Data Leakage Detection and Privacy

Bharat Bhargava , Professor in Computer Science Department, Purdue University

Project Description: Business-to-Business systems use service-oriented architecture (SOA) with decomposed business processes. Those processes (or services) can interact and share data among each other, including services from untrusted environments. Databases, associated with services, can be hosted by an untrusted cloud provider. Cloud platforms are vulnerable to large attack surface that could violate privacy of stored data shared with web services. Data owner needs to be sure that each service can access only those fragments of a database for which the service is authorized. Data privacy can be threatened by accidental data diffusion or intentional malicious data disclosures, including multiple collusive attacks on the network. Data leakages made by authorized insiders to unauthorized services need to be detected. In addition, encrypted search over encrypted database needs to be supported. PhD student, Denis Ulybyshev, will work on his thesis based on these ideas.

A cloud enterprise framework that ensures privacy – preserving data dissemination, based on role-based access control and on cryptographic capabilities of client's browser, on authentication method and on subject's trust level has been implemented. We are collaborating with Rohit Ranchal (our former Ph.D student), IBM. We have also discussed this research with Gang Ding (our former PhD student), Qualcomm. “WaxedPrune” (Web – based Access to Encrypted Data Processing in Untrusted Environments) prototype was implemented in collaboration with NGC and W3C / MIT. We collaborated with Donald Steiner, NGC. The prototype provides privacy – preserving dissemination of Electronic Health Records (EHRs) hosted by untrusted cloud provider. Modular software architecture and design incorporate database and information security principles. The prototype was demonstrated at Northrop Grumman company “Tech Fest 2015” and “Tech Fest 2016” exhibitions. Prototype demo video is available [2]. In 2015, we received CERIAS best poster (1 out of 43) award [5].

In order to ensure privacy – preserving data dissemination in untrusted clouds the framework relies on using an Active Bundle (AB). AB is a self-protecting structure that consists of key-value pairs in encrypted form, access control policies and policy enforcement engine (Virtual Machine). Each subset of data is encrypted with its own symmetric key. One unique feature of the system is that symmetric key is generated on-the-fly based on execution

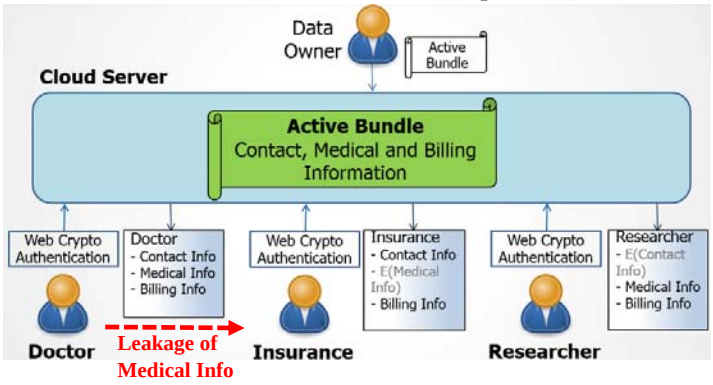

Fig.1. EHR Dissemination in Cloud (by Dr. Leon Li, NGC)

flow, depending on subject's role (e.g. doctor, insurance agent, researcher), set of access control policies, and AB modules and their resources [1]. Thus, key is not stored inside AB or on TTP. Service requesting data from AB needs to authenticate itself. Then symmetric decryption keys are generated to decrypt those data items for which the authenticated service is authorized, based on access control policies, stored in AB. Another feature is that, in addition to access control policies, data dissemination depends on context, on service attributes, such as trust level, level of cryptographic capabilities of client’s browser; authentication method (password–based vs. fingerprint. Data owner’s availability is not required. This approach, compared to Attribute-Based Encryption (ABE), has several advantages: a) It does not rely on Trusted Third Party (TTP) to issue keys for the recipient services and b) It supports complex policies that can be written in Java (AB implementation language), whereas ABE policies are expressed as boolean and threshold operations over a set of attributes

Research Tasks: *A. Data leakage* problem arises when authorized services share data behind the scene with an unauthorized service. As shown in Fig.1, Doctor, who has authorized for medical data, can by mistake leak it to Insurance service, who is not authorized for it. It is assumed that data leakages can occur in two forms:

A1. *The whole record (AB) having data in encrypted form got leaked.* In fig. 2, AB contains encrypted data **Enc[Data(D)] = {Enc$_{k1}$ (d$_1$), ... , Enc$_{kn}$ (d$_n$) }** and Access Control Policies **(P) = {p$_1$,.., p$_k$}**. If AB is sent by X to unauthorized service Y, Y would not be able to decrypt d$_1$ since AB data can only be extracted after authentication is passed and access control policies are evaluated. When Y tries to decrypt d$_1$, AB kernel will query CM to check whether Y is allowed to read d$_1$. In addition, digital watermarks [3], [4] embedded into data can be verified by web crawlers. If attacker uploads illegal content to publicly available web

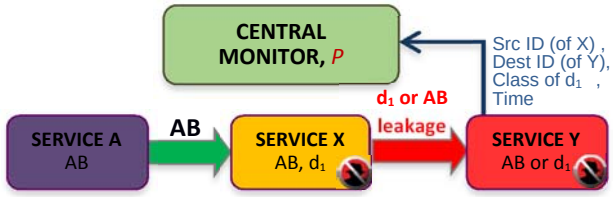

Fig.2. Data leakage detection by Central Monitor

hosting data can be scanned by web crawlers, then web crawler can verify the watermark and can detect copyright violation.

A2. *Plaintext got leaked.* Service X is authorized to read d$_1$ from AB and it may leak decrypted d$_1$ by taking a picture of a screen with d$_1$ and sending it to Y, who is not authorized to read d$_1$ via email. In this case, solution can rely on visual watermarks embedded into data. Provenance data [3], [4] is recorded to help investigating data leakage. Additionally, only part of data can be given to authorized service at first. Then service's trust level is being constantly monitored and re-computed by CM and if trust level goes down then the rest of the data will not be given to that service. CM detects anomalies in the system, based on service's characteristics such as CPU and memory usage, amount of uploaded / downloaded data, etc.

After leakage is detected, compromised role (e.g. Doctor) can be separated into two: *suspicious_role* and *benign_role*. New certificates will be sent to all benign users for benign role; new AB with new policies, restricting access to *suspicious_role* (e.g. to all doctors from the same hospital with a malicious one) will be created. Also, sensitivity level for leaked data items can be increased to prevent leakage repetitions. The idea is to provide anti – fragility and make system stronger against similar attacks. Leakage damage will be assessed.

B. In order to support *encrypted search* [7] over database of ABs, hosted by untrusted cloud, AB can contain extra – attribute, i.e. short abstract or keywords, used for indexing and searching. User's search query is stemmed, stopwords are removed and then it is converted to SQL query. Then SQL query needs to be converted into encrypted form. User-defined functions can be used. *CryptDB* [6] can be employed to support encrypted search queries over encrypted data. SWP crypto system is used to support SQL "LIKE" operator.

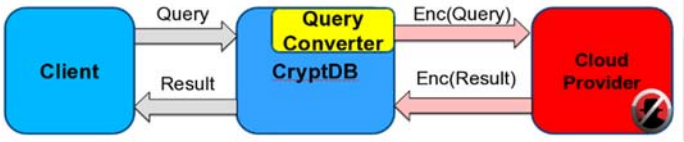

Fig.3. Encrypted search over encrypted data

Query example: **select prescription from EHR_DB where diagnosis LIKE '%insomnia%';**
Converted Query: **select C1 from T1 where ESRCH (Enc(diagnosis), Enc(insomnia)).**

Budget: We request an half time RA position and $2000 travel funds for visiting corporate partners and conferences.

References

[1] R. Ranchal, "Cross-domain data dissemination and policy enforcement," PhD Thesis, Purdue University, 2015

[2] D. Ulybyshev, B.Bhargava, "Secure dissemination of EHR," demo video https://www.dropbox.com/s/30scw1srqsmyq6d/BhargavaTeam_DemoVideo_Spring16.wmv?dl=0 , accessed: Mar.2017

[3] B. Bhargava, "Secure/resilient systems and data dissemination/provenance," NGCRC Project Proposal, CERIAS, Purdue University, Aug.2016

[4] Y.L. Simmhan, B. Plale and D.A. Gannon, "A survey of data provenance in e-science," SIGMOD Rec., 34(3):31–36, 2005.

[5] R. Ranchal, D. Ulybyshev, P. Angin, and B. Bhargava. "PD3: Policy-based Distributed Data Dissemination," 16-th CERIAS Security Symposium, Apr. 2015 **(Best poster award)**

[6] R. A. Popa, C. M. S. Redfield, N. Zeldovich, and H. Balakrishnan. "CryptDB: Protecting confidentiality with encrypted query processing". In ACM SOSP, 2011

[7] A. Arasu, S. Blanas, K. Eguro, R. Kaushik, D. Kossmann, R. Ramamurthy, and R. Venkatesan. "Orthogonal security with Cipherbase". In Proceedings of the 6th Conference on Innovative Data Systems Research, 2013