Privacy-Aware Large-Scale Virological and Epidemiological Data Monitoring

Constantinos Patsakis[1], Michael Clear[1], Paul Laird[1], Athanasios Zigomitros[2], Mélanie Bouroche[1]
[1] Distributed Systems Group, School of Computer Science and Statistics , Trinity College, Dublin, Ireland
[2] Institute for the Management of Information Systems, "Athena" Research Center, Greece &
Department of Informatics, University of Piraeus, Greece

***Abstract*—Modern mobile and wearable devices are enabling the realization of so-called ubiquitous computing. This provides citizens the technological means to contribute to urban management by becoming sensors within a smart city. Notwithstanding, the health sector is a very crucial factor for city management, imposing restrictions to the decisions directly or indirectly. The question that arises is given the current technological advances, could we collect health related data from citizens without violating their privacy? In this work we propose a methodology that can be used to allow citizens to send their data without disclosing their identity, while simultaneously enabling almost real-time urban-scale virological and epidemiological data monitoring.**

***Index Terms*—Privacy, Epidemiology, Data Monitoring, patient monitoring**

I. Introduction

Recent advances in telecommunications and hardware development have enabled users with many advanced and highly sophisticated gadgets to access remote services easily and securely, with minimal costs. Smartphones and tablets are two very well known examples that are conquering the market in the past few years. Nevertheless, numerous wearable sensors and gadgets are gradually being developed and being pushed into the market, enabling even more opportunities to developers, as well as more personalised services for the end users.

In the mean time, academia and industry are trying to provide in-house low-cost medical tests to enable simple medical tests to be performed remotely, from people in their houses. Several solutions that can be embedded in mobile phones or used as wearable devices have been developed, redefining drastically how medical examinations are made and delivered, along with their frequency and level of penetration. Such devices enable low-cost ubiquitous monitoring of patients with minimal intrusion. The significance of this achievement can be understood on two levels; the personal and the societal. On a personal level, patients can keep track of their medical conditions and detect symptoms in real-time, and even more interestingly, trace symptoms that could have been neglected due to their gradual development. Even more, patients could have access to personalised treatments, exactly specialized to their needs. From the societal perspective, such methodology will allow objective and indisputable translation of patients' symptoms, along with detection of unknown patterns and procedures. Moreover, real-time medical reports from patients can drastically improve the management of smart cities. From automating the procedure of routing emergency vehicles to reach patients that suffer heart attacks, to keeping track of which drugs are probable to be used in the near future in hospitals according to real-time epidemiological reports.

While all the above sound ideal, the ground truth is that patients and citizens would be very hesitant to share all this information. The major concern does not have to do with the physical invasiveness of these methods, but more with the privacy of their contributed data. The sensitivity of medical data and the condition of many patients due to their problems is definitely a big drawback towards their adoption in the near future.

Exploiting the capabilities of modern mobile devices allows users to perform medical tests and access to more personalised medical services, real-time monitoring and real-time reporting of medical emergencies. However, one could exploit this information for research and urban management to derive real-time knowledge. To achieve this goal, the users must share the gathered information with others. Due to the sensitivity of the information, many privacy issues arise, making users quite hesitant on whether they should share this information. The main research question considered in this work is how to aggregate sensor information.

While there are already some solutions in this field, the main drawback that most of them face is that they offer poor scalability. Moreover, if they have to be applied on a large scale, they require the presence of local aggregators, which should be regarded as trusted third parties. Thus, aggregation results will be disclosed to entities in the system beyond the target recipient such as a health authority. Therefore, for urban-scale reporting, local aggregators might be necessary, but they cannot be trusted. So the problem is: "How can we aggregate sensitive information such as medical measurements, in a scalable and private way?"

The main contributions of the article is the provision of a novel architecture that allows the ubiquitous monitoring of patients, which respects their right to privacy. Moreover, the proposed architecture is scalable, allowing the monitoring to be performed on an urban scale. The underlying protocol is one that has been recently proposed in the literature, and we adopt it in this work to illustrate how it can be exploited to allow virological and epidemiological data monitoring.

The next section presents two motivating scenarios that illustrate the need and probable applications of the proposed methodology. In section three we present the related work,

1063-7125/14 $31.00 © 2014 IEEE
DOI 10.1109/CBMS.2014.89

discussing the limitations of the current state of the art. The following section presents the proposed methodology, discussing its security and performance. Finally, the article concludes in section five.

II. Motivating scenarios

Scenario I: We assume that there is an easy and affordable gadget $\mathcal{G}$ that provides testing for virus $\mathcal{V}$. $\mathcal{V}$ might not cause big issues to the health of the general population, however, susceptible groups such as the elderly or people with heart diseases might face important issues with it. A typical example of $\mathcal{V}$ can be common flu. Moreover, $\mathcal{G}$ can connect to the Internet and report the result to a central server $\mathcal{S}$ that is monitoring the epidemiology results of area A. Alice would like to have a walk in A, however, since she has heard that there are many incidents of infection of $\mathcal{V}$, she is very rather hesitant on where to go, as she is suffering heart issues for many years. Alice would like to be able to see which parts of $\mathcal{A}$ are the ones with most infections in order to choose her destination, her route and mode of transportations.

Scenario II: We assume that there is an easy and affordable gadget $\mathcal{G}'$ that provides testing for virus $\mathcal{V}'$. Patients with $\mathcal{V}'$ suffer many heavy side effects and some of them might even face social discrimination due to their health conditions. For instance, $\mathcal{V}'$ could be HIV. The ministry of health would like to monitor the patients and their vital measurements, as this allows the ministry to track the epidemiology and allows researchers to extract valuable information about $\mathcal{V}'$. Given the sensitivity of realizing the health condition of the patients, the data could be collected from local authorities; nevertheless, they should not be able to link the data to individual patients.

III. Related work

The contribution of mobile phones in providing faster, better and more personalized medical services in terms of so-called m-health [1] is indisputable. In real world applications, even the use of mobile phones to send and receive exams via SMS is proving to be very beneficial. However, most of the applications are exploiting only a fraction of the possibilities that modern smart phones can offer. Going a step further, many researchers regard that they can be used to provide wireless ubiquitous computer patient monitoring. Several approaches towards this direction have been proposed, with many of them involving wireless body area networks [2], [3], [4], [5], [6]. Other approaches are more focused on the privacy perspective of the problem, targeting the monitoring of mobility of patients with mild cognitive impairment and dementia [7], [8].

Current hardware advances have enabled a wide range of tests to be made from users themselves by making use of their smartphones and some additional hardware extensions. Therefore, molecular microscopy can be made with low-cost processing on smartphones, enabling users to make a wide range of blood or urine tests [9], [10], [11], [12], [13]. Some of these solutions have already been exploited to create commercial products[1]

With one more layer of abstraction, the smart phones could use other sensors, apart from the embedded or attached ones to provide even more fine-grained health services. In this context, Solanas et al. are bridging the concepts of smart cities and e/m-health proposing the exploitation of smart city sensors, with what is defined as s-health [14]. However, apart from personal services, the results can be used to provide a better insight of the city to urban management stakeholders.

While providing the need for gathering such statistics is imminent, the sensitivity of the data imposes the data to be submitted anonymously, so that users are not "targetted". There are several ways to provide users with anonymity, nevertheless, each of them has its own limitations. A quite straightforward approach is using mix-networks, nevertheless, this introduces the presumption that the user trusts the proxy that his data will not be disclosed.

In order to provide anonymity to users a group key management scheme could be used. The drawback of this approach is that the provided anonymity would allow users to flood the service. This could be easily achieved since the key that they possess allows them to submit an arbitrary number of reports. Another fostered approach to monitor and interact with patients is through the use of medical online social networks. Using the concept of "Privacy as a Product" in the case of m-health, Patsakis and Solanas proposed a scheme where users could trade their personal medical information within health-related online social networks [15].

IV. The anonymous aggregation protocol

Patsakis et al. in [16] have proposed a protocol which allows users to perform private aggregation of their values. This protocol, from now on PCL, can be understood as an extension of the smart grid aggregation protocol of Kursawe et al. [17], from now on KDK. The protocol is proved to be secure in the semi-honest model under a standard cryptographic assumption (the Decisional Diffie-Hellman Problem). This means that users are assumed to be honest but curious; they will execute the protocol correctly but might collude in order to recover the input of others. Apart from its security, the protocol introduces several additional features compared to its predecessors.

The PCL protocol can be understood as another multi-round variant of KDK without pairings. The protocol allows a bounded number of rounds ℓ to be performed from the same public key information. However, ℓ depends on the acceptable collusion tolerance $t \leq n$. Both single-round and multi-round versions of KDK are t-private for any $t \leq n$, while PCL is t-private, for at most $\ell = \lfloor \frac{n-t}{2} \rfloor$ rounds. Concretely, for $t = n/3$ (Byzantine tolerance) and $n = 100$, 33 rounds can be executed before re-keying. One of the advantages of PCL is that it only relies on the DDH assumption. A cyclic group is one that is generated by a single element, called a *generator*. Here we denote by g such a *generator* for the group $\mathbb{G}$. Now $\mathbb{G}$ may be

[1] http://eyenetra.com/product-netrag.html, http://www.withings.com/bloodpressuremonitor, http://www.scanadu.com, http://en.sanofi.com/products/diabetes/diabetes.aspx

instantiated by different types of groups. In fact, multi-round KDK requires elliptic curves over finite fields due to its heavy reliance on pairings, making it far less efficient in practice than PCL. Additionally, the lack of pairings allows the adoption of PCL on devices with low processing capabilities. While we instantiate $\mathbb{G}$ in this work with a subgroup of $\mathbb{Z}_q^*$ (of prime order p), we note that an instantiation with an elliptic curve group would have offered improved performance.

KDK is based on a fixed matrix A with entries its entries being $\{-1, 0, 1\}$ and determining the exponents used to compute w_i. This matrix is used as follows: user U_i raises user U_j's public key u_j to the power $A_{i,j}$ when computing w_i. In KDK skew-symmetric matrix A is used, that is $-A = A^T$. The main idea in PCL is to generate a new skew-symmetric matrix $A^{(k)}$ in a deterministic manner for each round k, with the minimal communication effort. Furthermore, the matrix $A^{(k)}$ is chosen to have coefficients in $\mathbb{Z}_p$ instead of $\{-1, 0, 1\}$ in order to prove its security (recall that the group $\mathbb{G}$ we are using is of order p). We refer the reader to [16] for more details. Here we assume a function $\chi : \mathbb{Z}_p \times \mathbb{Z} \to \mathbb{Z}_p^{n\times n}$ that takes a random seed and a round number, and outputs a pseudorandom skew-symmetric matrix over $\mathbb{Z}_p$. Note that the seed can be predetermined or derived from the users' public keys. The main differences to single-round KDK are as follows: Let $s \in \mathbb{Z}_p$ be a seed deterministically derived form $u_1, \ldots, u_n$. In round k, user U_i computes $A^{(k)} \leftarrow \chi(s, k) \in \mathbb{Z}_p^{n\times n}$ and $w_i \leftarrow \prod_{j\in 1}^{n} u_j^{A_{i,j}^{(k)}} \in \mathbb{G}$. Then he computes $v_i^{(r)} \leftarrow w_i^{x_i} \cdot g^{m_i^{(r)}} \in \mathbb{G}$ and broadcasts it. To compute the aggregated value of round r, one has to compute the product $\prod v_i^{(r)}$. The values $w_i^{x_i}$ will cancel each other out, leaving the value $g^{\sum m_i^{(r)}}$, which for small values of $m_i^{(r)}$ can be easily brute forced, without disclosing though the value individual values of each user.

Extending the protocol even further, it can even support Blind Aggregation, that allows users to forward their values to a local aggregator in such a way that when he computes the aggregated result is obfuscated to him as well. Therefore, the users may use the aggregator as an intermediate between them and the data consumer, without the risk of disclosing the result to him. This allows the data consumer to have several aggregators between him and the users, which allows scaling, without the need to fully trust aggregators, decreasing significantly his exposure.

V. The proposed solution

The proposed solution is based on the *anonymous aggregation system* that was described in the previous section. In what follows, we will first introduce the main players of the scheme along with the prerequisites. Afterwards, we are going to discuss how the concept of *anonymous aggregation* can to be used and finally we are going to discuss the efficiency of the proposed model in terms of security, privacy and performance.

The main actors in our architecture and their roles are as follows: (i) **Health authority:** This is the entity that wants to monitor the epidemiology of a virus $\mathcal{V}$, such as the ministry of health, a hospital or a research institute. This entity is denoted by $\mathcal{H}$. (ii) **Users:** They are citizens equipped with a testing device $\mathcal{D}$ for virus $\mathcal{V}$ that can connect to the Internet to transmit the results. Each user is referred to as U_i. (iii) **Local aggregators:** Every area $\mathcal{A}$ has its own anonymous aggregator, where users of this area can connect. Each local aggregator has a registry where users can store and read data securely. The latter will help in minimizing the amount of exchanged messages. This entity is denoted by A_j. We assume that the tests are made periodically every $\mathcal{T}$ hours.

Given that we have a large number of users, each user U_i connects to the closest Aggregator A_j to send his result. Each aggregator, multiplies the users' submitted values, to compute an obfuscated version of the local sum. The latter is forwarded to $\mathcal{H}$ who recovers the actual local summaries and maps them to the according areas, as shown in Figure 1. Since the values that are sent from each device are bounded, their summary is also bounded and can be extracted from the PCL protocol.

Depending on the nature of the data and what kind of statistics have to be computed, apart from summaries, the distinct submitted values can also be extracted by making a slight change to the protocol. For instance, if we assume that the n users will submit values in the range $\{1, ..., k\}$, then the following changes can be made which facilitates extraction of the individual values of the users, although without allowing anyone to find a link between a given value and a user. The prime p (recall that this is the order of the group $\mathbb{G}$, and in our case the latter is instantiated by a multiplicative subgroup of $\mathbb{Z}_q^*$) is selected so that $\mathcal{P}(k)^n < p$, where $\mathcal{P}(i)$ denotes the $i-$th prime number. Then, instead of setting $v_i^{(r)} \leftarrow w \cdot g^{m_i^{(r)}} \in \mathbb{Z}_q$ as in the original protocol, users set $v_i^{(r)} \leftarrow w \cdot \mathcal{P}[m_i^{(r)}] \in \mathbb{Z}_q$. It is obvious that the result that the Health Authority will recover i s going to be a smooth number which is a product of small primes. This means that it can be trivially factored, recovering the individual values of the users, but in an anonymous manner.

The computational cost of the scheme can be easily computed. The cost of each aggregator is to perform a product of n elements $\mod q$. The cost of the Health Authority to recover each local product is one multiplication, one exponentiation and one inversion $\mod q$. Finally, the cost of the user is proportional to the amount of users that connect to the local aggregator. So for the initialization she has to perform one exponentiation $\mod q$. Afterwards, she has to perform $n-1$ multiplications and n exponentiations $\mod q$ (per round of aggregation). Instead of setting $\mathbb{G}$ to a multiplicative subgroup of $\mathbb{Z}_q^*$ (of prime order p), one could use an elliptic curve group so as to have an even more efficient implementation. However, for sake of continuity with the reported experimental results, we have not opted to do this for our implementation.

The use of PCL protocol, apart from diminishing the computational effort on both $\mathcal{H}$ and users, it minimizes the communication overhead as well. The latter is achieved as users do not have to exchange keys before each round. To further speed up the process, the local aggregator could offer users a registry to publish their values, and then broadcast the

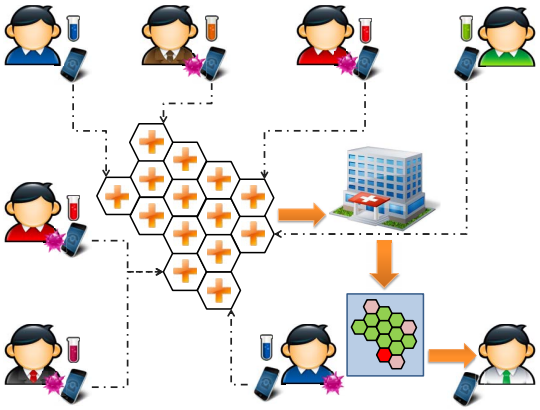

Fig. 1. The proposed solution.

TABLE I
EXPERIMENTAL RESULTS, TIME IN MS.

Prime	Users			
	100	500	1000	2000
512 bits	167	176	327	717
1024 bits	293	337	722	930

submitted values, minimizing the exchanged messages. Thus the amount of exchanged messaged, drops from quadratic to linear with respect to the number of users. These refinements allow devices with low processing resources to manage their computing resources more efficiently.

It is important to note that the scheme provides users with full privacy of their submitted value, which can only be lifted in case of full collusion of the other users. This is the case for 1 round of aggregation. If the users are satisfied that it takes t users to collude to violate their privacy, then $\frac{n-t}{2}$ rounds can be executed without re-generation of public keys in addition to the communication cost of transmitting them. We also have to note here that the submitted values are blinded to the local aggregator, since he is not aware of the decryption key. Additionally, even if some users decide to collude with him, the local sum cannot be deduced. Therefore, local authorities cannot derive the local sums without $\mathcal{H}$'s approval

An Android application to test the efficiency of this solution has been developed, where each user is submitting a value in the range of $[0-1000]$. The tests were executed on a Samsung GT-I9001 with Android 2.3.3. As already stated, we used the less efficient approach of the multiplicative group $\mod q$ without multi-threading. The experiments are summarized in Table I and clearly indicate that the computational cost is minimal for a modern smart phone. The results in Table I refer to the time taken to perform the computations involved in a single round of the protocol for a given participant.

VI. CONCLUSIONS

In this work we illustrated a solution which allows virological and epidemiological data monitoring that scales efficiently, even in urban scale, without compromising users's privacy. The computational and communication overhead of the solution is rather low, enabling the development of the scheme to many modern mobile devices and data reporting to almost real-time. The scheme realizes a very powerful data reporting tool, that can be considered very beneficial for both researchers and patients. Moreover, the reports can be exploited by authorities to provide more fine-grained daily services to their citizens, depending on their medical needs without "big-brother" phenomena.

REFERENCES

[1] R. S. Istepanian, E. Jovanov, and Y. Zhang, "Guest editorial introduction to the special section on m-health: Beyond seamless mobility and global wireless health-care connectivity," *Information Technology in Biomedicine, IEEE Transactions on*, vol. 8, no. 4, pp. 405–414, 2004.

[2] C. Otto, A. Milenkovic, C. Sanders, and E. Jovanov, "System architecture of a wireless body area sensor network for ubiquitous health monitoring," *Journal of Mobile Multimedia*, vol. 1, no. 4, pp. 307–326, 2006.

[3] S. Sneha and U. Varshney, "Enabling ubiquitous patient monitoring: Model, decision protocols, opportunities and challenges," *Decision Support Systems*, vol. 46, no. 3, pp. 606–619, 2009.

[4] S. Nelwan, T. Van Dam, P. Klootwijk, and S. Meij, "Ubiquitous mobile access to real-time patient monitoring data," in *Computers in Cardiology, 2002*. IEEE, 2002, pp. 557–560.

[5] E. Jovanov, A. Milenkovic, C. Otto, and P. C. De Groen, "A wireless body area network of intelligent motion sensors for computer assisted physical rehabilitation," *Journal of NeuroEngineering and rehabilitation*, vol. 2, no. 1, p. 6, 2005.

[6] S. Kumar, K. Kambhatla, F. Hu, M. Lifson, and Y. Xiao, "Ubiquitous computing for remote cardiac patient monitoring: a survey," *International journal of telemedicine and applications*, vol. 2008, p. 3, 2008.

[7] A. Solanas, A. Martinez-Balleste, P. A. Perez-Martinez, A. F. d. l. Pena, and J. Ramos, "m-carer: Privacy-aware monitoring for people with mild cognitive impairment and dementia," *Selected Areas in Communications, IEEE Journal on*, vol. 31, no. 9, pp. 19–27, 2013.

[8] C. Patsakis, "A cryptographic approach for monitoring patients with mild cognitive impairment and dementia," in *Computer-Based Medical Systems (CBMS), 2013 IEEE 26th International Symposium on*. IEEE, 2013, pp. 502–505.

[9] V. Oncescu, M. Mancuso, and D. Erickson, "Cholesterol testing on a smartphone," *Lab on a Chip*, 2014.

[10] H. Zhu, I. Sencan, J. Wong, S. Dimitrov, D. Tseng, K. Nagashima, and A. Ozcan, "Cost-effective and rapid blood analysis on a cell-phone," *Lab on a Chip*, 2013.

[11] A. F. Coskun, R. Nagi, K. Sadeghi, S. Phillips, and A. Ozcan, "Albumin testing in urine using a smart-phone," *Lab on a Chip*, vol. 13, no. 21, pp. 4231–4238, 2013.

[12] I. Navruz, A. F. Coskun, J. Wong, S. Mohammad, D. Tseng, R. Nagi, S. Phillips, and A. Ozcan, "Smart-phone based computational microscopy using multi-frame contact imaging on a fiber-optic array," *Lab Chip*, vol. 13, no. 20, pp. 4015–4023, 2013.

[13] H. Zhu and A. Ozcan, "Wide-field fluorescent microscopy and fluorescent imaging flow cytometry on a cell-phone." *Journal of visualized experiments: JoVE*, no. 74, 2012.

[14] A. Solanas, C. Patsakis, M. Conti, I. Vlachos, V. Ramos, F. Falcone, O. Postolache, P. Pérez-Marténez, R. Di Pietro, D. Perrea, , and M.-B. A, "Smart health: A context-aware health paradigm within smart cities," submitted manuscript. 2013.

[15] C. Patsakis and A. Solanas, "Privacy as a product: A case study in the m-health sector," in *Information, Intelligence, Systems and Applications (IISA), 2013 Fourth International Conference on*. IEEE, 2013, pp. 1–6.

[16] C. Patsakis, M. Clear, and L. Paul, "Anonymous aggregators and sensor aggregation on untrusted servers," *IACR Cryptology ePrint Archive*, vol. 2013, p. 661, 2013.

[17] K. Kursawe, G. Danezis, and M. Kohlweiss, "Privacy-friendly aggregation for the smart-grid," in *Privacy Enhancing Technologies*. Springer, 2011, pp. 175–191.