Data Exfiltration Techniques and Data Loss Prevention System

Conference Paper · December 2019

DOI: 10.1109/ACIT47987.2019.8991131

4 authors, including:

Hamzeh Kilani
Princess Sumaya University for Technology
4 PUBLICATIONS **3** CITATIONS
SEE PROFILE

Mohammed Nasereddin
Princess Sumaya University for Technology
3 PUBLICATIONS **3** CITATIONS
SEE PROFILE

Ali Hadi
Champlain College
37 PUBLICATIONS **146** CITATIONS
SEE PROFILE

Data exfiltration techniques and Data Loss Prevention system

Hamzeh AlKilani, Mohammed Nasereddin, Ali Hadi and Sara Tedmori
Department of Computer Science
Princess Sumaya University for Technology
Amman, Jordan
hamzeh_k8@hotmail.com, mohamedsgn8020@gmail.com, ali@ashemery.com, S.Tedmori@psut.edu.jo

Abstract— **One of the primary concerns of data security specialists is to mitigate insider threats and prevent data leaks. Often, unfortunately, insider threats go unnoticed. In most cases, the longer such activity goes unnoticed, the greater the resulting damages are likely to be. This paper provides an overview of the basic data exfiltration techniques that deal with file structures and were utilized in multiple scenarios in an attempt to bypass a Data Loss Prevention system. Details show which of the scenarios have been detected and which have not been by the Data Loss Prevention system. The paper also proposed solutions for the undetected scenarios.**

Keywords— Data Exfiltration, Data Loss Prevention, DLP, Insider Threat, Sensitive Data

I. INTRODUCTION

Organizations around the globe are collectively creating staggering amounts of data, which they need to store and protect against theft, loss, and misuse. Whether this data is stored in-house or on the cloud, it is subject to data leak threats. These threats can originate from malicious insiders who have authorized access to organization assets, or from malicious outsiders who are not authorized to access such assets **[1].**

Insider employees often cause some of today's most damaging data leak threats. In 2018, studies have shown that 53% of organizations confirmed insider attacks in the previous 12 months**.** In addition, 27% of organizations say that insider attacks have become more frequent **[2]**. Insider threats are often more severe than outsider threats, because they are usually more difficult to detect **[3]** and are time intensive **[4]**. The difficulty of detecting internal threats can be attributed to the nature of such threats, which are at many times the result of simple human mistakes that do not intentionally threaten the security of organizations data **[6].** The European General Data Protection Regulation (GDPR), which has become mandatory in 2018, has forced the deployment of exfiltration prevention mechanisms [**5**]. Hence, it is essential for organizations to take the necessary measures to protect against such threats in order to prevent their customers, suppliers and investors from losing confidence in them **[7].**

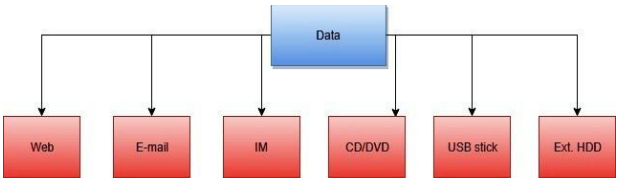

Figure 1. DLP monitored channels

Data loss prevention (DLP) came into use in 2006 and gained some popularity in 2007 **[8]**. DLP is a set of technologies, products, and techniques designed to help prevent sensitive information from leaving an organization. **[9].** A DLP system includes a set of rules and policies that classify data according to its type to ensure that it is not maliciously or accidentally shared. The system monitors end-user activities, data flow, as well as data sent over the network. If any suspicious activity is detected, a system alert is triggered. **[10]**

In order to detect confidential data in the organization, DLP systems are typically implemented using various mechanisms, such as Described Content Matching (DCM), exact data matching, indexed document matching, vector machine learning, content matching classifiers, and file type detection **[11].** Each type of detection mechanism provides unique capabilities to detect DLP violations in a message or document.

In this research, a DLP system dependent on DCM is used to test the scenarios. The system detects content and context using keywords, regular expressions, file properties in addition to other pre-defined data identifiers that are used to reduce false positives. Moreover, the proposed DLP system is capable of monitoring multiple destinations (channels) such as removable storage, CD/DVD, printer/fax, email, web protocols, and network shares **[12]** as shown in Fig. 1.

The rest of the paper is organized as follows: an overview of DLP related works is provided in section II. Section III describes the proposed methodology. Possible scenarios with their results are displayed in section IV. Finally, the conclusion is presented in section V.

II. LITERATURE REVIEW

Raman et al **[13]** described some problems of DLP solutions and identified three related challenges; namely, encryption, access control, and semantic gap. The authors stressed the importance of acting against these challenges. The

authors proposed Test Clustering and Social Media Analysis as potential DLP research directions. The authors also addressed the current state of DLP solutions that suffer from signature misuse detection techniques and do not have the capabilities to detect complex data leak scenarios. Finally, the authors encourage authors to focus on real-time data leak detection when developing DLP solutions.

Tahboub et al **[14]** presented an overview of the approaches used to protect data from external attacks, such as the use of intrusion detection systems, anti-malware, and firewalls. Such intrusion detection systems are devices or programs that monitor and control incoming and outgoing data transmitted over the network using a set of predetermined security rules (access rules) for the purpose of protecting the network from unauthorized access. The authors compared the afore mentioned approaches with a DLP system that protects sensitive data such as personal credit card data scores from internal threats. The authors also highlighted that DLP systems fail to read both encrypted files and data hidden within images. The authors proposed the development of algorithms to alleviate the protection level in DLP systems. They also proposed extending the system's services to smart phones.

Lopez et al **[15]** proposed an evaluation of DLP systems based on the criteria of defined sensitive information, evasion data file extension, DLP characteristics like performance, capabilities, user experience and policy violations related to sensitive data such as regular expressions, dictionary lists and fingerprinting, to give better understanding of the technical features of the DLP. The authors also explained the importance of DLP systems as enforcement tools to make sure that employees comply with the data security policy.

Praba et al **[16]** provided a survey about data leakage detection and prevention approaches. The authors discussed current DLP standards and the process of detection/prevention which contains the three phases of data collection, data analysis and the remedial action phase and how they correlate with each other. Furthermore, the authors summarized the techniques of data leak prevention and detection systems such as applying policy and access rights, cryptography approaches, watermarking techniques, behavioral and data analysis, test mining, etc. Finally, the authors identified seven main challenges that face DLP systems. The first challenge is the data transaction that may happen in many channels which are difficult to deal with. The second challenge is the data modification which may make it difficult to detect the partial and full data leakage. The third challenge is to determine and set appropriate access rights for specific users. The fourth challenge is the process of encrypting the data before leaking which is difficult to detect their content. The fifth challenge is to use the watermarking concept, because its contents are easily recoverable, causing many challenges. The sixth challenge, the scalability and integration for vast domain is certainly impossible, which creates a scalability issue. The last challenge is the problem of detecting data leakage from the log, which needs a complete learning process.

Kaur et al **[17]** described some of the challenges that need to be solved by DLP systems, such as encryption, access control, new data and customization, and social network (when a person belongs to one or more groups or when new groups are formed. This causes the disappearance of the old groups, so it is difficult to detect a person who leaks data to the outside or has limited data access). They also discussed the main DLP approaches that are used to build and characterize the model's specification, and learning approach as well as taking into consideration DLP limitations. The authors concluded their paper by offering several suggestions that will help DLP systems to detect and prevent data leakage. In addition, they proposed future activities that need to be followed to prevent data leakage in organizations such as using secured communication protocols, restricting the access to send emails externally for specific users and specifying what are the systems they can access and finally, to monitor organization data on employee's smartphones.

Polozova et al [**18**] defined the different terms related to data leak prevention technology and emphasized on the risk of malicious insiders and how the internal violators can act in collusion to bypass the defense of the DLP system. The authors proposed a solution to keep the organization's information secure by building and developing a threat model that takes into consideration the security of the DLP system itself from intentional and unintentional threats. The proposed threat model was built in line with GOST R 51275-2006 "Data Protection. Informatization object. Factors affecting the information. General provisions".

III. METHODOLOGY

As we have seen in the literature review section, most of related works are theoretical without addressing the real world experiences on the performance of the DLP System in different cases. This research presents the results of practical experiments that have been conducted on several scenarios to test and determine the ability of DLP systems in detecting leaked confidential data.

There are different types of computer file formats that can be used for different purposes. The most commonly used file formats are Microsoft Word documents (DOCX), Portable Document Format (PDF), Joint Photographic Experts Group (JPEG) **[19].** In this research, the authors applied data exfiltration techniques using both Microsoft office word and Adobe's PDF file format that are commonly used by organizations.

There are varying techniques in which data exfiltration can be conducted, over time these techniques are becoming more and more sophisticated in an effort to stay one step ahead of data security solutions. Some of the techniques are basic, while others are more advanced and require tools in order to be applied on the targeted files. In this research, the authors will apply a number of scenarios that deal with encryption, merge streaming, and editing file binaries. Follows is a description of the data exfiltration techniques that were used in this research for purposes of highlighting the weaknesses in the Symantec DLP version 14.6.

Nowadays, encryption is widely used to protect communication, personal and confidential information. In addition, encryption can also be used for unauthorized movement of data as shown in the first and second scenarios. In the first scenario, a Windows version of GNU Privacy Guard tool was used to encrypt the test files using the software Kleopatra; while in the second scenario, the files were encrypted using WinRAR 256 AES encryption technology.

In the third, fourth and fifth scenarios, merge streaming techniques were used to simply compress, change and manipulate the extensions of the test files.

In the sixth scenario, the split archive feature was used on the test files. The method of splitting an archive file is generally used to assist with large downloads by splitting the file into smaller pieces of a size that the user specifies.

In the seventh scenario, Command Prompt application which is available in most Windows operating systems was used to append the confidential test file into non-confidential file using TYPE command to display the contents of the file into another file.

In the eighth scenario, hex editor software was used to manipulate the file's binaries with the formats of (PDF, DOCX) using Winhex software. The process requires to open the test files in the hex editor and to delete the beginning of each file in a process referred to as deleting the magic number [**20**].

A. Implementation/Experimentation

Various types of test files can be used when evaluating DLP systems to check the detection features after utilizing the DLP rules and data identifiers. For this research, two sample file types were created (*confidential.pdf*) and (*Ratesheettest.docx*). The pdf file contains the word *confidential*, while the docx file contains rate sheet information like customer name, account number, address, and phone number. Two DCM rules were created, confidential tag rule and rate sheet rule that matches on defined keywords to detect content using key words, key phrases; regular expressions to detect characters, patterns, and strings; file properties to detect files by type, name, size.

It is worth mentioning that DLP solutions by default provide a variety of built-in rules and policies or they can be created from scratch by the administrator to protect user-defined confidential and classified documents. The DLP agent will trigger an alert on the test documents if there is an attempt to leak them outside the organization using any of the channels mentioned in Fig1.

B. Testing Environment

Although there is an open source DLP solution [**21**], a commercial solution Symantec DLP 14.6 [**22**] was used to test the techniques and it was configured on the monitor mode and not on the prevention mode which means that if any policy is violated, incident will be raised but there will not be any blocking. This way the security team can work in a retroactive way when receiving a DLP incident. The reason behind this is to minimize the negative impact on the productivity of people working in the organization especially if the solution is applied in the medium/large organization.

The machine used for testing has Windows 10 operating system, with no restrictions on copying data to removable storage device or sending emails to out of domain (external) recipients. In addition, the installed DLP agent has the capability of detecting the incident whether it was inside the organization network or outside the organization network.

IV. RESULTS

No.	Technique
1	Encrypt the files using GNU Privacy Guard (GPG).
2	Files added to archive file, then encrypt the file with password.
3	Compress the files into zipped file and change the extension to .docx (Merge Streams).
4	Change the extension of the files, compress them and then change the extension for the archive file to .docx.
5	Files renamed, and then compressed.
6	Add the files to archive and use the split archive option.
7	Use Type CMD command prompt to append the confidential file into non-confidential file.
8	Delete the magic number [19] of the files using Winhex.

Table 1. The eight scenarios used to test the DLP system.

In the first scenario, the files *confidential.pdf* and *Ratesheettest.docx* were compressed into the file *Encrypted privateK.zip*. Then, the file was encrypted into the file *privatek.zip.gpg* using GPG encryption by the software Kleopatra. After that, the encrypted confidential file was copied to an external removable storage. The encrypted file *privateK.zip.gpg* bypassed the DLP system and the file was not detected. However, the solution was to modify the detection rule on the file properties identifier to detect .gpg files. After that, the same file was copied to the removable storage but this time, the file was detected by the system.

In the second scenario, the confidential test files were added to an archive file, then set password option was used to encrypt the file, after encryption was complete, the file was sent to an external email address. This time the file was not detected by the DLP system. However, to detect password protected or encrypted files, the DLP admin must choose the Password-Protected ZIP Archive option and apply it on all DLP rules.

In the third scenario, the confidential files were compressed into a zipped file *Embeddedfiles.zip* Then the extension for the .zip file was changed to .docx (Merge Streams) and the file was sent to an external email address (Not the organization domain). The DLP agent detected both confidential files within the zipped file even when the zipped file extension was changed from.zip to .docx.

In the fourth scenario, the extension for both confidential files (docx, pdf) was changed to the .jpg extension. The files were then compressed into a zipped file and the extension for the zipped file was changed to .docx. The zipped file was sent to an external email address. In this scenario, the files also got detected even their extensions were changed to .jpg and the DLP system showed that it can identify the file's original format (PDF, DOCX).

In the fifth scenario, both confidential files were renamed, compressed and then sent to an external email, where they got detected by the DLP system. Also, the authors tried adding them to an archive file, renamed the files within the WinRAR software itself from .zip to .docx but the file got detected as well.

In the sixth scenario, the split zip file option was used to split the file into six smaller files after the confidential files were compressed into one archive file. However, the DLP system triggered an alert after sending the split files to the external email address.

In the seventh scenario, TYPE CMD command was used to append the file *confidential.pdf* to non-confidential file *template.docx*, the latter file was then sent to an external email address. Also this time, the file was also not detected by the DLP system.

In the last scenario and after deleting the magic number for the PDF file format, the DLP system did not detect the file after copying it to removable storage, but for the DOCX file format, the DLP system has detected the keywords without resolving the original file name as it was replaced by (word/document.xml).

V. CONCLUSION AND FUTURE WORK

In this paper, a set of exfiltration techniques that deal with file structure were presented, the testing process explained that there are valid techniques that could be used to bypass the Symantec DLP system like deleting the magic number in documents, encrypting the archive file with a password or using TYPE CMD command. Statistics show that there is an increasing number of reported data leak incidents around the globe and the importance of using a data loss prevention solution, a successful DLP deployment requires constant attention and it is very important to consider customizing the default rules and configuration before applying the DLP system as per the business environment and user behavior.

In our future work we plan to test the proposed scenarios on various DLP systems and to add other more techniques based on cryptography and steganography algorithms to check if we can bypass the DLP system.

REFERENCES

[1] Dark Reading, Data Breach Threats Bigger Than Ever, (2018, November 28) from https://www.darkreading.com/vulnerabilities---threats/data-breach-threats-bigger-than-ever/a/d-id/1333332

[2] CA Technologies, 2018 Insider Threat Report, (2019, February 22), from https://www.ca.com/content/dam/ca/us/files/ebook/insider-threat-report.pdf

[3] R. F. Trzeciak. 2017. SEI Cyber Minute: Insider Threats. (2018), from http://resources.sei.cmu.edu/library/asset-view.cfm?assetid=496626

[4] Claire Kirk, Ten Cybersecurity Tactics to Protect Against Insider Threats, (2018, October 22), form https://www.lightedge.com/blog/protect-against-insider-threats/

[5] https://eugdpr.org/the-regulation/ GDPR (April, 2019)

[6] Nena Giandomenico, Juliana de Groot, Insider vs. Outsider Data Security Threats: What's the Greater Risk?, (2019), from https://digitalguardian.com/blog/insider-outsider-data-security-threats

[7] Bahar Yasin, Zehra Bozbay, "The Impact of Corporate Reputation on Customer Trust", 16th International Conference on Corporate and Marketing Communications, At Athens, Greece (2011)

[8] Sans Institute 2008, Data Loss Prevention, (2018, December 2), from https://www.sans.org/reading-room/whitepapers/dlp/paper/32883

[9] Cisco, What is data loss prevention, (2018, December 3) https://www.cisco.com/c/en/us/products/security/email-security-appliance/data-loss-prevention-dlp.html

[10] Ellen Zhang, What is Data Loss Prevention (DLP)? A Definition of Data Loss Prevention, (2019), from https://digitalguardian.com/blog/what-data-loss-prevention-dlp-definition-data-loss-prevention.

[11] Symantec, (2019, March 13), from https://www.symantec.com/content/dam/symantec/docs/data-sheets/data-loss-prevention-solution-en.pdf, Page 2.

[12] Symantec, (2019, March 13) https://www.symantec.com/content/dam/symantec/docs/data-sheets/data-loss-prevention-solution-en.pdf, Page 3.

[13] Preeti Raman, Hilmi Güneş Kayacık, Anil Somayaji "Understanding Data Leak Prevention", Annual symposium on information assurance (Asia), June 7-8, 2011, Albany, NY.

[14] Tahboub R., Saleh Y., "Data Leakage/Loss Prevention Systems (DLP)", World Congress on Computer Applications and Information Systems (WCCAIS), 2014 IEEE.

[15] López G.; Richardson N.; Carvajal J. "Methodology for Data Loss Prevention Technology Evaluation for Protecting Sensitive Information", Revista Politécnica - Septiembre 2015, Vol. 36, No. 3.

[16] Mercy Praba, G. Satyavathy "A Technical Review on Data Leakage Detection and Prevention Approaches", Journal of Network Communications and Emerging Technologies (JNCET) Volume 7, Issue 9, September (2017).

[17] Kamaljeet Kaur, Ishu Gupta and Ashutosh Kumar Singh, "A Comparative evaluation of data leakage/loss prevention systems (DLPs)", 3rd International Conference on Artificial Intelligence and Soft Computing. (2017)

[18] Ekaterina Polozova, Nataliia Anashkina "Analysis of information security threats for developing DLP-systems", Production Engineering Archives 17 (2017) 25-28.

[19] What are the most common file types and file extensions?, (May, 2019) , from https://www.computerhope.com/issues/ch001789.htm

[20] What is magic number, (May, 2019) , from http://www.linfo.org/magic_number.html

[21] Comodo Dome DLP, (May, 2019) , from https://www.mydlp.com/

[22] What's New in Data Loss Prevention 14.6, (May, 2019) from https://www.symantec.com/content/dam/symantec/docs/data-sheets/whats-new-in-dlp14-6-en.pdf