Personal Data Management with Secure Hardware

The Advantage of Keeping your Data at Hand

Nicolas Anciaux
INRIA Rocquencourt
Domaine de Voluceau
Le Chesnay, FR

Benjamin Nguyen
Iulian Sandu-Popa
University of Versailles St-Quentin-en-Yvelines
45 Avenue des Etats-Unis
Versailles, FR

***Abstract*— How do you manage your personal data? More specifically, how do you keep a secret about your personal life in an age where your glasses record and share everything you sense, your wallet records and shares your financial transactions, and your set-top box records and shares your energy consumption? Currently, several decentralized alternative solutions are proposed, based on the emergence of trusted personal devices controlling the data at the edges of the Internet. In this seminar, we review existing solutions, present a functional architecture for such alternatives, expose recent techniques dealing with embedded data management and global query processing this architecture, and conclude by presenting existing and future implementations of this approach.**

Keywords—personal data management; embedded data management; decentralized architecture; privacy preserving data sharing.

"The following provisions apply to users and non-users who interact with Facebook outside the United States:
1) You consent to having your personal data transferred to and processed in the United States. (...)"

an extract from the Facebook Terms of Service

INTRODUCTION (15 MINS)

With the convergence of mobile communication, sensors and online social networks technologies, we are witnessing an exponential increase in the creation and consumption of personal data. Paper-based interactions (e.g., banking, health), analog processes (e.g., photography, resource metering) or mechanical interactions (e.g., as simple as opening a door) are now sources of digital data that can be linked to one or several individuals. This *personal* data is recognized by the World Economic Forum as a most valuable resource comparable to "*the new oil*" [1], creating an unprecedented potential for applications and business.

Until now, enthusiasm for these new opportunities has thwarted privacy concerns. Individuals conscientiously build Facebook pages, conduct their communications via Gmail, and send and receive megabytes of personal information to and from administrations or commercial services. Nevertheless, the risk of a backlash is growing as new devices and new services bring us closer to the dystopias described in science fiction literature. Current practices are often not compliant with basic privacy laws and directives. Data leaks are legion [20]. Worse, underlying business models are even *based* on breaches of users' privacy. Anyone may exploit weak privacy policies or cross-analyze sensed data with data conscientiously registered on social networks.

Many laudable projects cannot grow on such bases. Indeed, should humanitarian aid workers build their applications if they cannot provide privacy guarantees? Imagine an application whose goal is to help the homeless, by providing medical monitoring. Managing critical information on potentially discriminated people under weak privacy guarantees could be seen as too strong a danger to create this application. Many other applications, undoubtedly useful, but placing respect for human dignity and privacy upfront, are thus sometimes left by the wayside.

The nature of the solution is consensual: it is necessary to increase the control that individuals have over their personal data [2, 3, 4]. The World Economic Forum even claims that "*increasing the control that individuals have over the manner in which their personal data is collected, managed and shared will spur a host of new services and applications*" [1].

Centralized solutions, including emerging cloud-based personal data vault management platforms, trade security and protection for innovative services. At best, such approaches formulate sound privacy policies, but none of them propose mechanisms to automatically enforce them [5]. Even TrustedDB [6], which proposes tamper-resistant hardware to secure outsourced centralized databases, does not solve the two intrinsic problems of centralized approaches. First, users are hostages of sudden changes in privacy policies, their data can also be unexpectedly exposed by negligence or because it is regulated by too weak policies. Second, users are exposed to sophisticated attacks, whose cost-benefit is high on a centralized database.

Decentralized solutions are promising because they do not exhibit these intrinsic limitations. FreedomBox [7] is a good representative. Low-cost plug computers and open software are provided to users to enable anonymous and independent communication networks. The Personal Data Server (PDS) [8] project embeds a relational database in a tamper-resistant token at the user side. Privileges on user views can be granted and revoked by PDS holders and are enforced by other PDS holders. Many other projects and startups (e.g., Project VRM) are discussed in [9].

In this seminar, we advocate neither for nor against any type of application, be it the most trivial Facebook app, or a low-cost electronic health record for the homeless. However,

the advent of secure hardware embedded in many forms of personal devices, at the edges of the Internet, may cause a sea change for personal data management, by returning complete control of users on their data, anywhere and anytime.

The core of the seminar will be organized as follows:

In Section II, we review the recent initiatives pursuing the objective of reestablishing user control over their data by decentralizing this control in personal secure or trusted devices. Then, we discuss an abstract distributed architecture focusing on secure storing, managing and sharing of personal data, i.e., the *asymmetric architecture*. We then indicate the main challenges inherent to decentralized data management, and focus on embedded data management (Section III) and global query processing (Section IV). Regarding embedded data management, user control must be exercised within a user-constrained device (highly constrained when endowed with tamper resistant resources). We review the main attempts proposed in the literature and concentrate on those addressing the specific context of (secure) microcontrollers [19, 21, 22]. Regarding the problem of global processing, we present the difficulties to overcome to execute queries on a PDS based architecture, and illustrate it by focusing on data anonymization techniques and Group By SQL queries [13, 18]. Finally, we conclude the seminar by presenting concrete existing and future instances of decentralized privacy preserving data management architectures (Section V). We mainly focus on attempts and proposals targeting the social-medical, smart houses, and rural areas contexts.

We plan an overall duration of 2 hours. In the next sections, we present in more detail the key points of the seminar.

DECENTRALIZED ARCHITECTURES (30 MINS)

Recently, several decentralized architectures have been proposed for better user privacy and security [7, 9, 8]. We present them in detail in the seminar. We also discuss the functionality or usability that may be sacrificed for security and privacy, since decentralized personal data architectures are also severely judged [9]. However, among these projects, very few focus on the aspects related to data management. We describe in this section a global architecture enabling the secure management and sharing of personal data, while maintaining the full functionalities of a traditional database server.

The global architecture for managing personal data revolves around a key element, i.e., a trusted Personal Data Server (PDS). PDSs allow users to store, manage and share their personal data. Besides, an untrusted infrastructure for communication and other services is required to interconnect the PDS ecosystem. It is worth mentioning that one cannot assume that the PDSs are available at all times as a typical data server.

A trusted PDS (or secure token) can have different form factors (e.g., SIM card, USB token, Secure MicroSD [10]) and names (e.g., Personal Portable Security Device, Smart USB Token [11], Portable Security Token [10] or Secure Portable Token [8]). Despite the diversity of existing hardware solutions, a PDS can be abstracted by (1) a Trusted Execution Environment and (2) a (potentially untrusted) mass storage area. For example, the former can be provided by a tamper-resistant microcontroller, while the later can be provided by an SD card. An important observation is that it is very unlikely to tamper with the code executed by the secure token. In addition, the content of the mass storage area can be protected thanks to cryptographic protocols.

Due to their weak availability and modest computing resources, the secure tokens cannot implement all the expected functionalities on their own. First, they require external communication facilities (i.e., a mailbox-like service) to handle the asynchrony of their connections and allow inter-device communications. Second, they require a safe (i.e., highly available and resilient) external storage area to implement their own durability (i.e., being able to recover the content of a Personal Database in case of crash/loss/steal of a secure token. Third, they require a large temporary storage area to store the intermediate results produced by global processing functions. Since it seems a reasonable assumption, we suppose that these external storage and communication facilities are provided by so-called (unsecure) Supporting Services, which can be implemented in multiple ways (e.g., by a service provider in the Cloud). The main requirement for supporting services is the high availability of the services they provide. However, as any traditional server, a supporting service is prone to privacy violations due to negligence, internal and external attacks.

This architecture is distributed since the personal data is scattered in the PDS ecosystem. Also, this architecture consists, on the one hand of a very large number of low power, weakly connected but highly trusted secure tokens and on the other hand of highly available external storage and communication resources provided by untrusted supporting services. We call it the *asymmetric architecture*.

Therefore, the challenges of the PDS architecture are threefold. The first objective is to allow the development of new, powerful, user-centric applications, which require a well-organized, structured and queryable representation of user's data. Second, we want to provide the data holder with control over the sharing conditions related to her data and to provide the data recipient with certified information related to their provenance. Third, PDSs must provide traditional database services like durability, query facilities, transactions and must be able to interoperate with external sources. In the rest of this seminar, we focus on secure and constrained data management and query execution on an asymmetric architecture.

RESOURCE CONSTRAINED DATA MANAGEMENT (30 MINS)

The expected behavior of a personal data server in a highly distributed setting is to efficiently manage the personal data of its owner by acting as a trusted doorkeeper delivering only authorized data to each requester connecting to the personal device. The requester has to be authenticated at each access, so that privileges can be properly associated with users in order to protect data confidentiality and integrity. Fine-grained privileges are usually granted to users by means of views on the data, expressed by database queries (e.g., seeing the result of a select of an aggregate query can be authorized while the raw data targeted by that query remain hidden). Hence, for the PDS to respond to the privacy requirement, it needs to embed a

powerful query engine, in charge of extracting user views and checking the integrity of the data involved in the computation.

To enforce access control, the embedded code evaluating the authorized view should be evaluated into a trusted area. The trusted area can consist of a tamper resistant and secure microcontroller (SMCU) with an access to the storage area storing the data (e.g., a flash memory card or chip). However, any other form of hardware/software combination providing a certain degree of trust (e.g., a plug computer running open source code) can be employed.

Several studies have been proposed, suited to different categories of end-user devices. Well known state-of-the-art embedded DBMS products (e.g., SQLite, BerkeleyDB) and light versions of popular DBMSs (e.g., DB2 Everyplace, Oracle Database Mobile Server) target devices like smart phones and set top boxes, far more powerful than smart tokens. We focus on the devices providing the highest security guarantees, namely secure microcontrollers. Embedding data management functionalities into a SMCU is difficult because of contradicting hardware constraints: (1) existing SMCUs have a tiny RAM; (2) the external NAND flash memory storing the data badly supports random writes; and (3) that external flash memory does not benefit from the SMCU's tamper-resistance. These constraints lead to contradictory objectives: executing queries with acceptable performance with a tiny RAM entails indexing massively the users' data in Flash, while index updates generate fine-grained random writes, and then unacceptable NAND flash write costs and cryptographic overhead. Proposals dedicated to data management techniques embedded on SMCUs, e.g., [25], consider very small databases –hundreds of kilobytes–, and are not scalable.

Many studies (e.g., [23, 24, 26]) address the problem of storage and indexing in NAND flash, either by adapting the traditional B+-Tree structure by relying on a flash-resident log to delay the index updates [23, 24] or by adapting principles from log-structured file systems [26]. The gain in terms of flash write cost is linked to the size of the log and the log size itself determines RAM consumption, which is usually quite high, suited for relatively powerful devices like smart phones.

For a SMCU linked to a large personal database (storing potentially hundreds of thousands up to millions of documents), the RAM consumption and random write cost are always conflicting. Recent work [21, 22] that have been proposed in this context, help resolving the intrinsic conflict of low RAM/large Flash by organizing the complete database (raw data but also indexes, buffers and logs) into purely sequential Log Containers in flash, while still reaching scalability using an iterative reorganization of the database indexes. In this part of the seminar we review several techniques with a focus on those studies.

GLOBAL QUERYING (30 MINS)

The asymmetric architecture completely changes the paradigm of data centric applications. Indeed, most of these applications need to *holistically* process large quantities of user data in order to produce results. It is important to understand how it is possible to adapt existing applications to this architecture. Secure Multiparty Computation (SMC) solutions [16, 17], which could appear as generic candidates to solve the problem, are in fact not useable for data intensive applications since they do not scale. In this part of the seminar, we provide a thorough analysis of the impact of the asymmetric architecture when looking for solutions supporting such applications.

A centralized algorithm can indeed be transformed into an asymmetric one, if it is possible to divide it into 3 phases: 1) a collection phase where encrypted data is gathered, 2) a construction phase, where this encrypted data is managed by the supporting server and 3) a sanitization phase where this data is recombined and deciphered by the secure tokens.

As a first illustration [13], we study non-dynamic data applications, which only involve publishing user data. Obviously, since this data is private, it must be published using a privacy preserving method (i.e. an anonymization technique). Many different privacy preserving data publishing techniques exist, all linked to various anonymization models, ranging from simple and not very secure pseudonymization or k-anonymization techniques [14], to advanced differentially private [15] methods. In a traditional setting, all the data is centralized on a server, and an algorithm is executed. We show how it is possible to provide generic functions (called *asymmetric architecture primitives*) that can be used by a generic protocol to support many existing anonymization techniques. We also note the generality of these primitives, which will appear as building blocs in all global query processing problems. As a second illustration, we discuss the challenges of running any kind of SQL query on the asymmetric architecture [18]. Difficulties similar to those of PPDP arise when computing Group By queries.

CURRENT AND FUTURE INSTANCIATIONS (15 MINS)

In this part of the seminar, we present existing and future instances of the decentralized privacy preserving personal data management systems.

A. DMSP Experiment

The seminar includes a presentation of a field experiment conducted in the Yvelines District in France since the end of 2011, which we consider emblematic for the decentralized approaches considered here. The system architecture is designed to manage social-medical folders for dependent patients, and relies on (i) secure and portable personal data servers made of a SMCU and Flash memory which is provided to patients embedded under a USB key form factor, (ii) a smart badge composed of a storage and a smartcard reader to authenticate medical-social professionals provided to each professional, and (iii) a central server used to store (possibly encrypted) copies of the data. The personal folders are used to access medical folders without any internet connection. They implement access control rules and regulate data dissemination. The central server coupled with synchronization protocols, is used to achieve data durability, to securely share personal data over the internet, and to feed external processing chains. We emphasize what this architecture can bring in terms of privacy and functionalities.

B. *Human-Powerd Information System*

Several reports indicate clearly that Information and Communication Technologies can play a catalytic role in Least Developed Countries (LDCs), by helping to achieve primary education, reduce mortality, boost individual commercial initiatives, etc. Yet, for many LDCs' analysts, the deployment of e-services in a foreseeable future is facing many obstacles: lack of infrastructure, high deployment and maintenance cost, and reluctance of field workers due to privacy concerns. Without any global access to the Internet, providing personal digital records, data sharing and e-services is difficult. LDCs are thus, due to lack of infrastructure, left by the wayside.

We believe that PDSs can be leveraged to provide e-services in LDCs in the absence of any infrastructure. To this end, we introduce a new paradigm called Human-Powered Information System (HPIS), to provide e-services in LDCs. HPIS promotes the idea of a fully decentralized and participatory approach, where each individual implements a small subset of a complete information system. This paradigm builds upon the emergence of highly secure, portable and low-cost storage and computing smart tokens. A PDS will host the digital history of an individual, along with the capability to securely carry data between PDSs up to their destination or an Internet access point. Thanks to their PDSs, people will transparently perform data management and networking tasks as they physically move, so that global e-services are finally delivered by the crowd. HPIS has the potential to bring benefits to the individuals (e.g., communicate with friends) and to field workers (e.g., personal health/social folders and monitoring), as well as global benefits to the community (e.g., epidemiological studies and targeted information diffusion). We present the HPIS architecture and discuss the scientific challenges to be addressed before HPIS becomes reality.

C. *Trusted Cells*

Recently, AMD announced that it will incorporate a secure Trust Zone-based ARM processor [27] on its chips to be included into smart phones, set-top boxes and laptops. We argue that the advent of secure hardware embedded in all forms of personal devices, at the edges of the Internet, will trigger a sea change for Personal Data Services. Such secure tamper-resistant microcontrollers provide tangible security guarantees in the context of well-known environments. We can now imagine that whenever you take a picture, your smart phone securely contacts the personal services of all individuals in the frame of the picture, and automatically blurs the face of those who request it. We can also imagine that the GPS tracker in your son's car gives him detailed turn-by-turn guidance, but hides those details from your insurance company., only delivering aggregate road-pricing results.

We propose the Trusted Cells vision [12], i.e., PDS running on a large variety of secure devices to form a decentralized data platform. We illustrate how trusted cells can be used in the context of a wide range of application scenarios, describe the trusted cells architecture and discuss requirements and challenges for future research.

REFERENCES

[1] The World Economic Forum. Rethinking Personal Data: Strengthening Trust. May 2012.

[2] A. Pentland et al. Personal Data: The Emergence of a New Asset Class. World Economic Forum. January 2011.

[3] H. Nissenbaum, Privacy in context: Technology, policy, and the integrity of social life,"*Stanford Law Books*, 2010.

[4] S. Petronio, Unpacking the paradoxes of privacy in CMC relationships: The challenges of blogging and relational communication on the internet, *In Computer-mediated communication in Personal Relationships*, 2011.

[5] R. Agrawal, J. Kiernan, R. Srikant, Y. Xu: Hippocratic Databases. VLDB 2002: 143-154

[6] S. Bajaj, R. Sion: TrustedDB: a trusted hardware based database with privacy and data confidentiality. SIGMOD Conference 2011: 205-216

[7] FreedomBox: http://freedomboxfoundation.org/.

[8] T. Allard et al.: Secure Personal Data Servers: a Vision Paper. PVLDB 3(1): 25-35 (2010)

[9] A. Narayanan, V. Toubiana, S. Barocas, H. Nissenbaum, D. Boneh: A Critical Look at Decentralized Personal Data Architectures CoRR abs/1202.4503: (2012)

[10] Giesecke & Devrient, "Portable Security Token", http://www.gd-sfs.com/portable-security-token.

[11] Eurosmart. Smart USB token. White paper, Eurosmart, 2008, (10p).

[12] N. Anciaux, P. Bonnet, L. Bouganim, B. Nguyen, I. Sandu Popa, P. Pucheral. Trusted Cells: A Sea Change for Personnal Data Services, in "6th Biennal Conference on Innovative Database Research (CIDR)", Asilomar, États-Unis, 2013

[13] T. Allard, B. Nguyen, P. Pucheral: MetaP: Revisiting Privacy-Preserving Data Publishing using Secure Devices, in Distributed and Parallel Databases (DAPD), to appear

[14] L. Sweeney. k-anonymity: a model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems (IJUFKBS), 10(5):557-570, 2002

[15] C. Dwork. Differential privacy. In Proceeding of the 39th International Colloquium on Automata, Languages and Programming (ICALP), 2006

[16] A. C. Yao. Protocols for secure computations. In Proceedings of the 23rd Annual Symposium on Foundations of Computer Science (SFCS), 1982.

[17] O. Goldreich, S. Micali, and A. Wigderson. How to play any mental game. In Proceedings of the nineteenth annual ACM Symposium on Theory of computing (STOC), 1987.

[18] Q.-C. To, B. Nguyen, P. Pucheral, Secure global protocol for computing aggregate functions, in First Association of Vietnamese Scientists and Experts (AVSE) Doctoral Workshop, Cachan, 2012.

[19] C. C. Tan, B. Sheng, H. Wang, Q. Li: Microsearch: A search engine for embedded devices used in pervasive computing. ACM Trans. Embedded Comput. Syst. 9(4) (2010)

[20] 2011 Cost of a Data Breach Study : United States, Ponemon Institute LLC, march 2012.

[21] N. Anciaux, L. Bouganim, P. Pucheral, Y. Guo, L. Le Folgoc, S. Yin: MILo-DB: a Personal, Secure and Portable Database Machine. Distributed and Parallel Database Journal (DAPD), to appear, 2013.

[22] S. Yin, P. Pucheral, PBFilter: A flash-based indexing scheme for embedded systems, In Information Systems, Vol. 37(7), 2012.

[23] D. Agrawal, D. Ganesan, R. Sitaraman, Y. Diao, S. Singh, "Lazy-Adaptive Tree: An Optimized Index Structure for Flash Devices", PVLDB, 2(1), 361-372, VLDB Endowment, 2009.

[24] Y. Li, B. He, R. J. Yang, Q. Luo, K. Yi, "Tree Indexing on Solid State Drives", PVLDB, 3(1-2), 1195-1206, VLDB Endowment, 2010.

[25] C. Bolchini, F. Salice, F.A. Schreiber, L. Tanca, "Logical and Physical Design Issues for Smart Card Databases", ACM Transactions on Information Systems, 21(3), 254-285, 2003.

[26] Lim H., Fan B., Andersen D., Kaminsky M., "SILT: A Memory - Efficient, High-Performance Key-Value Store", SOSP, 2011.

[27] http://www.arm.com/products/processors/technologies/trustzone.php