2 Technologies for Security of the Internet

2-1 A Study on Process Model for Internet Risk Analysis

NAKAO Koji, MARUYAMA Yuko, OHKOUCHI Kazuya, MATSUMOTO Fumiko, and MORIYAMA Eimatsu

Security Incidents such as Scans and Probes, Computer Intrusions, Malicious Software (Viruses, Worms, etc), Computer Sabotage and Damages (DoS Attacks, etc.) are well recognized in our IT environments today. One of the key solutions to protect against the above security incidents and to minimize damages, "INTERNET RISK" should be efficiently observed by means of incident monitoring and analysis in conjunction with supports from ISPs (Internet Service Providers). This paper discusses the requirements of the incident analysis and proposes the model for Internet Risk Analysis. This activity was carried out with a support of Telecom-ISAC Japan.

Keywords
Network security, Data mining, Log analysis, Analysis model

1 Introduction

With the widespread use of broadband Internet, encouraged by the Japanese government's IT strategy, the information and communications infrastructure is becoming increasingly important throughout society. Ensuring the safety of this infrastructure is now a major issue.

If we are to promote a more sophisticated network infrastructure and further develop information and communications technologies, measures for network security and against cyber attacks are essential and urgently required. To implement such measures we have to analyze the state of various network risks so as to prevent or reduce such risks. To assess network risk, advanced incident analysis will be critical, including trend- and frequency-based analysis of security logs and network traffic data. These analysis results will then be applied to assess the severity and effects of network-related risks, and to facilitate the implementation of measures that will ensure information and communications infrastructure safety.

As a first step, we performed a study of a process model, including collection and storage of data, pre-analysis processing, and advanced analysis, all in collaboration with various Internet service providers (ISPs), with the ultimate aim of identifying risks with the absolute minimum delay.

This paper will discuss the methods of collection and storage of data, as well as this process model that has been subjected to pre-analysis processing and precise analysis. The paper will also discuss the development of a

specific system and collaborative efforts involving Telecom-ISAC Japan.

2 Background and research requirements

Technology for monitoring and analysis of various Internet logs is researched and commercialized using log data issued from an intrusion detection system (IDS) or firewall (FW). This chapter will describe current research trends and requirements in these areas.

2.1 Current activities related to log analysis

In many cases, log data from an IDS or FW is used to perform statistical analysis to determine the state of network risk. SOCs (security operation centers) of various security vendors currently analyze a wide range of logs and offer analysis results to users on a commercial basis. These vendors may perform statistical analysis or simple correlation analysis of these IDS/FW logs.

In performing statistical analysis, these systems use network parameters (port numbers, destinations, etc.) to analyze trends. When similar incidents occur, the systems perform correlation analysis as a filter to collate similar data so that unnecessary analysis can be avoided.

These activities are often carried out to mainly provide information of analysis results in understanding the spread of worms/viruses and in the detection of various intrusions.

Recently, groups including Japan's JPCERT/CC[1], IPA[2], @Police[3], and Telecom-ISAC Japan[4], in addition to the South Korean KRCERT, have been performing trend analysis through the wide deployment of IDSs to collect information on adverse incidents at disparate monitoring points. In this context, we identified the requirements of currently conducted log analysis, as described in the next subsection.

2.2 Requirements of network risk analysis

Taking into consideration of the current status of log analysis mentioned above, if Internet risk analysis is to be performed more accurately and effectively across a wider area, the following requirements must be met:

(1) Requirements concerning information to be monitored

The organizations mentioned above currently monitor IDS/FW incident logs. However, traffic information should also be collected in collaboration with ISPs, to permit integrated analysis of these two sources of information.

(2) Requirements concerning monitoring areas

Monitoring is currently limited mainly to the user side (within companies, among general users, etc.). To perform comprehensive monitoring of wider areas requires increased monitoring at the ISPs that provide network services.

(3) Requirements concerning high-capacity processing

When collecting traffic information in addition to incident logs, the amount of collected data will be enormous. Processing capacity must therefore be high in order to process this enormous amount of data efficiently. In this context we must consider the use of a filtering function to remove unnecessary information, as opposed to simply increasing processing capacity.

(4) Requirements concerning accuracy of analysis of monitored data

To analyze a large amount of monitored data quickly and accurately, we should consider performing analysis in a hierarchical manner or combining a number of analysis methods (such as code static analysis and dynamic analysis).

(5) Requirements concerning confidentiality of monitored data

Care is essential in handling monitored data, given the importance of protecting private information and in light of ISP restrictions concerning the confidentiality of collected data. In this context as well, we should con-

sider the use of a filtering function, which is an effective technical means of blocking non-essential information prior to analysis of collected data.

3 Process model of internet risk analysis

To address the risk analysis requirements described in Section **2**, we are designing an analytical process based on a specific system configuration. As shown in Fig.1, this risk analysis process model consists of the following:

3.1 Monitored data sources

Through collaboration with Telecom-ISAC Japan, which provides support for stable and safe ISP operations, incident logs are collected from IDSs or firewalls and traffic information is gathered from ISPs at disparate monitoring points. In addition to analysis of incident logs and traffic information, these two categories of monitored data are correlated for increased accuracy in analysis. The process model is then designed such that both offline static data and online dynamic data can be used. In the event of trouble anywhere within a given ISP's area, local (offline) static data is analyzed to resolve the problem.

3.2 Filtering / digest processing module

In this module, filtering or digest processing facilitates analysis of monitored data. Specifically, unnecessary information is filtered without deleting necessary data; the smaller resultant data set is thus easier to handle. Digest processing compacts data based on a range of parameters, including UnixTime, IP addresses (source/destination), and payload length (for TCP, UDP, and ICMP) and port numbers (source/destination) (for TCP and UDP). Additional parameters for TCP involve flags (IP/TCP) and HTTP methods, while ICMP also includes type and code parameters. To normalize and streamline large amounts of data by removing unnecessary payload, corre-

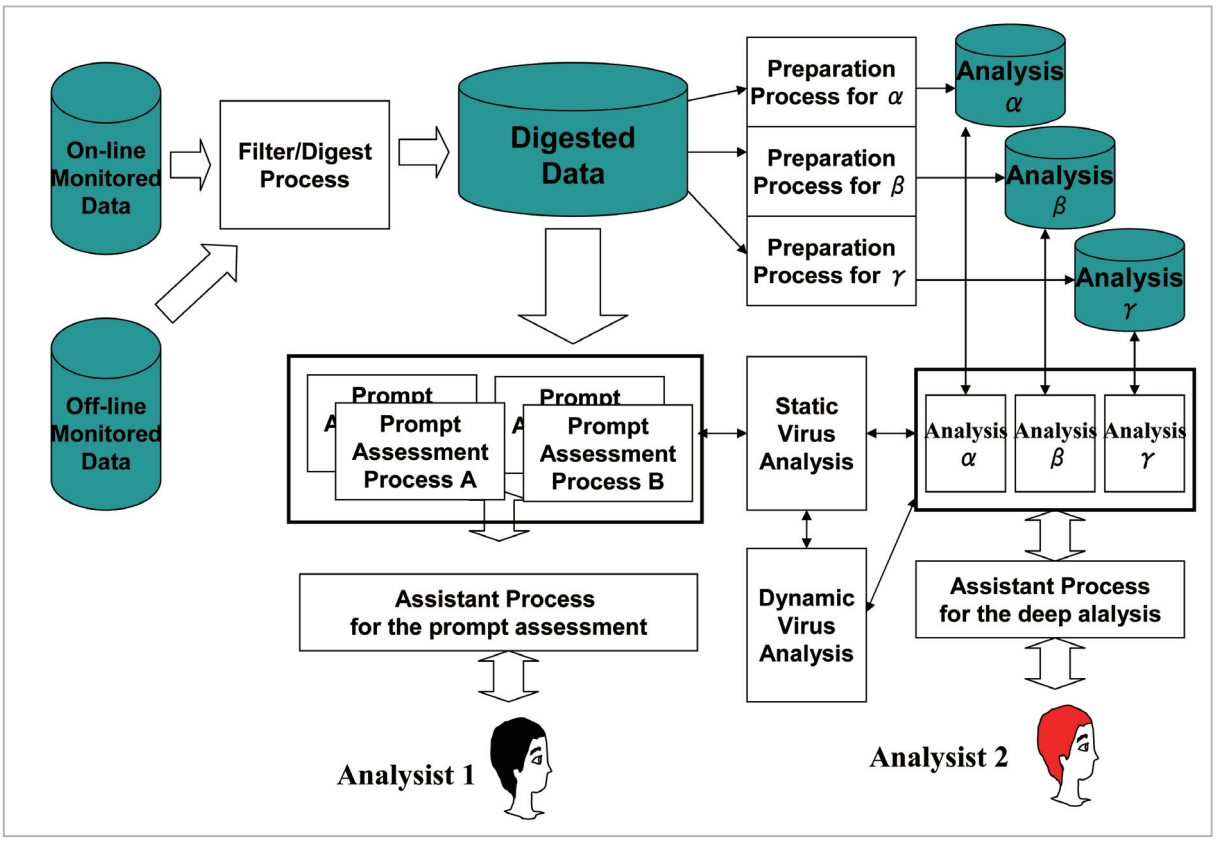

Fig. 1 Internet risk analysis model

lation processing is used as a filter, for example to determine the presence of a significant number of similar attacks. Digest processing, on the other hand, requires so-called "spiral" study of analytical methods and results (both primary and secondary). The aim of filtering and digest processing is to reduce data amounts by 90%.

3.3 Digest database

Following the filtering and digest processing mentioned above, the data is used in subsequent analysis (primary and secondary). A digest database consists of these data items, with parameters optimized for analysis. The database stores data issued from several monitoring points. For example, if the sizes of incident logs and traffic data are approximately 1 GB and 500 GB respectively per day, several terabits of data will be accumulated over the course of several days. Data storage periods and methods must therefore be taken into account.

3.4 Primary analytical processing module

The purpose of primary analytical processing is to assess the current or future state of Internet-related risks very rapidly. This module is thus designed to incorporate analytical processes in which accumulated digest data is analyzed quickly based on the required parameters, in what is referred to as "primary analysis". Specifically, initial statistical processing is performed based on parameters such as IP addresses, ports, destinations, and applications. The primary analysis results are presented to primary analysts (discussed in the next section) via dedicated display to enable further analysis utilizing different combinations of parameters with their support. If these analysts determine that urgent measures are necessary, the relevant ISP implements a primary response with the guidance of Telecom-ISAC Japan.

3.5 Primary analytical support module

In this module, primary analysts provide the primary analytical processing mentioned above. If based on primary analysis results these analysts determine that further statistical analysis is necessary using different combinations of parameters, a request for reanalysis is submitted to the analytical processing module to perform the necessary statistical analysis. The specific procedure is described in another paper[5].

3.6 Secondary analytical preprocessing modules

Secondary analysis is more accurate and detailed than primary analysis. Preprocessing is performed to convert data in the digest database into a form suitable for data mining in secondary analysis. Specifically, the digest data is compiled in various ways to create new data items for analysis. The formulation of these items clarifies the nature of the data in question and allows for greater efficiency in subsequent analysis.

Examples of preprocessing include (1) the summarization of data for a certain period and (2) data summary based on a key data item. As shown in Fig.2, we used an HTTP method to record the number of hits on a server per hour and plotted variations over time. The results clearly indicate the characteristics of a virus variant not otherwise detectable by simply observing the logs.

The key success factor for the secondary analysis depends on the preprocessing. Therefore, sufficient time must be spent preparing data items for use in analysis. A specific preprocessing module will be required for each of the secondary analytical methods.

3.7 Databases for secondary analysis

Each of the secondary analytical processes has a database to store analysis results. These databases are independent of the corresponding processes, although the databases are not required to be stored in physically separate modules.

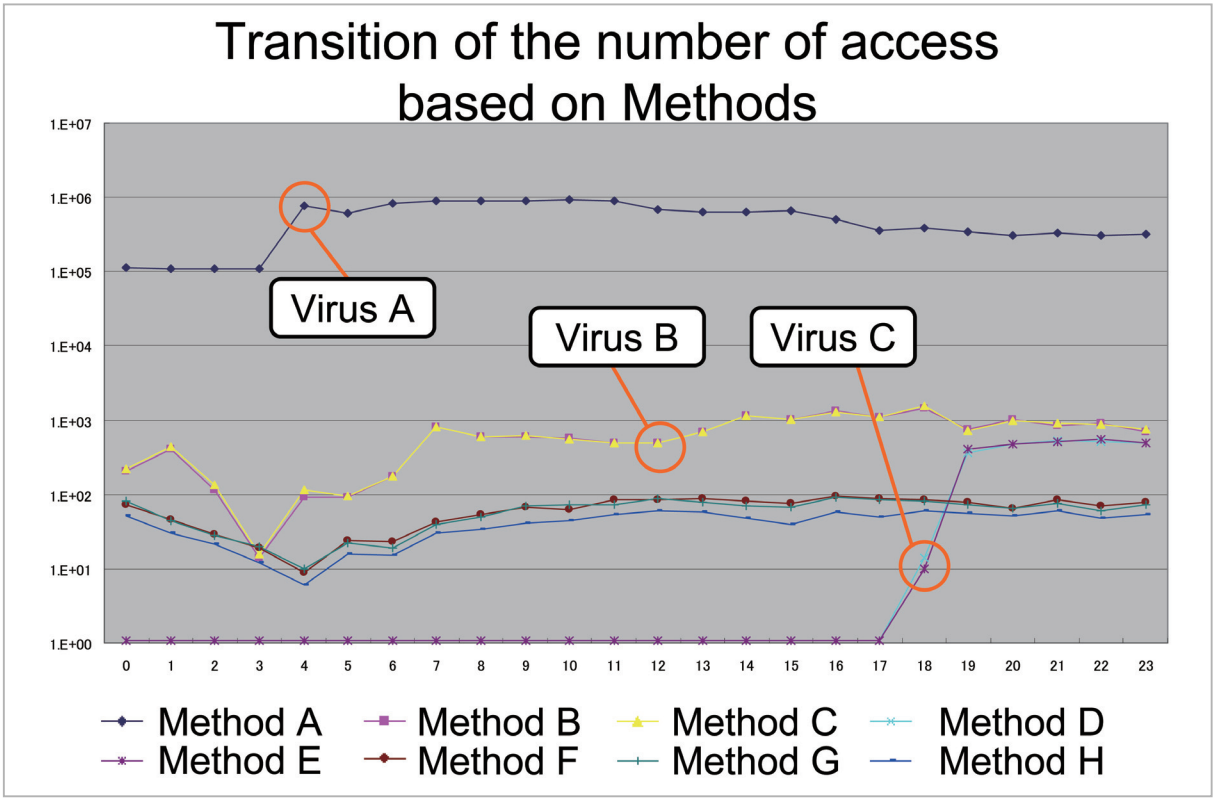

Fig.2 *Display of preprocessed data using HTTP method*

3.8 Secondary analytical processing module

The purposes of secondary analytical processing are to enable precise analysis (to detect slight variations in change points or feature points) and to assess the state of Internet risk in detail. Primary analysis cannot achieve these goals. However, instead of setting up given secondary analytical methods, our aim is to assess a variety of methods, which we will then continuously evaluate in terms of accuracy and effectiveness as we work to establish the most effective approach. We are currently evaluating what are referred to as the "anomaly detection" and "change point detection" methods. Both of these methods examine overall data patterns on a statistical basis, detecting deviations from a known statistical model[6]. These methods adapt to changes in data while discounting moderate amounts of past data, with the aim of detecting current anomalies in real time.

- Anomaly detection method

This method is designed to detect anomalies based on large amounts of access history data. Specifically, this method detects anomalous sessions (i.e., those with noticeable deviations from overall access patterns) and anomalous patterns. If sessions featuring an unknown pattern occur in clusters, the anomaly detection method will identify this as a new behavior pattern. Newly added or eliminated patterns can be identified by calculating, based on a dynamic model selection theory, an optimal number of time-variant behavior patterns[7].

This method is effective in detecting attacks by worms and vulnerabilities whose access destination or sequence differs noticeably from that of normal sessions, and also serves as a tool for identifying access patterns upon occurrence of DoS attacks or worms.

- Change point detection method[6]

This method is designed to detect sudden changes in time-series traffic data. Specifically, this method detects 1) outliers (values that deviate noticeably from overall patterns of time-series traffic) and 2) change points

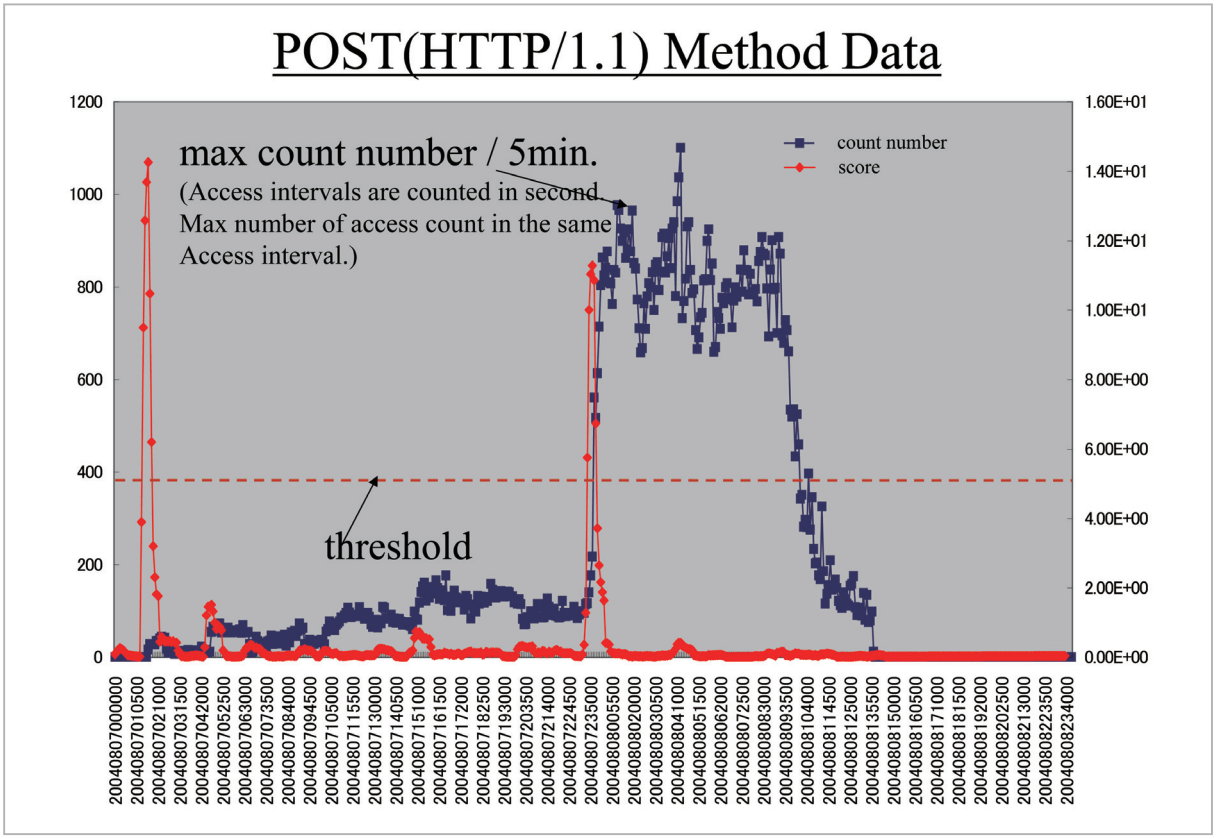

Fig.3 *Example of detection of a change point (Anntiny example)*

(points at which traffic suddenly changes). Focusing on these anomalies allows ready detection of any sudden increase in traffic due to DoS attacks or worms. Figure 3 shows an example of detection of a change point using a POST (HTTP/1.1) method in the case of the Anntiny worm.

3.9 Secondary analytical processing support module

In this module, secondary analysts provide support in the secondary analytical processing mentioned above. As in the case of primary analytical processing, if secondary analysts determine, based on secondary analysis results that further complex (integrated) analysis is required using different combinations of parameters, a request for reanalysis is submitted to the analytical processing module to perform the precise analysis required in secondary analysis.

3.10 Static analysis module

In this module, static analysis is performed of the code for a worm or virus, to verify programmed behavior in advance. The results of this analysis prove extremely useful as basic data in primary and secondary analytical processing.

3.11 Dynamic analysis module

We cannot analyze actual behavior using the static analysis mentioned above. Thus with this module, a worm or virus is run within a closed virtualized area to analyze actual behavior and to assess its effects on a network. The results gained from this module are particularly useful in secondary analysis. Dynamic behavior analysis enables verification of the results of this precise analysis in a spiral manner. The specific procedure involved is described in another paper[8].

4 Discussion

Up to this point, we have described a process model of Internet risk analysis. In this chapter, we will discuss the above model in relation to the log analysis requirements and the collaboration with Telecom-ISAC Japan.

4.1 Requirements

(1) Requirements for monitored information

In collaboration with Telecom-ISAC Japan, we are now working to establish a system to collect a maximum amount of traffic information and incident log data.

(2) Requirements for monitoring areas

Although we have not yet increased the number of monitoring points, we plan to provide monitors at several ISPs in collaboration with Telecom-ISAC Japan. We will install monitors at these ISPs as well as at remote end users to find the best way to collect data. In addition to efforts relating to data collection at monitoring points, we are now studying methods for the temporary store of this data.

(3) Requirements for high-capacity processing

Efficient processing and analysis of large amounts of log data represents an enormous challenge. While large amounts of data can be filtered on the monitoring or collection side, given the amount of data and the storage periods involved filtering would appear necessary on both sides, along with the normalization of a certain amount of data. The storage period for collected data must therefore allow for analysis of fairly extended log periods.

When collecting data online, digest processing should be performed sequentially (in quasi-real time), as opposed to gathering large amounts of data all at once. This processing should be performed in parallel. Moreover, considering the temporary nature of data storage at multiple monitoring points as mentioned above, concurrent parallel and distributed processing may be also required.

(4) Requirements concerning accuracy of analysis of monitored data

This model consists of two parts: primary analytical processing, in which different items are processed in parallel for rapid trend analysis; and secondary analytical processing, consisting of more precise analysis. Since digest data is used in both primary and secondary analysis, we made it accessible at both stages. We also designed the system such that the results of static analysis of worms are accessible at both stages. Moreover, to enable precise and detailed test-bed analysis, we investigated a combination of dynamic analysis and secondary processing.

(5) Requirements for confidentiality of monitored data

Care is essential in handling monitored data, given the importance of protecting private information and in light of ISP restrictions concerning the confidentiality of collected data. In this context as well, we should consider the use of a filtering function, which is an effective technical means to discard non-essential information prior to analysis of collected data.

As described above, we are collaborating with Telecom-ISAC Japan in many segments of Internet risk analysis. This analysis model meets virtually all of the requirements of the individual segments. Going forward it will be necessary to verify the performance of these segments in the context of specific implementations and evaluations.

4.2 Collaboration with Telecom-ISAC Japan

Telecom-ISAC Japan is a consortium dedicated to stable and safe ISP operation. Its roles include information monitoring at individual ISPs, formulation of countermeasures against adverse incidents based on analysis results, release of essential information to ISPs, and collaboration with counterpart organizations in other countries.

We decided to work with Telecom-ISAC Japan specifically in the area of analysis, and are in the process of devising a system in which we receive monitored data from ISAC and return precise analysis results.

5 Conclusion

It is not easy to determine the precise state of Internet risk accurately and quickly. However, when working with Telecom-ISAC Japan, a consortium of ISPs, it becomes possible to monitor the latest data and to perform analysis with precision and speed. Applying the basic requirements of risk analysis, we formulated the analytical model described in this paper after a great deal of trial and discussion. We believe that this model will provide the first step toward significant future developments in the field.

To increase the reliability of analysis, many tasks remain to be addressed: more consistent collection of monitored data, incorporation of improvements suggested by analysts (both primary and secondary), development of additional analysis methods, and automated handling of analysis results. With the increasing importance of security in ISP operations and the growing threats to Internet security, we intend to continue to focus efforts on research that will lead to safer, more secure Internet operations.

As a final note, we would like to express our sincere gratitude to Mr. Baba and Mr. Suzuki of Yokogawa Electric Corporation and to the members of Telecom-ISAC Japan for their valuable advice and assistance.

References

1 JPCERT/CC Internet Scan Data Acquisition System (ISDAS) http://www.jpcert.or.jp/isdas/

2 IPA http://www.ipa.go.jp/security/

3 Security Potal Site (@police) http://www.cyberpolice.go.jp/detect/observation.html

4 Telecom-ISAC Japan https://www.telecom-isac.jp/

5 F.Matsumoto, S.Bab, Y.Izawa, and K.Nakao, "The framework and prototype of a network security analysis tool for the Internet Services Provider", SCIS 2005.

6 K.Yamanishi and J.Takeuchi, "A Unifying Approach to Detecting Outliers and Change-Points from Non stationary Data", The Eighth ACM SIGKDD International Conference on Data Mining and Knowledge Discovery, ACM Press, pp. 676-68.

7 Y.Maruyama and K.Yamanishi, "Dynamic Model Selection and Its Applications to Computer Security", The IEEE Information Theory Workshop 2004, (http://ee-wcl.tamu.edu/itw2004/)

8 M.Fujinaga, K.Nakao, and M.Morii, "A testbed for virus dissection".

NAKAO Koji

Group Leader, Information and Network Systems Department

Information Security

OHKOUCHI Kazuya

Expert Researcher, Security Advancement Group, Information and Network Systems Department

Data Mining

MORIYAMA Eimatsu

Senior Researcher, Security Advancement Group, Information and Network Systems Department

Security, Mobile Communication

MARUYAMA Yuko

Expert Researcher, Security Advancement Group, Information and Network Systems Department

Data Mining

MATSUMOTO Fumiko

Researcher, Security Advancement Group, Information and Network Systems Department

User Interface, Security Log Analysis