Research on implicit transmission monitoring of files across business systems

Chunru Zhou*, Wenjing Zeng, Guo Wu, Junhao Li, Wenqiang Pan
Institute of Computer Application, China Academy of Engineering Physics, Mianyang, Sichuan, China

ABSTRACT

File transmission is a key concern of Intranet Security Monitoring. Compared with explicit file transmission within a single system, the implicit transmission of files across various business systems is a difficult problem in current Intranet Security Monitoring. In order to shield the differences of file operations in various business systems, explore the clues of implicit file transmission across different business systems, and realize the retrospective security analysis of complex scenarios related to file transmission, this paper proposes an effective technical framework. Firstly, it takes network flow as the main data source, aiming at different business applications and focusing on files, and extracts valuable business information from flow data. Secondly, it reorganizes key elements such as files, users, and terminals in business information, and uses the file transmission process as a link to form unified high-value clue data. Thirdly, it integrates multiple clues, designs and builds a unified file transmission graph data based on an open-source graph database. Finally, an example of security analysis on the implicit transmission of suspicious files based on the above is given, the results show that the organized file element graph data proposed in this paper can effectively describe the transmission of the same files across different business systems.

Keywords: File transmission, implicit transmission, intranet security monitoring, business system.

1. INTRODUCTION

In the intranet environment of large units, network security issues are mainly divided into two aspects: technical security and business security. At present, technical security problems are mainly solved by various network security products and equipment such as Firewalls, IPS, Antivirus Software, and Host Audit System[1], which performs good protective effect. However, business security problems do not have a broad range of common features. Due to the lack of targeted business scenario research in existing network security technologies, in different business networks, there are many business types, business logics, file types, and importance[2]. It is difficult to form a unified business security product, not to mention when faced with the confidentiality requirements of business work[3].

Security monitoring of file transmission is a key requirement for intranet business security. On one hand, the management and control capabilities of various business systems such as email, conference, production management, etc., mainly focus on the life cycle management of the file itself, providing functional guarantee for file services. Most of them can only support file transmission queries with clear log records, cannot natively support cross-business system association. On the other hand, although a large amount of business application data can be collected by relying on a centralized security monitoring platform (Security Operation Center, SOC)[4], the analysis logic of its alarm judgment rules is mostly one-way linear, cannot backtrack and deal with complex associations. There are still major deficiencies in the monitoring of abnormal file transmission scenarios across business systems, with a large time span and without explicit records; in addition, internal file operations in each business application system are different because of the large differences in software architecture and program modules, security analysts need to be deeply involved in the entire process of data collection, processing and correlation analysis, which greatly restricts the security analysis work of file transmission[5, 6].

2. CURRENT STATUS OF FILE TRANSMISSION MONITORING

2.1 File transmission classification from the perspective of security monitoring

The transmission of business files refers to the process of moving or copying a certain file (documents, drawings, models, audio and video, etc.) from one user terminal to another via online or offline methods. Online transmission is mainly

* zcr214@qq.com

Third International Conference on Computer Science and Communication Technology (ICCSCT 2022)
edited by Yingfa Lu, Changbo Cheng, Proc. of SPIE Vol. 12506, 125065L
 · 0277-786X · doi: 10.1117/12.2662635

through the business application system (such as mail, official documents, etc.), offline transmission is mainly through data ferry (such as direct transmission, media copy, etc.). According to the characterization of file transmission traces, it can be divided into two categories: explicit transmission and implicit transmission, as shown in Table 1.

Table 1. File transmission classification from the perspective of security monitoring.

	Explicit transmission	**Implicit transmission**
Transmission process	Continuous and complete	Scattered
Transmission trace	Recorded and easy to query	No direct record
File identification	Correct identification	No or wrong identification
File properties	Complete	Partly missing or unclear
Example scene	Normal mail sending; Internal file distribution after multiple approvals	Mail sending with sign modified Transferring within zip files

Explicit transmission generally means that the file properties and sign are correct, the transmission process is continuous and complete, and detailed business documents and log records are recorded, that is, the characteristic data of transmission traces are clear and verifiable, such as normal mail sending, internal file distribution after multi-level approval etc.

Implicit transmission is relatively complicated. Although the file has been processed by legal business processes, its transmission process may be discontinuous or even involves multiple businesses, and its attributes and sign may have been modified or may be entrained in other files, making the key identifiers of the files or the explicit representation of the transmission traces is not clear, and the effective information cannot be reflected in the audit function of the business system, but is hidden in the lower-level original data, which is not easy to extract. Sending emails, hiding important and sensitive files in compressed packages or other file attachments for transmission, etc. are all implicit transmission.

2.2 Files management in business systems

In actual work, the business application system and the file management system are separated as shown in Figure 1.

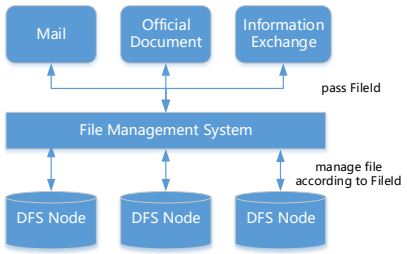

Figure 1. Current status of files management in business systems.

The file upload, download, storage and other related operations are mainly completed by the file management system. The business application does not need to care about the specific management method of the file, passing the file id for file interaction is all it has to do.

The business application system only manages the basic attribute information such as ID number, file sign, file format, file name and so on of the file attached to a business item, and cannot control the file entity; while the file management system only manages the file entity and does not grasp the business information. Such a decoupled software architecture is conducive to performing their own duties, and the management of business and documents is more concise and efficient, but there are also certain business security problems: the separation of business and files makes both parties only have their own relevant information, file management system cannot know the correspondence between the file and the business, and the business application system cannot know whether the file obtained through the file id is accurate.

For conventional explicit transmission security scene, such as “transmission of attachments with important and sensitive identifiers in ordinary mail”, through the formulation of business logic rules, the file can be identified and blocked immediately before it has been transferred; however, because the business application system cannot master the management of file entities, it is unware when the content of the file is modified, transferred, replaced, or other non-business normal operations. In addition, when there are multiple copies of the same file, and multiple files contain the same or similar content, the business application system can only distinguish them by the file attribute information such as ID and file name, but cannot identify whether there is a relationship between them. Therefore, it is difficult to realize security control through the business application system in complex implicit security scenarios such as multi-link transmission and multiple files association.

2.3 SOC alarm rules

At present, the most effective way to judge SOC alarms is to adopt linear logic rules, as shown in the left part of Figure 2. The data collection and processing to judgement of alarms are carried out in sequential flow. The generation of alarms mainly depends on the triggering of set conditions or statistical thresholds. The security and confidentiality scenarios can quickly and accurately detect alarm events. For example, for a period of time, the TCP (Transmission Control Protocol) network flow of a terminal is normal but the heartbeat flow packets of the host audit program are missing. By checking the statistics of flow packets, it can be determined that the host audit program of the terminal may be offline. In the aspect of explicit file transmission, file attributes or signs are clear such as “general mail transmits attachments with important and sensitive identifiers”, by checking the consistency of file sign and business sign, the linear logic judgment rules have been able to achieve fast and accurate monitoring[7].

However, when faced with scenarios related to business data with large time span, time sequence interleaving, business interleaving, and complex user relationships, especially when files are implicitly transferred, such as “transmission after sign modified”, linear logic does not have the ability to retrospectively analyze, its monitoring effect will be greatly weakened. When the associated business scope and time scope are further expanded, the linear logic rules themselves do not even have good preparation feasibility[8] as shown in the right part of Figure 2.

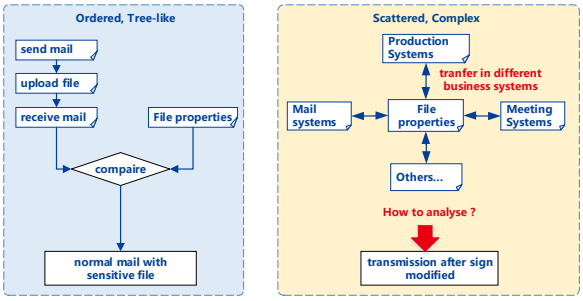

Figure 2. Application scenarios of current SOC alarm rules.

In addition, due to the large differences in the software architecture and program modules of file operations within each business application system, it is difficult to unify business-related data collection methods, data preprocessing, and key feature identification and extraction, and the security analysis work will be severely restricted.

3. OVERALL TECHNICAL FRAMEWORK

In order to solve the difference of file operations in various business systems and realize the retrospective security analysis of file transmission, especially implicit transmission, this paper takes files as the center for different business applications, extracts the effective business information in the original flow data, including files, users, and terminals. And it uses the file transmission relationship as a link to build a unified file transmission graph data. The differences between different business contents are shielded, and the analysis tasks of security user in security monitoring are focused on the design and implementation of security and confidentiality violation scenarios, instead of focusing on business system connection and underlying data processing, which can significantly improve analysis efficiency. The overall framework is shown in Figure 3.

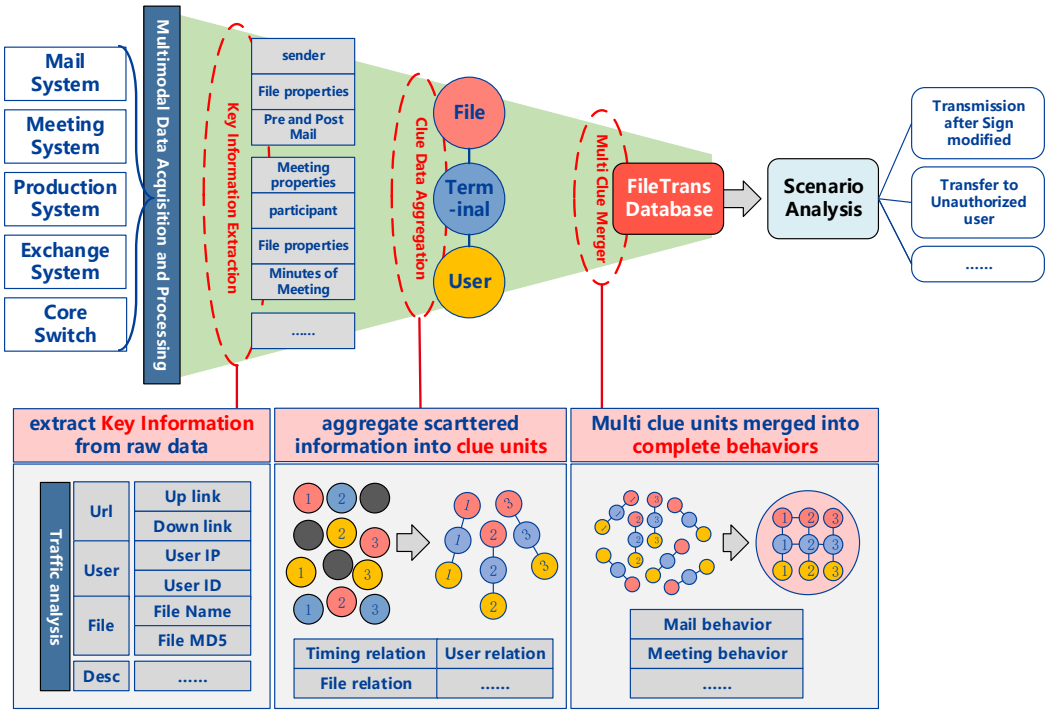

Figure 3. Overall technical framework.

4. BUSINESS-ORIENTED FILE TRANSFER BEHAVIOR DATABASE CONSTRUCTION

4.1 Extract key information

There are two main ways to obtain data related to file transmission: business document active pushing and bypass network flow monitoring. The former mainly relies on the business application system recording information of document, file, user and other data by itself when the business process is in progress, and pushed to the outside through negotiated network interface; the latter uses the method of replicating flow mirroring on the core switch to obtain all network data packets of the business process.

The way to actively push data is simple and direct, but it requires the business application system to do corresponding customized alteration, such as node recognition, data structure processing, and interface management. To a certain extent, it will have a certain performance impact on the original business process, which is not conducive to the efficient development of business work; on the other hand, due to the different nature of work, the business logic itself of each business application system is quite different, and it is difficult to provide a unified solution of data specification; in addition, for the business products purchased on the shelf and the business application system that has been online and running stably, it is not feasible to reform the function like this kind.

The network flow mirroring method needs to parse the flow data and extract the key information related to the file transmission, which is more complicated than the active push method in terms of data preprocessing. However, as a bypass technique, flow mirroring does not interfere with the operation of the business application system, does not need to be connected to business application software modules, and collects data from a single source, which is more feasible; in addition, network flow data is more complete and comprehensive, and conducive to the detailed grasp of security analysis, and can provide better data support for the extension of subsequent analysis and other security scenarios. Therefore, the key information extraction of file transmission monitoring in this paper mainly comes from flow mirroring, and the corresponding analysis is made according to the specific business content, that is, business flow analysis.

Business flow analysis is different from traditional network protocol analysis. Basic elements such as quintuple are no longer the object of attention, but more attention is paid to the execution of business matters. Taking the behavior of

sending emails as an example, business flow analysis needs to extract the entire process information of email event, from people create emails to recipients complete downloading attachments. Because the service execution has a certain time span, and the number of data packets involved in the core steps is not large, the traditional packet capture and full flow analysis will consume a lot of performance resources when processing a long time span. Therefore, the business flow analysis must be precise information extraction process after deeply integrated with the business, as shown below in Figure 4.

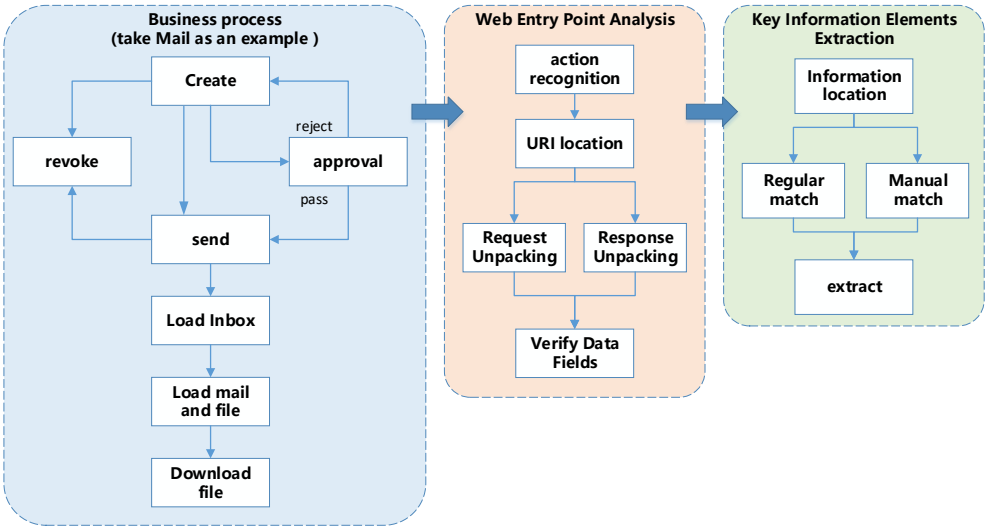

Figure 4. Business flow analysis.

Firstly, in order to locate the application service that may be related to file transmission, specific processes for business content must be descript. Secondly, it is neccessary to analyze Web entry points for each specific application service, identify business actions and locate the corresponding URI resource; through the constraints of the URI resource, business data packets can be accurately filtered from the massive flow data, further disassemble the request and response data packets, and determine the valid data fields. Finally, with regular matching or manual matching, the key information elements of file transmission in the packects will be extacted.

4.2 Clue data aggregation

In the monitoring of file transmission, files, users, and terminals are the focus of attention. Through the connection of business documents, the fragmented information obtained from business flow analysis is reorganized by the definition of "subject", "object" and "relationship" in the knowledge graph, to build the complete basic properties of each entity or relationship in the business process.

The information elements contained in the complete file, user, and terminal entity objects can be respectively represented as vectors: $f=\{f_1,f_2,\ldots f_i\}, u=\{u_1,u_2,\ldots u_j\}, t=\{t_1,t_2,\ldots t_k\}$, The number of elements included in the three are i,j,k . For a complete business document B of a business event, it may generate n valuable business flow data packet, represented as $B=\{b_1,b_2,\ldots b_n\}$, Each flow data packet b can be parsed to contain the above three different types of information elements, The x th packet is represented as $b_x=\{f_x,u_x,t_x\}$, and f_x,u_x,t_x can be $\varnothing$. To extract complete entity information from the flow data corresponding to the business document, f,u,t,b satisfies the following conditions:

$$\begin{cases} f = \sum_{f_x \in b_x} f_x \\ u = \sum_{u_x \in b_x} u_x, b_x \in B \\ t = \sum_{t_x \in b_x} t_x \end{cases}$$

The reorganization process of information elements can be regarded as three mapping processes such as $b \to f, b \to u, b \to t$. Each mapping extracts the information elements of the corresponding entity from the flow data packet b as shown in Figure 5:

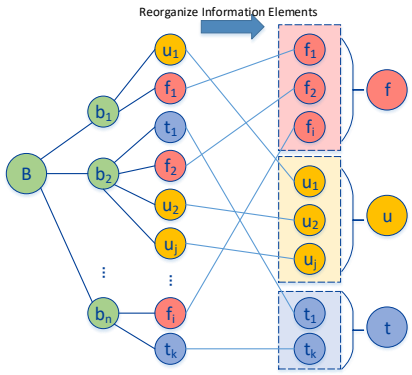

Figure 5. Reorganization of information elements.

In the actual intranet security monitoring, the relevant main information elements are shown in Table 2.

Table 2. Examples of main information elements related to document transmission.

Information elements	Type	Contents
u (users)	Entity	User id, name, unit, department, user level
f (files)	Entity	File id, file name, file level, Md5, file type, file size
t (terminals)	Entity	Ip, mac, terminal level, os type, physical location
b (business document)	Relationship	Business type, business id, business level, timestamp

4.3 Graph database constructed with multiple clue

The businesses involved in multiple data clues are different, but the connected nodes are always users, files and terminals. Different user may establish relationships by processing the same file, thus establishing relationships between different businesses. Using graphs to represent relationships is a more intuitive and clear way[9, 10]. In this paper, graph databases are used to integrate and store relational data in a unified manner.

The entities are users, files and terminals, and the association of the three mainly depends on business documents. In the actual flow analysis, it is found that information elements directly related to user entities may be missing in business documents, and file entities may not be directly related to user entities; asset entities contain IP information that can always be obtained in business documents, so its relationship with the file entity is relatively easy to build; part of the relationship between the asset entity and the user entity can be built in the business flow, and the missing part can be supplemented by the assets ledger. Therefore, the three types of entity relationships mainly include two types of relationships, namely “asset-user” and “asset-file”. According to the direction of file transmission, the “asset-file” relationship can be divided into

"upload" and "download". The relationships between entities in file transmission graph database can be represented as the following Table 3.

Table 3. The relationships between entities in file transmission graph database.

Object		Relationship		Subject
User	→	Workon	→	Terminal
Terminal	→	Upload	→	File
File	→	Download	→	Terminal

An example of an entity and its relationship is shown in Figure 6 below. This article selects the open-source Neo4j graph database, and according to the design of three types of entities, including user, files, and terminals and their relationships, graphs about file transmission can be constructed from business flow and stored in the graph database. As shown in Figure 7 below, it shows the transmission of 93 files in 3 departments, 28 users, and 28 terminals.

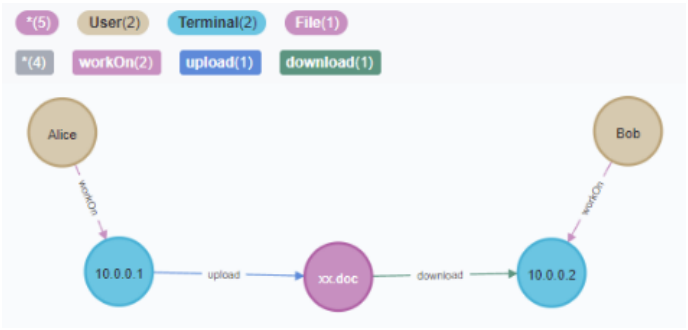

Figure 6. Example of entities and relationships in file transmission.

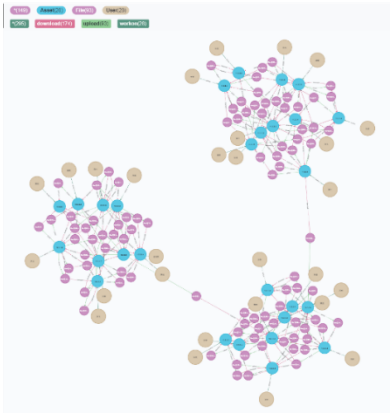

Figure 7. Graph database of file transmission.

5. DATA TYPE MAPPING BASED ON STRUCTURAL INFORMATION ENTROPY

The file transmission graph data can intuitively display the active location of files in the business, and the relationship between files, user, and terminals is closer to natural language description. For different security and confidentiality scenarios, under the condition of determining the file transmission relationship logic and attribute value range, the advantage of the graph database retrieval capability can be used to quickly form a security monitoring data model, and solidify it into a monitoring module in the SOC platform.

The construction of security monitoring scenarios based on file transmission graph data, as shown in Figure 8, is mainly divided into the following steps:

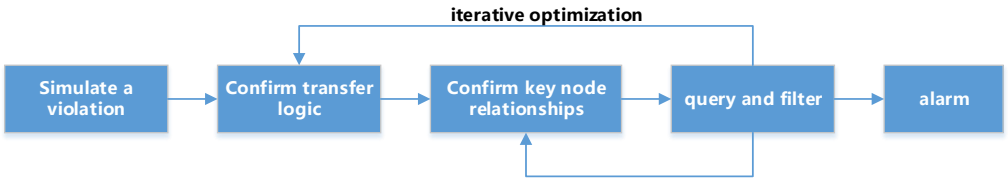

Figure 8. Security monitoring scenario constructing steps.

- Determine the security and confidentiality violations that may exist in the process of file transmission, and simulate illegal operations.
- Infer the possible logical sequence of the illegal transmission of the file according to the business operation.
- Locate key nodes and relationships, and determine the range of attribute values or attribute relationships.
- Construct and execute the search statement, and filter the data sets that meet the violations from the search results
- Iterate multiple times to optimize the retrieval conditions, finally lock the target user or terminal, and generate an alarm event.

Taking the scenario of **"File Transmission with Sign Modified"** as an example, the hypothetical violation scenario is described as follows: user Alice sends a sensitive file to users Bob and Cindy through encrypted communication email. User Bob's subsequent operations are normal, but user Cindy downloaded the file locally and modified the file name and its sensitive sign, then sent it to the user David by ordinary mail, just in order to avoid the trouble of approval. If the user David does not know that the file is a sensitive file, the file may be kept, processed and disseminated in accordance with ordinary files. Since the sensitive sign of the file is lost, the subsequent transmission of the file will no longer be strictly protected, that is an unacceptable risk.

In the above scenario, the two email behaviors themselves are legal, but in fact risky behaviors have already occurred. After analysis, we can know that there are 2 main clues in this scene:

Although the two files have different sensitive sign and file names, the content of the files has not modified, and the basic attributes of the files remain the same;

The transmission of files carrying sensitive sign is earlier in timing than ordinary file transmissions.

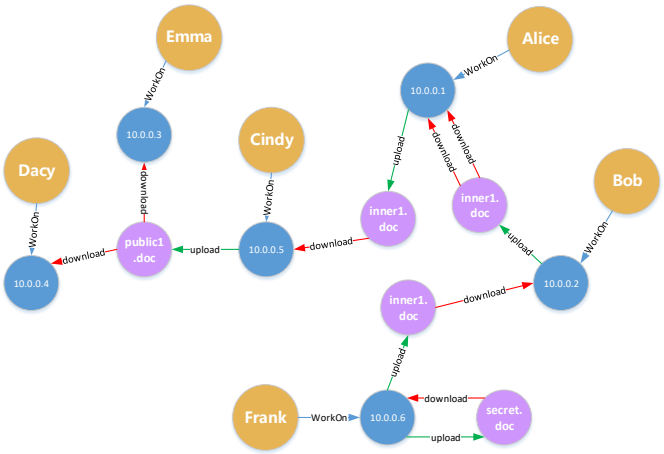

Figure 9. File transmission with sign modified.

Therefore, the basic logic of constructing a graph database retrieval statement is: "Query all file transmission behaviors → filter files with the same file attribute (mainly MD5) → filter the relationship that the file transmission of sensitive sign is earlier than the normal file transfer". The results obtained by executing the search are shown in Figure 9.

In the above figure, files with the same attributes may not be directly related, and the involved user may not be in the same unit or department. The transmission method may be through other business applications besides email, which is in line with some of the characteristics of implicit file transmission introduced ina this article. It can be seen that the file transmission monitoring technology in this paper can find security and confidentiality matters that cannot be covered by traditional security equipment and protective measures, and provide certain business security support capabilities.

Based on the transmission relationship of files, multiple scenarios such as "file knowledge scope monitoring", "community relationship monitoring" and "business application abnormal access monitoring" can also be constructed, which has a good ability to expand research support.

6. CONCLUSION

Files are the key protection objects of security and confidentiality in intranet. The file transmission security monitoring technology introduced in this paper is centered on files, extracts effective business information from the original flow data, builds a unified file transmission graph database, and conducts security analysis based on the transmission relationship. To a certain extent, it solves the restriction of security analysis caused by the differences of various business systems, makes up for the deficiency of non-retrospective analysis of online association rules, and has achieved remarkable results in actual security monitoring work.

The current monitoring capability for business security mainly lies in the detection of abnormal situations associated with normal business processes, and there are still many deficiencies: for example, hidden abnormal business processes cannot be found, and changes in file content cannot be monitored. Businesses are diverse, and business changes require corresponding manual adaptation of the business flow analysis module, which cannot be self-adapted. Further technical research in these areas will be carried out in the future.

ACKNOWLEDGEMENT

This work is supported by Defense Industrial Technology Development Program (JCKY2019602B013) and CAEP Foundation (CX20210011 C. R. Zhou), CAEP-ICA Foundation (No.SJ2021A03 P. Lu).

REFERENCES

[1] Sun, Z., Li, Y. and Zhang, W., "Research on the development trend and auditing mode of high security enterprise intranet security audit," IEEE 11th Int. Conf. on Advanced Infocomm Technology (ICAIT), 153-156 (2019).
[2] Belov, V. M., Pestunov, A. I. and Pestunova, T. M., "On the issue of information security risks assessment of business processes", XIV International Scientific-Technical Conference on Actual Problems of Electronics Instrument Engineering (APEIE), 136-139(2018).
[3] Jin, Q. and Wang, L., "Intranet user-level security traffic management with deep reinforcement learning," Int. Joint Con. Neural Networks (IJCNN), N-19787(2019).
[4] Cyril, O., "Cyber security operations centre: Security monitoring for protecting business and supporting cyber defense strategy," Int. Conf. on Cyber Situational Awareness, Data Analytics and Assessment (CyberSA), 1-10(2015).
[5] Grzegorz, K. and Krzysztof, J., "Complex networks monitoring and security and fraud detection for enterprises," IEEE 28th Int. Conf. on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE), 124-125(2019).
[6] Wang, X. S., Herwono, I., Cerbo, F. D., Kearney, P. and Shackleton, M., "Enabling cyber security data sharing for large-scale enterprises using managed security services," IEEE Conf. on Communications and Network Security (CNS), (2018).
[7] Krivo, A. and Mirvoda, S., "The experience of cyberthreats analysis using business intelligence system," Ural Symp. on Biomedical Engineering, Radioelectronics and Information Technology (USBEREIT), (2020).
[8] Elsayed, M. A. and Zulkernine, M., "Predict deep: Security analytics as a service for anomaly detection and prediction," IEEE Access, (8), 45184-45197(2020).
[9] Duan, Y. C., Shao, L. X. and Hu, G. Z., "Specifying architecture of knowledge graph with data graph, information graph, knowledge graph and wisdom graph," IEEE Int. Conf. on Software Engineering Research (SERA), 327-332(2017).
[10] Zhang, Z. P., "Graph databases for knowledge management," IT Professional, 19(6), 26-32(2017).