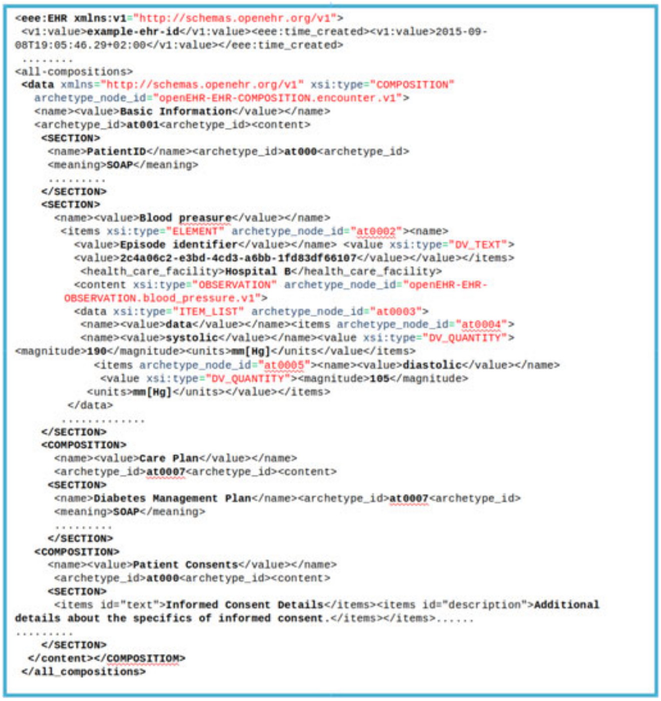
Our proposal is based on an Authenticated Dictionary structure (ADT), which enables to combine and group user's attributes from different information sources while preserving user's privacy. ADTs are data structures that support both update queries and tamper-evident membership queries. Thus, the use of ADTs structures to construct user' credentials offers desirable properties to preserve user's privacy in healthcare systems, since ADTs enable to prove the presence of an attribute without requiring to reveal any other attributes in structure.

The well-known Merkle's tree [10] was the first ADT structure. A Merkle tree is a binary tree where leaf nodes are labeled by the hashed values of the elements of a set, S, and internal nodes are labeled by the hashed values of concatenated labels of their children. The root value is then the label of the root node, and the proof that element e belongs to S consists of the labels of all sibling nodes on the path from the leaf node representing e to the root node. Hence, the main goals of Merkle's trees are the following: 1) to make one-time signature schemes feasible; and 2) to provide an efficient key management scheme that reduces the amount of public keys and their size.

Merkle trees can be generalized by a structure called "hash DAG", based on a directed acyclic graph [11], thus allowing to extend the original Merkle tree (a binary tree) to an *m-ary* Merkle tree. Hence, our proposal is based on an extended and unbalanced Merkle tree, because this structure enables richer clustering and node verification using a single signature. Thus, we offer potentially a large number of verifiable attribute combinations by means of a single verification tree that empowers the user to realize a **selective disclosure** of his information to the different entities. Skip lists are another kind of ADT structure introduced by W. Pugh as an alternative data structure to search trees [12]. The main idea is to add pointers to a simple linked list in order to skip a large part of the list when searching for a particular element. While each element in a simple linked list points only to its immediate successor, elements in a skip list can point to several successors.

III. Motivation

In order to show the benefits of our approach, in this section, we describe a potential use-case that can be realized by applying our proposal. Alice is a diabetic patient who also has hypertension and kidney problems. She finds difficult to manage her condition effectively. Let us assume a given domain, such as the State of California, where we have several healthcare communities in San Francisco and Los Angeles. As Alice



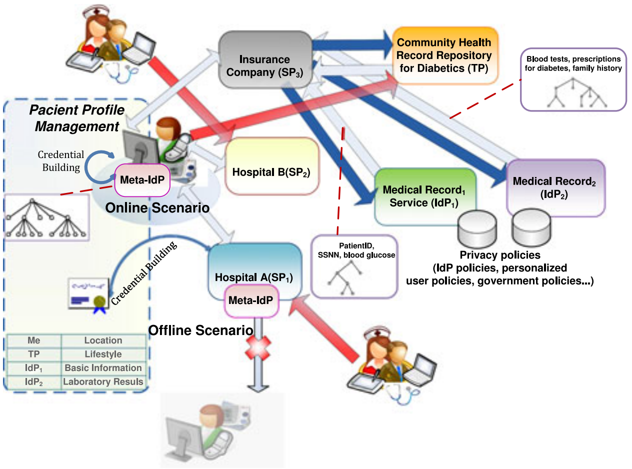
```
<eee:EHR xmlns:v1="http://schemas.openehr.org/v1">
 <v1:value>example-ehr-id</v1:value><eee:time_created><v1:value>2015-09-
08T19:05:46.29+02:00</v1:value></eee:time_created>
 .........
<all-compositions>
 <data xmlns="http://schemas.openehr.org/v1" xsi:type="COMPOSITION"
   archetype_node_id="openEHR-EHR-COMPOSITION.encounter.v1">
   <name><value>Basic Information</value></name>
   <archetype_id>at001<archetype_id><content>
    <SECTION>
     <name>PatientID</name><archetype_id>at000<archetype_id>
     <meaning>SOAP</meaning>
     .........
    </SECTION>
    <SECTION>
      <name><value>Blood preasure</value></name>
       <items xsi:type="ELEMENT" archetype_node_id="at0002"><name>
         <value>Episode identifier</value></name> <value xsi:type="DV_TEXT">
         <value>2c4a06c2-e3bd-4cd3-a6bb-1fd83df66107</value></value></items>
          <health_care_facility>Hospital B</health_care_facility>
         <content xsi:type="OBSERVATION" archetype_node_id="openEHR-EHR-
       OBSERVATION.blood_pressure.v1">
         <data xsi:type="ITEM_LIST" archetype_node_id="at0003">
          <name><value>data</value></name><items archetype_node_id="at0004">
          <name><value>systolic</value></name><value xsi:type="DV_QUANTITY">
<magnitude>190</magnitude><units>mm[Hg]</units</value</items>
            <items archetype_node_id="at0005"><name><value>diastolic</value></name>
             <value xsi:type="DV_QUANTITY"><magnitude>105</magnitude>
           <units>mm[Hg]</units></value></items>
        </data>
       .............
    </SECTION>
    <COMPOSITION>
      <name><value>Care Plan</value></name>
      <archetype_id>at0007<archetype_id><content>
     <SECTION>
      <name>Diabetes Management Plan</name><archetype_id>at0007<archetype_id>
      <meaning>SOAP</meaning>
      .........
     </SECTION>
   <COMPOSITION>
     <name><value>Patient Consents</value></name>
      <archetype_id>at000<archetype_id><content>
     <SECTION>
      <items id="text">Informed Consent Details</items><items id="description">Additional
details about the specifics of informed consent.</items></items>......
.........
    </SECTION>
  </content></COMPOSITIOM>
 </all_compositions>
```

Fig. 1. XML fragment of a patient's EHRs, which can be represented with a tree structure.

travels frequently, she has received healthcare in each of these communities. She is undergoing kidney surgery in the hospital in LA (hospital A) next month. The attending physician, Bob, will need to use her hospital information system to query across multiple domains for healthcare information about this patient (e.g., chronic conditions, critical diseases, past surgical, family history, laboratory results, blood glucose, blood pressure, etc.) On the other hand, in one of her visits to San Francisco, Alice was admitted to hospital B. Her doctor, Robert, advised Alice to subscribe to a diabetes management program offered by hospital B. As a part of the program, Alice wears a hospital-provided device that continuously monitors her activity level and calories burned, and installs software on her mobile phone. The software processes data it receives from the monitor along with contextual information such as Alice's location.

Furthermore, Alice decides to join a social network for diabetics, whose privacy settings enable her to share information with the group (e.g., her daily activity and food intake progress) and to allow complete access to her personal health information to her family members. Once a week, Alice records her weight, blood glucose and blood pressure, using devices that send the measurements wirelessly to her mobile phone. Due to her participation in the management program, Alice's insurance company offers to reduce her premium if she shows significant improvement in controlling her diabetes. In this dynamic scenario, different parts of Alice's medical history can be distinguished and merged as EHR profiles to construct an *M*-ary Merkle tree according to the openEHR Information Model specification [7] (see Fig. 1):

1) *Basic Information:* Patient ID, SSN, weight, blood glucose, blood pressure and blood group.



Insurance Company (SP$_3$)
Community Health Record Repository for Diabetics (TP)
Blood tests, prescriptions for diabetes, family history
Pacient Profile Management
Credential Building
Meta-IdP
Online Scenario
Hospital B(SP$_2$)
Medical Record$_1$ Service (IdP$_1$)
Medical Record$_2$ (IdP$_2$)
PatientID, SSNN, blood glucose
Privacy policies (IdP policies, personalized user policies, government policies...)
Credential Building
Hospital A(SP$_1$)
Meta-IdP
Offline Scenario

Me	Location
TP	Lifestyle
IdP$_1$	Basic Information
IdP$_2$	Laboratory Resuls

Fig. 2. IdM architecture for collaborative healthcare.

2) *Patient Preferences:* Alice has the choice to remain anonymous in the group for diabetics.
3) *Patient Consents:* Alice's husband can access to her complete Personal Health Information (PHR). To demonstrate progress, Alice must provide the insurance company access to certain parts of her health data. She instructs her PHR to provide aggregate information of her activity, diet and physiological parameters.
4) *Therapeutic Precautions:* it considers allergies (e.g., penicillin) and alerts.
5) *Lifestyle:* it includes exercise and food intake progress.
6) *Care Plan:* combinations of goals, targets, monitoring, education concerning the diabetes management plan.
7) *Laboratory Results:* for instance, blood tests.
8) *Prescriptions:* medication orders related to Alice's chronic conditions.
9) *Family History:* Alice's father died of myocardial infarction at 62.
10) *Physical Examinations:* observations appointment, admission and discharge at hospitals A and B.

IV. Privacy-Enhanced eHealth Through Adaptive Extended Merkle Trees

A. Architecture

Before explaining the mathematical model, it is necessary to describe the complete architecture in order to identify stakeholders that manage EHRs. We consider an Identity Management (IdM) architecture, based on our previous work in [13], with the following actors: 1) **Service Providers (SPs)**, which provide services to the end user and consume the identity data coalesced by the healthcare providers from several sources. For instance, this role is played by the insurance company in the use-case presented in Section III; 2) **Identity Providers (IdPs)**, which are entities issuing assertions about patient's medical records (e.g., hospital A and hospital B); and 3) **Users** who are the subjects of the assertions (e.g., Alice). Users have a particular digital identity and interact with SPs and IdPs (see Fig. 2).

The information sent to the healthcare providers may contain pieces of data stored in several identity providers and user devices. User's devices would act as an Identity Metasystem, meta-IdP [14], in order to provide interoperability, a consistent user experience and control of the information exchange. In this way, the role of the patient (Alice) is empowered letting her participate in the process. Thus, trust and access control decisions are no longer opaque to the user. The patient is given the ability to configure interactions with healthcare providers and third parties (e.g., the insurance company and the social network), by detailing which attributes may the healthcare providers take from her profile and which ones can be taken from an IdP. It is worth to note that, meta-IdPs can be also instantiated in the health care provider to cope with scenarios in which the patient is not online to accept the healthcare transaction. More technical details about the IdM architecture can be found in [13].

B. Proposed Adaptive Extended Merkle (AEM) Tree

We extend the traditional Merkle Tree structures in order to include privacy properties as mentioned in Section II-A, given the extremely sensitive nature of the information handled in eHealth environments; and also enhancements to meet the performance needs of these scenarios. The concrete advantages and underlying mechanisms in this proposed AEM Tree are:

A richer view of the EHRs by assembling different parts of medical records as profile groups and user's preferences: Patient's medical history or records do not have to follow a strict binary, ternary or quaternary structure. The AEM Tree enables to group information in a more flexible manner. Let us consider a patient's EHR example, in which the user has several contacts and critical diseases. To address the treatment of a specific disease, the intervention from 1 to N departments of different hospitals (e.g., surgery, chemotherapy) may be necessary. Each department may implement from 0 to N treatments and each treatment, has a date and may have from 1 to N participants (e.g., doctor, nurse, surgeon, etc.). To achieve a more flexible storage structure, we study and define a profile management based on a *M*-ary, adaptive and unbalanced Merkle tree. The tree is distributed suitably according to the frequency user's attributes are accessed. The set of attributes with similar access frequency may have semantic relationships, which will allow to build different profiles to be part of a patient's medical history. To this end, our proposal provides an algorithm for sorting the tree based on patterns of access according to the EHR Information Model [7]. Thus, the attributes frequently required will be placed closer to the root, whilst clinical data whose relevancy to the clinical care of the patient fades in time (e.g., most measurements made on the patients or in pathology) will be located in the lower levels.

Combining several sources of EHRs to be part of a single credential: We use the use case described above to illustrate the potential benefits of combining multiple sources of identity and selective information revelation in collaborative health care environments. In the scenario, solutions based on basic structures would require the healthcare provider must either see all of the claims or trust the providers of all information. This solution is not ideal from a security and privacy point of view. Hence, our approach includes an optional branch to some internal nodes of the full tree and it enables that healthcare providers (hospitals A and B) do not have access to all information about Alice. Healthcare providers are only responsible for claims related to their subject area. Furthermore, the used hash minimizes the need of individual verification of elements along a path and, instead, it would suffice with a root's hash check and the user only has to keep track of one credential. This also enables multiple attributes verification through a single verification tree without revealing information related to non requested attributes.

Adaptive performance: Considering the large information handled and the variability of data is much smaller than in the case of social networks and cloud computing scenarios, it is desirable to have an agile storage structure on read operations. In healthcare environments, response times of insertion or modification operations can be penalized in favor of applying more robust security and privacy mechanisms to protect sensitive information in accordance with the regulatory and legislative frameworks. Although the use of the Merkle Trees makes more difficult to add or update attributes without recomputing parts of the tree as well as changing the root itself, our work provides an algorithm to improve this aspect, by sorting the tree as we envision frequently accessed attributes to be closer to the root.

C. Mathematical Formalization

The purpose of this section is to describe how to handle patient's EHR profiles through a novel AEM tree to convey patient claims to other entities.

An AEM tree is essentially an *M-ary* and unbalanced tree, i.e., each node may have up to *M* maximum children. Thus, each node holds the hash of the concatenated values of its children nodes. Leaf nodes hold the identity attributes as well as other information as (e.g., node tag, semantic annotation, attribute value, attribute nature and type as *self-issued* - non verifiable - or *provider issued* - verifiable). Node's children influence the node hash and so does the node with its parent until the root node, so a large number of separate data can be tied to a single hash value (root node). In this way, given an attribute and its hash tree, if hashes related to the attribute are consistent until the root and the signature of the root node is valid, it is possible to verify that any of the leaf nodes of the tree are authentic without revealing any further data. Thus, selective patient's attribute disclosure and verification is achieved. To help the consumer to distinguish among different sources, the meta-IdP labels nodes by appending a bit to the end of the hash, true if the attribute is *provider-issued* (i.e., age, nationality), or false if the node attribute is *self-issued*.

AEM trees are constructed according to the following. A template node, named N, contains several attributes Att_n (of any nature) and children $N_1, \ldots, N_n$. Some node types may contain a Hash value H_N from a summary obtained from its attributes and related to its children. Moreover, they contain a node identifier N_{Id}:

$$N \leftarrow ([H_N], N_{Id}, Att_1, \ldots, Att_n), [\{N_1, \ldots, N_n\}] \quad (1)$$

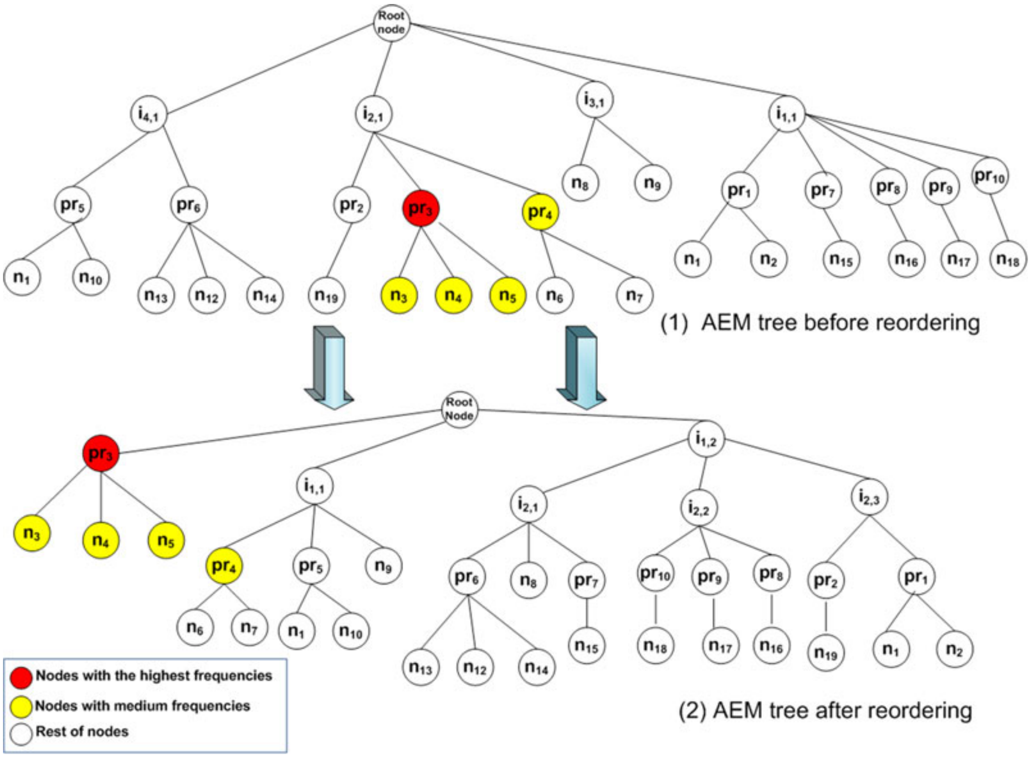


Fig. 3. Privacy framework based on extended Merkle trees for EHR profile management.

There are several node types: leaf nodes, named LN, contain attributes but no children;

$$LN \leftarrow (N_{Id}, Att_1, \ldots, Att_u), \{\} \tag{2}$$

profile nodes, named PN, contain attributes and descendants that should be kept together since they constitute a profile or set of interrelated claims;

$$PN \leftarrow (H_N, N_{Id}, Att_1, \ldots, Att_v), \{N_{p1}, \ldots, N_{pn}\} \tag{3}$$

inner nodes, named IN, are structural nodes containing no identity attributes but the necessary hash values to build a verification path from any leaf node to the root (that will be signed by the provider);

$$IN \leftarrow (H_N, N_{Id}, Att_1, \ldots, Att_w), \{N_{i1}, \ldots, N_{in}\} \tag{4}$$

the root node, named RN, contains several attributes including an identifier, a time stamp TS (generated during signature) and a signature Sig over the hash value related to its children. Its children contain, as well, a hash value related to their children, until a leaf node. In this way, a provider can certify all the data placing one signature in the root node over the hash value allowing the tree to be lopped by the meta-IdP removing branches without affecting the hash whenever hash values until the claim to be proven are known. Moreover, the root node has a set of special children nodes that contain pseudo identifiers $(P_{id1}, \ldots, P_{idn})$ that are randomly generated when the data structure is signed. Thus a signed tree can be used several times enabling unlinkability. Besides, due to selective disclosure properties of the AEM tree, a degree of unobservability is offered, since the meta-IdP allows the patient to handle health resources while keeping clear of other entities (e.g., the social network) have access to more information than it is necessary.

$$\begin{aligned} RN \leftarrow (&Sig, TS, H_N, N_{Id}, Att_1, \ldots, Att_n), \\ &\{N_1, \ldots, N_n\}, \{P_{id1}, \ldots, P_{idn}\} \end{aligned} \tag{5}$$

Fig. 3 illustrates Alice's medical history according to the use case described in Section III. Profile nodes corresponds to: physical examinations (denoted by pr_1), allergies (represented as pr_2), patient's basic data (pr_3 and pr_4), self-issued lifestyle attributes (pr_5), physiological parameters (pr_6), patient preferences (pr_7), events or conditions in Alice's family members (pr_8), laboratory measurements (depicted as pr_9) and information related to patient consent (pr_{10}). It must be noted that, nodes n_8 and n_9 store PHR data based on compliance with activity and diet. Leaf nodes n_8 and n_9 contain Alice's location information and her blood group, respectively. Finally, the rest of LN are children of the aforementioned PN.

Furthermore, our AEM tree annotates the query frequencies of its attributes, as a criteria for subsequent optimizations. In this way, the most frequent attributes (e.g., PHR data related to diabetes, Alice's basic information and her location) are placed in the upper levels of the AEM tree, making more efficient and faster the verification process.

Hence, the AEM tree can be dynamically optimized according to node frequencies given some structural constraints (tree depth and node children from 0 to M).

So, we denote by *m* the maximum degree of a node, i.e, the maximum number of branches that emanate from each node, the parameter *h* represents the AEM tree height. In addition, we use the term $L\%$ to represent the contribution percentage variance of the access frequency of the nodes remaining to be placed in the tree. Equations (6) define the set of possible AEM tree nodes (leaf, profile and inner nodes), (7) their query frequencies, and (8) denotes the maximum number of AEM tree nodes and the the maximum number of leaves, respectively:

$$N_{AEMTree} = \{N_1, N_2, \dots N_k\} \tag{6}$$

$$F = \{f_1, f_2, \dots, f_k\} \tag{7}$$

$$k = \sum_{i=1}^{h} m^i, \quad M = m^h \tag{8}$$

where,

$$f_1 < f_2 \dots f_{k-1} < f_k, k \geq M \tag{9}$$

As mentioned before, a sorting algorithm can be triggered to improve searches. The algorithm works as follows. *Step 1:* we order the set of AEM tree nodes containing data (LN and PN) by query frequencies in ascending order. *Step 2:* we take the p nodes, named P, that contribute ($Cvar_i$) to the $L\%$ of the frequency variance (var) of the remaining nodes (see (10), (11) and (12)).

$$P = \{p_1, p_2, \dots p_p\} \tag{10}$$

$$var = \sum_{i=1}^{RM} (f_i - \bar{f})^2 / RM, \quad std = \sqrt{var} \tag{11}$$

where RM are the nodes pending to be placed.

$$Cvar_i = (f_i - \bar{f})^2 / RM \tag{12}$$

Step 3: the algorithm iterates over the nodes in P. Until P is empty, we take the first node in P, p_i and check:

$$m^h - (m-1) > k - 1 \tag{13}$$

Equation (13) evaluates if p_i can be placed in this level (since it reduces the maximum number of leafs) leaving room for the rest of the nodes ($k-1$). If so, we place the node p_i, remove p_i from P and go back to step 3. Otherwise, *m* new internal nodes are added to the tree and go the next level. Finally, once nodes in P have been placed we move to step 1 where the variance and dispersity for the remaining nodes are recalculated and the following *p* nodes contributing the *L*% of the variance are chosen. This process is repeated until the number of nodes to place in the AEM tree is equal to zero.

It must be noted that, the sorting algorithm is also applied each time a node is inserted or updated in the AEM tree. As the new node does not have historical of access frequency, it will be placed at the "most disadvantaged" positions of the AEM tree, as happens in the real life situations when someone starts at the bottom and work her way up. If this new node is frequently consulted, it will prove itself and its position will improve. As regards the update attributes, whether a node is very frequently accessed when the proposed algorithm is applied, it will be in a good position. Otherwise, it will be located at the lower levels.

Fig. 3 illustrates an example of how our distribution algorithm works for Alice's AEM tree with parameters $m = 3$ and $h = 3$. The search algorithm looks for nodes upside-down and left to right, so after running the distribution algorithm, pr_3 and its children (n_3, n_4 and n_5) are put in a higher level and further to the left. The node called pr_4 is also relocated in a position that favours its search and verification when these attributes are shared with hospital A, B and the social network. It must be noted that, the $PNpr_{10}$ and pr_7 will be checked before disclosing attributes held by pr_3 and pr_4. The rest of the nodes are placed according the same criteria position respecting the parameters *m* and *h*.

V. Security and Privacy Considerations

The unlinkability and "partial anonymity" of the proposal stems from the corresponding IdPs services. Users claims, asserted by the IdPs, are only exposed according to the privacy rules and informed consents. When a restricted view is required, opaque parts of the EHR are incorporated and verifiable thanks to the hashes and the opaque and transient identifiers provided by the IdPs. Using Merkle Hash Trees to enforce privacy has been already explored in other works like [15]. Our searching and sorting algorithms may introduce information leakage suitable for a differential analysis. Let us consider a well informed attacker who performs selected searches to initiate new sorting of the AEM tree: measuring the sorting time, the attacker can perform estimations and even models of the attributes and relationships of parts of the EHR beyond her authorization. To prevent such privacy breaches, we propose to introducing random delays in the sorting algorithm. Besides the number of executions of the sorting algorithm should be limited within a given period of time.

VI. Evaluation Results and Discussion

We have developed a prototype in Java and conducted preliminary experiments on the performance of our data structure by generating sets of random requests with different probability distributions.

For this purpose, we have created synthetic data, which include quantitative (e.g., age, weight, blood glucose, blood pressure, etc.) and qualitative data (e.g., demographics and socioeconomic data, lifestyle choices, prescriptions, patient's preferences and consents, etc.) to evaluate different profiles. We perform two different tests for different tree structures: uniform (to represent a young and healthy person's medical history) in which each node (e.g., basic information and lab results, such as blood tests) has the same probability to be requested, and biased (patients' medical records with chronic diseases) in which few nodes (e.g., current medications and problems) have a high probability to be requested.

Likewise, in these experiments we have studied the behavior of the frequency-based adaptive distribution algorithm for different AEM tree sizes by modifying both their height and maximum number of children permitted per node. For each operation, the average search time, $\bar{ST}$, was computed over 400,000 trials. The experiment was conducted using a machine equipped with an Intel CORE i7 2760QM with 8 G of memory running at 4 GHz. Cryptographic hashing was performed using

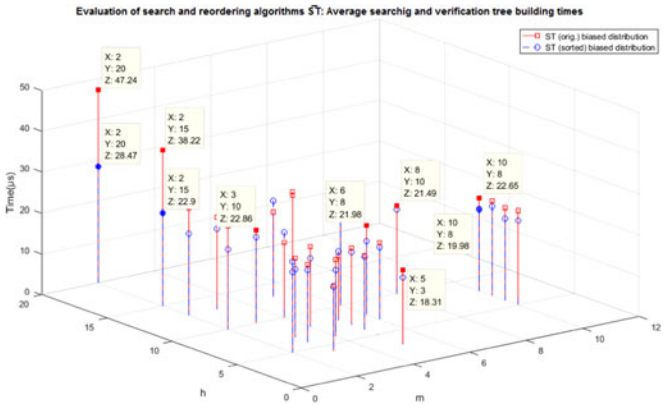


Fig. 4. The average searching and verification tree building times ($S\bar{T}$) globally for biased structures.

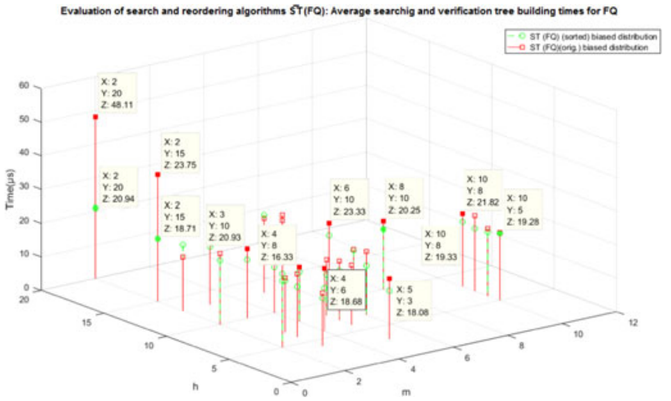


Fig. 5. The average search and verification tree building times ($S\bar{T}$) for the frequent queries (FQ - those that constitute the 50% of the queries).

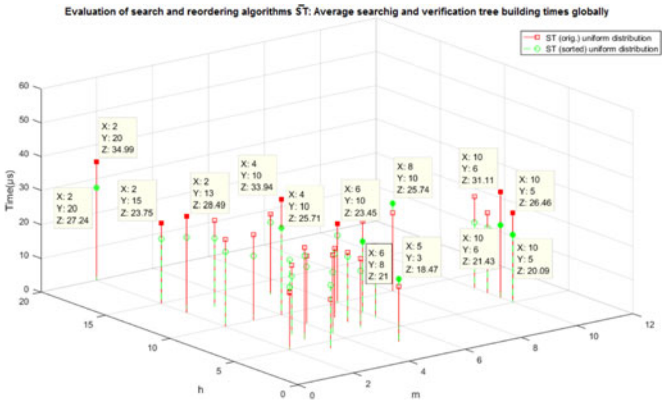


Fig. 6. The average search time and verification tree building times ($S\bar{T}$) globally for uniform structures.

the standard Java implementation of the SHA-256 algorithm. We have summarized the evaluation results in Figs. 4, 5 and 6 through 3D graphics that show the obtained times for searching and verification tree building. These times are represented by the Z axis, when different sizes of m-ary trees are used. To reflect the changes of the trees, the X axis, represents the number of maximum children that each node may have (denoted by m) and the Y axis pictures the different heights of the trees (parameter h) used for the experiments. For the results depicted in Figs. 4 and 5, we have used biased structures, whilst Fig. 6 presents findings for uniform structures. Figs. 4, 5 and 6 show in red the searching and verification tree building times for the different m-ary trees (the changes of m and h are represented in the axes X and Y, respectively) without applying the proposed sorting algorithm. The findings when the proposed sorting algorithm is executed are shown in Fig. 4 in blue and in Figs. 5 and 6 in green.

We have taken as reference binary trees ($m = 2$ and $h = 5$, 10, 15 or 20) and m-ary trees (e.g., $m = 4$, 5 or 6 and $h = 6$ or 8) and evaluated the average search time ($S\bar{T}$) and the average verification tree length ($V\bar{T}L$) for every node and for the set of nodes that are most frequently queried ($S\bar{T}$ (FQ) and $V\bar{T}L$ (FQ)). Note that, the $S\bar{T}$ includes searching and verification tree building times. In Fig. 4, we can appreciate that our algorithm reduces the total searching and verification tree building times over a 40.08% and 39.73% for $m = 2$ and $h = 20$ and 15, respectively. Moreover, the proposed algorithm also decreases the verification path length (this aspect is not represented in the above graphics). For the binary tree cases, the average verification path length before reordering are equal to 7.00 ($m = 2$, $h = 5$) and 13.05 ($m = 2$, $h = 15$), whereas the value of this parameter is reduced to 4.50 and 5.51 after running the algorithm, respectively.

Regarding the outcome of uniform query distribution test (see Fig. 6), the average searching and verification tree building times are slightly enhanced especially when m decreases and h increases (see the value of the Z axis for instance when $m = 2$, $h = 13$, 15 or 20). Finally, it is must be noted that, the time spent by the distribution algorithm (OT) is not significant when compared to the improvement over the total search time (Total ST). Although they are not shown in Figs. 4, 5 and 6, for instance, for $m = 2$, $h = 15$ and $m = 2$, $h = 20$, the OT are 1670.745 μs and 2711.752 μs, while the total search are 38,71 s and 43.57 s, respectively. For m-ary trees, the OT are 1531.469 μs and 1463.632 μs when $m = 6$, $h = 8$ and $m = 5$, $h = 6$; respectively.

VII. Related Work

As far as the related work is concerned, we found several research initiatives in the field. Many authors have suggested security and privacy as key issues to address in eHealth [16]–[19], but these issues as a whole have not yet been covered extensively for application scenarios. The focus is normally on security related issues in general wireless sensor networks.

Nowadays, several approaches to provide privacy-preserving techniques can be found in the literature [20]–[28]. Firstly, in attribute-based (ABE) encryption proposals each user has a set of attributes and access policies are defined to determine that the users with certain attributes are authorized to access the shared data. ABE cryptosystems [20] crowd in two categories: ciphertext-policy ABE (CP-ABE) [22] systems and key-policy ABE (KP-ABE) [21] systems. In the first, the users' secret keys are associated with sets of attributes, and a sender generates a ciphertext with an access policy specifying the attributes that the decryptors must have. Regarding KP-ABE solutions, the users' secret keys are labeled with access policies and the sender stipulates a set of attributes; only the users whose access policies match the attribute set can decrypt. In [29] authors suggest a multi-authority CP-ABE scheme to empower to the patient to associate an expressive access tree structure and on-demand attribute revocation. However, these ABE schemes require a priori access policies, which are not always available in EHRs because the policies to access health records are sometimes determined after key generation. [30] addresses this issue by considering a dynamic ABE paradigm, which provides a delegation mechanism that allows users to redefine the access policy and delegate a secret key without making the redefined access policy more restrictive. Nevertheless, how to construct fully secure hierarchical identity-based encryption systems in prime-order bilinear groups under simple assumptions remains as a challenging problem [31].

Secondly, cloud-based approaches as [23] and [24] propose privacy-aware schemes based on query authentication to enable data confidentiality, the query result integrity of sensitive data, secure storage and secure computation auditing. The work presented in [32] integrates a PRF-based key management for unlinkability, a search and access pattern hiding scheme based on redundancy for privacy-preserving data storage. This approach also combines ABE-controlled threshold signing with role-based encryption to provide access control and auditability. A signature algorithm that allows for controlled changes to the signed data is proposed in [33]. This work studies techniques that cryptographically link the integrity of the original and modified datasets for practical types of modifications such as redaction, pseudonymization and data deidentification.

Thirdly, when it comes to information disclosure, spatiotemporal cloaking and ADT-based approaches enable to preserve user's privacy. Spatial cloaking or perturbation allows to hide the participant location inside a cloaked region using spatial transformations, generalization, or a set of dummy locations in order to achieve location privacy [34]. In [35] authors propose a privacy-preserving emergency call scheme called PEC, enabling patients in life-threatening emergencies to fast and transmit emergency data to the nearby helpers via mobile healthcare social networks. Moreover, the PEC has been designed to withstand multiple types of attacks, such as identity theft attack, forgery attack, and collusion attack. However, this kind of works do not address privacy issues related to management of user profiles.

On the other hand, there are other approaches closer to our work. For example, some identity frameworks like U-Prove [36] allow selective disclosure of claims and pre-signed tokens that could be used when the entity responsible for issuing medical records is offline. Furthermore, there are other proposals such as identity agents [37], or veryIDX [38]. [37] proposes user-controlled identity agents, which allow defining in advance disclosure policies, monitoring credential usage, storing credentials based on a minimal disclosure scheme. The credentials are constructed using Merkle trees, but the details about how patient's attributes are built or can be shared by means of EHR standards are not provided. VeryIDX enables multi-factor identity credential verification, by using a cryptographic commitment and an aggregated zero-knowledge proof of knowledge (ZKPK).

Despite some current works [26], [27], [39] propose the use of authenticated dictionaries or opportunistic computing mechanisms to provide selective information disclosure, none of these works deals with neither building of the ADT structure based on EHR standards nor combining subtrees that allows claims from different sources to be in a single credential in order to make easier the tasks of management of patient attributes, profiles and preferences. In this context, typical research directions are related to development of more efficient and effective ADT-based structures and algorithms, in terms of storage overhead, times of signature generation and verification or query and update times. [26] proposes a signature scheme on the structure of the tree as defined by tree traversals (pre-order, post-order, in-order), that improves protection against information leakages. [39] describes a secure and privacy-preserving opportunistic computing framework, called SPOC, for m-Healthcare emergency. The authors introduce an attribute-based access control and a privacy-preserving scalar product computation technique that allows a medical user to decide who can participate in the opportunistic computing to assist in processing his overwhelming personal health information data. Eventually, in [27], the authors propose a multiway extension of the authenticated version of the skip-list data structure and study the authentication cost that is associated with this model when authentication is performed through hierarchical cryptographic hashing. However, due to the heterogeneity of data types in healthcare scenarios, this kind of structure requires a complex implementation.

VIII. Conclusions and Future Work

The collaboration between health stakeholders is currently an important challenge to reduce costs, as well as improving the quality of clinical practice and patient safety. Due to the significant amount of data that is stored or exchanged and the extremely sensitive information contained in EHRs, privacy is of paramount importance. However, this aspect has not been fulfilled by existing integrating healthcare enterprise solutions yet. We believe that, a flexible and efficient privacy-supporting mechanism to control the dissemination of patient's personal information in a seamless, interoperable and scalable manner is essential. We have addressed this problem by proposing a privacy-enhanced user profile management approach based on a novel Adaptive Extended Merkle structure that empowers the user role, by letting users to combine the sources of identity contained in different medical and community health records repositories with identity information stored in their personal devices. In addition, the AEM tree can store references to other health data, by governing access to them, but not necessarily holding them (e.g., test results that much space like CT, PET-CT, etc.). Moreover, in this work we have provided and evaluated the algorithm that allows to build enriched compositions of the patient's medical history and to sort the tree based on patterns of access compliance with open EHRs standards, which can contribute to facilitate its implementation in health information systems. The proposed AEM tree can be also used to enable audited access in other application scenarios, which include experiments in patients with complex diseases but requiring notification or confirmation to cross data, such as religion or race, etc. The evaluation results for different *m*-ary trees showed that the time spent by the distribution algorithm is not significant when compared to the improvement over the total search time for both biased and uniform structures. Now, we are working on studying in depth the AEM structure by considering issues related to its size and measuring its dynamicity cost. Further research is needed to test the integration of our profile management solution with cloud-based e-health scenarios, using images and genetic information, as well as more complex information models and archetypes defined by the openEHR Foundation.

References

[1] C. Baker *et al.*, "Wireless sensor networks for home health care," in *Proc. 21st Int. Conf. Adv. Inf. Netw. Appl. Workshops*, May 2007, vol. 2, pp. 832–837.

[2] U. Varshney, "Pervasive healthcare and wireless health monitoring," *Mobile Netw. Appl.*, vol. 12, no. 2/3, pp. 113–127, Mar. 2007.

[3] "Health level seven," Jan. 2016. [Online]. Available: http://www.hl7.org/

[4] "OpenEHR," Jan. 2016. [Online]. Available: http://www.openehr.org/

[5] "IHE IT infrastructure (ITI) technical framework volume 1 integration profiles. Revision 12.0," Jan. 2016.

[6] A. Pfitzmann and M. Hansen, "A terminology for talking about privacy by data minimization: Anonymity, unlinkability, undetectability, unobservability, pseudonymity, and identity management," Dec. 2009. [Online]. Available: http://dud.inf.tu-dresden.de/Anon_Terminology.shtml

[7] R. Chen, "EHR information model. Release 1.0.2," Jan. 2016.

[8] "ISOEN-13606," Jan. 2016. [Online]. Available: http://www.iso.org/iso/home.htm

[9] P. Gunn, A. Fremont, M. Bottrell, L. Shugarman, J. Galegher, and T. Bikson, "The health insurance portability and accountability act privacy rule: A practical guide for researchers," *Med. Care*, vol. 42, no. 4, pp. 321–327, 2004.

[10] R. C. Merkle, "A certified digital signature," in *Proc. 9th Annu. Int. Cryptol. Conf. Adv. Cryptol.*, London, U.K., 1990, pp. 218–238.

[11] T. Page, "The application of hash chains and hash structures to cryptography," Tech. Rep. RHUL-MA-2009-18, Aug. 2009.

[12] W. Pugh, "Skip lists: A probabilistic alternative to balanced trees," *Commun. ACM*, vol. 33, no. 6, pp. 668–676, Jun. 1990.

[13] R. Sánchez-Guerrero, F. Almenárez, D. Díaz-Sánchez, A. Marín, P. Arias, and F. Sanvido, "An event driven hybrid identity management approach to privacy enhanced e-health," *Sensors*, vol. 12, no. 5, pp. 6129–6154, 2012.

[14] "Microsoft's vision for an identity metasystem," White Paper, May 2005. [Online]. Available: http://msdn2.microsoft.com/en-us/library/ms996422.aspx

[15] R. Bazin, A. Schaub, O. Hasan, and L. Brunie, "A decentralized anonymity-preserving reputation system with constant-time score retrieval," IACR Cryptol. ePrint Archive, Rep. 2016/416, 2016.

[16] F. Kargl, E. Lawrence, M. Fischer, and Y. Y. Lim, "Security, privacy and legal issues in pervasive ehealth monitoring systems," in *Proc. 7th Int. Conf. Mobile Bus.*, Jul. 2008, pp. 296–304.

[17] M. Ameen, J. Liu, and K. Kwak, "Security and privacy issues in wireless sensor networks for healthcare applications," *J. Med. Syst.*, vol. 36, no. 1, pp. 93–101, Feb. 2012.

[18] J. Zhou, Z. Cao, X. Dong, X. Lin, and A. Vasilakos, "Securing m-healthcare social networks: Challenges, countermeasures and future directions," *IEEE Wireless Commun.*, vol. 20, no. 4, pp. 12–21, Aug. 2013.

[19] M. L. Braunstein, *Health Informatics in the Cloud*. New York, NY, USA: Springer-Verlag, 2013.

[20] A. Sahai and B. Waters, "Fuzzy identity-based encryption," in *Advances in Cryptology – EUROCRYPT 2005: 24th Annual International Conference on the Theory and Applications of Cryptographic Techniques, Aarhus, Denmark, May 22-26, 2005*. Berlin, Germany: Springer-Verlag, 2005, pp. 457–473.

[21] V. Goyal, O. Pandey, A. Sahai, and B. Waters, "Attribute-based encryption for fine-grained access control of encrypted data," in *Proc. 13th ACM Conf. Comput. Commun. Security*, 2006, pp. 89–98.

[22] J. Bethencourt, A. Sahai, and B. Waters, "Ciphertext-policy attribute-based encryption," in *Proc. IEEE Symp. Security Privacy*, 2007, pp. 321–334.

[23] L. Wei *et al.*, "Security and privacy for storage and computation in cloud computing," *Inf. Sci.*, vol. 258, pp. 371–386, 2014.

[24] M. Y. Miyoung Jang and J.-W. Chang, "A new query integrity verification method with cluster-based data transformation in cloud computing environment," *Int. J. Smart Home*, vol. 9, no. 4, pp. 225–238, 2015.

[25] Y. Tong, J. Sun, S. S. M. Chow, and P. Li, "Cloud-assisted mobile-access of health data with privacy and auditability," *IEEE J. Biomed. Health Informat.*, vol. 18, no. 2, pp. 419–429, Mar. 2014.

[26] A. Kundu and E. Bertino, "Structural signatures for tree data structures," *Proc. VLDB Endowment*, vol. 1, pp. 138–150, Aug. 2008.

[27] S. Dongwan, R. Lopes, and W. Claycomb, "Authenticated dictionary-based attribute sharing in federated identity management," in *Proc. 6th Int. Conf. Inf. Technol., New Gener*, Apr. 2009, pp. 504–509.

[28] D. A. G.-U. Layla Pournajaf, L. Xiong, and V. Sunderam, "Survey on privacy in mobile crowd sensing task management," Dept. Math. Comput. Sci., Emory Univ., Atlanta, GA, USA, Tech. Rep. TR-2014-002, 2014.

[29] H. Qian, J. Li, Y. Zhang, and J. Han, "Privacy-preserving personal health record using multi-authority attribute-based encryption with revocation," *Int. J. Inf. Security*, vol. 14, no. 6, pp. 487–497, 2015.

[30] B. Qin, H. Deng, Q. Wu, J. Domingo-Ferrer, D. Naccache, and Y. Zhou, "Flexible attribute-based encryption applicable to secure e-healthcare records," *Int. J. Inf. Security*, vol. 14, no. 6, pp. 499–511, 2015.

[31] A. Lewko and B. Waters, "Why proving HIBE systems secure is difficult," in *Proc. Adv Cryptol. EUROCRYPT 2014*, 2014, vol. 8441, pp. 58–76, 2014.

[32] Y. Tong, J. Sun, S. S. M. Chow, and P. Li, "Cloud-assisted mobile-access of health data with privacy and auditability," *IEEE J. Biomed. Health Informat.*, vol. 18, no. 2, pp. 419–429, Mar. 2014.

[33] S. Haber *et al.*, "Efficient signature schemes supporting redaction, pseudonymization, and data deidentification," in *Proc. 2008 ACM Symp. Inf., Comput. Commun. Security*, 2008, pp. 353–362.

[34] G. Ghinita, "Privacy for location-based services synthesis," *Synthesis Lectures Inf. Security Privacy, and Trust*, vol. 4, no. 1, pp. 1–85, Apr. 2013.

[35] X. Liang, R. Lu, L. Chen, X. Lin, and X. Shen, "PEC: A privacy-preserving emergency call scheme for mobile healthcare social networks," *J. Commun. Netw.*, vol. 13, no. 2, pp. 102–112, Apr. 2011.

[36] B. Stefan, "U-Prove technology overview," Mar. 2010.

[37] M. A. D. Mashima, D. Bauer, and D. Blough, "User-centric identity management architecture using credential-holding identity agents," in *Digital Identity and Access Management: Technologies and Frameworks*. Hershey, PA, USA: IGI Global, Dec. 2012.

[38] F. Paci, R. Ferrini, A. Musci, K. Steuer, and E. Bertino, "An interoperable approach to multifactor identity verification," *IEEE Comput.*, vol. 42, no. 5, pp. 50–57, May 2009.

[39] R. Lu, X. Lin, and X. Shen, "SPOC: A secure and privacy-preserving opportunistic computing framework for mobile-healthcare emergency," *IEEE Trans. Parallel Distrib. Syst.*, vol. 24, no. 3, pp. 614–624, Mar. 2013.

Rosa Sánchez-Guerrero, photograph and biography not available at the time of publication.

Florina Almenárez Mendoza, photograph and biography not available at the time of publication.

Daniel Díaz-Sánchez, photograph and biography not available at the time of publication.

Patricia Arias Cabarcos, photograph and biography not available at the time of publication.

Andrés Marín López, photograph and biography not available at the time of publication.