2010 Sixth International Conference on Intelligent Information Hiding and Multimedia Signal Processing

Privacy by Data Provenance with Digital Watermarking

A Proof-of-Concept Implementation for Medical Services with Electronic Health Records

Jérémie Tharaud[1], Sven Wohlgemuth[1], Isao Echizen[1], Noboru Sonehara[1], Günter Müller[2] and Pascal Lafourcade[3]

[1] National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, JAPAN
[2] University of Freiburg
Friedrichstr. 50, 79098 Freiburg i.Br., GERMANY
[3] Unviersity Joseph Fourier
2, avenue de Vignate, 38610 Gières, FRANCE
{tharaud, wohlgemuth, iechizen, sonehara}@nii.ac.jp
mueller@iig.uni-freiburg.de
pascal.lafourcade@imag.fr

***Abstract*—Security is one of the biggest concerns about Cloud Computing. Most issues are related to security problems faced by cloud providers, who have to ensure that their infrastructure is properly secure and client data are protected, and by the customers, who must ensure proper security measures have been taken by the provider in order to protect their personal data. When you move your information into the cloud, you lose control of it. The cloud gives you access to your data, but you have no way of ensuring no one else has access to these data. In this article, we propose an evaluation of a proof-of-concept implementation of a usage control system for an ex post enforcement of privacy rules regarding the disclosure of personal data to third parties. The system is based on cryptographic protocols and digital watermarking in medical services and electronic health records.**

***Keywords*—Cloud computing, security, cryptography, multimedia, watermarking.**

I. INTRODUCTION

Unlike the client-server model, cloud computing is a computing paradigm where massively scalable and often virtualized IT-related capabilities (resources, software, information, etc.) are provided on-demand as a service using Internet technologies to multiple external customers. Typical cloud computing providers deliver common business applications online that are accessed from another web service or software like a web browser, while the software and data are stored on the servers of the Cloud's service providers. However, the cloud appears to be a black box for the user. Disclosure of users' data to a cloud and, at the same time, the data federation at software service providers of the cloud facilitate the secondary use of personal data and stored digital content on a massively shared scale infrastructure for data analysis by third parties. Some of them, for instance advertising agencies aim to collect personal data in large databases to better target users and sell their services [1, 2]. Also, data could be disclosed to service providers who have their IT infrastructure in a country without any data protection legislation. Thus, these disclosures of personal data to third parties raise obvious privacy issues, since there is no usage control *a posteriori* of the disclosures of personal data. In practice, service providers publish their privacy policies as a part of their general terms and conditions. Users have to accept them and thereby give the service providers full authority to process their personal data. For instance, a provider of an electronic health record (EHR) datacenter collects health data from their users (patients) to share them among clinics, health insurance agencies, and pharmaceutical companies. These systems comply with the US American Health Insurance Portability and Accountability Act (HIPAA) [3] by letting users decide on the usage and disclosure of their medical data, e.g. x-ray images. However, they do not offer mechanisms to enforce the privacy policy rules.

In this article, we present a proof-of-concept implementation of our proposed usage control system for ex post enforcement of privacy policy rules regarding the disclosure of personal data to third parties. Section 2 introduces the concept of using data provenance (information to determine the derivation history) [4] for usage control, e.g. the characteristics of our preservation system called DETECTIVE [5] and its application to a concrete case study. Then section 3 describes the current state of implementation and presents the obtained results. Finally, section 4 gives an overview of future work.

II. SCENARIO

A. "Telemedecine" Case Study

The main application of the proposed DETECTIVE system is its use in medical services and electronic health records. We use data provenance to control the disclosures of personal data. The process consists of using copyright protection by labelling digital content (*Tag* protocol). Let us examine the following situation where a patient needs health assistance. When the patient goes to the clinic, the doctor takes an x-ray image. The image will be watermarked [6] with information related to the identity of the *data provider* (clinic), *data consumer* (data center), and the user. Then it is

978-0-7695-4222-5/10 $26.00 © 2010 IEEE
DOI 10.1109/IIHMSP.2010.130

IEEE computer society

disclosed it to a data center, provided that the patient has agreed to share personal data. To perform this, the patient uses his personal electronic health card. Thus, assuming the patient needs medical treatment abroad, a foreign clinic will retrieve a copy of the watermarked image of the patient from the data center. Also, this image will have additional information embedded: the new *data consumer* (the identity of the clinic abroad) and the new *data provider* (the data center). The sequence of *tags* for the same personal data constitutes a disclosure chain.

B. Privacy Risks

Privacy risks appear when additional disclosures of personal data are not permitted, e.g. when the DETECTIVE policy is not respected correctly. The three following attack scenarios can happen:

- *Unchanged data provenance*: data are sent from a provider to a consumer without changing the digital watermark
- *Modification of data provenance*: invalid values are tagged on the digital watermark
- *Removal digital watermark*: the watermark is removed.

Thus, one of the aims of the DETECTIVE system is to identify data providers who have violated the policy.

C. Controllable Disclosure of Data by DETECTIVE

Usage control by data provenance enables an ex post enforcement of obligations by identifying the last data provider. To proceed, the DETECTIVE system links the identities of data provider and consumer to data disclosure by cryptographic commitments and digital watermarking (*Tag* protocol). In the case of an audit (of a service provider), data provenance can be used to restore the information flow of personal data (*Verify* protocol), the verification being due to delegated rights as a private key called *watermarking key*.

III. DETECTIVE SYSTEM

A. Data Provenance by Digital Watermarking

We assume that the communicating parties (e.g. data consumers and providers) are authenticated within a public key infrastructure (PKI) and that users (patients) authenticate via a smartcard (electronic health card). Identities and access rights of all participants are certified by a certification authority (CA), which issues credentials. The main modules of the DETECTIVE system are the electronic health card, a terminal to read it, and the service provider's system. The terminal for the electronic health card consists of an authentication module for delegating the user's access rights to personal data to service providers, a DETECTIVE Policy Module for configuring privacy policies, a DETECTIVE Auditor Module for extracting and checking data provenance data of found personal data, and a cryptographic library as well as a digital watermarking library to detect and extract digital watermarks. The service providers' system differs from the user's system insofar as it has the DETECTIVE Signaling Module, a database for maintaining users' personal data (EHR), and a secure storage for the provider's secret/private cryptographic keys. The DETECTIVE Signaling Module embeds the data provenance information in the users' personal data to be disclosed. Regarding disclosure of personal data from one service provider to another, the data provider embeds data provenance information into the user's personal data to be disclosed. Afterwards, the tagged data is disclosed to the service provider acting as the data consumer. If subsequent disclosures are allowed by the privacy policy, every service provider in the disclosure chain will execute these steps with the successive data consumer. Regarding checking the user's data provenance, the user system has found the data or has it as a result of an audit. Afterwards, it starts the compliance check of the data's disclosures with the agreed-upon obligations on the basis of embedded data provenance information. After extracting all digital watermarks of the personal data under investigation, the DETECTIVE Auditor Module requests and checks the service providers' input to the data provenance information. The result of this check is either that "no violation has occurred" or "service provider XXX has violated the privacy policy".

B. Use Cases

We can distinguish four types of users for the system in accordance with the previous part: *patients* and *doctors* in medical services, *administrators* of the EHR, and *auditors* for auditing the system. The use cases of the DETECTIVE system for the *patient* are the authentication at the clinic in order to be treated by a doctor and to delegate his access rights to personal data. *Doctors* will treat the patient (for instance, take an x-ray picture), they can upload personal data to the EHR, and retrieve others in accordance with the principle of delegation rights. *Administrators* carry out the maintenance of the EHR and clinics' information systems. *Auditors* run the *Verify* protocol. They check that personal data have been disclosed in accordance with the policy. To perform this, they have the right to retrieve any watermarked data (string commitments) of service providers and consumers so that they reconstruct the disclosure chain. Since the system will be used by several users, it is advisable to develop a graphical user interface (GUI), easy-to-use, fast, and easily maintainable. Furthermore, most information systems of medical services use Microsoft's operating systems. As a consequence, we have decided to develop a Windows application for the system in Visual Basic .NET [8].

C. *Proof-of-Concept Implementation*

This subsection describes our current work on implementing the DETECTIVE system and the obtained results.

1. *Implementation Model*

Starting-up the Graphical User Interface of the DETECTIVE system leads to a login screen where the user of the system has to enter his/her name and password. In the domestic clinic, patients are also authenticated with an electronic health card. After this step, the medical treatment can begin. The doctor will take an x-ray image of the patient and send it to the data center. Figure 1 shows the results of data provenance embedding between the domestic clinic and the EHR.

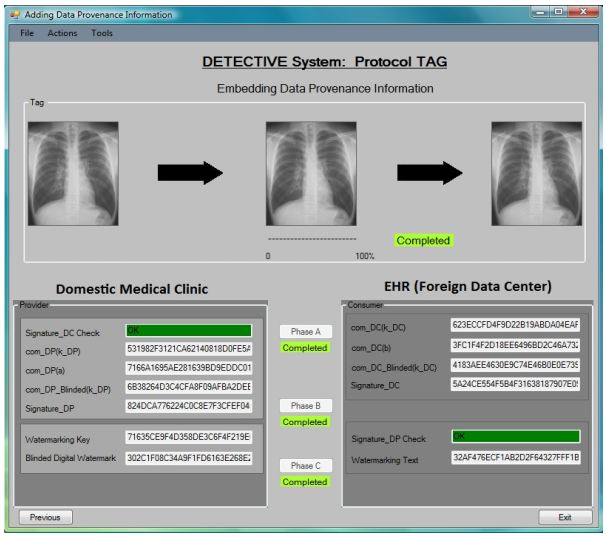

Figure 1. Embedding data provenance (local clinic and EHR)

This figure shows the different data exchanges between clinic and EHR. A final version of the system must obviously not show the cryptographic commitments of each part. The doctor can also retrieve a previously saved image from the data center. In this case, before the new disclosure, another watermark will be embedded into the image.

2. *Cryptographic Library for String Commitments*

The *tag* protocol of the DETECTIVE system is based on the Diffie-Hellman key exchange protocol and string commitments. During this phase, the watermark (the *tag*) is created by the service provider and embeds data on identities of the data consumer, provider, and user (obtained by string commitments between the two parties). To perform arithmetic operations (e.g. modular multiplication, random number generation, etc.) between huge integers (public/private keys) and strings, we cannot use the standard VB.NET cryptographic toolkit because it does not provide such methods (only common cryptographic protocols are available). For a first demonstration, we used a specific cryptographic toolkit called the Huge Integer Math and Encryption library (HIME), a standard Win32 direct link library which is compatible with the VB.NET framework, developed by the DevoTechs team [7].

3. *Digital Watermarking Library*

The aim of this part was to develop a graphical library that inserts (encrypted) contents on JPEG pictures, in particular a digital watermarking function which depends on a selectable key (the *watermarking key*). The watermark should be as follows:

- *Invisible*: since the doctor must be able to read the watermarked x-ray picture, even after multiple disclosures, and only those who know the watermarking key can retrieve the watermark
- *Robust*: the watermark must be difficult to remove from the picture without knowing the watermarking key. Thus, it must be resistant to geometrical attacks that do not prevent the reading but degrade the watermark (cropping, scaling, resampling, conversion, etc.).

The watermarked image is subjected to several operations. First, at each disclosure, another watermark is added, which must not affect the other inserted watermarks.

It also depends on a new watermarking key used to extract it. Second, for auditing the system, watermarks have to be easily extractable knowing the watermarking key, in order to reconstruct the disclosure chain. Lastly, since it can be extractable, it can also be replaced by a new one if the protocol asks for it.

4. *Feasibility Evaluation*

The evaluation aims to determine whether the DETECTIVE proposal is useful, applicable to its specific situation, and meets users' needs, especially if it fulfils the security needs. First, data protection implies secure storage to prevent leaks, especially in the EHR (integrity). Thus secondly, accessing records must be restricted to authorised users (authentication). In the DETECTIVE system, we can distinguish four types of users: 'patients' and 'doctors' in medical services, 'administrators' of the EHR, and 'auditors' for auditing the system. The DETECTIVE identity management system gives access to information and computing resources by using a public-key infrastructure (PKI) for anonymous credential (e.g. electronic health smartcard and Single Sign-On technology for the user). Third, since users do not take part in disclosing personal data, users give their agreement in advance (non-repudiation). This is made by the delegation protocol DREISAM [9]. Fourth, the sensibility of personal data imposes 192-bit cryptographic keys for symmetric cryptography operations and communications between medical services and EHR (integrity), and 1024-bit keys for asymmetric (NIST recommendation [10]). These properties can be checked by the auditor in the DETECTIVE system. Three types of

attacks (II.B) on the protocol can be generated; at each time the disclosure chain is reconstructed and displayed. The auditor is also a way of showing that the same number of digital watermarks as the number of disclosures of a disclosure chain can remain detectable and extractable for the user as well as keep the x-ray image usable for medical services. For instance, figure 2 shows the results of an execution of the *Verify* protocol by the *auditor* after an illegal disclosure of personal data. Here, the attack consisted of the foreign clinic sending personal data to a pharmaceutical company without changing the watermark (*unchanged data provenance* attack). The results of auditing an image found in the pharmaceutical company reveal that data consumer and provider commitments and identities are the same in the last two disclosures (disclosure 3 and disclosure 2), which prove that the foreign clinic had violated the protocol. This is the way the DETECTIVE system protects against such attacks.

IV. OUTLOOK

Security in small businesses or public institutions is a major concern for any economic policy. In particular, the emergence of cloud computing creates even more security threats. Understanding and studying these new threats requires new constructions to maintain and improve security. Thereby, our study targeting an application of cloud computing in medical institutions and electronic health records is entirely justified. Our future work will focus on evaluating the DETECTIVE system implementation. The feasibility evaluation will investigate attacks on the digital watermarking algorithm used in the DETECTIVE system and show if the same number of digital watermarks as the number of disclosures of a disclosure chain can be embedded into an x-ray image while keeping the digital watermarks detectable and extractable for an auditor as well as keeping the x-ray image usable for the medical institutions. Lastly, establishing a security model for cloud computing also forms a part of our future work.

ACKNOWLEDGMENT

This work has been funded by the German Academic Exchange Service (DAAD) and is a result of the Memorandum of Understanding between the National Institute of Informatics (Japan) and the Albert-Ludwig University of Freiburg (Germany) as well as with the National School of Applied Mathematics and Computer Science of Grenoble (ENSIMAG, France).

REFERENCES

[1] "Swamp Computing" a.k.a. Cloud Computing". Web Security Journal. 2009-12-28

[2] "Are security issues delaying adoption of cloud computing?" Eric Mandel, CEO of managed hosting services provider BlackMesh, 2009-04-27

[3] U.S. Department of Health & Human Services, Health Insurance Portability and Accountability Act of 1996 (HIPAA) Privacy Rule, http://www.cms.hhs.gov/HIPAAGenInfo (1996)

[4] Buneman, P., Khanna, S., and Tan, W-C., Why and Where: A Characterization of Data Provenance, ICDT 2001, LNCS Vol. 1973, Springer, pp. 316–330 (2001)

[5] S. Haas, S. Wohlgemuth, I. Echizen, N. Sonehara, and G. Mueller, "On Privacy in Medical Services with Electronic Health Records," Proc. of International Medical Informatics Association WG4 Security in Health Information Systems (IMIA SiHIS 2009), pp. 1-9, in CD-ROM (2009).

[6] Cox, I.J., Miller, M.L., Bloom, J.A., Fridrich, J., and Kalker, T., Digital Watermarking and Steganography, Morgan Kaufmann (2008)

[7] DevoTechs Team, Huge Integer Math and Encryption library V2.05.2, (2010)

[8] Microsoft MSDN Visual Basic Developer Centre, http://msdn.microsoft.com/en-gb/vbasic/default.aspx/ (2010)

[9] Camenisch, Jan; Lysyanskaya, Anna (2001). "An efficient system for non-transferable anonymous credentials with optional anonymity revocation". in Pfitzmann, Birgit. Advances in Cryptology — EUROCRYPT 2001. Lecture Notes in Computer Science. Springer. pp. 93–118. doi:10.1007/3-540-44987-6_. ISBN 978-3-540-42070-5

[10] NIST Draft Special Publication 800-131, "Recommendation for the Transitioning of Cryptographic Algorithms and Key Sizes" (2010)

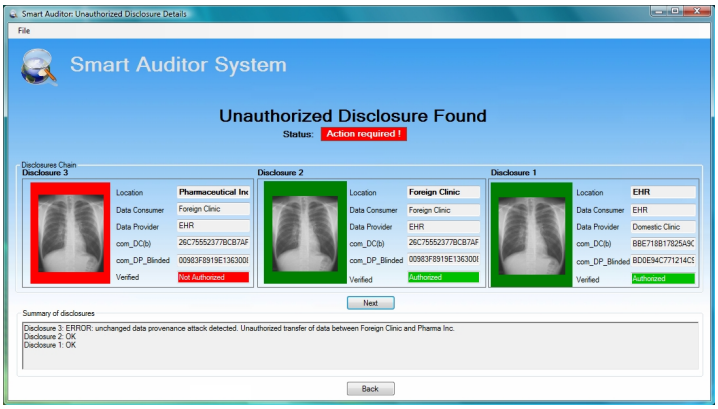

Figure 2. Result of an audit after an "unchanged data provenance" attack