Proceedings of
INTERNATIONAL CONFERENCE ON COMPUTING, COMMUNICATION AND ENERGY SYSTEMS (ICCCES-16)
In Association with IET, UK & Sponsored by TEQIP-II
29th -30th, Jan. 2016

PaperID: CSEIT11

HEALTHCARE DATA LEAKAGE DETECTION USING FUZZY FINGERPRINTS SIGNATURE TECHNIQUE IN CENTRALIZED PEER-TO-PEER NETWORK.

Sneha .S. Satpute
A.A. Manjrekar.
Department of Computer Science Department of Technology, Shivaji University Kolhapur, India

Abstract—Data is shared between different peers in Peer-to-Peer (P2P) network. Data leak instances has grown rapidly from different organizations, institutes. Sometimes there is need for handling the sensitive data to the third parties. So, there may be chances of data-leaks and it should be prevented. For preventing from data leaks or for detecting the data leakage the Data leakage Detection (DLD) systems can be used. But, the DLD may also try to gain the knowledge of sensitive data which is to be shared by the owner. So, there is need to provide security for sensitive data. The Proposed method proposes a Host-Specific system for healthcare data. To provide security, integrity and privacy to the sensitive data fuzzy fingerprints, watermarking and privacy customization techniques are used respectively. The secured privacy-preserved data is send to the DLD to prevent from data-leak. The proposed method provides higher security and integrity to the sensitive data for Healthcare Organizations in P2P network.

Keywords— P2P network, Data leaks, security, Message Digest 5, privacy

I. INTRODUCTION

The data leakage is one the big challenge for the many industries, organizations and institutes. Data loss has become frequent in the scenarios where the data gets exposed to public and it causes damages of millions of dollars to the different organizations. The human mistakes, misconfiguration of software, inadequate application functionality etc over the internet are some of the examples which causes information leakage. The data leakage happens every day when confidential business information such as customer or patient data, price lists, intellectual property and trade secrets, forecasts and budgets in spreadsheets are leaked out. Information leakage causes threats for different organization.

In P2P network [1], the peers are registered to a network. The registered peers to a P2P services can share and exchange information with other peers. They can exchange the information without servers using P2P networks. In P2P file sharing, internal privacy data leaks are frequent over the P2P network. The major security issue of the P2P file sharing system is the privacy of sensitive data from data leakage.

Healthcare organizations [11] maintain rich repositories of patient information and this data must be protected from both accidental and intentional disclosure. Patient Personally Identifiable Information (PPII), Protected Health Information (PHI), patient financial and insurance/payment information, and even sensitive documents such as physician peer reviews are often rich targets for medical identity theft. Additionally, when patient data need to be exchanged with business partners electronically, its security should be maintained.

Data leaks may occur due to many different reasons. E-mail, Web applications, Printed Documents, Instant messaging can be some of the data leak channels examples. In Day-to-day life most of the data is shared electronically. Most of the organizations and institutes also use these channels in their day-to-day life for exchanging or sharing their sensitive data. So, there is more possibility of data-leaks while sharing or transmitting the data in the network.

The DLD Systems are used for detecting the data leakage of the sensitive data. But the DLD may also attempt to gain the knowledge of the sensitive data as the Healthcare organizations do not want directly reveal the sensitive data to the provider. So, the sensitive data must be secured before sharing with the DLD systems to prevent from data leakage.

The proposed Host-Specific System protects sensitive data leakage. In Host-Specific System, the data owner of Healthcare system uploads the data. The uploaded data is preprocessed and the fuzzy fingerprints are created by the host. The K-Anonymity algorithm will be used for avoiding the data leakage. The fuzzy fingerprints are created using the MD5 algorithm. Watermarking will be used for more security as well as integrity of the data in the P2P network. In Watermarking technique, unique code is embedded in each copy which acts as a watermark. Privacy will be provided to the sensitive data by privacy customization.

II. RELATED WORK

The author's Xiaokui Shu and Danfeng (Daphne) Yao [2] proposed a technique of fuzzy fingerprints using Rabin Algorithm. In this paper, they presented a privacy preserving data-leak detection solution to solve the problem where a special set of sensitive data digests is used in detection.

Chaol-Joo Chae, YongJu Shin, Kiseok Choi, Ki-Bong Kim and Kwang-Nam Choi proposed a prevention method for data leakage in P2P networks[3]. The proposed system identifies the sensitive or private data, and removes the information by the privacy data leaking risk factor from the sharing file. The proposed method provides higher performance and security.

Xiaokui Shu, Jing Zhang, Danfeng (Daphne) Yao and Wu-Chun Feng [4] proposed a method called Rapid and Parallel Content Screening method which is used for

detecting transformed data exposure. There approach consists of sampling algorithm and a corresponding alignment algorithm working on preprocessed n-grams of sequences. The pair of algorithms computes a quantitative similarity score between sensitive data and content. This method is useful for big data analytics and for detecting transformed sensitive data exposure.

The Fang Liu, Xiaokui Shu, Danfeng (Daphne) Yao and Ali R. Butt [5] proposed a method for privacy preservation of sensitive data with MapReduce. MapReduce has the ability to arbitrarily scale and utilize public resources for the task, such as Amazon EC2. The data privacy protection is realized with fast one-way transformation. The transformation requires pre and post-processing by the data owner for hiding and precisely identifying the matched items, respectively. Before sensitive data is given to the MapReduce nodes for the detection, both the sensitive data and the content need to be transformed and protected by the data owner.

Jinhyunh Kim, Jun Hwang and Hyung-Jong Kim [6] proposed a Data Leakage Prevention (DLP) system. DLP system is the monitoring system that uses a network packet for monitoring the system's information. The DLP system may monitor some part of private information and it could a privacy violation. They considered two cases of privacy violation level. One is static privacy violation level which can be calculated simply using private data portion of monitoring target. The other is dynamic privacy violation level which is calculated using currently monitored private data portion. In addition, by removing some part of private keywords, they can increase the privacy violation level.

Sandip .A Kale and S.V. Kulkarni [7] proposed a data detection model for secure data transmission. Perturbation is a technique where data is modified. Before handling the data to the agents, the data is made "less sensitive" and then transmitted. They developed a model for assessing the guilt of agents. In this model, the objects are distributed among the different agents to improve our chance of identifying the leakers. Finally, they also consider the option of adding "fake" objects to the distributed set. Such objects do not correspond to real entities but appear realistic to the agents. This fake objects acts as a type of watermark for the entire set, without modifying any individual members.

Jason Croft Matthew Caesar proposed a network wide method for confining and controlling the flow of sensitive data in the network. There approach is based on black-box differencing. They run two logical copies of the network; one with private data scrubbed, and compares outputs of the two to determine if and when private is leaked [12]. They introduced the concept of a paired packet to allow both copies of the process to send data onto the network to allow the sharing of sensitive data within the confines of the network.

III. PROPOSED SYSTEM

The DLD system does not provide security and privacy of data. So, there is need of security, privacy and as well as integrity of the data which is to be shared in the P2P network. By providing solution to these problems, the issue of sensitive data leakage can be overcome.

The proposed system is the Host-Specific. In this system, the owner will upload the healthcare data. Then the preprocessing will be done in which the noise and null values will be removed. Data leakage will be avoided by using K-Anonymity [13] algorithm. For providing the security and integrity fuzzy fingerprints and watermarking techniques will be used. Privacy of most sensitive data will be maintained by privacy customization. The main objectives which will be covered in proposed system are as follows:

- To preprocess the uploaded data and protect the data using K-Anonymity algorithm.
- To generate signature based fuzzy fingerprints using MD5 algorithm.
- To maintain integrity and privacy using Watermarking and Privacy customization respectively.
- To detect data-leaks using DLD and evaluate results.

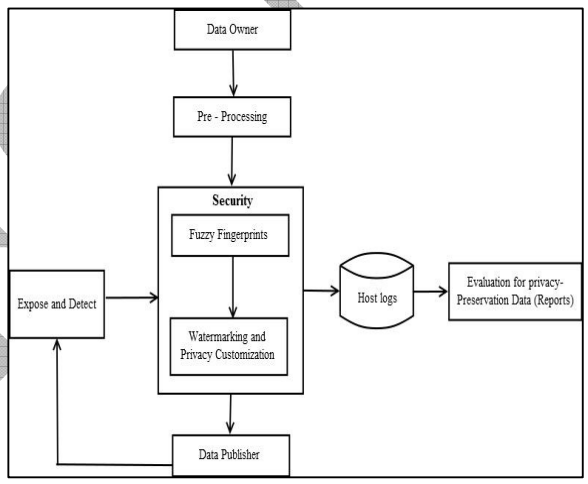

Fig 1.1
Phase-1: *Preprocessing*

The data owner uploads the sensitive data on the system shown in the fig 1.1.The uploaded data will be preprocessed as the data may come from many sources in the P2P network. In preprocessing the empty and noisy data will be removed. Then data leak firstly must be avoided. So, K-Anonymity algorithm will be used here for avoiding the data-leaks.

In K-Anonymity algorithm, some part of the data are suppressed and generalized so that the leaked data cannot be identified. Quasi identifiers are the identifiers for e.g. name, phone number, pin code etc which attacker may use it for identifying the sensitive data. For identifying accurate quasi identifiers, the data holder can identify attributes in his private data that may also appear in external information.

Definition: K-Anonymity

Let RT(*A1*,...,*An*) be a table containing attributes *A1*,……,*An* and *QIRT* be the quasi-identifier associated

with it. RT is said to satisfy *k*-anonymity if and only if each sequence of values in RT [*QIRT*] appears with at least *k* occurrences in RT [*QIRT*][8].

Table 1.1 [8]

Race	Birth	Gender	Zip	Problem
Black	1960	Male	12345	Asthama
Black	1975	Female	12341	Short breath
Black	1967	Male	12342	Hypertension
White	1964	Male	12345	Chest pain
White	1970	Female	12341	Hypertension
White	1972	Female	12342	Chest pain

Table 1.2 [8]

Race	Birth	Gender	Zip	Problem
Black	1960	Male	1234*	Asthama
Black	1975	Female	1234*	Short breath
Black	1967	Male	1234*	Hypertension
White	1964	Male	1234*	Chest pain
White	1970	Female	1234*	Hypertension
White	1972	Female	1234*	Chest pain

Example 1: Table adhering to *k*-anonymity, where k=2 and QI= (Race, Birth, Gender, ZIP) [8].

Table 1.1 provides an example of table T that consists sensitive data. Table 1.2 provides an example of a table T that adheres to *k*-anonymity. The quasi-identifier for the table is QIT= {*Race*, *Birth*, *Gender*, *ZIP*} and *k*=2. Therefore, for each of the tuples contained in the table T, the values of the tuple that comprise the quasi-identifier appear at least twice in T. That is, for each sequence of values in T[*QIT*] there are at least *2* occurrences of those values in T[*QIT*]. In particular, *t1*[QIT] = *t2*[QIT], *t3*[QIT] = *t4*[QIT], *t5*[QIT] =*t6*[QIT], *t7*[QIT] = *t8*[QIT] = *t9*[QIT], and *t10*[QIT] = *t11*[QIT] **[8]**.

Phase-II: Signature-Based fuzzy fingerprints generation.

In this stage, the fingerprints are generated from the processed data. The fingerprints are generated by the system using Message Digest5 (MD5) algorithm [9]. MD5 is an algorithm that is used to provide security .It is created through the creation of a 128-bit message digest from data input (which may be a message of any length) that is claimed to be as unique to that specific data as a fingerprint. This fingerprint is unique for each individual. MD5 processes a variable-length message into a fixed-length output (i.e.128 bits).

Fig 1.2 shows the working of MD5 algorithm. MD5 algorithm operates on 128-bit state. It is divided into four 32-bit words and denoted as A, B, C and D. These are initialized to certain fixed constants. The algorithm works on each 512-bit message block. Each block modifies its state while processing. The processing consists of four stages: rounds, non-linear function F, modular addition and left rotation.

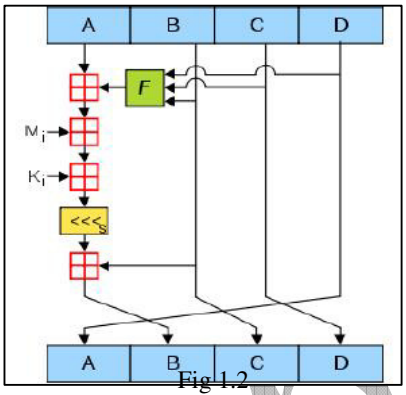

Fig 1.2

Phase-III: Providing Security and Privacy

It includes two stages: Watermarking and Privacy Customization. Watermarking is used for providing more security and integrity to the sensitive data. In watermarking there are two types: Visible and Invisible watermarking. Watermarks are easily visible in visible watermarking. Hence, cannot provide strong security. In invisible watermarking, the watermark is not visible to anybody. So, invisible watermarking will be used for providing security and integrity in proposed system. Invisible watermarking may contain use of unique code embedded with sensitive data which act as a watermark or it can use images etc.

Privacy customization is used when there is need of giving more privacy to the information or data. The Healthcare data may contain most private information which should not be leaked. Hence, for this privacy customization should be done.

It means privacy should be set for most private or sensitive data. It can be done by encryption with shared key mechanism.

Phase-IV: Data Leakage Detection.

The DLD is used for detecting the data leakages while sharing the sensitive data of Healthcare organizations in P2P networks. The secured and privacy preserved data is forwarded to the DLD system. The DLD system will detect the data leakage and will send the report to the host. The reports generated will show data leaked while sharing.

IV. APPLICATIONS

A) E-mail Spam Detection

In today's world, E-mail spam is very serious problem. As, we open our mail there are many e-mails related to jobs, offers, discounts related to products which are fake. They may lead to data leaks of our personal information. The spammers make money from these unsolicited messages. Hence, the personal information from these spammers should be prevented. The proposed system helps to overcome the problem. It checks the message header, content of the message and verifies it.

B) Online Transaction

Nowadays, online transactions are widely done on network. While transactions, personal information of user like debit card number, password etc is submitted over the network from merchant to third parties and from third party to bank's website. The information of user flows from one user to another user. Hence, the data-leaks may occur while transformation of sensitive data. The proposed system can help for detecting the data leaks over the network.

V. CONCLUSION

In P2P network data is shared between different nodes. Data Leakage is very important issue while sharing the data in network. So, the need arises of preventing the data from leakage. The proposed system will use fuzzy fingerprint technique for security, watermarking for checking data integrity and privacy customization for maintaining privacy. The DLD system extracts the leaked content while transfer of data over the network between peers. The proposed system helps from data leaks of Healthcare organizations in P2P networks while maintaining its security, privacy and integrity.

REFERENCES

[1] V.S.Motade and Prof. Deepak Gupta "Content Leakage Detection by Using Traffic Pattern for Trusted Content Delivery Networks" 2014.
[2] Xiaokui Shu and Danfeng (Daphne) Yao "Privacy-Preserving Detection of Sensitive Data Exposure" 2015.
[3] Cheol-Joo Chae & YongJu Shin & Kiseok Choi & Ki-Bong Kim & Kwang-Nam Choi "A privacy data leakage prevention method in P2P networks" Springer,2015
[4] Xiaokui Shu, Jing Zhang, Danfeng (Daphne) Yao and Wu-Chun Feng "Rapid and Parallel Content Screening for Detecting Transformed Data Exposure" 2015.
[5] Fang Liu, Xiaokui Shu, Danfeng (Daphne) Yao and Ali R. Butt "Privacy-Preserving Scanning of Big Content for Sensitive Data Exposure with MapReduce" 2015.
[6] Jinhyung Kim, Jun Hwang and Hyung-Jong Kim "Privacy Level Indicating Data Leakage Prevention System" 2012.
[7] Sandip A. Kale and Prof. S.V.Kulkarni "Data Leakage Detection" 2012.
[8] Latanya Sweeney "*k*-ANONYMITY: A model for protecting privacy".
[9] Sandesh D Manocharya1, A Prabhakar "Detection and Avoidance of Data Leakage" 2012.
[10] J. Croft and M. Caesar, "Towards practical avoidance of information leakage in enterprise networks," in Proceedings of the 6th USENIX Conference on Hot Topics in Security, serHotSec'11, 2011.
[11] Xiaokui Shu "http:// people.cs.vt.edu/~danfeng" 2012.
[12] Xiaokui Shu and Danfeng (Daphne) Yao "Data Leak Detection as a service" 2012.
[13] B.Kohil and K.Sashi2 "Data Leakage Detection using K-Anonymity algorithm" 2012.