Privacy – Preserving Detection Of Sensitive Data Using Vector Based Fuzzy Fingerprint

V.Elakkiya[1], S.Nithyadhevi [2],

[1]V.Elakkiya, PG Scholar, Dept. of computer science,
SINCET, Nagapattinam, Tamil Nadu, India.
[2]S.Nithyadhevi, Assistant Professor, Dept. of computer science
SINCET, Nagapattinam, Tamil Nadu, India.

---***---

***Abstract* - *Generally Commercial based business application is hosted in third party server, which servers resources are depended on outsource providers. All kind of business information are outsourced and stored in remote server based on privacy and reliability. The mutual agreement between business partners and outsource providers are written on e-document and manual progress. The agreement terms and conditions are stabilizing the privacy about business information of the commercial application. The people who is working in the outsource provider place may use influence or misuse the system resource for their own gain. Secretly, they will reveal privacy data and they sell into some other person. The proposed system provides protection for information privacy of outsource third party data, and leads to detect the leakage of data at any party location. Existing techniques like watermark, duplication, etc help to detect information leakage in major percentile. But every time manual / systematic process for integrating identity information with outsourced data is over weight of normal work. Business people cannot maintain or track the information identity before they do outsource. So far, this kind of techniques helps minimum range of probability on leakage detection.*

***Key Words*: Data-leak detection (DLD), Risk Based Security (RBS), HTTP, NIDS.**

1. INTRODUCTION

Typical approaches to preventing data leak are under two categories host-based solutions and network-based solutions. Host-based approaches may include encrypting data when not used detecting stealthy malware with antivirus scanning or monitoring the host and enforcing policies to restrict the transfer of sensitive data. These approaches are complementary and can be deployed simultaneously. It presents a network-based data-leak detection (DLD) solution that complements host-based methods. Network-based data-leak detection focuses on analyzing unencrypted outbound network through deep packet inspection or information theoretic analysis. In addition to provide the regular networking, computing, or storage services, network or cloud providers may introduce additional security protection for their customers. For their customers, these add-on security services such as data-leak detection are attractive, as they may have a lower cost compared to building in-house security management of their own. Thus, one may outsource the data-leak detection to a DLD provider. However, the data owner may not allow the DLD provider to access the sensitive data.

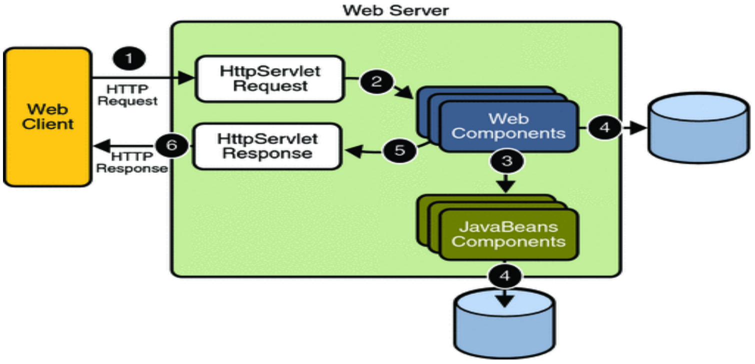

Fig 1.1: Architectural View of web security

To overcome existing issues, proposed system offer feature like security constraint before processing incoming request, Complex encryption system before store and decryption after retrieval information form back end server. Proposed system has implementation of AES, and domain ads on service, policy based request filter system. Those are performing individual action on every stage of this outsource business. All incoming request are easily traced and direct access on hosted server are prohibited. The technical challenge is that the detection algorithm needs to provide guarantees on the secrecy of customers' sensitive data while still enabling the provider to identify signs of data leak. This problem of the lack of support for privacy-enhancing data-leak detection has not been systematically addressed in the security literature. In this paper proposed design, implement, and experimentally evaluate a client technique that enhances the data privacy during the data-leak detection operations. Our method is based on a fast and practical one-way computation and does not require any expensive cryptographic operations. In the academic section available academic studies have been clustered into various categories according to the nature of the leakage and the protection provided. Next, the main data leakage scenarios are described, each with the most relevant and applicable solution or approach that will mitigate and reduce the likelihood or impact of data leakage.

Data-leak detection solution which can be outsourced and be deployed in a fully honest and reliable detection environment. The design, implements, and evaluates our policy based request filter, base information encryption that enhances data privacy during data-leak detection operations. Data leak approach is based on a fast and practical one-way computation on the sensitive data. It enables the data owner to securely delegate the content-inspection task to DLD providers without exposing the sensitive data. Using data detection method, the DLD provider, who is modeled as an honest-but-curious adversary, can only gain limited knowledge about the sensitive data from either the released digests, or the content being inspected. Using fuzzy techniques, an Internet service provider (ISP) can perform detection on its customers' traffic securely and provide data-leak detection as an add-on service for its customers. In another scenario, individuals can mark their own sensitive data and ask the administrator of their local network to detect data leaks for them.

2. LITERATURE SURVEY

2.1 A.Z.Broder,et.al.,[1], proposed that fingerprinting scheme is a method for implementing fingerprints using polynomials over a finite field. The basic idea is that the file system computes the cryptographic hash of each block in a file. To save on transfers between the client and server, they compare their checksums and only transfer blocks whose checksums differ. But one problem with this scheme is that a single insertion at the beginning of the file will cause every checksum to change if fixed-sized (e.g. 4 KB) blocks are used. So the idea is to select blocks not based on a specific offset but rather by some property of the block contents. Any hash function could be used to divide a long file into blocks (as long as a cryptographic hash function is then used to find the checksum of each block): but the Rabin fingerprint is an efficient rolling hash, since the computation of the Rabin fingerprint of region B can reuse some of the computation of the Rabin fingerprint of region A when regions A and overlap. The challenges are currently, the new pointer blocks always hold a fixed number of scores. When a new block is inserted in an existing hash tree is split in the new archive, it now changes all pointer blocks to right of it in the hash tree.

2.2 A.Z.Broder et.al.,[2],exposed that the classic analysis of hashing schemes often entails the assumption that the hash functions used are random. More precisely, the assumption is that keys belonging to a universe U are hashed into a table of size M by choosing a function h uniformly at random among all the functions. File access patterns should also be taken into consideration. Whole file content and fixed size blocking strategies present the disadvantage that file updates may lead to the recompilation of SHA-1 digests for large amounts of data. There are lots of string matching algorithms that are faster than O(n+m)It's practically as slow as brute force matching and it requires additional space. Rather simple, but currently the size of a sub tree is always stored in 8 bytes. This grows the size of one pointer from 20 bytes to 28 bytes. The lowest-level pointer blocks, which point to blocks of at most 32KB, will be the most numerous. Data security is a company imperative. Seeks to limit the distribution of personally identifiable information consistent with the nature and sensitivity of the information. It strives to accurately report information in its products.

2.3 S.Gervand,et.al.,[3], proposed that a bloom filter is a simple-efficient randomized data structure for representing a set in order to support membership queries. bloom filters allow false positives but the space savings often out of weigh this drawback when the probability of an error is made sufficient low. Bloom introduced bloom filters in the 1970's and ever since they have been very popular in database applications. Expressed that produces accurate correction results with much less memory compared with previous solutions. The algorithm, named Bloom-filter-based Error correction Solution for high-throughput Sequencing reads (BLESS), uses a single minimum-sized Bloom filter, and is also able to tolerate a higher false-positive rate, thus allowing us to correct errors with a 40× memory usage reduction on average compared with previous methods. With the aims of providing a unified mathematical and practical framework for them and simulating their use in future application.

2.4 Jung .J et.al.,[4] proposed that modern personal computational devices, from desktops to mo- bile computers, smart phones and consumer electronics, run a wide variety of applications that send user information over the network to other parties. The goal of our work is to develop general tools and techniques that enable consumers and their agents to discover leaks of personal information in applications. Our approach is to employ a conceptually straightforward testing and analysis methodology that we refer to as differential black-box fuzz testing. The challenges solution are By comparing the set of network traces from the same test in-put, Privacy Oracle identifies byte segments that remain un-changed for the given usage of the application .Privacy Oracle runs the target application with slightly modified test inputs (username = alice in T3), and collects net- work traces. Application programs track users' inputs and change their behavior in an attempt to pro- vide a more convenient service. The host operating system of the target application may independently generate packets during a test. This extraneous traffic can pollute output data sets, misleading analysis results.

2.5 M.O Kajorath et.al.,[5],exposed that the move toward large-scale deployment of electronic commerce is hampered by privacy concerns of potential customers. To appease such concerns, enterprises publish privacy statements that promise fair information practices. These problems are amplified if personal data is used not only by the enterprise that collected the data, but also by secondary users such as partner organizations, and government agencies. These flows of data are complex. A privacy control model should reflect that there are three entities that influence authorizations. The security administrator protects the interest of the organization. The challenges are Data security is a company imperative. Seeks to limit the distribution of personally identifiable information consistent with the nature and sensitivity of the information. It strives to accurately report information in its products.

2.6 J.Li,et.al.,[6], proposed that traditional searchable encryption schemes allow a user to securely search over encrypted data through keywords and selectively retrieve files of interest, these techniques support only exact keyword search. In this paper, for the first time we formalize and solve the problem of effective fuzzy keyword search over encrypted cloud data while maintaining keyword privacy. Fuzzy keyword search greatly enhances system usability by returning the matching files when users' searching inputs exactly match the predefined keywords or the closest possible matching files based on keyword similarity semantics, when exact match fails. In our solution, exploit edit distance to quantify keywords similarity and develop two advanced techniques on constructing fuzzy keyword sets. A server that communicates with a target application can unexpectedly change its behavior during test runs, which is beyond our control. During installation and initial use, an application may prompt for user preferences or for personal data such as name, email address, organization, gender, and zip code. In our solution, exploit edit distance to quantify keywords similarity and develop an advanced technique on constructing fuzzy keyword sets, which greatly reduces the storage and representation overheads.

2.7 M.O Rabin,et.al[7],explained that Fingerprinting random polynomials scheme based on arithmetic modulo in irreducible polynomial with coefficients in Z. This paper presents an implementation and several application of this scheme that take considerable advantage of its algebraic properties. The purpose of this paper is to propose an advanced fingerprint-based indoor localization scheme for wireless sensor networks (WSN) to improve the accuracy. Many localization methods have been introduced for WSN systems using wireless signals mainly divided into two categories, which are range-based and range-free. As wireless ranging is not reliable in indoor environment due to multipath fading and other attenuations, range-free solutions. random and unpredictable human presence and movement cause certain level of interference to reduce the accuracy of indoor localization. In this paper, LWMA scheme is introduced to filter interfered RSSI values and contribute to algorithm recover back to no interference condition. The experiments show that the scheme can improve the average positioning accuracy results. missing values are scattered across the data set, minimum, maximum and average similarity coefficients are a simple means of visualizing the effects of missing data on tree structure. Our approach indicates the range of values that a data set containing missing data points might generate, and forces the investigator to consider the effects of missing values on data interpretation.

2.8 X.Shu, et.al.,[8] proposed that a network-based data-leak detection (DLD) technique, the main feature of which is that the detection does not require the data owner to reveal the content of the sensitive data. Instead, only a small amount of specialized digests are needed. Our technique – referred to as the fuzzy fingerprint – can be used to detect accidental data leaks due to human errors or application flaws. The privacy-preserving feature of our algorithms minimizes the exposure of sensitive data and enables the data owner to safely delegate the detection to others. Its main feature is that the detection can be performed based on special digests without the sensitive data in plaintext, which minimizes the exposure of sensitive data during the detection. To be deployed by an individual on a home network inter- face, or to be deployed by a company that provides net- work security services for home networks. Provides a natural platform for conducting data-leak detection by cloud providers as an add- on service. May involve additional cost, probably on a subscription basis. Initial cost may be high; ongoing management may require dedicated resources, so ongoing costs may also be high. The sensitive data is accidentally leaked in the outbound traffic by a legitimate user. A rogue insider or malicious and stealthy software may steal sensitive personal or organizational data from a host.

2.9 K.Xhu,et.al.,[9],explained that malicious software stealthily downloaded from the Internet has been the leading infection vector, accounting for 53% of all incidents in 2008.Malware may be delivered stealthily through a networked application such as a browser, a peer-to-peer file sharing client, or a chat application. Spyware may be bundled with files shared through P2P networks. Web browser is the most common vehicle for a host to contract malware. 10% of the websites were found to contain drive-by-download exploits Drive-by-download (DBD) attacks exploit software or design vulnerabilities in a browser or its external components, and stealthily fetch executables from remote malware-hosting server without the user permission. Conventional signature-based techniques may not be effective against zero-day exploits or code with sophisticate obfuscation. In comparison, host-based detection approaches are much more feasible against drive-by download attacks and the onset of infection in general. In this paper, demonstrate the feasibility and quality of such a host-based monitoring framework. All sensitive information that is introduced into the system in the automated tests is marked as a taint source. The test engine in Panorama allows us to perform the analysis of samples and the detection of malicious code with- out human intervention.

2.10 H.Yin, et.al.,[10] exposed that Malicious software creeps into users' computers, collecting users' private information, wrecking havoc on the Internet and causing millions of dollars in damage. Surprisingly, even software provided by reputable vendors may contain code that performs undesirable actions which may violate users' privacy. Malware detection and analysis is a challenging task, and current malware analysis and detection techniques often fall short and fail to detect many new, unknown malware samples. Current malware detection methods in general fall into two categories: signature-based detection and heuristics- based detection. We observe that numerous malware categories, including spyware, key- loggers, network sniffers, stealth backdoor, and root kits, share similar fundamental characteristics, which lies in their malicious or suspicious information access and processing behavior. This approach is less of portability. In addition to mapping instructions executed on the processor to operating-system processes, they are interested in obtaining more information when data is exchanged be- teen the memory and hardware devices. The products are designed to only look for, detect and protect your website against known attacks. Signature engines report many false positives.

3. SYSTEM ANALYSIS

3.1 Existing System

Existing system lets to detect and prevent data leaks require a solutions for outsourcing environment, where all business information are outsourced in terms of application hosting. The environment have front end and back end mode, there single model available for data-leak detection, data confinement, stealthy malware detection, and policy enforcement. Network pocket inspection techniques provide Network data-leak detection (DLD) by performing deep packet inspection (DPI) and searches for any occurrences of sensitive data patterns. All kind of protocol based communication has the TCP/IP as base for the communication. So header information section will have much section to have Meta data about pocket information.

3.1.1 Drawback Of Existing System

As the Internet grows and network bandwidth continues to increase, administrators are faced with the task of keeping confidential information from leaving their networks. In response, researchers have created data loss prevention systems that check outgoing traffic for known confidential information. These systems stop naive adversaries from leaking data, but are fundamentally unable to identify encrypted or obfuscated information leaks.

3.2 Proposed System Analysis

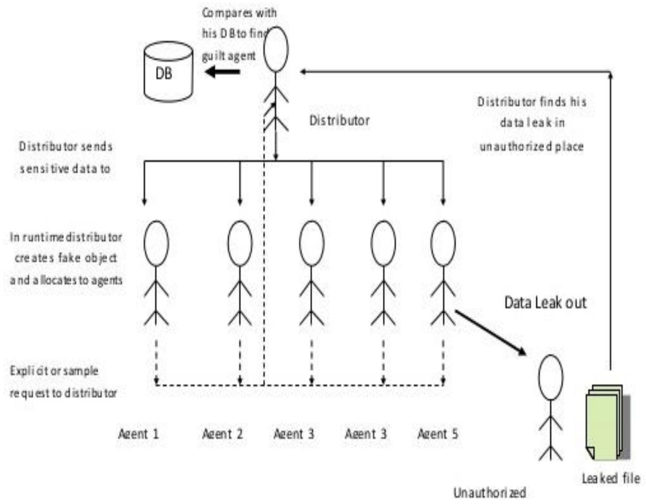

Fig 3.2: Architectural View of Proposed System

Data-leak detection solution which can be applied on outsource provider place, where outsourced data can be deployed in a fully honest detection environment. Design, implement, and evaluate our policy based filter system, and base encryption system technique that enhances data privacy during data-leak detection operations. Fuzzy approach is based on a fast and practical two-way computation on the sensitive data .which means data stored in cipher text in base system, and incoming request and outgoing response are deeply investigated. It enables the data owner to securely delegate the content-inspection task to DLD providers without exposing the sensitive data.

Using detection techniques, an Internet service provider (ISP) can perform detection on its customers' traffic securely and provide data-leak detection as an add-on service for its customers. Host server logs are incoming request resource identity and prefix the response content for the all request based on comparison on existing activity log. Rather than making privacy content with business information, the new system looks out into the request limit and response content.

3.2.1 Advantages

New system describe own privacy-preserving data-leak detection model for preventing inadvertent data leak in network traffic. Proposed model supports detection operation delegation and ISPs can provide data-leak detection as an add-on service to their customers using our model. Every business records are converted and stored in terms of cipher text in base information system. For this implementation AES (Advanced encryption system) technique applied. Data leak detection system and perform extensive experimental evaluation on internet surfing traffic of **100** users, and also N number simulated request is prohibited.

4. MODULES

4.1 Network security privacy model

4.2 Security goal threat model

4.3 Experimental evaluation module

4.4 Domain add-on module

4.5 Policy filter module

4.1 Network Security Privacy Model

Network-accessible resources may be deployed in a network as surveillance and early-warning tools, as the detection of attackers are not normally accessed for legitimate purposes. Techniques used by the attackers that attempt to compromise these decoy resources are studied during and after an attack to keep an eye on new exploitation techniques. Such analysis may be used to further tighten security of the actual network being protected by the data's. Data forwarding can also direct an attacker's attention away from legitimate servers. A user encourages attackers to spend their time and energy on the decoy server while distracting their attention from the data on the real server. Similar to a server, a user is a network set up with intentional vulnerabilities. Its purpose is also to invite attacks so that the attacker's methods can be studied and that information can be used to increase network security.

4.2 Security Goal and Threat Model

DLD provider from gaining knowledge of sensitive data during the detection process, we need to set up a privacy goal that is complementary to the security goal above. We model the DLD provider as a semi-honest adversary, who follows our protocol to carry out the operations, but may attempt to gain knowledge about the sensitive data of the data owner. Our privacy goal is defined as follows. The DLD provider is given digests of sensitive data from the data owner and the content of network traffic to be examined. Data present a privacy-preserving DLD model with a new fuzzy fingerprint mechanism to improve the data protection against semi-honest DLD provider. We generate digests of sensitive data through a one-way function, and then hide the sensitive values among other non-sensitive values via fuzzification.The sensitive data is sent by a legitimate user intended for legitimate purposes. In this paper, we assume that the data owner is aware of legitimate data transfers and permits such transfers. So the data owner can tell whether a piece of sensitive data in the network traffic is a leak using legitimate data transfer policies.

4.3 Experimental evaluation module

Data-leak detection solution can be outsourced and be deployed in service provider's host that enhances data privacy during data-leak detection operations. Our approach is based on a fast and practical two-way computation on the sensitive data (SSN records, classified documents, sensitive emails, etc.), which are policy based filter and base encryption. It enables the data owner to securely delegate the content-inspection task to DLD providers without exposing the sensitive data. Using our detection method, the DLD provider, who is modeled as fully honest gain knowledge about the sensitive data from either the released digests, or the content being inspected. Using fuzzy techniques, an Internet service provider (ISP) can perform detection on its customers' traffic securely and provide data-leak detection as an add-on service for its customers. We set up a networking environment in Virtual Box, and make a scenario where the sensitive data is leaked from a local network to the Internet. Multiple users' hosts are put in the local network, which connect to the Internet via a gateway . Multiple servers and an attacker-controlled host are put on the Internet side. The gateway dumps the network traffic and sends it to a DLD server/provider .Using the sensitive-data identity defined by the users in the local network, the DLD server performs off-line data-leak detection.

4.4 Policy Filter Module

Domain-guard Data Loss Prevention (DLP) solution, a comprehensive data leakage protection solution, helps organizations effectively manage the threats and risks to organizations' information, and safeguard what organizations value. It carries rich and matchless features to help organizations against data security threat, and prevent accidental and malicious data leakage. It effectively controls and protects sensitive data as it leaves, moves and transforms through your IT environment. With it, organizations can centrally monitor how the data is being used, records data operations and prevents data leaking out via various channels such as removable storage devices, email, instant messaging, and printing. IP-guard enables you to get a grip on your critical information, avoid costly loss and safeguard intellectual property. Distributor data will contain some additional hidden information, which is added to the real info for law enforcements. (Key information) Read and write permission for folder and file level on webs server Data base columns are encrypted using AES (Advanced encryption system).Request and response filter are applied in lifecycle of the web services.

4.5 Domain Add-On Module

For every individual domain able to add this filter option as Add on service. After enabling the Add on service, Agent cannot access the local web service related files and resource available under a domain will be accessed via web server request only. Distributor and other allowed end user will have separate UI for accessing web service resources. For all kind of data transfer the communication stream should offer confidential data on the servers, for new and previous domain. Policy filter entries are applied while initiate the service begins. General request and response are filtered based on IP identity / and preventing manual file access with the public place of resources of a domain.

5. TECHNIQUES

5.1 FUZZY FINGERPRINT METHOD AND PROTOCOL

Data leak describe technical details of our fuzzy fingerprint mechanism in this section. The DLD provider obtains digests of sensitive data from the data owner. The data owner uses a sliding window and Rabin fingerprint algorithm to generate short and hard to-reverse (i.e., one-way) digests through the fast polynomial modulus operation. The sliding window generates small fragments of the processed data (sensitive data or network traffic), which preserves the local features of the data and provides the noise tolerance property. Rabin fingerprints are computed as polynomial modulus operations, and can be implemented with fast XOR, shift, and table look-up operations.

The Rabin fingerprint algorithm has a unique min-wise independence property, which supports fast random fingerprints selection (in uniform distribution) for partial fingerprints disclosure. The shingle-and-fingerprint process is defined as follows. A sliding window is used to generate q-grams on an input binary string first. The fingerprints of q-grams are then computed. A shingle (q-gram) is a fixed-size sequence of contiguous bytes. For example, the 3-gram shingle set of string abcdefgh consists of six elements {abc, bcd, cde, def, efg, fgh}. Local feature preservation is accomplished through the use of shingles. Therefore, our approach can tolerate sensitive data modification to some extent, e.g., inserted tags, small amount of character substitution, and lightly reformatted data. Fuzzy introduce a fuzzy hash function h_ϕ to compute a fuzzy-fingerprint for a given document d □ D. As reference similarity function ϕ in Property 1 the well-known cosine measure along with the vector space model is employed.

While most fingerprint approaches rely on the original document *d*, from which substrings are selected and given to a mathematical function, our approach can be developed simplest from a document's vector space model d. The key idea behind $h\,\phi$ is an analysis and comparison of the distribution of the index terms in d with respect to their expected term frequency class [see 2]. In particular, we abstract the concept of term frequency classes towards prefix frequency classes, by comprising index terms into a small number of equivalence classes such that all terms from the same equivalence class start with a particular prefix. E. g., there might be the equivalence class of terms whose first character is from the set {"a", "A"} or, as the case may be, the equivalence class of terms whose first character is from the set {"x", "X", "y", "Y", "z", "Z"}. Based on the analysis of extensive corpora, a standard distribution of index term frequencies can be computed, and, for a predefined set of prefixes, the a-priory probability.

5.2 IDENTITY BASED ENCRYPTION TECHNIQUE

S(λ) → (PP,MK) output params, PP, and master-key, MK

K(MK, ID) →d_{ID} outputs private key, d_{ID}, for ID

E(PP, ID, m) → c encrypt m using pub-key ID (and PP)

D(d_{ID}, c) → m decrypt c using d_{ID}

IBE "compresses" exponentially many PKs into a short PP

6 .CONCLUSION

The proposed system of the data leakage system is improved in two ways to prevent information leakages. In the new system probability of leakage prevention rather than detection is increased due to high security of hosting environment. As an Add on component for a domain, the new implementation encapsulates high security in terms of privacy of information and leakages prevention. Hypertext based request are intercepted and filtered then forwarded to right resource for processing the request. All the response content retired from the database are encrypted using Attribute based encryption, which is implemented in Base encryption system, exists as one of major component at outsource provider. As the result of the new system, probability of information leakage is reduced and data identification is increased.

7. FUTURE ENHANCEMENT

Proposed fuzzy fingerprint, a privacy-preserving data-leak detection model and present its realization. Using special digests, the exposure of the sensitive data is kept to a minimum during the detection. It conducted extensive experiments to validate the accuracy, privacy, and efficiency of our solutions. For future work, to plan and focus on designing a host-assisted mechanism for the complete data-leak detection and prevention for large-scale organizations.

REFERENCES

[1] Broder A [1993], "Some applications of Rabin's fingerprinting method," in Sequences II. New York, NY, USA: Springer-Verlag, 1993, pp. 143–152.

[2] Broder A, Charikar M, Frieze M [2000], "Min-wise independent permutations," J. Comput. Syst. Sci., vol. 60,no. 3, pp. 630–659, 2000.

[3] Gervand S, Ahmadi M, [2013], "Bloom filter applications in network security: A state-of-the-art survey," Compute. Netw.vol. 57, no. 18,pp. 4047–4064, Dec. 2013.

[4] Jung J,Greenstein B, Wetherall D[2008], "Privacy oracle: A system for finding application leaks with black box differential testing," in Proc. 15th ACM Conf. Comput.Commun.Secur. pp. 279–288.

[5] Kajorath M.O,[2002], "A privacy policy model for enterprises," in Proc. 15th IEEE Comput. Secur. Found. Workshop, Jun. 2002,pp. 271–281.

[6] Li J, [2010], "Fuzzy keyword search over encrypted data in cloud computing," in Proc. 29th IEEE Conf. Comput. Commun., Mar. 2010, pp. 1–5.

[7] Rabin M.O, [1981], "Fingerprinting by random polynomials," Dept.Math.,Hebrew Univ. Jerusalem, Jerusalem, Israel, Tech. Rep. TR-15-81, 1981.

[8] Shu X, Yao D[2012], "Data leak detection as a service," in Proc. 8th Int.Conf. Secur. Privacy Commun. Netw., 2012, pp. 222–240.

[9] Xhu K,Yao D,[2011], "Detecting infection onset with behavior-based policies," in Proc. 5th Int. Conf. Netw. Syst. Secur.Sep. 2011, pp. 57–64. [15].

[10] Yin H,Song D,[2007]," "Panorama:Capturing system-wide information flow for malware detection and analysis," in Proc. 14th ACM Conf. Comput.pp. 116–127.