Prediction of certain network events based on active data flow

Jun Fang, Xiaodong Wong

Abstract—**This paper was originally submitted to Xinova as a response to a Request for Invention (RFI) on cyber-attack forecasting, detection, and stabilization methods. In this paper, what is proposed is a method of predicting network events by sending traceable particles and analyzing the flow of such data.**

In more detail, this describes how a sensitive file can be broken into fragments that are distributed throughout a network system, and where the fragments are "moved" continuously within the system in a random manner. Unauthorized access to these fragments may actively verify a security situation.

I. Introduction

PROBLEMATIC network locations and security conditions always present certain statistical regularities. The background of the present invention can be summarized as two aspects:

In order to improve system's wellbeing, the National Cybersecurity and Communications Integration Center (NCCIC) of the US Department of Homeland Security (DHS) proposed that with dynamic resilient defense achieved by controlling changes across multiple dimensions of a system, the uncertainty of and the complexity of access to the system can be increased to raise the cost of unauthorized user performing detection and unauthorized access to the system and narrow the time window of unauthorized access so as to achieve systematic improvement of security. Based on this idea, a sensitive file is broken into fragments that are distributed throughout a network system, and the fragments are "moved" continuously within the system in a random manner, thereby improving the security of file storage. On this basis, statistics of network locations and background conditions in which security threats occur are collected to obtain a distribution map of the threats, so as to further predict future threats.

Second, given that misjudgments occur in the defense system, false alarm rate of intrusions can be very high. With regard to this matter, when an unauthorized access is found after the dynamic distribution of said file, active verification as to whether the unauthorized access actually occurred can be made by intentionally transferring file fragments to the alarmed area where the suspected unauthorized access occurred so as to eliminate false alarms as much as possible, thereby improving the accuracy of early warning. This approach of actively responding to unauthorized access is very unique.

Combining the design of the above two aspects, the idea of the present invention can be vividly analogous to the immune system of the human body composed of white blood cells, which move free within the body under normal circumstances, but aggregate, alert and activate the immune system to take antiviral actions once a virus intrusion is found. It is the goal of the present invention to inject such "white blood cells" (i.e., particles of sensitive files) into the network to predict problems.

II. Summary of the Invention

Problems that occur in a network system always meet certain technical conditions, and therefore different network locations are associated with varied security. Future threats can be predicted by statistical analysis of locations where security problems occur even if no prior knowledge of system security problems is available. In order to obtain security in different locations of the network, sensitive files are cut into particles which are put into continuous random motion among different network locations, and statistics of unauthorized access to file particles in different locations are collected. When a suspicious unauthorized network access is detected, the file particles are intentionally transferred to an area where an alarm of predicted unauthorized access occurs so as to actively verify whether the unauthorized network access actually occurred, thereby improving the accuracy of prediction and finally obtaining a distribution of situations of security threats. Prediction of future security threats is made based on the obtained statistical information.

III. Detailed Explanation

Below describes each of the conceptual components.

Technical Principles

1. Dynamic drifting of file particles

According to the idea of dynamic resilient defense in section 1.2, sensitive files (e.g., documents, images, videos, etc.) are divided into a number of small data particles individually stored in different network storage nodes, and storage locations of those particles are changed continuously and randomly (by driving the data particles into a "drifting state" similar to Brownian motion of pollen grains), thereby obtaining a dynamic changing map of the distribution of data storage as shown in Fig. 1 below.

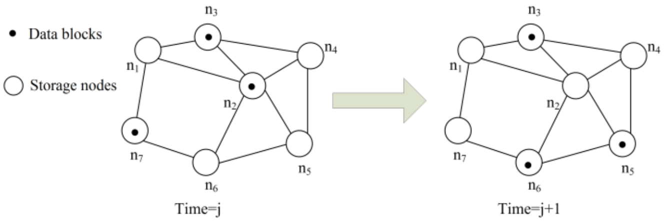


Fig. 1: schematic of random movement of file particles

When an unauthorized user (malignant) tries to detect, access and corrupt a data particle, information such as the location and time of the network node where the particle is situated is recorded. Security of properties of network locations are obtained by statistical calculation for a period of time, and locations where future security threats may possibly occur are predicted.

2. Active verification of threats of unauthorized access

If the abovementioned dynamic drifting mechanism of file particles detects unauthorized access somewhere in the network, in order to reduce the false alarm rate, the present invention provides a mechanism to actively direct more file particles to the location where unauthorized network access is detected (provided that main file information will not be divulged). Following the logic of the unauthorized user, he/she will certainly take further actions to access the information without authorization to expand the result, and will therefore try to exploit the newly added file particles furthermore. The defender can effectively verify the unauthorized access based on recurrence of such unauthorized access.

3. Centerless distributed design

In order to improve the robustness and security of the system, the security detection mechanism devised in the present invention adopts a centerless distributed structure. The method adopted is to let the network nodes to announce their trust relationships and form trust and dependency among them, such that tampering of information will be detected immediately. This centerless distributed structure has the property of non-repudiation by being able to ensure mutual trust among the nodes and decentralize system risks.

Technical advantages. The method avoids complex analysis of unauthorized access to the network itself, but instead correlates the prediction of unauthorized network access to the statistic regularities of unauthorized access to sensitive data particles, and is therefore able to resolve the problem of predicting current unauthorized access without introducing more advanced technologies.

Advantages of the present invention are as follows:
1. Technology cost of prediction is reduced;
2. The present invention is universally applicable;
3. The present invention makes use of the technical advantage of dynamic defense so that the security of the flowing sensitive data itself is guaranteed.
4. The approach of security testing through active verification improves the accuracy of prediction;
5. A centerless distributed structure is adopted to bring high credibility to the system;
6. Other advantages, including: the ability to develop upper-level applications, etc., are provided.

In summary, the present solution has provided a necessary and feasible invention with great potential of market value.

With the method of the present solution, sensitive files are cut into small particles, and those particles are switched between a random motion state and an active motion state based on unauthorized access in a network. Statistics of the security of network locations is collected by observing and counting unauthorized access to the particles, and prediction of network threats is made on the basis the statistical data. A schematic of the whole process is as follows:

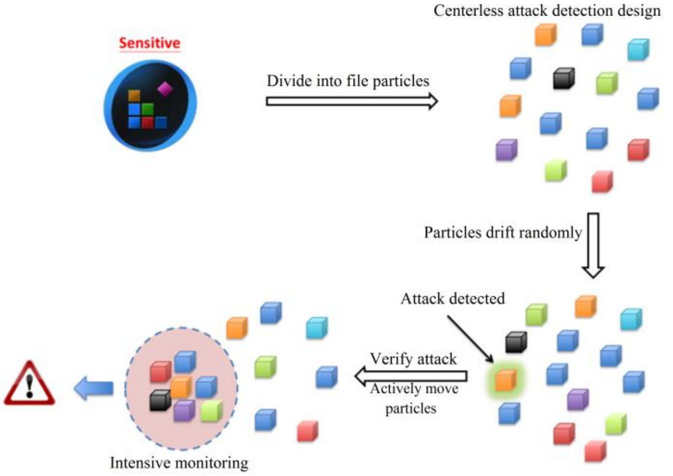


Fig. 2. Schematic of the described method

In the present invention, sensitive files are firstly divided into file particles, and those file particles are randomly moved among network nodes. Detection of unauthorized access to the file particles adopts a distributed design. If a certain file particle is detected to be suspiciously accessed without authorization, a number of other file particles are moved to the node where unauthorized access occurred so as to intensively monitor the node and thereby verify the unauthorized access. If the verification passes certain conditions, a security alert is issued.

The specific process of the method is shown in the figure below.

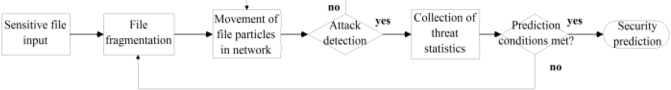

Sensitive file input
File fragmentation
Movement of file particles in network
no
Attack detection
yes
Collection of threat statistics
Prediction conditions met?
yes
no
Security prediction

Fig. 3. Flowchart of the method.

The specific implementation of the method comprises the following steps.

1. File input.

A to-be-protected file containing sensitive information is imported into the system.

In order to make the sensitive file to be tempting to attract an unauthorized user, the file is selected in the following way (but not limited thereto).

Sample files that have been exposed in unauthorized access events are collected, from which attribute features thereof are extracted. Similarity (or correlation) between the sample files and candidate files is calculated, and a number of candidate files with the largest similarity or correlation value are taken as bait files.

$$r(S, d) = \mathrm{Max}(\sum_{i} \mathrm{sim}(s_i, d))$$

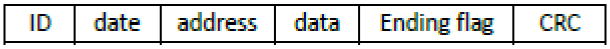
where $S \in s_i$, S is the sample set, and d is a candidate file.

If the system resources are sufficient, all the files that are desired to be protected can be treated by the following steps at a certain calculation and storage cost. In the actual implementation, files can be chosen with a certain probability so long as the accuracy of security prediction can be improved.

2. File fragmentation

The above files are divided into a number of data particles.

Suppose that D is a target file and $p_1,\cdots, p_i,\cdots, p_n$ are the divided data particles (where n is an integer, and n>>1), the size of p/ must ensure that no meaningful information can be obtained from the plaintext thereof. The format of p/ is defined as follows:



ID	date	address	data	Ending flag	CRC

- wherein ID represents an identifier (or serial number) of p/, such that management nodes
- identify each data particle by its ID and perform data reconstitution upon request from a user;
- date represents the time at which p_i is created and the time it arrives at the next node, and serves as timestamps for determining unauthorized access;
- address represents the IP address of the storage node in which p_i is located;
- data represents the content of p_i;
- ending flag indicates the end of the transmission of p_i (0 indicates there are subsequent data packets, whereas 1 indicates no subsequent data packet is present); and
- CRC can be error-checked by hashing.

The above parameters can be classified as fixed values and variable values: the fixed values including ID and data remain constant throughout the lifecycle of each data particle; while the variable values including date, address, ending flag and CRC are always changing after the data particles begins to drift.

For the balance between security and attraction, the file is fragmented in the following way:

The information entropy of the data portion is kept within an interval, i.e., [h2, h7], where h2 and h7 are the lower bound and upper bound of entropy, respectively. That is, the information entropy of the data portion has to attract a hacker's interest while not divulging too much information. This
can be achieved by control the length of information content in the data portion.

The entropy is calculated as follows:

$$\mathrm{H(data)} = \sum \log_2(p_i)$$

where p_i is the probability value of words/symbols/identifiers in the data portion.

3. Movement of file particles

After the file fragmentation, each data particle needs to randomly and irregularly jump among different storage nodes, i.e., perform data drifting. Storage locations of the data particles are different in different periods of time (the allocation of storage locations is random).

4. Detection of unauthorized access

The status of unauthorized access to the file particles themselves is detected in every intermediate state during the movement of the file particles.

Integrity: whether there is any addition, deletion or modification to a particle.

Availability: whether a particle is quarantined, interrupted or delayed.

Controllability: whether password protect is compromised.

Credibility: whether the trusted attributes of the data are compromised.

If the unauthorized access involves a certain file particle, the location where the access occurred is recorded into threat statistics.

Distributed detection nodes (can coincide with existing storage nodes) are deployed in the network while the file particles are flowing among different nodes during their movement. The file attributes are detected using the detection nodes to find out unauthorized access behavior.

A file particle f is taken as an example.

Integrity: the hash value of the data particle is detected, and data analysis is performed when the hash value does not match the hash value of intercepted data. If the probability of occurrence is greater than the bit error rate of the communication channel, unauthorized access by a hacker is confirmed.

The integrity of f is recorded using f^a, such that $f^a = 0$ if the integrity is not compromised, or $f^a =1$ otherwise.

Availability: data particles within a segment of the network are sampled per unit time. If the delay of the sampling exceeds a threshold t_u, the availability is considered compromised by the hacker, and thereby unauthorized access can be confirmed. The availability of f is recorded using f^b, such that $f^b = 0$ if the availability is not compromised, or $f^b = 1$ otherwise.

Controllability: the nearest password authentication interface is checked to count the number of erroneous password input. Password threats are cumulated according to routing distance attenuation.

The controllability of f is recorded using f^c, such that the controllability of f^c is divided into 2^n levels, with level 0 representing the highest controllability.

Credibility: the number of pieces of redundant information and copies in the transmission of data particles, such that unauthorized access by the hacker is confirmed when the redundancy exceeds a set value.

The credibility of f is recorded using f^d, such that the credibility of f is divided into 2^m levels, with level 0 representing the highest controllability.

The security $\{f^a, f^b, f^c, f^d\}$ of f is stored in distributed nodes of the network to create a distributed and shared trust relationship, such that tampering of information will be detected immediately.

5. Active response to unauthorized access

The movement of each one of the file particles in step 3 is random with equal probability under normal circumstances; however, when a suspected unauthorized access event is detected in step 4, the present invention further verifies the unauthorized access in a unique way by actively responding to the event. In particular, the randomness of the movement of the file particles is tuned to actively move file particles near a location where the unauthorized access event is detected to that location without drastically changing the performance and parameters of the system.

Taking the location of node A as an example:

Before an unauthorized access event is detected, if the probability that any file particle will flow through A is: p, then after the detection of unauthorized access to A, the probability that any file particle will flow through A is adjusted to p', where p'>p.

As described above, once the unauthorized user successfully completes an unauthorized access, he/she will certainly take further actions to access information without authorization in the same location to expand the result, and will therefore try to access the newly added sensitive file particles furthermore. This just provides a more concrete basis confirming the detected unauthorized access event in step 4.

6. Security prediction

When the statistics of threats obtained in step 4 satisfy certain statistical conditions, for example, the number of threats reaches 100,000 and the statistical time reaches 48 hours, all of the data are aggregated for analysis of high-frequency events (for example, through machine learning, but not limited thereto) so as to predict the security for a future period of time.

]The method provided in the present invention performs real environment testing and collects statistical information based on security threats to sensitive data without considering the specific causes of the security threats. In this way, complex technical problems of security traceability are avoided by using only the statistical results, making the method a currently viable solution to security prediction. Moreover, sensitive files are broken into data particles so that information is well protected, making unauthorized users unable to obtain complete information.

IV. Conclusion

Sensitive files are divided into particles which organize a totally decentralized network. These particles drift randomly in the network. Suspicious attacks are verified by actively and directly moving some particles to the attacked point. It is analogous to the reaction to virus of white blood cells in the immune system. This decentralized and dynamically network based security system is much safer and more robust comparing with existing centralized and stable ones.

Our predication system reacts quickly to cyber attacks and predication accuracy is significantly improved using active moving mechanism.

Acknowledgment

This research was originally submitted to Xinova, LLC by the author in response to a Request for Invention. It is among several submissions that Xinova has chosen to make available to the wider community. The author wishes to thank Xinova, LLC for their funding support of this research.

More information about Xinova, LLC is available at www.xinova.com.