Tutorial: Managing Personal Data with Strong Privacy Guarantees

Nicolas Anciaux[*,#]
[*]INRIA Paris-Rocquencourt
78153 Le Chesnay Cedex, France
(+33) 1 3963 5635
Nicolas.Anciaux@inria.fr

Benjamin Nguyen[*,#]
[#]PRiSM Lab., U. Versailles St-Quentin-en-Yvelines
45 Avenue des Etats-Unis, 78035 Versailles, France
(+33) 1 39 25 40 48
Benjamin.Nguyen@prism.uvsq.fr

Iulian Sandu Popa[*,#]
(+33) 1 39 25 40 85
Iulian.Sandu-Popa@prism.uvsq.fr

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*" [24], 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, send and receive megabytes of personal information to and from administrations or commercial services. However, the loss of privacy has severe consequences. The PRISM affair is unveiling a situation only reached in the worst dystopias of science fiction literature. Current practices are often not compliant with basic privacy laws and directives. Data leaks are legion [21]. 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 companies are now concerned by the potentially negative impact of an increasing exploitation of the data of their users. In contexts like smart cities, smart home and smart energy, the trend is to return the personal data to the *users* rather than to a *central server*, and to enable personal data services with a better form of usage control and user consent. Many laudable projects as well, which place respect for human dignity and privacy upfront, are left by the wayside. For example, social workers avoid building digital data services for discriminated people, because managing critical information on potentially discriminated people under weak privacy guarantees is seen as too strong a danger.

The nature of the solution is quite consensual: it is necessary to increase the control that individuals have over their personal data [18, 19, 20]. 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*" [24].

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 [2]. Even TrustedDB [9], 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 benefit/cost ratio is high for a centralized database.

User centric and decentralized solutions are promising because they do not exhibit these intrinsic limitations. This is a sea change for personal data management, where the control over personal data is pushed to the edges of the Internet, within sensors acquiring the data and in a variety of user devices endowed with a form of trust. The FreedomBox [12] was a pioneer attempt in this direction. Relying on low-cost plug computers and open software, FreedomBox enables anonymous and independent communication networks between users. The Personal Data Server (PDS) [3] project embeds a personal database in a tamper-resistant token on the user side, such that the PDS holders can grant and revoke privileges on views computed with the data, rather than exporting the raw data itself to a central server. Many other initiatives e.g., Project VRM, are currently investigating this approach [17], as well as major companies like Mozilla which supports the CozyCloud.cc solution based on a personal server user side and on open source software to manage personal data under the control of its owner.

This tutorial reviews several existing solutions going in this direction, presents a functional architecture encompassing these alternatives, and exposes the underlying techniques and open issues dealing with user centric and decentralized data management platforms. More precisely, the tutorial will be organized as follows. In a first part, 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 [3, 11, 12, 13, 17]. We discuss an abstract distributed architecture focusing on secure storing, managing and sharing of personal data, i.e., the asymmetric architecture. Then, we indicate the main challenges inherent to decentralized data management, with a focus on client-side data management and global query processing. In a second part, we explore data management techniques exercised within a trusted device at the client side. We review the main attempts proposed in the literature

10.5441/002/edbt.2014.71

and concentrate on those addressing the specific context of microcontrollers equipping e.g., sensors and mobile phones (SIM cards) [1, 7, 15, 16, 23, 26, 27, 29]. In a third part, we investigate the problem of performing global processing without any compromise on data privacy. We present the difficulties to overcome to execute privacy preserving computations on populations of personal devices, and illustrate it by focusing on Group By SQL queries and Privacy Preserving Data Publishing [4, 10, 14, 22, 25, 28]. In a fourth part, we conclude the tutorial by presenting existing and future instances of decentralized privacy preserving data management architectures [5, 6, 8]. We mainly focus on attempts and proposals targeting social-medical, smart houses, and rural areas contexts.[1]

ACKNOWLEDGMENTS

This work was partially supported by KISS ANR-11-INSE-005, INRIA CAPPRIS, and DMSP-CG78 grants.

REFERENCES

[1] Agrawal, D., Ganesan, D., Sitaraman, R., Diao, Y., Singh, S. 2009. Lazy-adaptive tree: An optimized index structure for flash devices. In *PVLDB* 2(1).

[2] Agrawal, R., Kiernan, J., Srikant, R., Xu, Y. 2002. Hippocratic Databases. In *VLDB*.

[3] Allard, T. et al. 2010.Secure Personal Data Servers: a Vision Paper. In *PVLDB* 3(1).

[4] Allard, T., Nguyen, B., Pucheral, P. 2013. METAP: revisiting Privacy-Preserving Data Publishing using secure devices. In *Distributed and Parallel Databases* (to appear).

[5] Anciaux, N., Bonnet, P., Bouganim, L., Nguyen, B., Sandu Popa, I., Pucheral, P. 2013. Trusted Cells: A Sea Change for Personal Data Services. In *CIDR*.

[6] Anciaux, N., Bouganim, B., Delot, T., Ilarri, S., Kloul, L., Mitton, N., Pucheral, P. 2014. Folk-IS: Opportunistic Data Services in Least Developed Countries. In *PVLDB* 7(1).

[7] Anciaux, N., Bouganim, L., Pucheral, P., Guo, Y., Le Folgoc, L., Yin, S. 2013. MILo-DB: a personal, secure and portable database machine. In *Distributed and Parallel Databases* (preprint).

[8] ARM TrustZone technology. http://www.arm.com/products/processors/technologies/trustzone.php

[9] Bajaj, S., Sion, R. 2011. TrustedDB: a trusted hardware based database with privacy and data confidentiality. In *SIGMOD*.

[10] Dwork, C. 2006. Differential privacy. In *Automata, languages and programming*.

[11] Eurosmart. 2008. Smart USB token. White paper.

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

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

[14] Goldreich, O., Micali, S., Wigderson, A. 1987. How to play any mental game. In *ACM Symposium on Theory of Computing*.

[15] Li, Y., He, B., Yang, R. J., Luo, Q., Yi, K. 2010. Tree indexing on solid state drives. In *PVLDB* 3(1-2).

[16] Lim, H., Fan, B., Andersen, D. G., Kaminsky, M. 2011. SILT: A memory-efficient, high-performance key-value store. In *ACM SOSP*.

[17] Narayanan, A., Toubiana, V., Barocas, S., Nissenbaum, H., & Boneh, D. 2012. A critical look at decentralized personal data architectures. In arXiv preprint arXiv:1202.4503.

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

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

[20] Petronio, S. 2011. 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*.

[21] Ponemon Institute LLC. 2013. Cost of a Data Breach Study.

[22] Sweeney, L. 2002. k-anonymity: A model for protecting privacy. In *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10(5).

[23] Tan, C. C., Sheng, B., Wang, H., Li, Q. 2010. Microsearch: A search engine for embedded devices used in pervasive computing. In *ACM Transactions on Embedded Computing Systems* 9(4).

[24] The World Economic Forum. 2012. *Rethinking Personal Data: Strengthening Trust*.

[25] To, Q.-C., Nguyen, B., Pucheral, P. 2014. Privacy-Preserving Query Execution using a Decentralized Architecture and Tamper Resistant Hardware. In *EDBT*.

[26] Tsiftes, N. and Dunkels, A. 2011. A database in every sensor. SenSys 2011: 316-332

[27] Wang, B. and Baras, J. S. 2013. HybridStore: An Efficient Data Management System for Hybrid Flash-Based Sensor Devices. EWSN 2013: 50-66

[28] Yao, A. C. C. 1982. Protocols for secure computations. In FOCS.

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

Nicolas Anciaux is a researcher at INRIA Paris-Rocquencourt, France. He received his Ph.D. from University of Versailles in 2004 and was in 2005 and 2006 a researcher at University of Twente, Netherlands. His main areas of interest are core database systems, embedded databases, database security and privacy.

Benjamin Nguyen is Associate Professor at University of Versailles St-Quentin (UVSQ), member of the CNRS PRiSM Lab and INRIA Secured and Mobile Information Systems (SMIS) team. He received his Ph.D. from University of Paris-XI in 2003, joined UVSQ in 2004 and INRIA-SMIS in 2010. His current research topics revolve around privacy protection in data centric applications, personal data over-exposure and anonymization.

Iulian Sandu Popa is Assistant Professor in Computer Science at the University of Versailles Saint-Quentin (UVSQ) and member of INRIA-SMIS since 2012. He received his Ph.D. in Computer Science from UVSQ in 2009. His main research interests are embedded database management systems, spatiotemporal databases, and mobile data management, with a particular interest in topics revolving around privacy and personal data management.

[1] An extended four pages version of this tutorial description can be found at: http://www-smis.inria.fr/edbt2014/