An Approach for Preserving Privacy and Protecting Personally Identifiable Information in Cloud Computing

Rohit Ranchal
Department of Computer Science
Purdue University
West Lafayette, IN, 47907
rranchal@purdue.edu

Bharat Bhargava
Department of Computer Science
Purdue University
West Lafayette, IN, 47907
bbshail@purdue.edu

Lotfi Ben Othmane
Department of Computer Science
Western Michigan University
Kalamazoo, MI, 49008
lotfi.benothmane@wmich.edu

Leszek Lilien
Department of Computer Science
Western Michigan University
Kalamazoo, MI, 49008
leszek.lilien@wmich.edu

Anya Kim
Naval Research Laboratory
4555 Overlook Ave. S.W.
Washington, DC, 20375
anya.kim@nrl.navy.mil

Myong Kang
Naval Research Laboratory
4555 Overlook Ave. S.W.
Washington, DC, 20375
myong.kang@nrl.navy.mil

ABSTRACT

Privacy and security in cloud computing is an important concern for both the public and private sector. Cloud computing allows the use of internet-based services to support business process and rental of IT-services on a utility-like basis. While cloud computing offers a massive concentration of resources, it poses risks for privacy preservation. The expected loss from a single breach can be significant and the heterogeneity of "users" represents an opportunity of multiple, collaborative threats.

Problems associated with trusted 3rd party managed Cloud Computing stem from loss of control, lack of trust (mechanisms) and multi-tenancy. Identity management (IDM) is one of the core components in cloud privacy and security and can help alleviate some of the problems associated with cloud computing. Cloud computing requires a user-centric access control where every user's request for any provider is accompanied with the user identity and entitlement information. The system creates digital identities for its users, and protects the users' Personally Identifiable Information (PII). User identity has identifiers or attributes that constitute PII, which identifies and defines the user. The identity is portable although tied to a domain. This user-centric approach gives the users the ultimate control of their digital identities.

A review of the available privacy-enhancing solutions shows that there is a lack of standard system that address all the privacy issues in cloud computing. Cloud computing can benefit from the owner-centric mechanism for protecting privacy of sensitive data throughout their entire lifecycle.

We discuss and propose approaches for privacy preservation in the cloud that does not use a trusted third party. The components of the proposed approach are: (i) use of active bundle—which is a middleware agent that includes data, privacy policies and a virtual machine that enforces the policies and use a set of protection mechanisms (i.e., integrity check, apoptosis, evaporation, decoy) to protect itself, as a container for PII; (ii) use of active bundle to mediates interactions between the user and cloud services using user's privacy policies; and (ii) use of predicate over encrypted data computing when negotiating a use of a cloud service.

Categories and Subject Descriptors

K.6.5 [Computing Milieux]: Security and Protection

General Terms

Management, Design, Security

Keywords

Active Bundles, Cloud Computing; Identity Management (IDM); Personally Identifiable Information (PII); Privacy-enhancing Technologies (PET); Privacy; Security

1. INTRODUCTION

1.1 Overview of Cloud Computing

The growing popularity, continuing development and maturation of cloud computing services is an undeniable reality. Cloud computing services are available in different areas, like document processing websites, navigation websites, data storage sites, multimedia sites including audio, video, photos, tax preparation sites, personal health record websites, social networking sites. Any information stored locally on a computer can be stored in cloud, including word processing documents, spreadsheets, presentations, audio, photos, videos, records, financial information, appointment calendars, address books, and more. Cloud computing offers an immense concentration of resources. However, it spawns huge risks such as: (i) expected losses from a single breach can be significantly large; and (ii) the heterogeneity of "users" represents an opportunity of multiple collaborative threats. A cloud service provider is like *third party* that maintains information about, or on behalf of, another entity. Whenever an individual, a business, a government agency, or other entity shares information in the cloud, privacy or confidentiality questions may arise [2].

Problems associated with Cloud Computing stem from loss of control, lack of trust and multi-tenancy. These problems exist mainly in third party management models. Consumer's loss of control is due to the fact that cloud providers host data, applications and resources. User identity data, access control rules, security policies are stored, managed and enforced by cloud providers. Consumer relies on the providers to ensure data security and privacy. Trusting a third party requires taking the risk of assuming that the trusted third party will act as it is expected (which may not be true all the time). Tenants

(cloud service users) share a pool of resources and may have opposing goals. If tenants cannot be trusted, they need to be isolated with some level of guarantee. In a third party managed model, service providers (e.g., Google and Amazon) manage and control various aspects of the cloud. The main problems associated with such a model are:

1. *Loss of control:* Data, applications, resources are located with service provider. The cloud handles user identity management (IDM) as well as user access control rules, security policies and enforcement. The consumer has to rely on the provider to ensure data security and privacy, resource availability, monitoring of services and resources.

2. *Lack of trust*: Trusting a third party requires taking risks. Basically trust and risk are opposite sides of the same coin. Some monitoring or auditing capabilities would be required to increase the level of trust.

3. *Multi-tenancy*: Consumers are tenants sharing a pool of resources and may have opposing goals. There may be conflicts between tenants' opposing goals. There is a need to provide a strong degree of separation between tenants.

1.2 Privacy in Cloud Computing

Privacy is the ability of an individual or group to seclude themselves or information about themselves and thereby reveal themselves selectively. Privacy in cloud computing is defined as the ability of a user or a business to control what information they reveal about themselves over the cloud (or to a cloud service provider,) and the ability to control who can access that information [3].

Numerous existing privacy laws impose the standards for the collection, maintenance, use, and disclosure of personal information that must be satisfied even by cloud providers [2,3]. The United States has several privacy laws applicable to particular types of records or businesses. Some of these laws establish privacy standards that have bearing on a decision by a business to use a cloud provider. Others laws do not. Some laws specifically allow a business to share personal information with another company that provides support services to the business. For example, *the Gramm-Leach-Bliley Act* restricts financial institutions from disclosing a consumer's personal financial information to a non-affiliated third party. Disclosure to a service provider is generally not restricted. However, the terms under which information is disclosed and the rights acquired by service providers could make a difference to the legality of the disclosure or subsequent use. The same conclusion applies to video rental records protected by the *Video Privacy Protection Act* and to cable television subscriber records protected by the *Cable Communications Policy Act*. These particular laws may not directly prevent the use of a cloud provider [2]. Due to the nature of cloud computing, there is little or, sometimes, no information available in a cloud to point out where data or information is stored, how secure is it, who has access to it, or if it is transferred to another host or if the host can be trusted [2,3]. The location of information in the cloud may have significant effects on the privacy and confidentiality protections of information and on the privacy obligations of those who process or store the information [2]. When users store their data with programs hosted on someone else's hardware, they lose a degree of control over their sensitive information. The responsibility for protecting that information from attackers, hackers and internal data breaches then falls into the hands of the hosting company rather than the individual user. Government investigators trying to subpoena information could approach that company without informing the data's owners. Some companies could even willingly share sensitive data with marketing firms. So there is always a privacy risk in putting your data in someone else's hands. Obviously, the safest approach is to maintain your data under your own control. For instance, a United States cloud provider of services to a firm or an individual may itself subcontract to or avail itself of the service of another cloud provider. That second-degree cloud provider may be located in another country or another state in the United States. The user may be unaware of the existence of a second-degree provider or the actual location of the user's data. Indeed, it may be impossible for a casual user to know in advance or with certainty which jurisdiction's law actually applies to information entrusted to a cloud provider. These uncertainties complicate the ability of a user to determine the protections that apply to data entrusted to a cloud provider [2]. For some types of information and some categories of cloud computing users, privacy and confidentiality rights, obligations, and status may change when a user discloses information to a cloud provider. Procedural or substantive barriers may prevent or limit the disclosure of some records to third parties, including cloud computing providers.

There are many questions that need to be answered. Does the user or the hosting company owns the data? Can the host deny a user access to their own data? If the host company goes out of business, what happens to the users' data it holds? And, most importantly from a privacy standpoint, how does the host protect the user's data?

The consumer has to disclose his Personally Identifiable Information (PII) (information that can be used to uniquely identify, contact, or locate a single person) to use a cloud service. It becomes even more complex when cloud service provider use services from other providers to provide a service so there may be a chain and tracking the distribution of PII may not be simple. Cloud computing has significant implications for the privacy of personal information as well as for the confidentiality of business and governmental information. With the continued expansion in cloud computing, present and potential privacy and confidentiality consequences deserve a more careful look. Faced with these issues, the major problem regarding privacy in cloud computing is how to secure personal data or information from being used by unauthorized users (including other tenants), how to prevent attacks against privacy such as identity theft, even when a cloud provider cannot be trusted, and how to maintain control over private information. The concept of handing sensitive data to another company is a concern. Is data held somewhere in the cloud as secure as data protected in user-controlled computers and networks? Privacy and security can only be as good as its weakest link. Cloud computing can increase the risk that a security breach may occur. Knowing who has their personal data and how it is being accessed, and the ability to maintain control over it prevents privacy breaches of PII, and could minimize the risk of identity theft and fraud [3].

1.3 Identity Management in Cloud Computing

A cloud user has to provide sensitive personal information (e.g. name, home address, credit card number, phone number, driver's license number, date of birth etc) while requesting services from the cloud. This leaves a trail of Personally Identifiable Information (PII) that can be used to uniquely identify, contact, or locate a single person, which—if not properly protected—may be exploited and abused [3].

The traditional model of application-centric access control, where each application keeps track of its collection of users and manages them, is not acceptable in cloud-based architectures [2,4]. A set of solutions for IDM exists, e.g. OpenID [5], Microsoft's Windows CardSpace [6] and PRIME [7].

The identity management systems in the cloud are more complex than traditional web-based systems and consumers hold multiple

accounts with the service providers. Strong authentication is a critical element in securing of cloud users, and identity management (IDM) is the key to cloud security. More details are present in [3].

1.4 Contribution and Paper Organization

The contribution of this paper is to propose an approach for identity management in the cloud that does not use a trusted third party. The components of the proposed approach are: (i) use of active bundle—which is a middleware agent that includes data, privacy policies and a virtual machine that enforces the policies and use a set of protection mechanisms (i.e., integrity check, apoptosis, evaporation, decoy) to protect itself, as a container for PII; (ii) use of active bundle to mediate interactions between the user and cloud services using user's privacy policies; and (iii) use of predicate over encrypted data computing when negotiating a use of a cloud service.

The paper is organized as follows. Section 2 presents a motivating scenario for identity management in cloud computing. Section 3 discusses related work. Section 4 presents the research problem. Section 5 presents our proposed approach for protecting PII in Cloud computing. Section 6 concludes the paper.

2. MOTIVATING SCENARIO FOR IDENTITY MANAGEMENT IN CLOUD COMPUTING

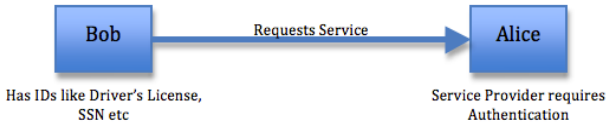


Figure 1: User-Service Provider interaction

To use a cloud service, a user needs to authenticate himself to the other party. The user provides some of his associated information, which uniquely identifies him to the other party (i.e. service provider). This is user's PII, commonly known as identity information. The identity information provides some assurance to the service provider about the user's identity, which helps him to verify whether to permit the user to use a service or not. Since identity information is personal and unique to the user, if misused, or if the user's privacy is compromised, it can lead to serious crimes involving identity theft [2,7]. The purpose of an Identity Management System is to decide upon the disclosure of this information in a secure manner.

Fig. 1 shows an example of authentication that uses PII. In the figure, Bob wants to use a service. Bob needs to authenticate to the service provider[1] but doesn't want to disclose his identity data. Bob has to disclose personal information, which uniquely identifies him to the service provider. The main problem is to decide on, which information the user *should* disclose? and how to disclose it?

Almost all online activities today—such as sending emails, filing tax declarations, managing bank accounts, using e-commerce applications, connecting to a company intranet, and meeting people in a virtual world require the user to provide sensitive personal information (e.g., the name, home address, credit card number, phone number, SSN etc.) when requesting services from the other party. This leaves a trail of PII that can be used to uniquely identify, contact, or locate a single person [3].

[1] The authentication could be performed by the service itself or delegated to another service provided by the service provider.

Fig. 2 shows the traditional and available identity management systems (which is used for some cases as authentification system), which depend on the trusted third party management models.

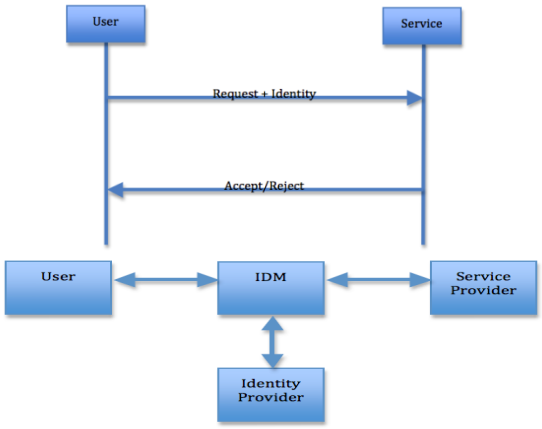


Figure 2: Authentication using Third Party Identity Management

The identity management systems in the cloud are more complex than traditional web-based systems because consumers can hold multiple accounts with the service providers. The traditional model of application-centric access control, where each application keeps track of its collection of users and manages them, is not feasible in cloud-based architectures. Sharing the user space across applications, which can lead to data replication, making mapping of users and their privileges a herculean task, further complicates this. The traditional model requires the user to remember and maintain multiple accounts/passwords. [3,4].

Strong authentication is a critical element in providing security to cloud users, and *IDM* is the key. Since various services can use different IDMs, several IDMs based on various technologies must inter-operate and function as one consolidated body over a cautiously shared user space. Hence, IDM in cloud-based projects brings about a new dimension that traditional IDMs cannot meet. More details are present in [3,4].

Identity theft thrives in a distributed environment if everyone's digital identity is distributed among many entities—such as the government, credit card companies, cell phone providers, hospitals and other organizations. A standard user-centric privacy enhancing system with IDM at its heart is required for protecting users' privacy [3,4,8].

The owner of data, including PII data, needs to be responsible for his privacy in cloud computing. The owner needs technical controls supporting this challenging task. For instance, a user-centric identity management system should allow the users to create and manage their digital identities, and allow them to authenticate themselves in a way that does not reveal their actual identities to vendors, service providers, etc even if these third parties collude to collect personal information.

IDM is the central component in cloud privacy and security. The system creates digital identities for the users, and protects the users' Personally Identifiable Information (PII). User identity has identifiers or attributes that constitute PII, which identifies and defines the user. The identity is portable although tied to a domain. Cloud requires a user-centric access control where every user's request for a service from any provider is accompanied with the user identity and entitlement information [3,4]. Problems associated with Cloud Computing stem from loss of control, lack of trust and multi-tenancy.

These problems exist mainly in third party management models. Privacy in cloud computing is an important concern. It cannot be used for storing and processing data and applications if it is unsecure.

3. RELATED WORK

The growing popularity of cloud computing increased the awareness of the privacy threats arising from its widespread use. A few already existing platforms, such as OpenID [5], Microsoft Windows CardSpace [6] and PRIME (Privacy and Identity Management for Europe) [7] are not widely accepted, recommended and commercialized due to the lack of a complete solution for protecting the privacy and the lack of standardization of their underlying protocols. Most cloud vendors have simplified proprietary IDMs with shortcomings that have to be well understood.

In the following we describe set of solutions for IDM. The solutions are: OpenID [5], Microsoft Windows CardSpace [6], and PRIME [7]. We describe these in brief. More details are present in [3].

3.1 OpenID

It is a standard used by many platforms that help users manage their multiple identities, by creating an openId (a single username/password). Its goal is aiding users in managing their various digital identities by using a single account/ID, providing them with a greater control over who to share their personal information with. The limitation is that the solutions that use the standard work only for the websites which support openId.

Products like openID have been termed "phishing heaven" due to its susceptibility to phishing attacks and Social Engineering; a malicious attack can be easily set up to lure users into entering their authentication information at a website that poses as an openID provider website.

The attack is as fellows. Whenever a user types a website's name in an Internet browser, DNS handles the translation of the name to a machine-usable IP address. Even though SSL certificates are used to verify the website identity, there is no assurance that the DNS name corresponds to the information displayed on that site. A phishing attacker might use a certificate issued for a DNS name that he owns, and craft his website to look like an authentic website the user is trying to access; then the user can be prompted to provide her credentials (known as Social Engineering). In this way, the client runs the risk of being spoofed and rerouted to a malicious cloud. User privacy becomes completely compromised [9].

3.2 Microsoft Windows CardSpace

This software can be used by Windows' applications for management of multiple digital identities belonging to the user. It has already come under scrutiny for its security issues. For example, in the default scenario for the CardSpace framework, the *Identity Provider* (IdP) is aware of the identities of the *relying parties* (RP) to which the user attempts to log in. Accordingly, the IdP can learn about the behavior of users on the web, this breaks the *Law of Directed Identity* mentioned in the Laws of Identity for IDMs [10]. The Law states that assertions made by a user using digital signatures should not make it possible for the RP or IdPs to trace or correlate these assertions to the real identity of the user or the claimant, i.e. assertions should not turn into identifiers [11].

3.3 PRIME

PRIME (Privacy and Identity Management for Europe) project produced privacy architecture, and a prototype and various application scenarios [7].

Wide scale or real life implementation has failed due to the lack of standardization of protocols. We plan to test the PRIME IDM system called IDEMIX on Yahoo cloud to learn from experience and understand all the shortcomings of PRIME when applied to cloud environment.

4. RESEARCH PROBLEMS

We describe in this section an example of PII leaks. Then, we discuss the common characteristics of existing solutions for IDM. Next, we provide selected research problems that we are investigating of which we describe our approach to solve in this paper.

4.1 AT&T's iPad

A security vulnerability in the way AT&T set up its network allowed hackers to capture the email addresses of 114,000 iPad owners [12]. The breach was a pretty basic. Indeed, if you fed an iPad ID number to a script that was publicly available on AT&T's website, it returns the email address associated with that ID. The hackers quickly decided to test numerous likely IDs. Evidently, they got back the email addresses of the owner of those iPads, including those of notable people in industry, media and politics, along with some in the military and other government agencies. The list included New York City Mayor Michael Bloomberg, Diane Sawyer of ABC News, and White House Chief of Staff Rahm Emanuel, E-mail addresses of users at the Army, the Defense Advanced Research Projects Agency, the Federal Aviation Administration, the Federal Communications Commission, the Justice Department and NASA. The breach did not result in damage to the iPad users because the only thing exposed were e-mail addresses, along with the users' ICC identification numbers, which authenticate them on AT&T's network but people whose email addresses were disclosed could receive an increased spam or phishing attacks.

4.2 Characteristics of Existing Solutions for IDM

Different solutions use different ways of sending Identity Information for negotiation. The common ways are:

- *Token/Pseudonym*: Identity providers act as security token services to integrate attributes and pseudonyms into the token issuance mechanism, which could be used to provide authentication.
- *Identity Information in clear plain text*: To use some services, PII is disclosed in clear plain text for e.g. SSN during tax preparation or filing, generating credit reports etc.

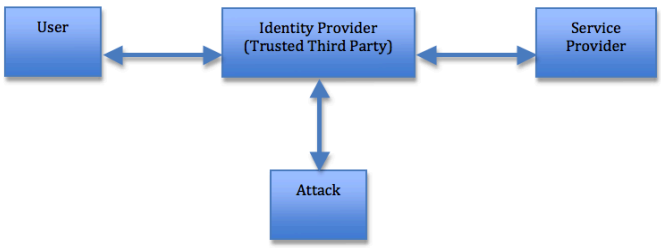


Figure 3. Architecture of Existing Identity Management Systems

These solutions have two characteristics, which are:

1. *The use of Trusted Third Party*. The major issues for adopting such approach for cloud computing are: (i) the trusted third party (it could be a cloud service located at the cloud provider) and the service provider may be the same. Therefore the trusted third party may not be an undependant-trusted entity anymore; (ii) it is a centralized approach. But if the Trusted Third Party is compromised; all the PII of its users is compromised as well.

2. *They do not support untrusted hosts.* The client application that holds the PII must be executed on a trusted host such that the host does not extract the PII. A host that has a malware that may extract the PII and may send it to the malware owner or do other malicious activity is not trusted. If we assume that most hosts used by regular users have malwares (which is true) on them, then these applications should not be used.

4.3 Selected Research Problems

The research problems are:

1. *Authenticate without disclosing data (unencrypted data):* When a user sends the identity information to get authenticated for a service, it may encrypt the data. However, before this information is used by the service provider, it is decrypted such that the service provider can use it. But as soon as the information is decrypted it becomes prone to attacks.This is particularly of a concern if the provider decides to store this information.
2. *Use service on untrusted hosts (hosts not owned by user):* The available IDM solutions need the user to be on a trusted host for using the IDM system or service. They do not allow usage of IDM on untrusted hosts like public host. With the advances in cloud computing where data may reside anywhere in the cloud, this issue needs to be addressed.
3. *Minimize risk of disclosure during communication between user and service provider (protect from Side Channel and Correlation Attacks):* Data needs to be protected from disclosure. In the scenario of cloud computing, this becomes even more important where the sensitive data may be held by a service provider and it is transmitted to another service provider (as a subcontractor), to use the service.

5. PROPOSED APPROACH FOR PROTECTING PII IN CLOUD COMPUTING

In this section we describe our proposed approach, which is based on identity management using Active Bundle scheme (allowing users to use the service on untrusted hosts) and computing predicate over encrypted data.

5.1 Characteristics of the Proposed IDM

1. *Ability to authenticate without disclosing unencrypted data.*This can be achieved by using predicate over encrypted data. We are evaluating the algorithm for this. E.g., no disclosure of passwords or decryption keys.
2. *Ability to use Identity data on untrusted hosts.* It has a self-integrity check to find if the data is tampered. If the integrity is compromised, it will destroy the data itself by doing apoptosis or evaporation to protect it from falling into wrong hands.
3. *Independent of third party*. This prevents from correlation attacks and side-channel attacks since the exchange of data from Active Bundle to host is local to the host.

5.2 Use of Predicate with Encrypted Data and Multiparty Computing Approach

In this approach we propose the use of a predicate encryption schema and secure multi-party computing for disclosure of partial information (used for giving answer to predicate) and encrypted PII based on the requests of the service provider.

Shamir [14] proposes threshold secret sharing. First, a secret data item D is divided into n pieces $D1,..., Dn$. Then, a threshold k is chosen so that: (i) to recover D, k or more of any Di's are required; (ii) using any $k-1$ or fewer Di's leaves D completely undetermined.

Ben-Or, Goldwasser and Wigderson [15] define a protocol for computing a function f by n players (entities who have share of the secret data). The function is specified as n programs, where each player uses one program only, and the n players can together compute f collaboratively.

A predicate encryption schema is an encryption schema that allows computing predicate with encrypted data. Using a predicate encryption, Alice computes predicate such as "(email sender = Bob) and (date in [2006, 2007])", over encrypted data [13].

There are some encryption schemas that provide such property like the one mentioned in [13] that we show in Fig. 4. In this schema, Alice uses a *Setup* algorithm to generate a public key *PK* and its associated secret key *MSK*. Next, Alice can use the public key to encrypt (using algorithm *Encrypt*) her PII to output *CT*. Then, she may store *CT* on an untrusted host such as in cloud. She may also publish the public key so it could be used to encrypt data that she can access (someone can send her an encrypted e-mail). When Alice has a predicate *f* that she wishes to compute on her encrypted data, she uses her *keyGen* algorithm to compute a token. The *KeyGen* algorithm uses input *PK*, *MSK* and *f* and outputs TK^f. Alice, can give her token TK^f to the host who computes it on the encrypted data *CT* and returns the result to Alice. Alice receives $f(x)$.

1.	Setup	PK, MSK
2.	Encrypt(PK, X)	CT
3.	KeyGen(PK,MSK, f)	TK^f
4.	Query(PK, CT,TK^f)	f(x)

Figure 4. Public-key Predicate Encryption Schema

We note that algorithm *KeyGen* uses the secret key as input. Therefore Alice can use it to generate a token TK^f given a predicate *f*. Alice can give the token to the host while preserving the privacy of her PII and secret key. If Alice gives the algorithm and secret key to the host then, the schema is not secure.

This observation allows us to conclude that Alice cannot use only a predicate encryption schema as a mechanism in Identity Management System for assuring to Alice that Bob has the required identity information that she specifies in a predicate.

As an alternative to the use of trusted third party, we propose the use of computing predicate using encryption schema and secure multi-party computing. In this approach the secret key *MSK* is split between *n* parties using Shamir's technique [14]. Then, the algorithm *keyGen* is provided to *n* parties and computed as specified in Ben-or *et al*. protocol [15]. The parties collaboratively compute *KeyGen* using their shares of the secret key, predicate *f* and the public key and outputs TK^f.

The algorithm *Query* takes as input *PK*, *CT*, TK^f and outputs f(x) which is the answer to the predicate.

5.3 Use of Active Bundle Scheme for IDM

In the following we provide an overview of the active bundle schema, and the use of the active bundles schema for IDM.

5.3.1 Overview of the Active Bundles Schema

An *active bundle* includes sensitive data, metadata, and a virtual machine [16]. Fig. 5 shows the general structure of an active bundle. Sensitive *data* constitutes content to be protected from privacy violations, data leaks, unauthorized dissemination, etc.

Metadata describes the active bundle and its privacy policies. The metadata includes (but is not limited to) the following components (details available in [16], [17]): (a) *provenance* metadata; (b) *integrity check* metadata; (c) *access control* metadata;

(d) *dissemination control* metadata; (e) *life duration* value; (f) *security* metadata (including: *security server id*; *encryption algorithm* used by the VM; *encrypted pseudo-random number generator*; *trust server id* used to validate the trust level and the role of a host; and *trust level threshold* required to access data in an active bundle); and (g) other application-dependant and context-dependant metadata.

Virtual machine (VM) manages and controls the program code enclosed in a bundle. The main VM functions include (a) enforcing bundle access control policies through *apoptosis*, *evaporation*, or *decoy* actions (e.g., disclosing to a guardian only this portion of data that the guardian is entitled to access); (b) enforcing bundle dissemination policies; and (c) validating bundle integrity. We are working on providing security against attacks by using an *obfuscated virtual machine* (*OVM*).

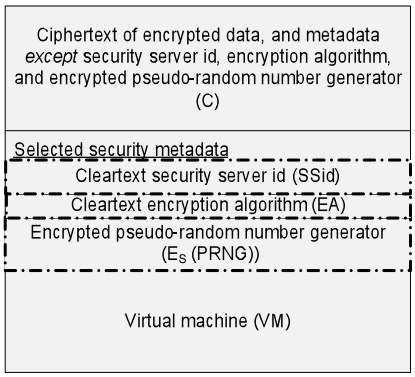


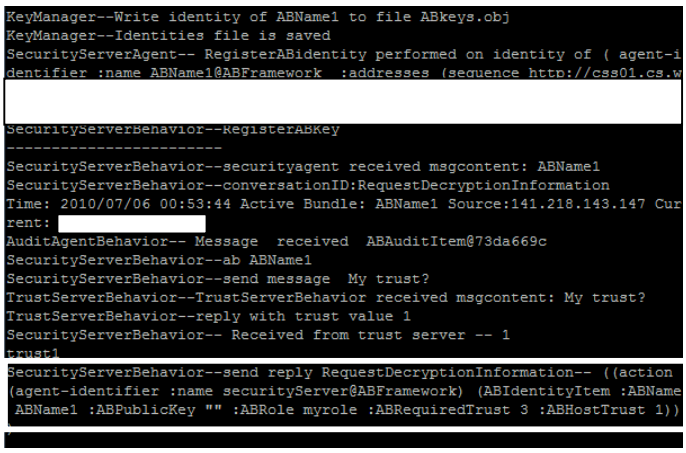
Figure 5. Structure of an active bundle with security metadata emphasized [13].

5.3.2 Using Active Bundles for the IDM

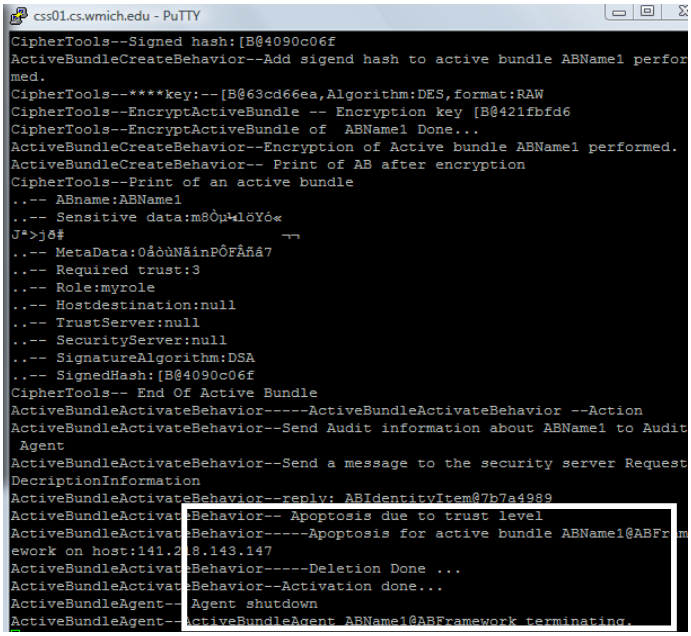
The Components of an active bundle for IDM System are:

1. *Identity data:* The data used during authentication, getting service, using service (i.e. SSN, DOB). This data will be packed inside the active bundle but should it be encrypted or not? The answer depends on how the Active Bundle is created. In the complete approach that we discuss in this section, the data is encrypted.
2. *Disclosure policy:* These constitute rules for choosing Identity Data from a database of Identities. For instance, if some particular Identity data has been used for a particular service then same data needs to used and disclosed every time for that service, there is no need to disclose another piece of identity data to that service.
3. *Disclosure history:* This can be used for logging and auditing purposes and selecting the Identity data to be disclosed based on previous disclosure.
4. *Negotiation policy:* We are evaluating different algorithms for data disclosure. These are mainly based on Predicate Data, Selective Disclosure, Zero Knowledge Proofing.
5. *Virtual Machine:* This contains the code/algorithm for protecting Identity data on untrusted hosts. This implements Active Bundle properties.

An active bundle could be send from a source host to a destination host. When arriving at a foreign host, an active bundle ascertains the host's trust level through a trusted third party [16]. Using its access control policies, it decides whether the host may be eligible to access all or part of the data and become a guardian, and what portion of sensitive data can be revealed to it. The remaining data (not to be revealed) is *evaporated* as specified in the access control policies, diminishing the value of data. We consider a number of different metrics for adaptive control of the degree of evaporation, including trust-based metrics.



a) Log of Security Server Agent. The text encircled with white shows that security server sends a message to the active bundle to state that the required trust level to access the active bundle is 3 and the host's trust level is 1. The message also does not give the correct decryption key.



b) Log of activation of the active bundle. The text encircled with white shows that the active bundle apoptosize and delete itself.

Figure 6. Log of the activation of an active bundle. The active bundle is apoptosize because the trust level of the host is less than the required trust level to access the active bundle content.

An active bundle may realize that its security or privacy is about to be compromised, e.g., it may discover that its self-integrity check fails, or the trust level of its guardian is too low. In response, the bundle may choose to commit *apoptosis*, that is perform atomically a clean self-destruction, that is, self-destruction that is complete and leaves no trace usable for an attacker. In this paper we omit the details of the active bundle schema, which are discussed in [16]. We currently developed a prototype for the active bundle schema [18]. Figure 6 shows an example of the activation process of a generic purpose active bundle when it is activated at a destination host.

5.4 Advantages of Proposed Approach

The main advantages of the proposed approach are:

1. It is independent of the usage of trusted third party so it is saved from the correlation attacks and side channel attacks since the exchange of data from Active Bundle to host is local to the host.
2. It has the ability to authenticate without disclosing unencrypted data. This prevents unnecessary data disclosure.
3. It has the ability to use Identity data on untrusted hosts. If the data reaches an unintended destination or the integrity is tampered with, it will destroy the data itself by doing apoptosis or evaporation to protect it from falling into the wrong hands.

5.5 Resilience of the Proposed Approach to Attacks

A system based on the proposed approach is independent of the usage of trusted third party. This reduces the risks of correlation attacks and side channel attacks within the cloud.

Correlation attacks in IDM are attacks where an entity acquires a set of data enough when correlated to identify physically an entity such as a person. Approaches that use a trusted third party increase the risk of correlation attacks on an entity's PII. Approaches that do not use a trusted third party reduce the risk of such attacks.

Ristenpart *et al.* [19] demonstrated that Amazon cloud is prone to side-channel attacks and it would be possible to steal data, once the malicious virtual machine is placed on the same server as its target. It is possible to carefully monitor how access to resources fluctuates and thereby potentially glean sensitive information about a victim. Though, they point out, that there are a number of factors that would make such an attack significantly more difficult in practice.

Approaches that use a trusted third party increase the risk of Side-Channel attacks on an entity's PII. Approaches that do not use a trusted third party such as the one that we use reduce the risk of such attacks.

The proposed solution is prone to attacks such as attacks performed by malwares. An execution of an active bundle on a host may be altered by a malware so that it can have access to unauthorized data. The active bundle may also be not executed at all at the host of the requested service. In this case its data is not disclosed but the user is denied access to the service that he requests.

6. CONCLUSION

With the immense growth in the popularity of cloud computing, privacy and security has become an important concern for both the public and private sector. It is very likely that users end up having multiple identities in multiple service providers' security repositories, multiple credentials and multiple access permissions with different services provided by different service providers. There is a strong need for an efficient and effective privacy-preserving system which is independent of a trusted third party and is able to unambiguously identify users that can be trusted both across the Web and within enterprises and protect their Personally Identifiable Information (PII). Identity management (IDM) is one of the core components in cloud privacy and security and can help alleviate some of the problems associated with cloud computing.

We discuss and propose approaches for privacy preservation in the cloud without the use of trusted third party. The components of the proposed approach are: (i) use of active bundle—which is a middleware agent that includes data, privacy policies and a virtual machine that enforces the policies and use a set of protection mechanisms (i.e., integrity check, apoptosis, evaporation, decoy) to protect itself, as a container for PII; (ii) use of active bundle to mediates interactions between the user and cloud services using user's privacy policies; and (ii) use of predicate over encrypted data computing when negotiating a use of a cloud service. Future work includes development of a prototype of the proposed approach. The goal is to prove effectiveness of the proposed identity management system.

7. REFERENCES

[1] NIST.2010. http://csrc.nist.gov/groups/SNS/cloud-computing/

[2] Gellman, R. 2009. Privacy in the Clouds: Risks to Privacy and Confidentiality from Cloud Computing. World Privacy Forum.

[3] Angin, P., Bhargava, B., Ranchal, R., Singh, N., Lilien, L., Othmane, L. B. 2010. A User-Centric Approach for Privacy and Identity Management in Cloud Computing. IEEE International Symposium on Reliable Distributed Systems. To appear in.

[4] Gopalakrishnan. A.2009. Cloud Computing Identity Management. SETLabs Briefings. Vol 7. http://www.infosys.com/research/

[5] OPENID. 2010.http://openid.net/

[6] OWASP. 2010.http://owasp.org/

[7] Hubner, S.F. 2008. PRIME, https://www.prime-project.eu/

[8] Identity Theft Primer. 2005.Libery Alliance Whitepaper, http://www.projectliberty.org/

[9] Sample, C. Kelley, D.2009. Cloud Computing Security: Routing and DNS Threats. http://www.securitycurve.com/wordpress/

[10] Camerson, k. 2008. Proposal for a Common Identity Framework: A user-Cenctric Identity Metasystem, http://www.identityblog.com/

[11] Alrodhan,W.A.,Mitchell, C.J. Improving the Security of CardSpace, EURASIP Journal on Information Security Vol. 2009, doi:10.1155/2009/167216

[12] AT&T iPAD. 2010. http://www.techdirt.com/articles/20100609/1604379757.shtml

[13] Shi, E. Evaluating Predicates over Encrypted Data. Ph.D. Thesis. Oct. 2008. Carnegie Mellon University, Pittsburgh, PA.

[14] Shamir, A. How to Share a Secret. In Communications of the ACM, vol. 22(11), 1979, pp. 612–613.

[15] Ben-Or, M., Goldwasser, S. and Wigderson, A. Completeness theorems for non-cryptographic fault-tolerant distributed computation. Proc. of the Twentieth Annual ACM Symposium on Theory of Computing, Chicago, IL, May 1988, pp.1-10.

[16] Othmane, L. B., Lilien, L. Protecting Privacy in Sensitive Data Dissemination with Active Bundles. Proc. 7th Annual Conference on Privacy, Security & Trust (PST 2009), Saint John, New Brunswick, Canada, August 2009.

[17] Lilien, L., Bhargava, B. A Scheme for Privacy-preserving Data Dissemination. IEEE Trans. on Systems, Man and Cybernetics, Part A: Systems and Humans, Vol. 36(3), 2006.

[18] Othmane, L. B., Lilien, L., Bhargava, B., Ranchal, R., Azarmi, M., Salih, R. An Active Bundle Prototype for Protecting Privacy in Data Dissemination. Paper in preparation.

[19] Ristenpart, T., Tromer, E., Shacham, H., Savage, S. Hey, You, Get Off My Cloud: Exploring Information Leakage in Third-Party Compute Clouds. Proc. 6th ACM conference on Computer and Communications Security, Chicago, IL, 2009.