sustainability

MDPI

Article

Research on Behavior-Based Data Leakage Incidents for the Sustainable Growth of an Organization

Jawon Kim [1], Jaesoo Kim [2] and Hangbae Chang [3,*]

[1] Department of Convergence Security, Chung-Ang University, Seoul 06911, Korea; jjawon@cau.ac.kr
[2] TL/IT Security Team, SK Hynix, Icheon, Gyeonggi-Do 17336, Korea; jaesoo777.kim@sk.com
[3] Department of Industrial Security, Chung-Ang University, Seoul 06911, Korea
* Correspondence: hbchang@cau.ac.kr

Received: 4 June 2020; Accepted: 31 July 2020; Published: 2 August 2020

Abstract: With the continuously increasing number of data leakage security incidents caused by organization insiders, current security activities cannot predict a data leakage. Because such security incidents are extremely harmful and difficult to detect, predicting security incidents would be the most effective preventative method. However, current insider security controls and systems detect and identify unusual behaviors to prevent security incidents but produce many false-positives. To solve these problems, the present study collects and analyzes data leaks by insiders in advance, analyzes information leaks that can predict security incidents, and evaluates risk based on behavior. To this end, data leakage behaviors by insiders are analyzed through an analysis of previous studies and the implementation of an in-depth interview method. Statistical verification of the analyzed data leakage behavior is performed to determine the validity and derive the levels of leakage risk for each behavior. In addition, by applying the N-gram analysis method to derive a data leakage scenario, the levels of risk are clarified to reduce false-positives and over detection (i.e., the limitations of existing data leakage prevention systems) and make preemptive security activities possible.

Keywords: insider threat; data leakage; security data analysis

1. Introduction

Security attacks that threaten the wellbeing of organizations are changing in various ways, including cyber-attacks. While cyber-attacks originating from outside an organization occur mainly to acquire important data, security attacks can also come from inside the organization. According to HfS Research statistics from a survey on data leaks administered to executives in corporate security departments, 69% had experienced insider leaks, and 57% had experienced outsider leaks [1]. For example, in the case of Google's autonomous vehicle project Waymo, after the main employee who executed the project left the company, he founded a startup company and sold the trade secrets of his previous company to other companies. This exemplifies how mainstream security attacks have changed from being caused by outsiders to being caused by insiders, but the countermeasures implemented by organizations have not evolved from existing cyber-attack frameworks to adjust to this change. Most organizations rely on the installation of security systems such as intrusion detection systems and control removable storage media to engage in security activities around the boundaries. However, there is a limit to detecting and preventing data leakages by insiders because the criteria are ambiguous for determining whether an act in the system is legitimate or not. Furthermore, data leakage by insiders is characterized by a high magnitude of damage because insiders know important data about the organization, and these data are freely accessible to the attackers [2]. To minimize damage to the organization, it is most effective to prevent security incidents caused by insiders and to predict and manage security risks in advance. In this study, to secure the sustainable growth of the organization, we collect in advance the signs of data leakage by insiders of the organization, analyze the signs of data leakage so that the accidents can be predicted, and attempt to derive the levels of risk according to those signs.

This study consists of six sections. First, the introduction (above) explains the research importance and background. In Section 2, we consider the characteristics of data leakage incidents and the limitations of existing data leakage prevention solutions. In Section 3, we explain the research methodology, which was intended to overcome the limits of existing data leakage prevention solutions. Through in-depth interviews and an analysis of earlier studies, we derive the data leakage behaviors from insider attacks. In Section 4, through a security expert survey, we statistically verify the derived data leakage behaviors and apply the N-gram method—a text mining analysis method—to derive the data leakage scenarios. In Section 5, a summary of the research results is given, which continues into Section 6, where we discuss the research contents, contributions, and future work.

2. Data Leakage Accidents and Protection Technology

2.1. Characteristics of Data Leakage Accidents

Most of the documents produced by organizations today are in the form of electronic files; thus, Information Communication Technology (ICT) technology is located in the organization's infrastructure. In addition, various industries have launched strategies to digitize data that previously existed in writing through digital transformation. According to a Smart Insight report, 76% of organizations have implemented digital transformation [3]. Through digital transformation, it becomes more convenient to use data for various reasons, such as effective operation management and productivity improvement. Due to their focus on using the data generated by introducing digital transformation, organizations have been limited in their data protection activities. However, when data in an electronic file format is leaked, it is difficult to identify the leaker due to the anonymity that is possible in cyberspace. Although organizations establish and operate regulations for preventing data leakages, current security policies have limitations in implementing effective security activities. Previous studies [4,5] that analyzed the limitations of existing security policies showed a strong impetus to protect the boundaries of organizations but an inability to prevent security attacks that occur and are fused. Further, given the rise of security incidents caused by humans, the shortcomings of human management illustrate the ineffectiveness of existing security policies.

Leaking data in the form of electronic files does not mean the original data loses its form or content. If an electronic file is captured and leaked, the organization will have difficulty determining if information leakage has occurred because the file remains present despite the primary data in the file being leaked. In particular, there is difficulty in determining whether the actions performed by insiders are for business execution or data leakage. As shown in Figure 1, the EKRAN statistics show that more than 40% of organizations take several years to detect data leakages involving insider threats [6].

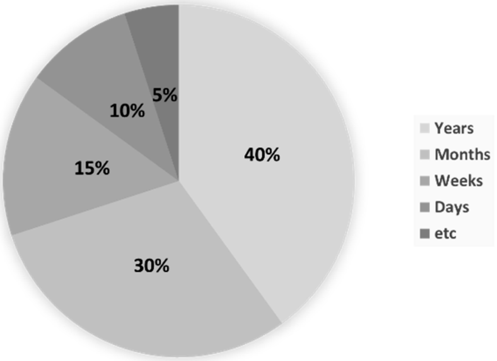


Figure 1. Insider threat-related data breach detection times.

In addition, security incidents have the characteristics of hidden crimes [7] and have a significant impact on corporate image, such as decreasing the organization's credibility when notifying the public of the security incident. Hence, most corporations do not count the number of internal data leakage incidents, so the various statistics are not representative of reality. Due to the nature of these internal data leakage incidents, statistics examining the number of security incidents are gradually decreasing, but the scale of damage caused by such incidents is steadily increasing. When important business-critical data are leaked, the damage is great. However, if the leaked information cannot be recovered or repaired, the damage is more severe. Hence, prevention before the occurrence of an incident is more important than detecting the moment of occurrence.

2.2. Security Technologies for Data Leakage Protection

To minimize the occurrence of and damage from information leakage incidents, organizations are presently attempting to develop and introduce various security technologies. There are many types of security technologies, which are difficult to classify, as many provide key functions that overlap. However, this study intends to classify these technologies into two types according to their security purpose: (1) security technologies aimed at prevention and (2) security technologies aimed at detecting incidents through continuous monitoring. We reconstructed the data leakage prevention framework proposed by Alzhrani [8] according to the criteria established in this study, as shown in Figure 2. Security technology for prevention can be subdivided into leakage control technology that grants responsibility to users and access control technology that limits user access. Security incident detection through continuous monitoring includes technology that detects abnormal behaviors and continuously monitors and filters the content generated and transmitted in the business. These security technologies are also capable of managing and detecting the behaviors of the user and each terminal, as well as behaviors through unreliable channels.

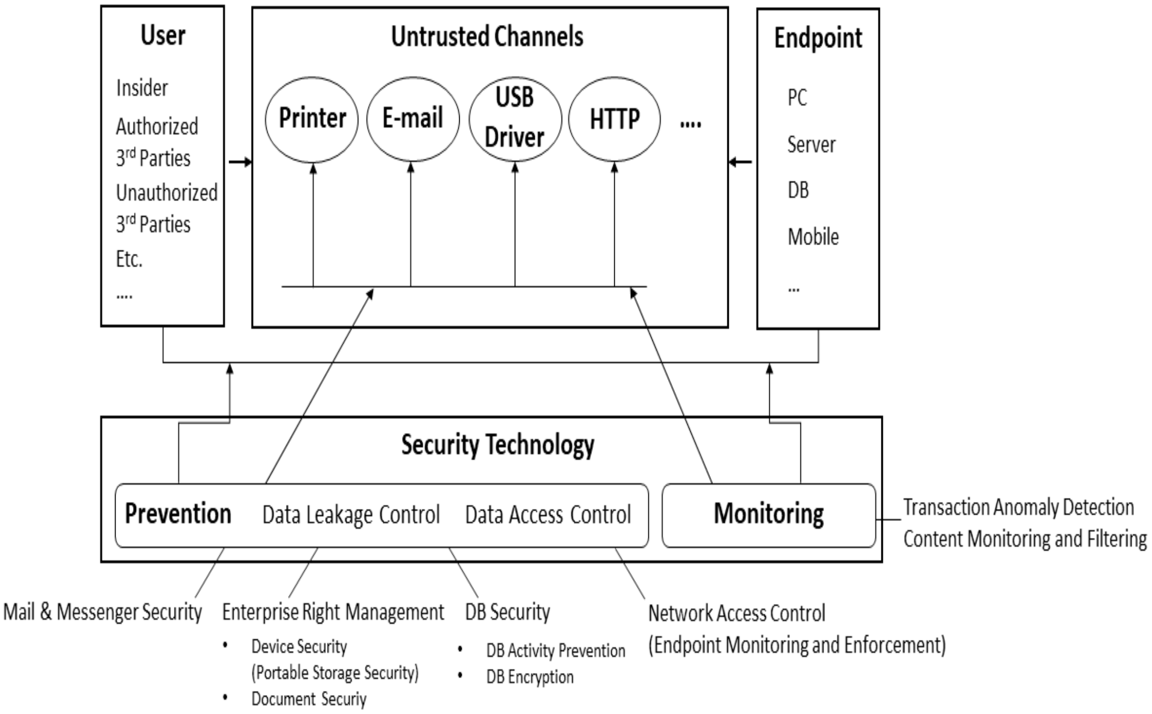


Figure 2. Security system classification based on the security purpose.

As shown in Figure 2, all security technologies classified and identified by the security objectives focus on detecting and identifying the point at which a data leakage occurs, while the ability to predict security incidents before a data leakage incident occurs is lacking [9]. For this reason, some security technologies have been developed to predict data leakage security incidents. However, these technologies predict security incidents by detecting abnormal behaviors rather than by identifying and detecting leakages. Since it is impossible to determine whether a behavior performed by an organization member is for work or for data leakage purposes, behaviors are classified into normal or abnormal, where the latter is judged as data leakage. For example, a behavior is considered abnormal if an employee goes to work early because there is more work than usual. This causes problems because the rate of over-detection is too high when identifying all abnormal actions as information leakage actions. To solve this problem, the present study increased the predicted level of data leakage security incidents by identifying and detecting the behaviors that actually cause information leakages, not simply abnormal behaviors inside the organization, and, ultimately, the number of information leakage security incidents.

3. Collection and Analysis of Data Leakage Incidents

3.1. Research Methodology

In this research, the methodology presented in Figure 3 was used to collect data leakage behaviors and analyze the risk of behaviors. The data leakage behaviors were primarily investigated through an analysis of previous research related to data leakage incidents. Following the investigation, we conducted in-depth interviews with security managers who had experienced incidents in their organizations and collected data leakage behaviors. To perform statistical verification of the collected data of the leakage behaviors, a validity assessment targeted to security experts and a risk assessment by behaviors were conducted. We used the N-Gram analysis method for data leakage behaviors that were statistically verified to design hypothetical scenarios with data leakage potential. The application of statistical methods to elicit survey targets, processes for collecting information leakage behaviors, and the risk of information leakage are detailed in Sections 4 and 5, respectively.

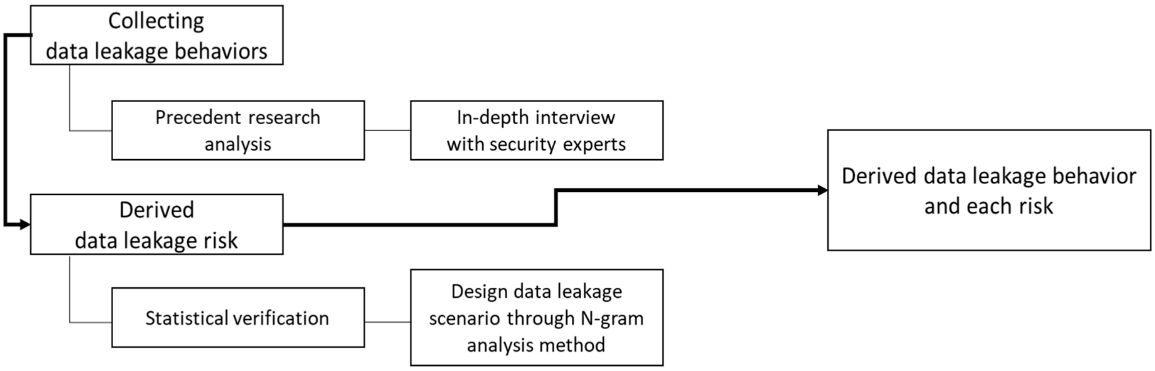

Collecting data leakage behaviors
Precedent research analysis
In-depth interview with security experts
Derived data leakage risk
Statistical verification
Design data leakage scenario through N-gram analysis method
Derived data leakage behavior and each risk

Figure 3. Research methodology.

3.2. Colloection of Data Leakage Behaviors

To collect the data leakage behaviors, we analyzed prior research. We focused on previous studies that used scenarios of security incidents involving insider threats [10–16]. As mentioned in the previous studies, the analysis of leaked behaviors through past research has limitations because only a small proportion of security incidents and data leakage behaviors are disclosed due to the hidden nature of the crime. Therefore, in this study, the in-depth interview method was used to collect the data leakage behavior that occur in industrial sites. The in-depth interview method is an interview method in which a small number of interviewees related to the study are interviewed using questions and answers based on previously prepared questions.

Since the in-depth interviews were conducted among a small number of experts related to the research, it was important to select the interviewee targets before starting the research. In this study, using the purposive sampling method, security experts with experience in data leakage security accidents were selected as targets from among relevant professionals who had performed security tasks for more than 10 years. The five security expert groups (in which the security teams belonged to companies that had experienced security incidents—32 of the security team members) selected for the survey were interviewed over three months, from May to July 2019. Prior to conducting the in-depth interviews, it was not possible to disclose the affiliations or for the interviewee to be investigated because each organization's security tasks were confidential. In addition, the main cases of security incidents experienced by the expert groups were determined, and the major data leakage incidents were partially determined. The main contents of the in-depth interview are shown in Table 1.

Table 1. Major contents of the in-depth interview.

(Respondent) Code	Responded Data Leakage Security Incidents (Number of Cases)	Major Contents of the In-Depth Interview
A	1	- Data leakage incidents due to the downloading of key documents (65 secret/confidential documents) through a VPN (Virtual Private Network) - Forensic investigation of the terminals (PCs, laptops) used for downloading the documents - Checking the commuting time - Looking into the transportation records of the laptop - Inquiring into storage registration - Investigating SW (software) that bypasses security policy - Inquiry and confirmation via the print history of the documents
B	1	- Data leakage incidents through abnormal access and a fabricated inspection of the internal system - Accessing and inquiring into personnel information through abnormal URL (Uniform Resource Locator) access to the internal system - Bypassing the security system by hacking the in-house system's source code (mandatory training, changing the - Password-change date, hacking a camera by disabling the mobile security application)
C	22	- Transporting key documents through external private mail - Sending major documents outside the company - Downloading key documents by bypassing the VPN
D	1	- Outflowing the key documents using an external private mail and terminating the download (47) - Downloading major documents to a private hard disk and deleting them (11,852) - Deleting documents on a private PC
E	1	- Turning off the encryption of key documents for pre-retirees - Data leakage accidents through large-scale email transmission - Leaked data and the illegal download of primary information into a data processing unit (2 PCs)/portable storage (2 external hard drives, 1 USB)

Table 1 summarizes the main contents of the interview by dividing by respondent code for each in-depth interviewee. The main contents of the in-depth interviews listed in this table were summarized based on the main leaking behaviors and the paths in which the data leakage security incidents occurred. Investigations were also used to confirm the facts after data leakage security incidents. As a result of the in-depth interviews, a total of 26 data leakage security incidents were investigated. After the detailed leakage routes of security incidents were revealed, all investigation methods and measures were explored. After collecting the data leakage behaviors, as shown in Table 2, both the data leakage behaviors from previous research analyses and those investigated through our in-depth interview methods were synthesized. Table 2 is categorized according to the types of comprehensive data leakage behaviors into four major categories and 10 classifications of medium importance according to the behavior type. Each subdivision includes the data leakage behaviors that were investigated through the in-depth interview method and an analysis of previous studies. A list of previous studies is shown in Table A1 of Appendix A and mainly refers to prior research on detecting information leakage by insiders. Based on previous research on insider threats, the importance of insider threat detection technology was largely confirmed. Moreover, we referenced studies related to insider leakage detection that applied mining and analysis techniques using machine learning, the RNN algorithm, and big data analysis.

Table 2. Collection of data leakage behaviors by type and the results of the interview analysis.

Type	Division	Subdivision	Source
Installation	Change (install) operating system environment	Change internal IP address	A2, A3, A7
		Local administrator privilege escalation	A1
	Install illegal SW (software) (declassify security environment)	Install virtual machine SW (VMware, virtual box)	A1, A7, B
		Install SW that records the computer's real data log (Key Logger)	A1, A7, B
		SW installation to disable the security environment (bypassing the security SW)	A2, A5, B
		Install an unauthorized web browser	A1, A2, A5, A7, A9, B
Manipulate	Manipulate behavior of (key) document file	Capture key documents (screen capture + file capture)	A2, A3, B
		Rename key documents	B
		Change (modulate) the filename extension of key documents	B
		Compress key documents into fragmented files + encryption files	B
		Unencrypt key documents (document security SW)	A2, A3, B
		Divide key documents into fragmented files	B
		Decode the Water Mark on files and print	B
		Save expired key documents	A1, A2, A9
		Delete the key document file usage log (inspection + modification + print + transmission)	A1, A2, A9
Access	Overuse of key files	Over-inspection of key documents (periodical inspection)	B
		Use of forbidden query (SELECT * FROM)	B
		Over-load key documents (file download)	B
	Change access authority	Inspect key documents by abnormal access (change access control list + change approval line→self approval)	B
		Change the level of documentation (high level→low level + for reading→ for output + short inspecting period→long inspection period)	B

Table 2. *Cont.*

Type	Division	Subdivision	Source
	Illegal access (In → Out)	Officially transfer documents to an authorized external site for a long period and arbitrarily leak them	B
		Access an unauthorized website (external sites + external job search site + external information exchange website)	A1–A3, A6–A8, B
		Leak bypassing data through external wireless network (proxy server, smart phone tethering)	A5, A8, A10, B
		Transfer data between the internet and intranet	B
		Use of an unauthorized shared folder	A2, A5, A7
		Access through numerous IDs in a specific computer	A1, A8, A10
		Use of retirees' IDs outside of normal business hours	A1
	Illegal access (Out → In)	Receive key documents by accessing them through a Virtual Private Network (i.e., through encrypted tunneling)	A1, A7, A8, A10, B
		Illegal access through a Virtual Private Network by using (passing through) an insider's computer with (IT) administrator rights	B
		Access via Microsoft's Remote Desktop Protocol	B
Send	Illegal file transmission using email	Use of key words in email contents	B
		Use of unauthorized external commercial email server (not using the authorized in-house email server)	A1, B
		Forward emails using other (internal) members' private computers	A6–A10
		Send emails while attaching larger-sized files than usual	A1–A3, A6–A10
		Multiple (internal) employees (two or more) forward an email by attaching related files	A6, A8, B
		Send emails through an external commercial email server to an insider for malicious purposes	A1–A4, A8, A10
		Accessing corporate partners' personal email accounts and transferring important files; then, transferring files after re-signing in	B
	Illegal use of portable storage	Copy key documents using universal serial bus memory	A1–A3, A6, A7, A9
		Duplicate key documents to a hard disk	A1, A2, A6, A7, A9
		Duplicate key documents to a CD	A1–A3, A6, A7, A9
		Copy key documents to a smart phone	A1–A3, A6, A7, A9, B
	Arbitrary transfer data processing unit	Alteration or desorption of portable storage (hard disk drive, solid state drive)	A6, A8, B
		Arbitrary transfer of laptops (including tablets)	B
		Exceeding the transfer period of the authorized laptop (including a tablet)	B
		False reporting of the loss of the data processing unit	B

* Source A is a data leakage behavior obtained through analysis of previous studies (Table A1), and B is the data leakage behavior determined through an in-depth interview.

4. Deriving Data Leakage Risk by Conducting Statistical Verification

4.1. Validity and Risk Evaluation

Statistical verification was carried out to conduct a validity analysis and risk assessment of the information leakage behaviors derived from the previous research findings and in-depth interviews. The questionnaire was used to verify the data statistically, and the survey was conducted among 76 security experts with more than 10 years' experience in performing security tasks and data leakage experiences involving leakage behaviors. The survey was conducted using a 5-point Likert scale. In this survey, a questionnaire was conducted based on the subdivision of the data leakage behaviors and the interviewee's experience. The risk degrees of the proposed data leakage behaviors were then assessed. From the survey, only data leakage behaviors with an average score of 3.5 or higher were extracted, while those with an average of less than 3.5 were rejected. The derived statistical average represents the degree of risk for the data leakage behavior. In total, 15 data leakage behaviors with a risk of 3.5 or higher were derived, as shown in Table 3. We also calculated the standard deviation. The most dangerous data leakage behavior is "Receive key documents by accessing them through a Virtual Private Network (through encrypted tunneling)," which means downloading important documents through a network that bypasses both inside and outside of the organization. The data leakage behavior with the lowest risk was identified as "Divide key documents into fragmented files" and leaking data.

Table 3. Research results for suitability and validity.

Data Leakage	Average (Score)	Standard Deviation	Priority
Change internal IP address	3.7	0.95	14
SW installation to disable the security environment (bypassing the security SW)	4.53	0.63	3
Rename key documents	3.76	0.88	11
Change (modulate) the filename extension of key documents	4.3	0.8	6
Unencrypt key documents (document security SW)	4.21	0.74	9
Divide key documents into fragmented files	3.59	0.88	14
Decode the watermark on files and print	3.71	0.97	12
Access an unauthorized website (external site + external job search site + external information exchange website)	4.3	0.81	6
Receive key documents by accessing them through a Virtual Private Network (through encrypted tunneling)	4.68	0.56	1
Use an unauthorized external commercial email server (not using the authorized in-house email server)	4.39	0.74	5
Forward email using another internal member's private computer	3.7	0.87	14
Send email through an external commercial email server to an insider for a malicious purpose	4.63	0.57	2
Duplicate the file to portable storage (USB, hard disk)	4.49	0.74	4
Copy key documents to a smart phone	4.09	0.89	10
Arbitrary transfer of a laptop (including a tablet)	4.22	0.88	8

In addition to deriving the degree of risk, the Cronbach's alpha coefficient was assessed to confirm the reliability of the survey. The reliability coefficient of Cronbach's alpha is a value that expresses the internal consistency of a survey and determines whether the test items are composed of homogeneous factors based on the average correlation between the variables within a survey. If the same concept is based on the assumption that the results will be similar when measured by different independent measurement methods (for example, when conducting a questionnaire survey), the same answer will be given after repeatedly asking the same question in different ways. The Cronbach's alpha coefficient is reliable when it is 0.8 or higher. This study determined a high reliability of 0.893.

4.2. Design of the Data Leakage Scenario through the N-Gram Analysis Method

One of the key components of text mining is representing documents. For more effectively representing documents, the Bag-of-Words (BoW) model is commonly used [17]. This model denotes the presence or absence of a particular word in the document, and provides a simple check of the frequency of words in the document [18]. However, the BoW model does not consider the words' sequence in the sentence while N-gram analyzes the sentence based on words' sequence. N-gram is one of the text mining methods based on words' sequence. There are other methods to text classification such as Naïve Bayes, Support Vector Machines, etc. Using them alone, their performance varies greatly depending on the model variant, features used and task/dataset. However, when they were used with N-gram, the text classification results always show improved results because bigram can identify sequence [19]. From existing studies, it can be seen that they commonly use a clustering method for those with similar properties and identify similarities among the scenarios in each cluster for designing a scenario [20]. Prior to selecting an analysis method, this research conducted a pilot test by using the Naïve Bayes analysis method and Clustering method. Table 4 is the result of Naïve Bayes method and Clustering method.

Table 4. The result of Naïve Bayes and Clustering Method.

Scenario Design Method	Derived Scenarios
Naïve Bayes	① Unauthorized accessing through Virtual Private Network ② Printing important file
	① Unauthorized accessing through Virtual Private Network ② Taking device (laptop/storage device) out
Clustering	① Printing important file ② Attaching important file ③ Taking device (laptop/storage device) out
	① Sending e-mail from other seats ② Connecting on Smartphone

However, in the case of the Naïve Bayes analysis method, a scenario with an unrelated action regarding a technology incident was produced, such as printing important files after obtaining unauthorized access through a Virtual Private Network and taking a device (laptop/storage device) after obtaining unauthorized access through a Virtual Private Network. In the case of the Clustering method, a leakage scenario with a lack of correlation occurred when taking portable device out after printing an important document then, attaching the file and connecting on a smartphone after sending an e-mail from other seats. Accordingly, it was confirmed that both mentioned analysis methods were inappropriate and, therefore, we proceeded with the use of the N-gram analysis method. We used N-gram to derive a scenario by considering a sequence of behaviors. The N-gram is a method of predicting a certain word that is most likely to follow another word by representing a relationship of currently recognized words. The approach of the N-gram analysis technique is to consider a word partially, not as a whole, when conducting a vast amount of text analysis. The "N" in N-gram refers to deciding how many words to set as the standard for analysis in consideration of some words. An "N-gram" is defined as a consecutive sequence of n items. The whole text is broken down into word units of n items and is regarded as a single token. For instance, the N-gram for each n in the sentence "An adorable little girl is spreading smiles." [21] presents the process of sorting out by the size of n item in the case of applying the N-gram analysis method regarding a given sentence. In this method, when the value of n item is 1, it is called a Uni-gram, a Bi-gram for 2, a Tri-gram for 3, a 4-gram for 4, and so on. When n item is 1, it refers to one consecutive word sequence and is broken down one by one as per spacing words. In the case of 2 being the n item, which means two consecutive word sequences, the result is a two-word pairing. Thus, it is confirmed that the word unit of consecutive word sequences grows by the increase of n items.

Predicting a following word in a sequence of words in the Language model using the N-gram method depends only on n item-1. For example, when predicting a following word after the sentence "An adorable little girl is spreading," using a language model of 4-gram, in which the n item is 4, the three preceding words—little girl is—which refers to n item-1, is considered. Assuming that the phrase "girl is spreading" appears 1,000 times in vast amount of text analyzing N-gram language models, given that "girl is spreading insults" appears 500 times and "girl is spreading smiles" appears

200 times, the logic is that the probability of "insults" to follow "girl is spreading" is 50%, whereas it is 20% for "smiles." In this way, according to probabilistic choice, the N-gram analysis method considers "insults" as a more appropriate word to be followed by "girl is spreading."

$$\begin{aligned} \mathrm{P}(\mathit{insult}|\mathit{girl\ is\ spreading}) &= 0.500 \\ \mathrm{P}(\mathit{smiles}|\mathit{girl\ is\ spreading}) &= 0.200 \end{aligned} \tag{1}$$

In this research, we proceed to draw an information leakage scenario by applying the N-gram analysis method to the information leakage. A total of 15 acts of information leakage, which verified a fitness (risk) evaluation through survey, were designated as the target of the N-gram analysis. Information leakage scenarios (a list of information leakage) were applicated by conducting N-gram analysis between information leakage acts along with the order of the analysis method shown in Figure 4. For the first stage, all 15 data leakage behaviors that were evaluated for suitability and validity (=degree of risk) were parameterized by code, as were the scenarios of actual data leakage behaviors, as confirmed through the previous in-depth interview method. Thereafter, when n is 2, a scenario analysis is performed between two data leakage behaviors, and when n is 3, a scenario analysis is performed between three data leakage behaviors. In the last stage, we attempt to design a data leakage scenario through N-gram analysis and to identify the data leakage scenario that has the greatest risk according to the risk of each information leakage activity.

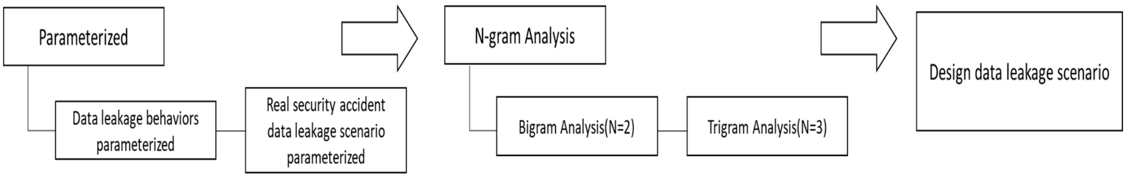
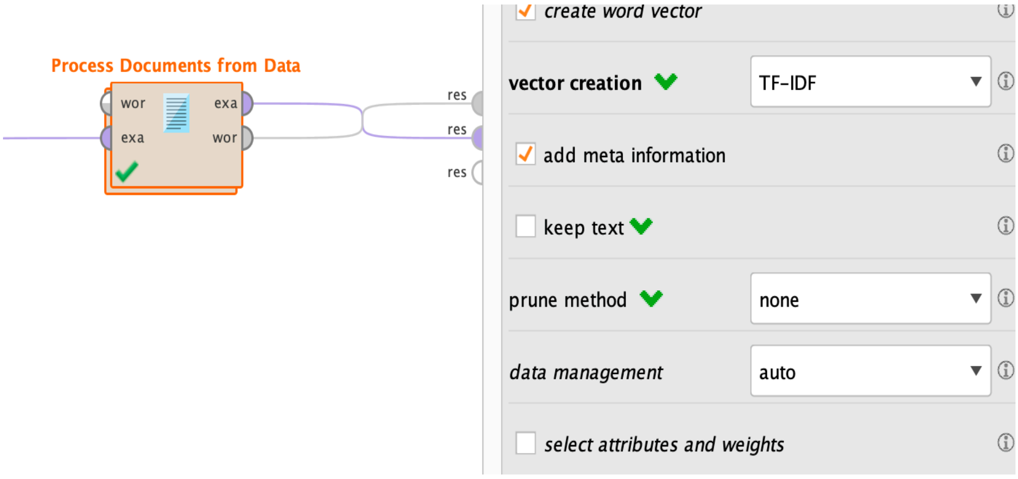

Parameterized
Data leakage behaviors parameterized
Real security accident data leakage scenario parameterized
N-gram Analysis
Bigram Analysis(N=2)
Trigram Analysis(N=3)
Design data leakage scenario

Figure 4. N-gram analysis methodology.

N-gram was conducted using the statistical analysis tool RapidMiner Studio version 7.2, which is specialized in predictive data analysis. For analysis, the TF-IDF (Term Frequency-Inverse Document Frequency) feature selection method was applied. The TR-IDF feature selection method means that the frequency of specific behaviors is divided by the frequency of appearance of all actions in n behavior sets like as below Figure 5.



☑ *create word vector*

vector creation ✓ TF-IDF

☑ add meta information

□ keep text ✓

prune method ✓ none

data management auto

□ *select attributes and weights*

Figure 5. Setting the TF-IDF Value.

5. Result of the Research

In this study, data leakage scenarios were designed through bi-gram and tri-gram analyses. In general, it is recommended to set the size of n not to exceed 5 [21,22]. However, in this study, an n value of 4 or more indicates a sparse matrix, from which it is difficult to derive meaningful results. When confirming the perplexity value [21] that can confirm the accuracy of the N-gram analysis result, we confirmed a perplexity value of 223 for the bi-gram and 158 for the tri-gram. The smaller the perplexity value is, the more accurate the analysis result of the data will be [21]. When n is 4, the perplexity value is 110, which is a high value. However, the four data leakage behaviors of the N-gram-analyzed data leakage scenario dataset can be considered as unreliable perplexity values due to insufficient data in the sequence [21]. Thus, we set the size of n to 2 and 3. The bi-gram analysis, classified by the sum of the risks of the two data leakage behaviors and the scenarios with more than eight points, is shown in Table 5.

Table 5. The results of the bi-gram analysis.

ID	Data Leakage Behavior 1	Data Leakage Behavior 2	Risk
2-gram-001	SW installation to disable the security environment (bypassing the security SW)	Receiving key documents by accessing them through a Virtual Private Network (through encrypted tunneling)	4.53 + 4.68 = 9.21
2-gram-002	Receiving key documents by accessing them through a Virtual Private Network (through encrypted tunneling)	Use of unauthorized external commercial email (not using the authorized in-house email server)	4.68 + 4.39 = 9.07
2-gram-003	Un-encrypt key documents (document security SW)	Receive key documents by accessing them through a Virtual Private Network (through encrypted tunneling)	4.21 + 4.68 = 8.89
2-gram-004	Use of unauthorized external commercial email (not using the authorized in-house email server)	Duplicate key documents to portable storage (USB, hard disk, CD)	4.39 + 4.49 = 8.88
2-gram-005	Un-encrypt key documents (Document security SW)	Use of unauthorized external commercial email (not using the authorized in-house email server)	4.21 + 4.39 = 8.6

The five data leakage scenarios were derived. The main peculiarities of the bi-gram analysis results are the behaviors of “Use of unauthorized external commercial email (not using the authorized in-house email server)” and “Receiving key documents by accessing them through a Virtual Private Network (through encrypted tunneling)”, which correspond to scenarios 3 and 5. This means that if the behaviors of illegally using an external commercial mail service to bypass the network through a Virtual Private Network and receiving a key document occur together with other data leakage behaviors, it is likely that a data leakage security accident will occur and, therefore, the data leakage behaviors need to be carefully assessed. As another key point of the analysis result, the case of “Receiving key documents by accessing them through a Virtual Private Network (through encrypted tunneling)” after the “SW installation to disable the security environment (bypassing the security SW)” was analyzed as the scenario with the highest risk (9.21).

The scenarios with more than 13 points from the tri-gram analysis, classified by the sum of the risks of the three data leakage behaviors, are shown in Table 6. Compared with the bi-gram analysis results, the main peculiarities of the tri-gram analysis results are “Receive key documents by accessing them through a Virtual Private Network (through encrypted tunneling)” and “Use of unauthorized external commercial email (not using the authorized in-house email server),” which are also major data leakage behaviors. The above two behaviors showed a high risk in the bi-gram analysis. Thus, the probability of a data leakage security accident is very high when these two behaviors occur in succession.

Table 6. The results of the tri-gram analysis.

ID	Data Leakage Behavior 1	Data Leakage Behavior 2	Data Leakage Behavior 3	Risk
3-gram-001	Un-encrypt key documents (Document security SW)	Receive key documents by accessing them through a Virtual Private Network (through encrypted tunneling)	Use of unauthorized external commercial email (not using the authorized in-house email server)	4.21 + 4.68 + 4.39 = 13.28

6. Conclusions and Future Work

In this study, we conducted research to prevent behavior-based data leakage to facilitate sustainable growth of the organization. In detail, data leakage behaviors and risks were derived to predict data leakage security incidents by collecting and analyzing the behaviors of data leakage to detect threats from organization insiders. To derive the risk, we surveyed and analyzed the data leakage behaviors of insiders by analyzing previous research and conducting an in-depth interview and then conducted a survey and statistically verified the analyzed data leakage behaviors. Moreover, the statistically verified data leakage behaviors were analyzed through the N-gram methodology. The data leakage behaviors considered in this study were determined through an in-depth interview with security experts who have experience in actual security accidents. These behaviors were then classified and identified as actual data leakage behaviors rather than abnormal behaviors. As the data leakage behaviors and scenarios were statistically verified through an additional survey of security experts, and data leakage scenarios were derived using the N-gram methodology, the behaviors and scenarios proposed in this study are different from those derived via previous detection methods using abnormal behaviors. As a result of this study, we can overcome the limitations of security solutions that use previous detection methods and thereby predict security accidents. Furthermore, the research results are expected to ameliorate the organization's risk of data leakage and contribute to the organization's sustainability. The main limitation of this study is that it failed to apply the derived data leakage behaviors and scenarios to the actual industry; thus, the reliability of the results could not be confirmed. Therefore, future work should develop an automated data leakage behavioral analysis tool and determine its applicability by analyzing and applying it to actual industry.

Author Contributions: Conceptualization, H.C.; methodology, H.C.; validation, J.K. (Jawon Kim); formal analysis, J.K. (Jaesoo Kim); investigation, J.K. (Jawon Kim); resources, J.K. (Jaesoo Kim); data curation, J.K. (Jawon Kim); writing—original draft preparation, J.K. (Jawon Kim) and H.C.; all authors have read and agreed to the published version of the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: This paper was supported by the Korea Institute for Advancement of Technology (KIAT) grant funded by the Korea Government (MOTIE) (P0008703).

Conflicts of Interest: There is no conflict of interest.

Appendix A

Table A1. The list of previous research.

No.	Contents
A1	Ha, D.; Kang, K.; Ryu, Y. Detecting Insider Threat based on Machine Learning: Anomaly Detection Using RNN Autoencoder. *J. Korea Inst. Inf. Secur. Cryptogr.* **2017**, *27*, 763–773. [23]
A2	Lee, J.; Kim, I. Detecting Abnormalities in Fraud Detection System through the Analysis of Insider Security Threats. *J. Soc. for E Bus. Stud.* **2019**, *23*, 153–169. [24]
A3	Kim, H. A Study on Method for Insider Data Leakage Detection. *J. Inst. Internet Broadcasting Commun.* **2017**, *17*, 11–17. [25]
A4	Noh, J.; Park, D. Forensic Evidence of Cyber Attack (APT) and Spear Phishing Scenario. International Information Institute (Tokyo). *Information* **2017**, *20*, 5601–5606. [26]
A5	Son, Y.; Kim, I. A Study on the Customized Security Policy for Effective Information Protection System. *J. Korea Inst. Inf. Secur. Cryptol.* **2017**, *27*, 705–715. [27]
A6	Kim, A.; Oh, J.; Ryu, J.; Lee, J.; Kwon, K.; Lee, K. SoK: A Systematic Review of Insider Threat Detection. *J. Wirel. Mob. Netw. Ubiquitous Comput. Dependable Appl.* **2019**, *10*, 46–67. [28]
A7	Oh, J.; Kim, T.; Lee, K. Advanced insider threat detection model to apply periodic work atmosphere. *TIIS* **2019**, *13*, 1722–1737. [29]
A8	Jiang, J.; Chen, J.; Gu, T.; Choo, K.; Liu, C.; Yu, M.; Weiqing, H.; Prasant, M. Anomaly Detection with Graph Convolutional Networks for Insider Threat and Fraud Detection. In Proceedings of the MILCOM 2019—2019 IEEE Military Communications Conference (MILCOM), Norfolk, VA, USA, 12–14 November 2019; pp. 109–114. [30]
A9	Lv, B.; Wang, D.; Wang, Y.; Lv, Q.; Lu, D. A hybrid model based on multi-dimensional features for insider threat detection. In Proceedings of the International Conference on Wireless Algorithms, Systems, and Applications, Tianjin, China, 20–22 June 2018; Springer: Cham, Switzerland, 2018; pp. 333–344. [31]
A10	Singh, M.; Mehtre, B.; Sangeetha, S. User Behavior Profiling using Ensemble Approach for Insider Threat Detection. In Proceedings of the 2019 IEEE 5th International Conference on Identity, Security, and Behavior Analysis (ISBA), Hyderabad, India, 22–24 January 2019; pp. 1–8. [32]

References

1. Accenture and HfS Research. *The State of Cybersecurity and Digital Trust 2016*; Accenture: Dublin, Ireland; HfS Research: Cambridge, UK, 2016.
2. SEI Cyber Minute: Insider Threats. Available online: http://resources.sei.cum.edu/library/asset-view.cfm?assetid=496626 (accessed on 30 March 2020).
3. Smart Insight. *Managing Digital Marketing in 2020 Research Report*; Smart Insight: Leeds, UK, 2019.
4. Soomro, Z.A.; Shah, M.H.; Ahmed, J. Information security management needs more holistic approach: A literature review. *Int. J. Inf. Manag.* **2016**, *36*, 215–225. [CrossRef]
5. Al-Dhahri, S.; Al-Sarti, M.; Ahmed, J. Information security management system. *Int. J. Comp. Appl.* **2017**, *158*, 29–33. [CrossRef]
6. Dupuis, M.; Khadeer, S. Curiosity killed the organization: A psychological comparison between malicious and non-malicious insiders and the insider threat. In Proceedings of the 5th Annual Conference on Research in Information Technology, Boston, MA, USA, 28 September–1 October 2016.
7. Insider Threat Report. Insider threat report. Insider threat related data breach detection time. In *Insider Threat Report: Executive Summary*; Verizon business ready: New York, USA, 2019.
8. Parrot, A.; Bechhofer, L. *Acquaintance Rape: The Hidden Crime*; Wiley: New York, NY, USA, 1991.
9. Sun, N.; Zhang, J.; Rimba, P.; Gao, S.; Zhang, L.Y.; Xiang, Y. Data-driven cybersecurity incident prediction: A survey. *IEEE Commun. Surv. Tutor.* **2018**, *21*, 1744–1772. [CrossRef]
10. Alzhrani, K.; Rudd, E.M.; Boult, T.E.; Chow, C.E. Automated big text security classification. In Proceedings of the 2016 IEEE Conference on Intelligence and Security Informatics (ISI), Tucson, AZ, USA, 28–30 September 2016.

11. Kim, K.; Kim, J. A Study on analyzing risk scenarios about vulnerabilities of security monitoring system: Focused on information leakage by insider. In Proceedings of the International Workshop on Information Security Applications, Jeju Island, Korea, 23–25 August 2018; Springer: Cham, Switzerland, 2018.
12. Shin, H.J.; Kim, M.H. A detection method of data leakage by cooperation of insiders. *Int. J. Appl. Eng. Res.* **2017**, *12*, 13321–13327.
13. NIST. *Computer Forensic Reference Data Set, Data Leakage Case*; NIST: Gaithersburg, MD, USA, 2018.
14. Prozorov, A. *10 Most Widespread Staff Errors behind Data Leakage*; InfoWatch Analytical Center: Moscow, Russia, 2013.
15. Bromiley, M. *Defend Your Business against Insider Threats*; SANS Institute Information Security Reading Room, Sans Institute: Boston, MA, USA, 2019.
16. Hyosun, Y.; JunDuk, K.; Sujin, K. A study on improvement for analyzing the security check items in security review of mobile financial services. *Korea Assoc. Ind. Secur.* **2017**, *7*, 129–158.
17. Zhao, R.; Mao, K. Fuzzy bag-of-words model for document representation. *IEEE Trans. Fuzzy Syst.* **2017**, *26*, 794–804. [CrossRef]
18. Radovanović, M.; Ivanović, M. Text mining: Approaches and applications. *Novi Sad J. Math.* **2008**, *38*, 227–234.
19. Wang, S.I.; Manning, C.D. Baselines and bigrams: Simple, good sentiment and topic classification. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, Jeju Island, Korea, 8–14 July 2012; Association for Computational Linguistics: Stroudsburg, PA, USA, 2012; Volume 2, pp. 90–94.
20. Mandelli, D.; Yilmaz, A.; Aldemir, T.; Metzroth, K.; Denning, R. Scenario clustering and dynamic probabilistic risk assessment. *Reliab. Eng. Syst. Saf.* **2013**, *115*, 146–160. [CrossRef]
21. Jun, W. Natural Language Processing Using for Deep Learning Method. 2020. Available online: https://wikidocs.net/book/2155 (accessed on 17 July 2020).
22. Chelba, C.; Norouzi, M.; Bengio, S. N-gram Language Modeling Using Recurrent Neural Network Estimation. Google Technology Report. 2017. Available online: https://arxiv.org/abs/1703.10724 (accessed on 17 July 2020).
23. Ha, D.; Kang, K.; Ryu, Y. Detecting Insider Threat based on Machine Learning: Anomaly Detection Using RNN Autoencoder. *J. Korea Inst. Inf. Secur. Cryptogr.* **2017**, *27*, 763–773.
24. Lee, J.; Kim, I. Detecting Abnormalities in Fraud Detection System through the Analysis of Insider Security Threats. *J. Soc. E Bus. Stud.* **2019**, *23*, 153–169.
25. Kim, H. A Study on Method for Insider Data Leakage Detection. *J. Inst. Internet Broadcast. Commun.* **2017**, *17*, 11–17.
26. Noh, J.; Park, D. Forensic Evidence of Cyber Attack (APT) and Spear Phishing Scenario. *Int. Inf. Inst. (Tokyo) Inf.* **2017**, *20*, 5601–5606.
27. Son, Y.; Kim, I. A Study on the Customized Security Policy for Effective Information Protection System. *J. Korea Inst. Inf. Secur. Cryptol.* **2017**, *27*, 705–715.
28. Kim, A.; Oh, J.; Ryu, J.; Lee, J.; Kwon, K.; Lee, K. SoK: A Systematic Review of Insider Threat Detection. *J. Wirel. Mob. Netw. Ubiquitous Comput. Dependable Appl.* **2019**, *10*, 46–67.
29. Oh, J.; Kim, T.; Lee, K. Advanced insider threat detection model to apply periodic work atmosphere. *TIIS* **2019**, *13*, 1722–1737.
30. Jiang, J.; Chen, J.; Gu, T.; Choo, K.; Liu, C.; Yu, M.; Weiqing, H.; Prasant, M. Anomaly Detection with Graph Convolutional Networks for Insider Threat and Fraud Detection. In Proceedings of the MILCOM 2019–2019 IEEE Military Communications Conference (MILCOM), Norfolk, VA, USA, 12–14 November 2019; pp. 109–114.
31. Lv, B.; Wang, D.; Wang, Y.; Lv, Q.; Lu, D. A hybrid model based on multi-dimensional features for insider threat detection. In Proceedings of the International Conference on Wireless Algorithms, Systems, and Applications, Tianjin, China, 20–22 June 2018; Springer: Cham, Switzerland, 2018; pp. 333–344.
32. Singh, M.; Mehtre, B.; Sangeetha, S. User Behavior Profiling using Ensemble Approach for Insider Threat Detection. In Proceedings of the 2019 IEEE 5th International Conference on Identity, Security, and Behavior Analysis (ISBA), Hyderabad, India, 22–24 July 2019; pp. 1–8.