On Protecting Privacy in the Cloud

Mahmoud Barhamgi[1], Arosha K. Bandara[1], Yijun Yu[1], Khalid Belhajjame[2], Bashar Nuseibeh[1,3]

[1] *The Open University, Milton Keynes, UK*

[2] *Paris-Dauphine University, France*

[3] *Lero – The Irish Software Research Centre, Limerick, Ireland*

Cloud computing has now emerged as popular computing paradigm for data storage and computation for enterprises and individuals. Its major characteristics include the *pay-per-use* pricing model, where users pay only for the resources they consume with no upfront cost for hardware/software infrastructures, and the capability of providing *scalable and unlimited storage and computation resources* to meet changing business needs of enterprises with minimal management overhead [1]. The cloud, however, presents a major limitation to enterprises and individuals who move to public clouds: they lose control over the systems that manage their data and applications, leading to increased security and privacy concerns [2,3,4].

In this article, we examine cloud privacy concerns, and provide an overview of current and emerging solutions for protecting privacy of data and applications deployed in the cloud. Based on this, we suggest a set of recommendations for practitioners and researchers to improve privacy protection of cloud users.

What is privacy?

Roger Clarke (2006) [5] provides a broad definition of privacy spanning four dimensions:

- ***i.*** *Privacy of personal information*, sometimes referred to as "*data privacy*". This covers the right to control when, where, how, to whom, and to what extent an individual shares their personal information, as well as the right to access personal information given to others, to correct it, and to ensure it is safeguarded and disposed of appropriately;
- ***ii.*** *Privacy of personal behavior,* which covers the right of individuals to keep any knowledge of their activities from being shared with others;
- ***iii.*** *Privacy of personal communications,* which covers the right to communicate without undue surveillance, monitoring, or censorship;
- ***iv.*** *Privacy of the person*, which is concerned with the integrity of an individual's body. It covers such things as requiring the person's consent before applying medical treatments or taking samples of body fluids.

This article concerns the first dimension in a cloud computing setting; to ensure that the additional actors introduced by the use of the cloud, do not abuse cloud users' data.

Cloud Privacy Concerns

Understanding privacy concerns in a cloud computing model requires understanding the key stakeholders involved in its data lifecycle:

- ***Data owners*** are the entities whose data is stored in the cloud. For examples, in a critical application domain such as healthcare, they include patients when healthcare providers outsource the storage of their medical databases to the cloud, but also the physicians and the hospitals whose medical and financial practices can be inferred by analyzing the outsourced data. The main concern of data owners is to protect their data and identities against unauthorized access or use;
- ***Data consumers*** are the persons who query the data for various reasons, such as physicians who consult the medical data of patients for treatments, researchers who query the patients' data to determine the side effects of a given medicine. Data consumers may also have privacy concerns, for example, a researcher who is working

on a new scientific invention may require his or her identity and queries (about his or her research) to be protected; and

- ***Service or cloud providers*** include all IT staff required to run and manage cloud services, including databases, servers, networks and applications software.

We illustrate two key privacy concerns for data owners and consumers in the cloud using examples from the healthcare domain.

Accidental or deliberate data disclosure: One source of privacy concerns is the cloud's administrators, who may be external entities for both data owners and consumers. They may accidentally or deliberately disclose the data, with unwelcome consequences. For example, in the healthcare domain, sensitive information such as patients' illnesses, unsuccessful medical interventions of healthcare professionals, and ongoing inventions of medical researchers can be identified and revealed with irreversible damage to patients, physicians and researchers. This information could be abused by different bodies, including employers, insurance companies and competitors. Indeed, the mere existence of such data may make some cloud administrators vulnerable to corruption or blackmail.

Of course traditional, non-cloud information systems have also been vulnerable to such privacy concerns, such as healthcare provider who store their medical data on their premises. However, the privacy risks and effects are significantly amplified by using the cloud: when a cloud managing the medical data of several healthcare providers is compromised, it is the entire medical history of a patient, aggregated across multiple healthcare providers, that is at stake, not only one single healthcare episode at a specific healthcare provider. In many ways, data is the currency of the 21st century and cloud-based data stores are the bank vaults, making them an increasingly preferred target for both malicious insiders and external attackers due to the collective value concentrated in one logical location.

Beneficial or harmful data mining: Medical histories of patients, aggregated across multiple healthcare providers that use the same cloud, may be mined to infer new knowledge that is beneficial to society as whole and to individual patients. For example, the predisposition of a certain category of people to a particular disease could be determined by analyzing common features of those suffering form that disease; healthcare professionals with illegal or criminal practices (e.g., physicians undertaking unusually high numbers of abortions, serial-killing nurses, etc.) can be identified, and so on. These medical histories may also be analyzed for malicious purposes, such as deliberate undermining of a physician's reputation by aggregating the number of his/her unsuccessful medical interventions, determining cheap and expensive healthcare providers for advertisement purposes, etc. An important privacy management aim, therefore, is to prevent cloud insiders from performing such harmful actions.

Current practices for privacy protection in the cloud

Adherence to agreed privacy policies is a de facto norm for addressing the above privacy concerns. Typically, an organization that outsources all or part of its information system to an external cloud signs a service level agreement with the cloud provider. The agreement defines, among other things, a privacy policy prescribing where and how the organization's data is stored, processed and used (i.e. accepted and prohibited uses) by the cloud service provider. It may also define some privacy related measures and technical controls to be applied on the cloud side, such as the vetting of employees, breach notification, isolation of tenant applications, and the use of products certified to meet national or international standards. However, with the lack of physical control by cloud users over data storage, and the absence of standardized and mature techniques for monitoring how data is accessed, processed and used inside the cloud, it is harder to verify a cloud's compliance with such privacy policies.

Emerging Solutions

We examine below some emerging approaches for enhancing the privacy of data owners and consumers in the cloud. We classify them, according to the kind of techniques they employ, in Table-1 into four categories: *Encryption*, *Trusted Computing*, *Private Information Retrieval (PIR)*, and *Intention hiding*.

Encryption. Encryption is a viable technique for protecting sensitive data from malicious cloud insiders. However, it also makes it difficult for the cloud to process queries on data on behalf of users. This limitation is being addressed by emerging approaches:

- ***Homomorphic Encryption***: Fully Homomorphic Encryption (FHE) [7] is a promising technique that enables cloud servers to perform computations on encrypted data, without decrypting it. Although this is still

prohibitively expensive to be applied to real-world applications, this is an area of ongoing research and development.

- ***Partial Homomorphic Encryption***: Also known as *Somewhat Homomorphic Encryption (SHE)* [8], it allows the cloud to perform only a limited number of operations on encrypted data, leading to improved performance. Even though SHE schemes are less powerful than FHE schemes, they can already be used in many real-world applications in the medical, financial, and advertising domains.
- ***Efficient Query-specific Encryption***: This kind of encryption techniques [6,9] are designed to allow the cloud to efficiently execute specific classes of queries such as *keyword search queries*, *range queries*, and *aggregation queries* without decrypting the data. For example, for answering (multi-dimensional) range queries, a combination of encryption and data partitioning techniques can be used to organize data elements into 'buckets' that can be stored safely in a public cloud [9]. The solution requires clients (i.e. data owners) to participate in the query processing by selecting relevant buckets (to be retrieved from the cloud) and filtering away false positives. Similarly, a technique for keyword queries proposed by Cao, et al [6] allows the cloud to rank a set of encrypted documents based on how well they match a set of encrypted keywords.

Trusted Computing: An interesting alternative to expensive encryption is to store and process plaintext data inside secured-hardware containers deployed in untrusted clouds. Computation inside such trusted hardware is orders of magnitude cheaper than any equivalent cryptography performed on untrusted clouds, despite the overall greater acquisition cost of secure hardware. Cipherbase [12] is an example of a database system that relies on secure cryptographic co-processors and FPGA boards to process privacy sensitive data. The user can tune the hardware to provide different privacy/performance trade-offs. Its performance is practical for a wide range of applications, for example, when all data is strongly encrypted, the performance drops by only one order of magnitude compared to when data is processed by a conventional database system operating on plaintext data.

Efficient Private Information Retrieval (PIR): The idea behind PIR techniques is to execute private queries on a remote server without letting the server learn anything about executed queries or their results. They can therefore be used to address the privacy concerns of data consumers. Although original PIR solutions were too computationally expensive to be practical, their efficiency can be substantially improved by applying their expensive cryptographic operations only on a subset of the data elements that contain queries' answers instead of the entire database [10].

Intention hiding techniques: Another new trend for protecting the privacy of data consumers is to hide the intention behind their queries by changing the query plans [14]. Such approaches are motivated by the observation that different (but equivalent) plans for the same SQL query may reveal vastly different information about the user's intention. Therefore, they exploit query optimization techniques to accommodate the privacy constraints and preferences of users and produce privacy-aware query execution plans.

Table 1: Emerging solutions for privacy protection in the cloud

<table>
<tr><th colspan="2">Approach</th><th>Supported Queries</th><th>Privacy Strength</th><th>Limitations</th></tr>
<tr><td rowspan="4">Encryption [6,7,8,9]</td><td>Homomorphic Encryption [7]</td><td>All types of queries</td><td>Strong</td><td>− Impractical for real-life applications due to prohibitive computation cost.</td></tr>
<tr><td>Partial Homomorphic Encryption [8]</td><td>Keyword search queries</td><td>Strong</td><td>− Not all queries are supported,
− Only practical for applications with moderate dataset sizes.</td></tr>
<tr><td rowspan="2">Query-specific Encryption [6,9]</td><td>Keyword search queries [6]</td><td>Strong</td><td>− Substantial computation overhead for data owners,
− Not all queries are supported.</td></tr>
<tr><td>Range queries [9]</td><td>Practical,
Tuneable to achieve a privacy/efficiency trade-off.</td><td>− Part of query computation overhead is shifted to data owners,
− The cloud's computation power is under used,
− Support to range queries only.</td></tr>
<tr><td colspan="2" rowspan="2">Trusted computing [12,13]</td><td>All types of queries [12]</td><td rowspan="2">Practical,
Tuneable to achieve a privacy/efficiency trade-off.</td><td rowspan="2">− Expensive secure hardware,
− Requires secret key handover from user to trusted hardware,
− Not all queries are supported (in [13])</td></tr>
<tr><td>Range and keyword queries [13]</td></tr>
</table>

Efficient PIR [10]	Range and join queries	Practical, Tuneable to achieve a privacy/efficiency trade-off.	- Part of query computation overhead is shifted to data owners,
Intention hiding techniques [14]	Join queries	Good for hybrid clouds	- Users are required to have a good knowledge of the servers involved in processing their queries.

A Call to Action

Table-1 demonstrates that there are many solutions with different trade-offs in terms of privacy protection, performance, computation overheads to users, and the range of supported queries. Given this rich landscape of solutions, we suggest the following sets of actions to improve the privacy protection in the cloud, aimed at practitioners and researchers.

For practitioners

- *Match requirements to solutions*. Cloud application developers should select the solution that most closely satisfies the requirements of their applications. For example, for document-oriented applications, such as email management and medical records privacy protection, encryption techniques for keyword search queries would be more secure than policy-based solutions, more efficient than full encryption techniques, and less expensive than trusted computing techniques. On the other hand, Online Transaction Processing (OLTP) applications, such as banking, require high performance and throughputs, so trusted computing solutions would be more appropriate;
- *Audit and Verify*. Cloud and service providers should offer data owners and consumers verifiable auditing mechanisms that would discourage malicious cloud insiders from data abuse. Cloud users should continually analyze the security and privacy controls of a cloud provider and verify that their security and privacy requirements are met.

For researchers

- *In the short term*, efforts are needed to improve the visibility of users inside cloud computing. Solutions are needed to allow cloud users to monitor, audit and control, with minimal overhead, their data flows, as well as to measure how well a cloud provider adheres to its stated privacy policies.
- *In the long term*, with data encryption research making promising progress, efficiency of query computation over encrypted data should be improved to make it practical for real OLTP applications.

Conclusion

Although privacy is regarded as a significant hurdle for wide adoption of cloud computing, it can also become one of its key selling features if addressed properly. Users involvement in protecting their privacy, by enabling them to control and verify how their data is stored, accessed and exploited, is essential for even greater adoption.

References

[1] European Commission Survey, *"Cloud computing - Statistics on the use by enterprises*", November 2014. http://ec.europa.eu/eurostat/statistics-explained/index.php/Cloud_computing_-_statistics_on_the_use_by_enterprises.

[2] International Telecommunication Union ITU, "*Privacy in Cloud Computing*", Technology Watch Report March 2012. http://www.itu.int/dms_pub/itu-t/oth/23/01/T23010000160001PDFE.pdf (accessed on October 2015).

[3] ENISA European Network and Information Security Agency, "*An SME perspective on cloud computing"* (survey), 2009. https://www.enisa.europa.eu/act/rm/files/deliverables/cloud-computing-sme-survey.

[4] Mark Dermot Ryan: Cloud computing privacy concerns on our doorstep. Communications of ACM 54(1): 36-38 (2011).

[5] Roger Clarke, "What's Privacy?" 2006. Available at: http://www.rogerclarke.com/DV/Privacy.html (accessed on October 2015).

[6] N. Cao, C. Wang, M. Li, K. Ren, W. Lou, "*Privacy-Preserving Multi-Keyword Ranked Search over Encrypted Cloud Data*", IEEE Transactions on Parallel & Distributed Systems, vol.25, no. 1, pp. 222-233, Jan. 2014.

[7] C. Gentry, A. Sahai, B. Waters: Homomorphic Encryption from Learning with Errors: Conceptually-Simpler, Asymptotically-Faster, Attribute-Based. CRYPTO, 75-92, 2013.

[8] S.Q. Ren, B. Hong, M. Tany, "*Homomorphic Exclusive-Or Operation Enhance Secure Searching on Cloud Storage*". CloudCom 2014:989-994.

[9] B. Hore, S. Mehrotra, M. Canim, M. Kantarcioglu: "*Secure multidimensional range queries over outsourced data"*. VLDB J. 21(3): 333-358 (2012).

[10] S. Wang, D. Agrawal, A. El-Abbadi: "*Towards practical private processing of database queries over public data"*. Distributed and Parallel Databases 32(1): 65-89 (2014).

[11] S. Brezetz, G.B. Kamga, A. Guesmi: "*End-to-end privacy policy enforcement in cloud infrastructure"*. CLOUDNET 2013:25-32.

[12] A. Arasu, K. Eguro, M. Joglekar, R. Kaushik, D. Kossmann: "*Transaction processing on confidential data using cipherbase*". ICDE 2015: 435-446.

[13] S. Bajaj, R. Sion: TrustedDB: "*A Trusted Hardware-Based Database with Privacy and Data Confidentiality"*. IEEE Trans. Knowl. Data Eng. 26(3): 752-765 (2014).

[14] N. Farnan, Ting Yu: PAQO: "*Preference-aware query optimization for decentralized database systems"*. ICDE 2014: 424-435.

Acknowledgements. This work is supported, in part, by SFI grant 13/RC/2094, ERC Advanced Grant 291652 (ASAP), and QNRF grant NPRP 05-079-1-018.

Mahmoud Barhamgi *is an associate professor at Claude Bernard University (France) and a researcher at the Open University, UK, whose research focuses on Privacy preservation in SOA, Web and Cloud environments. Contact him at mahmoud.barhamgi@open.ac.uk.*

Arosha K. Bandara *is a Senior Lecturer at the Open University, UK, whose research focuses on engineering adaptive security and privacy mechanisms in ubiquitous computing systems. He also leads the OU's Introduction to Cyber Security MOOC. Contact him at a.k.bandara@open.ac.uk.*

Yijun Yu *is a senior lecturer at the Open University, UK, with research interests in engineering automated software tools to solve fundamental and practical problems in the research areas of quality requirements in general, and security and privacy in particular. Contact him at y.yu@open.ac.uk.*

Khalid Belhajjame *is an associate professor at the University Paris-Dauphine. His research interests lie in the areas of information and knowledge management, where he has published over 60 papers. You can contact him at Khalid.Belhajjame@dauphine.fr.*

Bashar Nuseibeh *is a professor of computing at the Open University, UK and professor of software engineering at Lero, Ireland. His research interests are in software engineering, security and privacy, and adaptive systems. He received a PhD in software engineering from Imperial College London, and holds a Royal Society-Wolfson Merit Award. Contact him at b.nuseibeh@open.ac.uk.*