Peer-to-Peer File Sharing Communication Detection System Using Network Traffic Mining

Satoshi Togawa[1], Kazuhide Kanenishi[2], and Yoneo Yano[3]

[1] Faculty of Management and Information Science, Shikoku University,
123-1 Furukawa Ojin-cho Tokushima 771-1192, Japan
doors@shikoku-u.ac.jp
[2] Center for Advanced Information Technology, University of Tokushima,
2-1 Minami-Josanjima Tokushima 770-8506, Japan
marukin@cue.tokushima-u.ac.jp
[3] Institute of Technology and Science, University of Tokushima,
2-1 Minami-Josanjima Tokushima 770-8506, Japan
yano@is.tokushima-u.ac.jp

Abstract. In this research, we have built a system for network administrators that visualize the Peer-to-Peer (P2P) file sharing activities of network users. This system monitors network traffic and discerns traffic features using traffic mining. This system visualizes the P2P file sharing traffic activities of an organization by making the processing object not an individual user but a user group. The network administrator can comprehend the P2P sharing activities of the organization by referring to the map. This system extracts a traffic feature from captured IP packets that the users communicated. Afterwards this system creates a traffic model. The features of the traffic model are emphasized by weighting. After that, the traffic model is visualized by a Self-Organizing Map. The network administrator is assisted in understanding users' P2P file sharing communication behavior by this feature map. The administrator can then respond to the situation. As a result, we think we can assist the monitoring operation and network administration.

Keywords: Traffic Mining, Incident Response, Administrator Assistance, Peer-to-Peer Detection.

1 Introduction

Today, Peer-to-Peer (P2P) applications have become on the Internet. It is applied in the field of file sharing, VoIP and groupware. Especially, a lot of file sharing software has been designed on the P2P communication model. If Internet users want to get various kinds of data, the users can easily obtain various files and data using P2P file sharing software. The file content often includes music, movies and so on. However, because most of these files are extracted from music CDs and DVDs protected by copyright law, it is not appropriate to exchange these files. Moreover, popular P2P file sharing applications such as WinMX, Winny and Share need a huge bandwidth because these applications send and receive large amounts of data. As a result, regular communications are obstructed by P2P applications.

M.J. Smith, G. Salvendy (Eds.): Human Interface, Part I, HCII 2007, LNCS 4557, pp. 769–778, 2007.

In addition, a virus that causes data compromise has been generated for the P2P file sharing network. There are many cases of classified information being compromised because of these viruses. These viruses give out sensitive information stored in the user's computer using the publication function on the P2P file sharing application. When an organization's member uses a P2P file sharing application, there is a risk that the organizations security will be compromised. If certain classified information leaks, the data will pass along from place to place in the P2P network.

For example, many data compromise accidents occurred in Japan at 2006. A large amount of military intelligence and investigative information were leaked from Self-Defense Forces and police departments. In addition, a lot of companies leaked customers' private information. These accidents are extremely serious and can result in customers losing trust in an organization.

There are limitation techniques which use packet filters to limit illegal traffic that deviates from the policy established by the company or university. Filter technology, which synchronizes with the packet filter definition, is installed on the firewall, and illegal traffic that does not conform to the site policy are not forwarded to users. However, it is difficult to keep the filter definition perfectly set because the default destination port of each P2P application is different. Moreover, P2P applications such as Winny and Share select the destination port dynamically. This P2P application's traffic limitation can be impossibly strict. Therefore, P2P traffic cannot be limited only by filter technology based on packet filter definitions.

On the other hand, the network administrator can use a specialized firewall system to limit P2P traffic. These P2P traffic limitation techniques are based on the signature information that is extracted from illegal traffic. It is completely blocked when a signature matches a traffic pattern. However, when a signature does not match the traffic, illegal traffic is not restricted. Accordingly, the network administrator has to understand the P2P application's behavior in the organization's network traffic. If the administrator can understand the P2P application's behavior, the administrator can usually ascertain problems at an early stage.

At the present time, if an administrator wants to understand P2P application activities, protocol analysis can be used. However, this method is very labor intensive, and these methods only provide basic information like IP address/port number level classification. The network administrator really wants a result that shows where the P2P file sharing application is used?

For these reasons, we have developed a traffic visualization system for P2P communication detection and administrator assistance. This system provides a feature map of traffic behavior made up from results of network traffic mining. This system assists the monitoring operation of the administrator by showing the feature map that this system presents. As a result, we think that we can assist the monitoring operation of the administrator.

We pay attention to the traffic that the organization users send and receive. These features are extracted from this traffic using traffic mining. The features are the source/destination IP addresses and source/destination TCP (or UDP) port number and TCP flags. In addition, we pay attention to the results of the DNS query from internal clients. We found that when the P2P nodes try to find other nodes, the DNS query amounts are less than normal DNS queries such as Web browsing. We use these DNS query features for discerning P2P application behavior. Moreover, this system acquires the packet occurrence frequency and yield. Consequently, a traffic model is generated from the feature and packet occurrence frequency and a result of the DNS query.

The method of generating the model is the Vector Space Model. The similarity problem between traffic features is replaced with a cosine measure between vectors. Weighting is added to the obtained traffic model to emphasize feature quantity. Afterwards, a feature map is generated by using Self-Organizing Maps (SOM) from the traffic model. This algorithm maps multi-dimensional vectors on a two-dimensional plane.

This map shows an administrator which computer communicated to other computers and the volume of the communication. It expresses not only the summarized traffic amount but also each traffic type and behavior. It can be said that the feature map is a result of traffic mining from the users' traffic, and the administrator is assisted in understanding the organization's traffic behavior by this feature map.

In this paper, we proposed a system framework of traffic visualization for P2P communication detection, and we show a configuration of the prototype system. Next, we show the results of experimental use and examine these. Finally, we describe future study, and we show conclusions.

2 Assisting the Detection of P2P File Sharing Traffic

2.1 Framework of P2P File Sharing Traffic Detection

Fig. 1 shows a framework of administrator assistance for P2P file sharing traffic detection. We assist the monitoring and detecting operation of the network administrator by providing the traffic behavior of the organization users.

We paid attention to traffic between the internal site and the Internet. All users' traffic passes a gateway in the internal site. We collect all IP packets that pass the gateway. In addition, traffic features are extracted from collected IP packets.

In addition, we paid attention to the result of a DNS host queries from internal DNS servers to external DNS servers. Generally, the P2P nodes information is distributed without hostname (Fully Qualified Domain Name). In the result, the DNS host query amounts of the internal P2P nodes are less than normal applications. If the traffic feature of one host has different from other hosts, and that host's DNS host query amount is less than other hosts, it has high probability of the P2P node.

Consequently, a traffic model is generated from extracted features that the users communicated and the results of DNS host queries. The method of generating the traffic model is a Vector Space Model. As a result, the similarity problem between source IP addresses is replaced with a cosine measure between feature vectors. Weighting is added to the obtained traffic model to emphasize feature quantity.

A series of processing described here are traffic mining. Because, the feature related to P2P file sharing communication is extracted from all captured traffic by the

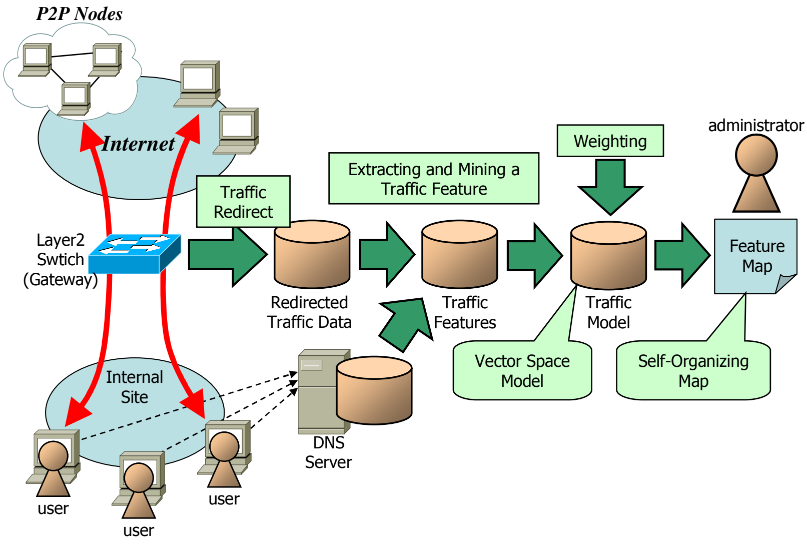

Fig. 1. Framework of administrator assistance for P2P files sharing traffic detection

series of processing. Moreover, extracted and emphasized features are stored to the traffic model. This model adapt to traffic feature visualization.

Afterwards, a feature map is generated by a Self-Organizing Map (SOM). SOM is an algorithm to map multi-dimensional vectors on a two-dimensional plane. As a result, this map expresses the typical source IP addresses that the users communicated.

The administrator gets a bird's-eye view of the organizations communication activities by referring to the map. Therefore, the administrator is assisted in understanding P2P traffic behavior by this feature map.

2.2 P2P Communication Model

Fig.2 is a hybrid type P2P file sharing architecture. Generally, hybrid P2P architecture has a central server which keeps all Meta information such as kind of files and file names. It is directory of user identities and index of resources on the P2P community. If the administrator wants to limit the use of hybrid P2P file sharing communication, it only has to block off the path to the central server.

However, this limitation technique is ineffectual for pure P2P architecture. Fig. 3 is a pure type P2P file sharing architecture. This architecture does not have central server. All information of sharing resources is stored to the some node. In this result, index information of sharing files is distributed on the pure P2P community, it is difficult to block off the path to the resources of sharing information.

Therefore, if an administrator wants to limit pure type P2P file sharing communications, the administrator must keep monitoring users' communication activity of organization.

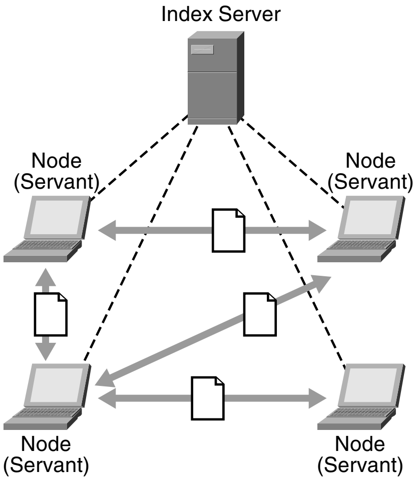

Fig. 2. Hybrid P2P communication model

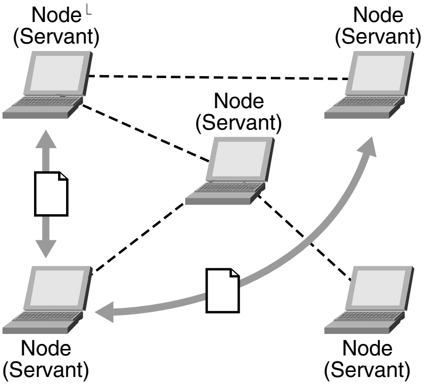

Fig. 3. Pure P2P communication model

2.3 Exploratory Experiment and Result for Traffic Feature Extraction

We made an exploratory experiment to clarify a feature of DNS host query by the P2P node. Especially, we want to clarify a DNS host query feature of the pure P2P nodes.

The P2P file sharing application was installed to experimental computer, and the experimental computer was used for 20 minutes with P2P application. After that, we generated general Web browsing traffic with other computers. Then both traffic was monitored and compared for this exploratory experiment.

Table 1. Measuring Results of pure P2P communication

Amount of sending IP packets	22,235
Amount of destination IP addresses	415
Appearance Ratio of TCP PUSH flag	35.2%
Appearance Ratio of DNS host query	Less than 0.1%

Table 2. Measuring Results of general Web browsing communication

Amount of sending IP packets	6,416
Amount of destination IP addresses	42
Appearance Ratio of TCP PUSH flag	9.2%
Appearance Ratio of DNS host query	91.8%

Table 1 and Table 2 show the measuring results of the exploratory experiment.

First of all, we can find a difference of the amount of sending IP packets. The case of pure P2P communication model much than Web browsing case per same measurement time. It is about 3.5 times larger than Web browsing case. What that means is that P2P communication model makes a lot of connections between internal node and P2P nodes on the Internet.

And then, it is understood that the appearance ratio of DNS host queries by P2P communication is remarkably low. In most situations, P2P nodes information is provided without Fully Qualified Domain Name. In that result, it is provided only IP addresses. Therefore, DNS hostname resolution is not required to make the connections between both P2P nodes. When the connection between P2P nodes is generated, the DNS host query is hardly generated. In that result, we can find striking difference of an appearance ratio of DNS host query between both communication models.

We can find small disparity of an appearance ratio of TCP PUSH flag between both communication models. And this feature is variable in amount. When we use appearance ratio of TCP PUSH flag, we have fear of erroneous decision for detecting P2P communication.

In this result, we think important features for detecting P2P communication are the appearance ratio of DNS host query and the amount of sending IP packets.

3 System Configuration

We show the configuration of proposed system in Fig. 4. This system has 5 modules that includes a “Traffic Collection Module”, “Traffic Analysis Module”, “DNS Query Analysis Module”, “Modeling Module” and “Visualization Module”. A detailed description of each module is provided below.

3.1 Traffic Collection Module

IP packets that users of organization sent and received are redirected by layer2 switch with port mirroring function. Traffic Collection Module accepts the redirected IP packets from layer2 switch. In addition, an Ethernet adapter configuration of this system is set to promiscuous mode. Because, this module have to accept all related IP packets. The accepted IP packets include normal traffic and illegal traffic, and all accepted IP packets are passed to the Traffic Analysis Module.

3.2 Traffic Analysis Module

This module attempts selection for all accepted IP packets. First of all, an administrator gives the IP address information of internal site servers to this module. And source traffic of the internal servers is dropped from all accepted traffic by using the given IP address information. Next, this module attempts to select traffic features from selected IP packets.

This module analyzes a packet field of selected IP packets, and some feature extracted from selected packets. The features are the source/destination IP address, and the source/destination TCP PORT number and TCP flags status. At the same time, each packets occurrence rate is calculated and stored. All extracted and calculated features are passed to the Modeling Module with other features generated from DNS Query Analysis Module.

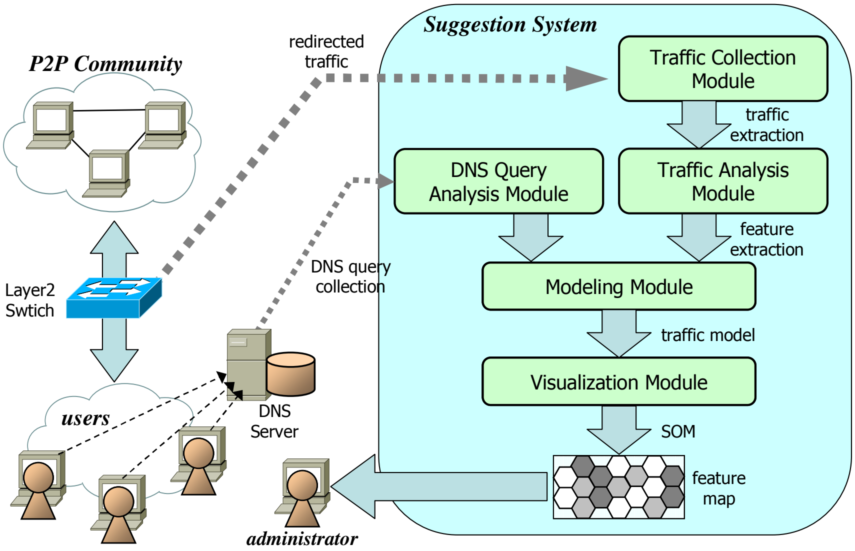

Fig. 4. System Configuration of Proposed System

3.3 DNS Query Analysis Module

The DNS server processes a DNS host resolution requests that was required from internal users. This module collects the results of DNS host resolution and requested client information from DNS server's log. It is selected excluding the incomplete results of DNS request. All extracted complete results of DNS hostname resolution are passed to the Modeling Module.

3.4 Modeling Module

This module generates a traffic model which is defined by the Vector Space Model. One source IP address corresponds to one multi-dimensionally composed vector, and each element of the multi-dimensional vector stores a number of destination IP address and the destination PORT number. We call this multi-dimensional vector a "feature vector". The number of feature vectors is the same as the total number of extracted source IP addresses. The set of these feature vectors becomes the traffic model.

The weighting process done to the feature vectors emphasizes the characteristics of the traffic model according to the occurrence rate with which the source IP address and the PUSH flags appear. As a result, if the module discovers frequency appearing source IP address, it is possible to find the P2P packet spreader host. When the weighting process is finished, the traffic model is passed to the Visualization Module.

3.5 Visualization Module

This module visualizes and making the feature map from the obtained traffic model. The Self-Organizing Map is used as a visualization method in this module. The source IP addresses of the processing object are self-organized by the SOM algorithm.

This results in a well-consolidated visual output that allows the administrator to get bird's-eye view of internal users' P2P communication activities.

4 Experimental Use and Result

4.1 Experimental Environment

This system was tested to confirm its effectiveness. We collected traffic from users belonging to one organization on November 20th 2006.

The amounts of observed data extracted from collected IP packets and generated feature vectors are presented in Table 3.

Table 4 shows the computer's specification of experimental use.

Table 3. Amount of Observed Data

Data Type	Amounts
Observed Data	1,423,592
Generated Feature Vectors	16,356

Table 4. Specification of the Experimental Use

CPU Specification	Intel Pentium4 3.2GHz
System Memory Capacity	1Gbytes
HDD Capacity	300Gbytes
Operating System	Linux (kernel 2.4.18)

4.2 Feature Map

Feature maps were generated once an hour in the experimental period. The traffic data was put into the system. The input packet amount was about 3,200,000 packets and the source IP address count that the system extracted was 720. Each feature map that the system presented had 320 elements, and each element corresponded to a summarized source IP addresses. The number of source IP addresses that the system extracted before clustering was 720. Therefore, the source the source IP addresses appearing in the map where communicated many times related with P2P communication by the computers.

We show the feature map in Fig. 5. This map is one period of the all generated feature maps. All display with application name on the map is marked by the hand for explanation.

We can find two large clusters on the map. These clusters are related to P2P type application's communication. The lower right cluster is related to P2P file sharing application. It is completely detected to P2P file sharing applications traffic. Unfortunately, the other one is not P2P file sharing application. This is related to Skype. However, Skype is based on the P2P type architecture. These application have a same behavior on the Internet, because each application has a lot of nodes on the P2P community.

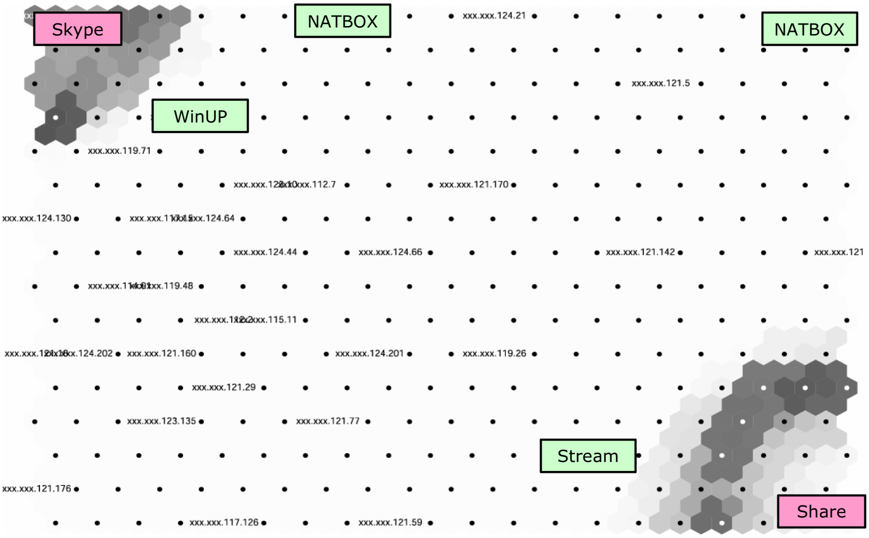

Fig. 5. Feature Map

As a result, we think that we can make complete to detect the P2P file sharing communication. The network administrator is assisted to detect the P2P file sharing traffic using this feature map. We think that the appearance ratio of DNS host query is especially effective for the feature map generating.

5 Conclusion

In this paper, we proposed a traffic visualization system for P2P communication detection. And we explained a configuration of the prototype system. And, we shown the results of experimental use and examine.

This system extracts records of P2P communication activities from the collected IP packets and the collected DNS query results. In addition, this system provides a feature map for the administrator. We developed a prototype system and experimented to confirm its effectiveness. It was shown that an administrator could inspect the results of the feature map.

References

1. WinMX Web site, http://www.winmx.com/
2. Winny Web site, http://www.geocities.co.jp/SiliconValley/2949/
3. BitTorrent Web site, http://bittorrent.com/
4. Skype Web site, http://www.skype.com/
5. Kohonen, T.: Self-Organizing Maps, 3rd edn. Springer, Heidelberg (2001)

6. Togawa, S., Kanenishi, K., Yano, Y.: WAVISABI: Users Activity Visualization System for Administrator Assistance based on Web Browsing Behavior. IPSJ Journal, Information Processing Society of Japan 46(4), 985–994 (2005)
7. Yoshida, K., Katsuno, S., Ano, S., Yamazaki, K., Tsuru, M.: Stream Mining for Network Management. IEICE Trans. Communication E89-B(6), 1774–1780 (2006)