Data loss prevention by using MRSH-v2 algorithm

Basheer Husham Ali, Ahmed Adeeb Jalal, Wasseem N. Ibrahem Al-Obaydy
Computer Engineering Department, College of Engineering, AL-Iraqia University, Iraq

Article Info

Article history:

Received Mar 6, 2019
Revised Dec 15, 2019
Accepted Jan 11, 2020

Keywords:

Approximation matching
Data breach
DLP
Mrsh-v2
Sensitive data

ABSTRACT

Sensitive data may be stored in different forms. Not only legal owners but also malicious people are interesting of getting sensitive data. Exposing valuable data to others leads to severe Consequences. Customers, organizations, and /or companies lose their money and reputation due to data breaches. There are many reasons for data leakages. Internal threats such as human mistakes and external threats such as DDoS attacks are two main reasons for data loss. In general, data may be categorized based into three kinds: data in use, data at rest, and data in motion. Data Loss Prevention (DLP) are good tools to identify important data. DLP can do analysis for data content and send feedback to administrators to make decision such as filtering, deleting, or encryption. Data Loss Prevention (DLP) tools are not a final solution for data breaches, but they consider good security tools to eliminate malicious activities and protect sensitive information. There are many kinds of DLP techniques, and approximation matching is one of them. Mrsh-v2 is one type of approximation matching. It is implemented and evaluated by using TS dataset and confusion matrix. Finally, Mrsh-v2 has high score of true positive and sensitivity, and it has low score of false negative.

Corresponding Author:

Basheer Husham Ali Al-Mafrachi,
Computer Engineering Department,
College of Engineering, AL-Iraqia University,
Baghdad, Iraq.
Email: Basheer.husham@aliraqia.edu.iq

1. INTRODUCTION

In the past, the electronic communication among people was so difficult. They were depended only on post office to exchange or share data or package. However, modern technology has fixed this problem. Different kinds of devices are available nowadays such as laptops, computers, iPad, iPods, mobiles, etc. The internet contributes in pushing wheel of development. By using internet, people can do several kinds of activities to serve their needs such as shopping using online websites, sending e-mails among them, studying online, electronic reservation and so on. Public and private organizations, universities, hospitals, or/and companies store, send, or receive different kinds of data.

These data may be texts, books, images, or videos. They may be represented in different kinds of form such as txt, pdf, doc, gif, jpg, mp4, xls, ppt, or exe. These data can be stored in different kinds of devices, online storage, or cloud computing. Most of these forms contain sensitive information such as full names, full home addresses, social security numbers, mobile phones, e-mail addresses, credit card numbers, or date of birth. Not only legal owners but also malicious people are interesting of getting these data to serve their needs. Malicious people have used advanced activities to get these information by many ways such as virus programs, junk mails, spywares, or ransomwares.

Exposing these valuable data to others leads to severe dangerous. Customers, organizations, and /or companies lose their money and reputation due to data breaches. They lose billions of dollars, reputations, brands. There are many examples about that. One example is that more than 650 malicious attacks were reported before 2013. Based on report of Open Security Foundation (OSF), more than one thousand incidents of data loss happened during 2013 [1]. Moreover, more than two hundred million of login information were compromised before 2012 based on Symantec company report [2]. In addition, more than 165 million of documents were stolen during 2011 according to Verizon company [3]. Finally, more than 60 million of sensitive files were breached from Sony company, and more than 5 million of login data were exposed for LinkedIn users during 2012 [1]. Because of these dangerous, researchers have designed many useful tools against data breaches. Data Loss Prevention (DLP) considers as one kind of data breach prevention. It is used to identify and prevent important data to fall in the wrong hands [4]. This tool can be used to protect sensitive data at rest which is data stored in end user or system such as hard disk or/and removable disk. It can also be used to identify, monitor, or protect data in use which is data transmitted in the network. Protecting data in motion is also one goal of DLP [5].

There are many kinds of security tools other than DLP that can be used in this regard such as Intrusion Prevention Systems (IPS) and Intrusion Detection Systems (IDS). Although they consider as security tools, they have main difference. DLP is responsible only for capturing and identifying sensitive information. However, IPS and IDS is in charge of all kinds of dangerous or threat that may face data. DLP has two main phases to identify important or sensitive information. The first one is generating fingerprints or predefined patterns which are based on extracting certain features from known files. The second is comparing the fingerprints of new files with the existed fingerprints that are derived from the first stage. Finally, if result of comparison was positive, detected file may be encrypted, blocked, removed, or transferred to safe place [6, 7].

Furthermore, DLP tools was first used in 2006. They are not a final solution for data breaches, but they consider good security tool to eliminate malicious activities and protect sensitive information [8]. In the market, DLP may have other names such as information monitoring and prevention, information loss protection, data analyzing and prevention, and/ or data leakage prevention [5]. There are many famous companies have developed DLP tools. For example, Palo Alto Network company produced a security tool to protect data in motion by analyzing, monitoring, and detecting sensitive information that are transmitted in the wire or wireless. AmXecure Company released security design to identify data leakage that is called PrivacyID [7]. RSA and McAfee have the best DLP security tool to identify information leakage as stated in [9, 10]. Finally, Websense and McAfee organizations designed DLP tool that has three stages which are data control, data endpoint, and data identification [11, 12]. Finally, the rest of the paper is organized as the following: section II explained the popular reasons that lead to data leakage. Furthermore, section III presented categories of where sensitive data can be stored. Moreover, solution methods are presented in detail in section IV. The implementation and evaluation of mrsh-v2 in section V and VI respectively. Finally, conclusion is presented in the in the last section.

2. POPULAR REASONS FOR DATA LOSS

Famous organizations and companies lose their reputational and billions of dollars due to sensitive information breaches. There are two primary causes of data breach: external and internal threat.

2.1. Internal threats

First of all, internal threats consider as a primary cause of data leakage. Human mistakes are at the top of this type of threat. There are many mistakes that can be done by people. For instance, negligence of employees who leave their computers, mobiles, iPads, or other devices in public transportation or places such as restaurant, markets, and/ or stores cost their companies a lot of money if these devices fall in the wrong hands. Removable devices such as flash disk or disk driver that are left in internet café are consider from this type of threat. More than forty percent was the rate of data loss due to stolen computers. As in [13], these kinds of problems lead to almost more than forty five percent of data leakage in the healthcare sector [14]. According to the statistics, human errors was the main causes of information leakage during 2014 [15].

According to the last survey that was accomplished by the Group of Healthcare Organizations and Ethics Association which represent a collection of expert researchers in this major, they stated that almost forty percent of data breaches incidents happened due to misplaced documents such as important files. They also showed that almost 30 percent of data leakages incidents caused by lost removable devices [16]. As stated also in [16], counsel public office for Massachusetts state obligated Goldthwait which is one of the healthcare charging company to pay more than one hundred and thirty thousand bucks as a fine. This is done because police officers found USB disk belong to one employee of their company that contains more than 65000 sensitive records that are related to their patients in public trash.

In addition, texting messages and e-mails is another form of information leakage. Staff who work in different kind of places such as hospitals, schools, universities, companies, or/ and organizations may send important documents that contain sensitive data outside boundaries of their workplace by using texting applications such as Gmail, Yahoo, and/or Hotmail. They can use other kinds of communication application such as Viber, Facebook, WhatsApp, Telegram, or Instagram. They may send that to wrong destination and these types of errors named as miscellaneous mistakes [14]. Finally, this can lead to data breaches. Furthermore, staff of certain kind of organizations may leave certain job and start a new job with another. They may take sensitive data of their previous job with them and expose that to others or their new job. According to the statistics, almost fifty percent of staff take confidential data of their previous job when they leave [5]. Atul Malhotra was in charge of giving important data from his previous job which is IBM to his new one which is HP [5].

Staff of certain companies believe that their important documents can be demolished when they deleted them. In other hands, these files can be recovered by using certain programs and can be used against their companies unless provider remove them permanently. For instance, every device such as computers, scanners, printers, or phones have memory inside of them. The data in memory can be recovered even if users remove them. Finally, using weak algorithms help employees who have bad intention to take data. Incorrect setting for organization system induces attackers to do their malicious activity such as unauthorized access to the system according to Version report in 2008 [17, 18].

2.2. External threats

Not only internal threats but also external threats have high effect on data breach. Attackers would like to have data to exploit their owner in order to induce them to pay money for them. Therefore, they have developed sophisticated application to deceive users of large organizations such as healthcare companies, financial, other business organizations. Point-of-Sale intrusions (POS) were one type of external threats.

POS is kind of malicious activities that can take and gather people visa, credit, or dept cards data at checkout market. Sensitive data such cards numbers, expiration date, and passwords can be collected in one file by attackers. According to the statistics that have been done, almost more than 10 gigabytes had been breached at checkout of Target markets which are series of largest stores in USA. Almost forty million of credit, dept, visa cards data were stolen at checkout when customers swipe their cards at checkout. Seventy people names, date of birth, and address were also stolen due to that malicious acts [14, 19]. Finally, not only Target stores but also Neiman Marcus markets was targeted by POS attacks. Two thousand credit card data were breached by this attack [19].

Moreover, another kind of external dangerous is crimeware. Attackers may deceive users to install malicious applications in their electronic devices without their knowledge in order to get their data. Attackers can do that by sending a poison e-mail to victims that contain malicious applications. Victims may also visit compromised websites by mistake. As soon as victims download and install the malicious application in their devices, attackers can spy, monitor, or phishing data [20]. Finally, there many other external threats lead to data breaches. One of them is Distributed Denial of Services (DDoS) attacks. Statistics showed that DDoS has increased in the last decade [14]. DDoS attacks overload servers with very large number of packets to shutdown services to legitimate users. Based on report of Worldwide Infrastructure Association, seventy five percent of bank data leakage were caused by DDoS attacks [21].

3. PROTECTION OF IMPORTANT INFO

Data is very important not only for legitimate owners but also for attackers. Many DLP vendors presented threats that face data wherever stored. Vonto is one of them. It stated that almost two out of eight hundred texts that are transmitted in the air may comprise sensitive data. Two out of hundred network messages imply private data. Vonto mentioned also in their report that eight out of ten companies lost their data stored on their laptops. In addition, two out of four companies lost their information stored on removable drivers [8]. Important data can be stored in different kind of categories such as data in motion, data at rest, and data in use. Finally, in the next three subsections, these categories are going to be presented in detail.

3.1. Data in motion

Data in motion is information that are transmitted in the network whether through wireless or wire. This data may be electronic books, word documents, excel documents, pictures, videos, voices, or power point slides. Attackers may intercept these communications between two legal users in order to steal people data. They can do that by developing malicious applications. They also may do that by exploiting vulnerabilities in network algorithms, network application or protocols. DLP solutions are very important to monitor network packets that are transferring in the channel. DLP provides feedback or report to the admin of network if there are any malicious activity. Therefore, administrators can take action such as blocking, filtering, encryption to the data.

3.1.1. Network monitor

DLP can do many tasks as mentioned earlier. One of them is network monitor. Passive monitor and active monitor are two kinds of network monitor in general. In specific, monitor of DLP considers as a passive one. Active monitor can deploy in both client and server sides. Server side can get full report about transmission medium such as number of packets, loss of packets, throughput, bandwidth, protocol type, delay time, source IP address, destination IP address, destination port address, and source port address. On other hands, passive monitor may be as a device that can be deploy only in one side (either on client side or on server side) [22]. It is used as packet sniffer to gather information in order to analyses or examine packets or flows to identify malicious activities. This can increase level of confidentiality, performance, and efficiency of the network. DLP likes Intrusion Detection System (IDS) in identifying malicious act and notifying person who in charge of network. On other hands, IDS is not designed to avoid data leakage [23]. Finally, network monitor of DLP can be deployed near router or switch that connected all devices together to control all incoming and outcoming packets [24].

3.1.2. FTP protocol and E-mail

E-mail is the most common way to send data through the internet among people around the world. Users can transfer different kinds of files, and it is on top priority of data leakage. Users can receive poison links, photos, or other documents. They also may get executable program such as botnet. As soon as users download these files in their devices, their important data may be in danger. Aa a result, DLP solution is an important way to protect e-mail contents by few methods.

One way is that some DLP tools obligate e-mail applications to attach small files size instead of large files size. This would guarantee that employee can not sent large size of important data to external parties. Furthermore, DLP tools can also notify persons who in charge of network when attacks happen. Finally, data can be encrypted or even blocked if there are any suspicion by using DLP tools.

Important documents can be sent by using another method which is file transfer protocol (FTP). This protocol face security challenges. Data may be changed or breached on server side when malicious persons invade this protocol. For instance, important data related to American army that are available in Iraq were exposed by Associated Press because of lake of security for FTP protocol [25]. Another consequence of FTP vulnerability happened when almost eight thousands of records that belong to SAIC were exposed as stated in [25]. DLP tools face FTP problems, but it is not enough to eliminate this problem. Therefore, Managed File Transfer (MFT), which can be used to transfer documents safely, may work with DLP to decrease dangerous of FTP completely [26].

3.1.3. Filtering, bridge, and blocking solutions

Another action that can be taken after detection of data leakage by using DLP is blocking data from being breached. Packets that carry data which are not identified as sensitive can be passing through the DLP. However, those packets that has important information can be blocked from passing through. Blocking, filtering, installing bridge are all ways that can be used in this regard.

First of all, bridge is a network device that can be used to connect computer devices to form a network. It can be used also to connect internal with outside networks. Bridge device can do deep content analysis for the packets that passing through it and stop transferring packets in case of finding important data [5]. This device collects incoming packets to form flow table. Each flow is a group of packets that has same characteristics such as IP source address, IP destination address, Mac source address, and Mac destination address. Bridge depends mainly on recording and storing incoming destination Mac addresses and source Mac addresses in its table. After a while, this device can get enough information to decide which traffic may be blocked based on Mac addresses. This leads to increase security level [18].

Proxy server is device that has high characteristic level such as high microprocessor speed, high random memory access (RAM), and large amount of storage space. This device may deploy between internal network and external network to do deep inspection. DLP can get packets that passing through proxy server for analysis. Internet Content Adaptation Protocol (ICAP) that run in the proxy can send a copy of packets flow to identify sensitive information. Finally, DLP can do full inspection and analysis for incoming packets as mentioned earlier. It can take action when sensitive data detected such as blocking, filtering, and encryption. This tool may break the communication between two sides. For example, this can be done by sending packet that called TCP Rest (RST) to break the connection [5, 27].

3.2. Data at rest

Data at Rest is kind of inactive data that may be not be used at the time in the system. It may be stored in different kind of forms such as word document, spreadsheet excel, electronic books (pdf), power point slides (ppt), videos, voices, images, or other kinds of file types. These data can be stored in different kinds of devices such as laptops, computers, database of servers, workstations, internal or external hard drives, tape drivers, cloud system, phones, iPads, or iPods [4].

3.3. Data in use

The best description for data in use is that any kind of data that may be used when system is active. In other words, all data that users can deal with when they use their devices. For instance, data that are available in Random Access Memory (RAM) may consider as data in use. This is because RAM is empty when system is off. However, as soon as users start up system and run programs, data can be uploaded to RAM. Another example, Central Processing Unit (CPU) has memory that contains a few kinds of small registers. Each register has special tasks such as storing temporary results, pointers, memory location addresses, number of increments, and number of decrements. These data consider as data in use which are active during system work. Data stored in off-line memory that is used when system is active such as DVD, CD, and Blue ray consider as data in use. Data stored in floppy disk, internal hard drive, external hard drive, removable disks are all kinds of data in use if they used in time of system execution. data stored in office application such as word, excel, power points, or outlook are all type of data in use. Data that are written in terminal execution of programs such as Java, C, C++, Python, or MATLAB may be also from this type [4, 5]. DLP is a good solution for this kind of data. For example, applying constraint on machines to prevent data loss is one solution of DLP. Putting limitation on using programs that let users transmitting important information outside their devices. Finally, DLP tools obligate employers to have limited access to data content in order to eliminate data leakage [5].

4. TECHNIQUES FOR DLP

DLP tools may use different and various methods to do data content analysis. According to SANS Institute in [28], seven types of methods can be used to implement DLP tools. First of all, the most popular type that can be used to implement DLP tools called regular expression or rule based. This method is based on looking for specific sensitive information such as social security numbers, users' names, e-mail addresses, dept or visa card numbers, or phone numbers. This technique is suitable for identifying patients' records, employees' records, electricity bills, phones' bills, hospital bills, or bank statements. However, it is not appropriate for detecting images, videos, or voices. False positive rate is going to be high. In other word, amount or rate of data that are detected by using this method and do not match with original sensitive data is high.

In addition, another method for identifying important data is database fingerprinting. This method is based on looking for a collection of important information with the data that are available in the database. This group might be any set of data such as dept card numbers and full names, full names and phone numbers, or e-mail addresses and card numbers. This method is fitting on identifying a selection of sensitive data. Amount or rate of data that are detected by using this method and do not match with original sensitive data is high. In other words, this method produces low false positive rate. However, this approach has problem that is similar to the previous one which is not suitable for identifying unstructured data such as video, image, or voices.

Moreover, another approach that used for DLP is exact file matching. This method is based on finding fingerprint signatures to document that has important data. the next step is comparing these fingerprint signatures with new files to find matching. It can be used with all document's kinds. It also has low false negative rate. Mrsh-v1, ssdeep, and sdhash algorithms can be used to implement this method. Furthermore, partial document matching is another technique. This approach is searching for incomplete or complete matching with sensitive data. Rolling hash method is an example about this approach. It is effective in detecting text data. However, it is not suitable for videos, photos, or voices.

In addition, another technique is statistical method. This method is based on mathematical equations and statistics. Bayesian approach can be used to implement this method. This method is suitable for very big dataset. However, it generates high false positive and high false negative. It also required massive dataset to produce accurate results.

Conceptual/ lexicon method is another technique that can be used to implement DLP. This approach is a collection of rules and dictionaries that can be used to find suspicious behavior and detecting important data. This method is appropriate for detecting sexual harassment, private trading by using work account, and illegal practice of stock exchanges. However, it generates high false positive. Finally, the last method is categories. This technique is a collection of previous methods that can be used to implement a model for DLP. For example, looking for specific e-mail address, specific username, and one full name. Finally, exact file matching is the best approach according to reports [28]. Thus, mrsh-vs is an algorithm that implemented in the next section to show the capabilities of this method.

5. IMPLEMANTATION

As mentioned earlier that approximation files matching is one technique that can be used to identify data breaches. It works on identifying leakage files such as pdfs or words. This technique has two phases. The first one is generating fingerprints or signatures which are based on extracting certain features from known files. The second is comparing the fingerprints of new files with the existed fingerprints that are derived from the first stage. The results of comparison fall within the range of 0 and 100. In other words, when the probabilistic results of comparing two files are high, the files are similar to each other. However, when these ratios are low, files are not similar to each other. This paper implemented and evaluated the Mrsh-v2 (Multi-resolution similarity hashing) which is one algorithm is based on approximation matching technique [6, 29].

In this paragraph, first stage of this algorithm is explained in detail. The input that contains sequence of bytes (b, b2,.., bn) can be divided and grouped into several window. Each window has 7 bytes in length. Then, these windows (w1,w2,…,wm) are grouped by using idea of rolling hash algorithm. This idea is simply based on removing the first element from old window and inserting a new element to form a new chunk. The size of elements in chunk (end of chunk) is determined by calculating pseudo random function PRF (for each chunk) and chunk size (c). if PRF (for certain element) == -1 mod c, this means stop adding new elements to the chunk. Otherwise, adding new element is keep going to form the chunk. Each chunk then is hashed by using FIN algorithm. The main goal is to have 160 bytes for the chunk size and 0.5 for the fingerprint size [6]. The idea of bloom filter was used to implement the second phase. Bloom filter is the method for answering set membership. It depends on set of values (input) and independent hash functions. Let us imagine that there are set of elements denoted by E, and all elements (n) are set initially to be false. In addition, independent hash function (Hf) that may contain a set of hash functions (Hf1,Hf1,…,Hfn) gives us random numbers between 0 and n-1. For each element e in the set of E, result of Hf(e) is going to be the index number (position) of E, and we are going to set to true the element of that position.

The second stage is comparing the fingerprints of new files with the existed fingerprints. To implement this stage, two bloom filters are used E1 and E2. e1 and e2 is going to be number of bits that are set to be true in E1 and E2 respectively. The number of bits that set to be true in common is (k=e1∩e2). To compare E1 and E2, k is going to be compared with certain kind of score (S). In such a way, if k is larger than S, then the similarity score is high. Otherwise it is going to be 0. S can be computed based on the maximum (Max) and minimum (Min) number of bits that interfere by chance between E1 and E2. Therefore,

$$S = \beta * (Max - Min) + Min \tag{1}$$

where β =0.3 based on better experiment. Where Max can be defined as in (2):

$$Max = min\,(e1, e2) \tag{2}$$

However, Min can be calculated as in (3).

$$Min = n * (1 - q^{(Hf * e1\$)} - q^{(Hf * e2\$)} + q^{Hf\,(e1\$ + e2\$)}) \tag{3}$$

Where n is the number of bits of bloom filter and Hf is the size of hash functions as mentioned earlier in the description of bloom filter. e1$ is amount of element for Bloom filter E1, and e2$ is the number of elements for E2. Finally, q is the rate that some kind of bit still false or zero in the Bloom filter when we add a new element and it can be defined as in (4).

$$q = (1 - 1/n) \tag{4}$$

Therefore, we can use (5) to find the similarity ratio (Sim) between two Bloom filters:

$$Sim = \begin{cases} 0, & k \le S \\ \dfrac{100(k - S)}{Max - S}, & Otherwise \end{cases} \tag{5}$$

6. EVALUATION

Mrsh-v2 algorithm was evaluated by using confusion matrix [29, 30]. Confusion matrix has many kinds of metrics that can be used in the evaluation. True positive (TP), false positive (FP), false negative

(FN), true positive rate (TPR) or sensitivity or recall, false positive rate (FPR) or fall-out or probability of false alarm, and false negative rate (FNR) or miss rate are some of these metrices that were used. TP means the file identify correctly by the algorithms. In other words, the identified file is similar to the one that is stored in database. Sensitivity or recall is the amount or rate of files that have similarity with files that are existed in the database. FP means that mrsh-v2 detects files, but they are not existed in the database. Fall-out or probability of false alarm is the amount or rate of files that are detected by the algorithm and do not have a corresponding one in the database. FN means that mrsh-v2 identifies files that are not available in the database. Miss rate is the number of files that are detected by this algorithm, but they are not available in the database.

Mrsh-v2 was evaluated in the network environment and by using TS dataset. TS dataset which is publicly available online. This dataset contains several kinds of files such as pdf, exe, doc, gif, xls, ppt, and txt. Approximately three thousand files from TS dataset were transferred in network by generating almost two hundred and ninety thousand packets. 249670 out of 290314 packets that carried different kinds of files which are available in the dataset were detected correctly by mrsh-v2 algorithm as shown in Table 1. However, 40643 packets that transferred different kinds of files were identified by mrsh-v2, but they are not available in the dataset. In other hand, 29031 packets that have files were not discovered by mrsh-v2 although those files are available in the dataset. Finally, the reset of Table 1 shows the amount of packets for all kinds of files in details based on the term FP,TP, and FN.

Finally, 0.85 out of 1 was the value of TPR or sensitivity as shown in Figure 1. This means that there is a high amount of different kinds of files was detected correctly by mrsh-v2. However, 0.14 and 0.1 out of 1 was the amount of FPR and FNR respectively. This means mrsh-v2 produces low amount of errors in general for different types of files. In specific, Mrsh-v2 algorithm detects jpg, gif, and pdf files correctly because TPR is high, and FNR and FPR are low as shown in Figure 1.

Table 1. Values of FP, TP, and FN

File Type	FP	TP	FN
jpg	31934	278701	8709
gif	29031	281604	5806
doc	40643	214832	72578
xls	52256	235154	52256
ppt	40643	267088	20321
pdf	34837	278701	11612
txt	40643	252573	34837
exe	52256	243863	43547
total	40643	249670	29031

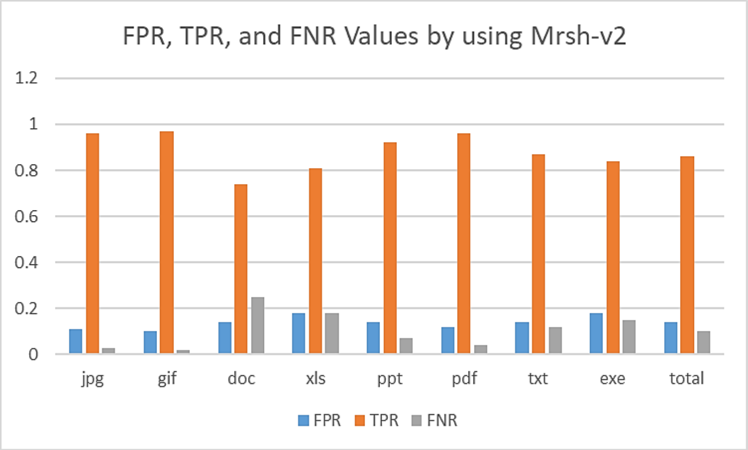

Figure 1. FPR, TPR, and FNR values by using Mrsh-v2

7. CONCLUSION

It is very significant to protect data whatever they store. Data can be stored in three main different kinds: data in motion, data at rest, and data in use. Data is important not only for legitimate owners but also for attackers. Data Loss Prevention (DLP) are good tools to identify sensitive data. DLP can do analysis for data content and send feedback to administrators to make decision such as filtering, deleting, or encryption. There are many kinds of DLP techniques. The best one is approximation files matching. Mrsh-v2 algorithm considers as an example about this approach. This algorithm was implemented and evaluated by using publicly available TS dataset. Confusion matrix results showed that this algorithm has high TP, and TPR. In other hands, mrsh-v2 has low FP, FN, FPR, and FNR.

REFERENCES

[1] "2019 Data Breach Investigations Report," verizon, 2019. [Online]. Available: https://enterprise.verizon.com/resources/reports/2019-data-breach-investigations-report.pdf.

[2] D. Antoniades, *et al.*, "Accurate Traffic Categorization," *Proceedings of IST Broadband Europe*, pp. 1-6, 2006.

[3] W. Ashford, "DDoS is the Most Common Method of Cyber-Attack on Financial Institutions," 2016. [Online], Available: https://*Computer Weekly.com,*

[4] F. Breitinger, and I. Baggili, "File Detection on Network Traffic Using Approximation Matching," *Journal of Digital Forensics, Security & Law,* vol. 9, no. 2, pp. 23-35, 2014.

[5] J. Beeskow, "Reducing Security Risk using Data Loss Prevention Technology," *Journal of The Healthcare Financial Management Association,* pp. 108-112, 2015.

[6] F. Breitinger and H. Baier, "Similarity Preserving Hashing: Eligible Properties and a New Algorithm MRSH-v2," *International ICST Conference on Digital Forensics and Cyber Crime,* pp. 167-182, 2012.

[7] B. Blevins, "Best of Data Loss Prevention," *Information Security,* pp. 13-15, 2014.

[8] A. Burroughs, "Data Breaches Cause Worry," *Smart Business Orange County,* 2015.

[9] A. Cecil, "A Summary of Network Traffic Monitoring and Analysis Techniques," *Computer Systems Analysis,* pp. 4-7, 2006.

[10] S. Gumaste, *et al.*, "Proxy Server Experiment and the Behavior of the Web," *International Journal of Advanced Research in Computer Science,* vol. 4, no. 1, pp. 84-87, 2013.

[11] "Data Loss Prevention Keeping your Sensitive out of the Public Domain," *EY*, 2011. [Online]. Available: http://www.ey.com/Publication/vwLUAssets/EY_Data_Loss_Prevention/$FILE/EY_Data_Loss_Prevention.pdf

[12] L. Grandia, "Nine Key Cyber Threats Identified in Verizon Data Breach Report," *Health Management Technology,* vol. 35, no. 6, 2014.

[13] "The Practical Executive's Guide to Data Loss Prevention," *Whitepaper*, pp. 2-17, 2019. [Online]. Available: https://cdw-prod.adobecqms.net/content/dam/cdw/on-domaincdw/brands/forcepoint/whitepaper-practical-executives-guide-data-loss-prevention-en.pdf,

[14] "Internet security threat report," *2019trends*, *Symantec, Inc.,* vol. 24, 2019. [Online]. Available: https://www.symantec.com/content/dam/symantec/docs/reports/istr-24-2019-en.pdf.

[15] J. Jaeger, "Human Error, Not Hackers, Cause Most Data Breaches," *Compliance Week,* vol. 10, no. 110, pp. 56-57, 2013.

[16] R. Hiesh, "Improving HIPAA Enforcement and Protecting Patient Privacy in a Digital Healthcare Environment," *Loyola University Chicago Law Journal,* vol. 46, no. 1, pp. 175-223, 2014.

[17] "McAfee Total Protection for Data Loss Prevention," *Solution brief, McAfee*, 2019. [Online]. Available: https://www.mcafee.com/enterprise/en-us/assets/solution-briefs/sb-total-protection-for-dlp.pdf.

[18] H. Tuttle, "Hacking Away at the Bottom Line," *Risk Management,* 2014.

[19] L. Musthaler, "The True Cause of Data Breaches," *NetworkWorld Asia,* pp. 6-6. 2008.

[20] S. Naidu, "Data in Motion-Securing Businesses on the Go," *SDA Asia Magazine,* pp. 46-48, 2009.

[21] N. Wynne and B. Reed, "Magic Quadrant for Content-Aware Data Loss Prevention," *Gartner Research,* 2013.

[22] A. Papadogiannakis, *et al.,* "Improving the Performance of Passive Network Monitoring Applications using Locality Buffering," *IEEE Xplore,* pp. 151-157, 2007.

[23] K. Košt'ál, *et al.,* "Management and Monitoring of IoT Devices Using Blockchain," *The Association of Digital Forensics, Security and Law (ADFSL),* pp. 1-12, 2019.

[24] L. Strauss, "Data breach study: Criminal attacks now leading cause," *J. of health care compliance,* pp. 61-63, 2015.

[25] "Survey Cites Human Error as Biggest Cause of Data Breaches," *Magazine Article-Information Management,* 2015, [online]. Available: https://www.questia.com/magazine/1G1-436228015/survey-cites-human-error-as-biggest-cause-of-data.

[26] "The GoAnywhere book of secure file transfer project examples," *GoAnywhere manag. file transf.,* pp. 1-33, 2019, [online]. Available: https://www.infosecurityeurope.com/__novadocuments/585230?v=636906728355170000.

[27] J. Wu, *et al.,* "Keystroke and Mouse Movement Profiling for Data Loss Prevention," *Journal of Information Science and Engineering,* pp. 23-42, 2015.

[28] R. Mogull, and LLC. Securosing, "Understanding and Selecting a Data Loss Prevention Solution," *Technicalreport, SANS Institute,* 2007.

[29] V. Gupta, "File detection in network traffic using approximate matching," *MS Thesis. Institutt for Telematikk,* 2013.

[30] K. Ting, "Confussion Matrix," *Encyclopedia of Machine Learning and Data Mining. Springer, Boston,* 2017.