July 16, 1999
RT0325
Network 6 pages

Research Report

Study on an Inter-Organization Activity based on the Internet Traffic

Takayuki Kushida

IBM Research, Tokyo Research Laboratory
IBM Japan, Ltd.
1623-14 Shimotsuruma, Yamato
Kanagawa 242-8502, Japan

Research Division
Almaden - Austin - Beijing - Haifa - India - T. J. Watson - Tokyo - Zurich

Study on an Inter-organization Activity based on the Internet Traffic

Takayuki Kushida
IBM Research, Tokyo Research Laboratory
E-mail: kushida@trl.ibm.co.jp

Abstract

A traffic analysis on the Internet has an advantage for obtaining the characteristics of transferred packets. There were many studies to understand the characteristics of the Internet traffic with mathematical statistical approach. The approach of this study is different from previous studies. It focuses on studying the traffic analysis with the framework of knowledge extraction. The measurement results of traffic capturing on the backbone network is applied to the inter-organization activity on the Internet. It shows that the proposed framework can be useful for the experiment results.

1. Introduction

There are many applications on the Internet as a communication media. The Internet becomes a common infrastructure for our daily activity. In this results, the amount of traffic on the Internet is exponentially increasing recently [1]. In the future, most of transactions for the inter-organization and intra-organization will be used on the Internet. As the usage of the Internet is increasing, the traffic analysis has an important rule for both the economical and the social activity.

The Internet traffic is aggregated with IP (Internet Protocol) packets. Each packet has the different source and destination address in the packet header. If these packets are gathered and analyzed by certain parameters which are the volume, the granularity of the time, the IP addresses and the transport ports, the characteristics of traffic is obtained with the results of the mathematical analysis.

There are so many transferred IP packets on the Internet, and we can measure the number of packets and the volume which is obtained with the sum of IP packet length. If these packets are analyzed with the source and the destination of IP addresses, the traffic flow between hosts is provided. If packet header information on TCP/UDP is analyzed, the application traffic of the network is understood with the ports number and the state of the transport protocol. If the results of the traffic analysis is provided, the basic information for the Internet is obtained.

The traffic measurement and analysis for Internet traffic is mainly focused on the characterization of the traffic with mathematical statistical method. The empirical method is also used for obtaining the characteristics of traffic [9]. The results of previous studies of traffic analysis were not directly related to the activity of inter-organization or intra-organization. They usually described a general model for the characteristics of the Internet. There is no common procedure to acquire the activity of the inter-organization and intra-organization from the Internet traffic.

This paper describes the framework of the knowledge extraction from Internet traffic, and the focused area of the investigation is the inter-organization activity which is analyzed with transactions on the Internet. The main difference from previous studies is that the obtained results in this study is directly related to the daily activity. The characteristics of the inter-organization activity will be used to understand the economical behavior and the human behavior. The rest of this paper is constructed as follows: related work of traffic analysis for previous studies, the method for studying inter-organization activity, the results from experiment and discussion and future direction of this approach.

2. Related work

Previous studies of the network traffic have been usually based on either the statistical or the empirical approach of the analysis. In early days in the traffic analysis of a packet switched network, it was assumed that the characteristics of Internet traffic was based on the Poisson distribution, which is the similar characteristics as the distribution of telephone calls. The statistical distribution of the Internet traffic was stud-

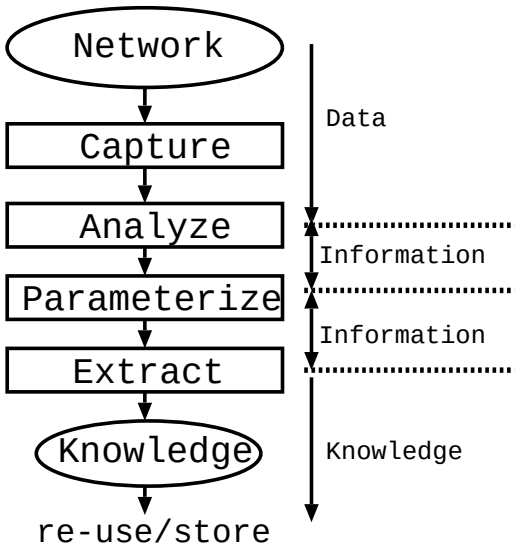

Figure 1. Workflow for understanding Internet traffic

ied for fifteen wide area traffic traces, and it showed that the wide area traffic is less persuasively Poisson distribution[2]. Due to the failure of the Poisson distribution for the wide area traffic, the new model was needed. The study proved that the characteristics of traffic both local area network and wide area network showed statistically self-similar manner[3]. The self-similar traffic pattern is one in which there is no natural length of a burst at every time scale ranging from a few milliseconds to a few minutes or even hours, and in which similar-looking traffic bursts are evident.

The characteristics of the WWW traffic on the Internet is also accompanyed with the self-similar [4]. In this result of the long-term traffic gathering, the traffic has the characteristics of self-similar manner which is related to document sizes of contents, caching effect, and size of file transfer.

A traffic analysis to construct the model of "cascades" which are also known as multiplicative processes was studied with the wavelet-based analysis and inference tools for the wide area network [5]. They proved that the measured network traffic well conforms to an underlying cascade construction and identify aspects of network traffic where its multiactive properties could be examined in detail.

In addition to these previous statistical studies, there are studies for the development from traffic data and the evaluation of the traffic measurement system. They have the important rule for traffic studies. The architecture and the implementation of the passive measurement system for Internet traffic was described in [7], and it could measure protocol performance metrics with the passive monitoring method. It is useful to check the validity of Internet applications at the protocol level and to show the characteristics of Internet traffic at the protocol level. It was well applied to the traffic analysis on the Internet.

On the other hand, the security demand of the Internet is rapidly increasing for using critical applications in business. The method of extracting the characteristics of the traffic is currently focused on checking the security which is the intrusion detection on the Internet. These applications of the security tool for traffic measurement take the critical mission which is to protect information asset from intrusion and attacking of unknown users[8]. The technology of security analysis is in general to monitor danger or unknown incoming packets from the Internet, and evaluates these packets whether the attacking and the intrusion is begun. There are some variations and patterns for this kind of the activity, and detection programs for the activity could automatically detect the characteristics with limited information in packets, and immediately report to the operator with priority level information. The data of the characteristics for the activity has already stored in the component of the program before it was installed in the system. Although the technology is specially customerized for the security purpose, it can be applied to check the activity of inter-organizaiton.

3. Method

In this paper, we define the workflow for the knowledge extraction in figure 1. Figure 1 shows four different tasks to extract knowledge from raw packets on the Internet. These tasks are called "Capture", "Analyze", "Parameterize" and "Extract".

For logging method of packets on the network, there are three methods for the measurement point. These are defined as follows: (1)server logging, (2)client logging and (3)intermediate logging. They are explained as follows:

- Server logging
  This method is based on capturing information of transaction in applications on servers. In this method, the granurality of the logging information therefore depends on the function of the server software. For example, the message exchange server such as the SMTP server has the logging capability for access information to the server from both local users and remote hosts, and store logging information to the file. The ad-

vantage of the method is that the logging capability is usually included in the server software. The disadvantage is that if there are multiple servers at the same service, all servers have to keep logging information for all transactions and anlayze them among servers with synchronization.

- Client logging
  This method is based on capturing event information for transactions in client software. For example, the WWW client software can intercept all events of transactions in the client software and store them in the file. The advantage is that logging information can be precisely monitored and the granularity of these events is smaller than the server logging. The disadvantage of this method is that the event monitor function has to load in the target client. If there are many clients that are needed to monitor, the monitoring becomes the huge task. There is no central management point for logging of multiple clients. As a results of the analysis for the client logging, user's behavior in detail will be understood. It is useful to profile and categorize the client.
- Intermediate host logging
  This method is based on monitoring transactions at the intermediate host which is located between the client and the server. The monitor of the network traffic is usually categorized into this method. Logging information to monitor traffic is exactly all transactions between the client and the server even if there are multiple servers for the same service. The capability of the central monitor can be provided with this method. The advantage of the method is that inter-organization and intra-organization transactions can be captured, and the behavior of all transactions on the network can be obtained by the analysis of transport protocol. The disadvantage of the method is that the measurement point should be carefully located to obtain all transactions of the inter-organization activity.

The intermediate host logging method is applied to study on the inter-organization activity.

3.1. Operation of the traffic analysis

A workflow of the traffic analysis is as follows: (1) to capture packets on the network, (2) to store these packets to the file and (3) to analyze these packets with different parameters. The key feature of the traffic analysis would be the procedure (3) which could parameterize tacit information in all packet headers. Lists of values obtained with parameters are essential information for the analysis results. Graphical representations with various forms are effective to understand the results of traffic data. The analysts of the network traffic will extract a basic feature of the current network status with graphical representations.

We defined three methods to understand the Internet traffic. They are described to categorize the analysis of Internet traffic as follows:

- Manual operation
  The manual operation is that there is no automatic processing to obtain the results of the traffic analysis. Analysts which have the specific skill and the expertise are needed to manipulate the results with graphical representation to make the proper decision. At least, they have tools to manipulate the traffic.
- Semi-automatic operation
  The semi-automatic operation is defined that the human operation is needed to know the specific information on the network and to extract a knowledge for the operation. This operation is the primary challenge of the study to obtain the specific knowledge from traffic.
- Full automatic operation
  A program automatically obtains the specific knowledge without any human intervention. The higher abstract level of knowledge will be obtained from the traffic, and the system can predict the future trend of the traffic. The operation would be an ideal operation, but it is not available today.

3.2. Extract useful information

The measurement method of the study is based on monitoring transferred packets on the intermediate host. Extracted information from capturing IP packets is parameterized in follows values: (1) capture time (in second), (2) length of IP packets (in bytes) which is information on packet header, IP address pair, transport protocol number and TCP/UDP port pair and (3) application protocol data.

The capture time is defined at the time of capturing the packet, and the granularity of the time is in milli second because the measurement period is rather long.

There are two IP addresses, which are source and destination addresses, in the packet header. They are directly binded to the domain name which can be obtained by the access to DNS (Domain Name System).

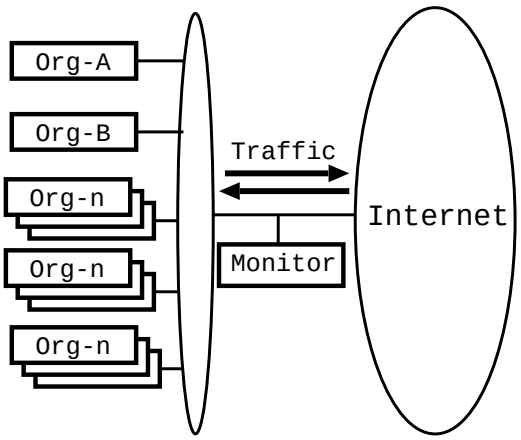

Figure 2. Configuration of the measurement system

The domain name is belong to the organization to which the packet is originated and is destined.

Transport protocols, such as TCP and UDP include the pair of port numbers in their packet headers. If the port number is under 1024, it are called the well-known port which usually represents the application service on the server. The application program listens incoming requests from the client to check the well-known port. When the well-known port number is obtained, information on the application service is provided. The application for the well-known port is defined by Internet Activity Board (IAB).

4. Measurement and results

This section describes the measurement system and the results with the system. The results shows the translation to information from the raw packet, and the effort to extract knowledge from information. In this study, the backbone traffic was evaluated with the framework which described in previous section. The detail of the measurement system for the backbone traffic was reported[6]. Figure 2 shows the configuration of the measurement system. All transferred packets from organizations to the Internet can be captured at the NOC LAN. In this figure, organizations are connected to the backbone network and the backbone network is connected to the rest of the Internet. All traffic to the Internet can be monitored at the "Monitor" point in figure 2.

Captured data of packet header is that the capture time(32 bit), the source and the destination IP address(32 bit), the source and destination ports(16 bit). The total length of captured data for a IP packet needs 16 bytes. The number of packets were approximately 2,500,000 for the measurement in this study. The volume of measurement for an hour is approximately 15 Mbytes, and therefore the volume of one day would be approximately 360 Mbytes. Although the accumulation of packet header on the backbone network is rather large, the detail analysis of these packet headers takes time. When only a few fixed parameters are needed for the analysis, the measurement data can be deleted after the parameterized data is extracted.

Figure 3 shows TCP traffic on the backbone network with volume. Since more than 94 % in IP traffic occupied with TCP traffic, UDP traffic could be omitted from this study because of less traffic. In this results, there is a definition of the threshold for traffic volume to analyze the traffic of the further study. For example, if UDP traffic volume is more than 20 % for the entire traffic, the further analysis will be done. In this case, as a knowledge, the percentage of traffic can be used for the threshold of the analysis.

Figure 4 shows the distribution of TCP volume for the bandwidth, and the TCP traffic has the bandwidth between 100 KB/s and 225 KB/s. The figure shows that the TCP traffic usually falls in the range of the bandwidth. Information on the range from the experiment are stored for the future comparison as a reference information.

In the measurement, after the IP packet header is examined, both source and destination IP addresses are changed to domain names with access to DNS. The part of domain name depends on the organization name: e.g., The domain name, '.ibm.co.jp" designates IBM Japan, and ".trl.ibm.co.jp" is Tokyo Research Laboratory. This is the simple procedure for the conversion of IP address to domain name which shows the organization name. The matching rule of this procedure is the simple, and in this results the domain name is provided from IP address. This conversion is the part of the procedure for data to information.

Figure 5 and 6 show that organization activity was divided into applications on TCP. The activity categorized into WWW, FTP and other applications. In these figures, other applications in the activity of organization A has different characteristics than that of organization B. In figure 6 there are sharp peaks for the measurement, but in figure 5 there is no such a peak. The characteristics of peaks can be stored as information.

Figure 7 and 8 show the volume for target organizations. Both organization A and B exchange IP packets with other organizations, and these figures show the order of the activity with other organizations. Each organization has different activity for other organizations, and the number of the organization is different.

Organization A has 94 different organizations and organization B has 72 different organizations. Information is characterized for the activity on the organization.

5. Discussion and Conclusion

There is the discussion of ambiguity about the relation between the domain name and the organization name. It is ambigious because the relation of them is only binded to DNS. We have to get the authorized data of the relation. The relation between port number and application has the same issue. For example, the WWW server proxy often accept the request on the port number 8080 which isn't well-known port.

The study is only focused on packet header information for capturing packets. In addition to packet header information, the application data in the packet can be used for the analysis of the activity. For example, every transactions in SMTP connection has the application data for the message sender and the message receiver, and that for WWW connection also has the data for the URL request and the reply with contents. More information will be provided by analyzing the application data.

If additional information from other sources is related to the target traffic at the measurement period, the knowledge from the traffic analysis is increased with this information. For example, if there is the multimedia experiment which consumes the high bandwidth of the network and the traffic measurement system knows this information, the volume of normal state is calculated by subtracting the total volume. The extraction of the related information for the organization such as the event data is needed to know the activity of the organization. The results is more precises than ones which are obtained only from information on the network.

This paper describes the framework of extraction of knowledge from the Internet traffic, and it is applied to analyze the inter-organization activity on the Internet.
The author thanks Dr. J-K. Hong of Tokyo Research Laboratory for supporting this study.

References

[1] K. Claffy, G. Miller, and K. Thompson, "The Nature of the Beast: Recent Traffic Measurements from an Internet Backbone", In Proc. of INET98, (Geneva, Switzerland), July 1998.

[2] V. Paxson and S. Floyd, "Wide-Area Traffic: The Failure of Poisson Modeling", In Proc. of ACM SIGCOMM '94, 1994.

[3] W. E. Leland, M. S. Taqqu, W. Willinger and D. V. Wilson, "On the Self-Similar Nature of Ethernet Traffic", *In Proc. of ACM SIGCOMM '93*, (Ithaca, NY), 1993, pp. 183–193.

[4] M. E. Crovella and A. Bestavros, "Self-Similarity in World Wide Web Traffic Evidence and Possible Causes", In Proc. of ACM SIGMETRICS '96, (Philadelphia, PA, USA), 1996, pp. 160–169.

[5] A. Feldmann, A. C. Gilbert and W. Willinger, "Data networks as cascades: Investigating the multifractal nature of Internet WAN traffic", *In Proc. of ACM SIGCOMM '98*, ACM SIGCOMM, (Vancouver, BC, Canada), Sep. 1998.

[6] T. Kushida, "The Traffic Measurement and The Empirical Studies for the Internet", *In Proc. of GLOBECOM '98* , IEEE COMSOC, (Sydney, NSW, Australia), Nov. 1998.

[7] G. B. Malan and F. Jahanian, "An Extensible Probe Architecture for Network Protocol Performance Measurement", *In Proc. of ACM SIGCOMM '98*, ACM SIGCOMM, (Vancouver, BC, Canada), Sep. 1998, pp. 215–227.

[8] Paxson, V., "Bro: A System for Detecting Network Intruders in Real-Time", In Proc. of the 7th USENIX Security Symposium, (San Antonio, TX, USA), January 1998.

[9] V. Paxson, J. Mahdavi, A. Adams and M. Mathis, "An Architecture for Large-Scale Internet Measurement", *IEEE Communications Managzine*, August 1998, pp. 48–54.

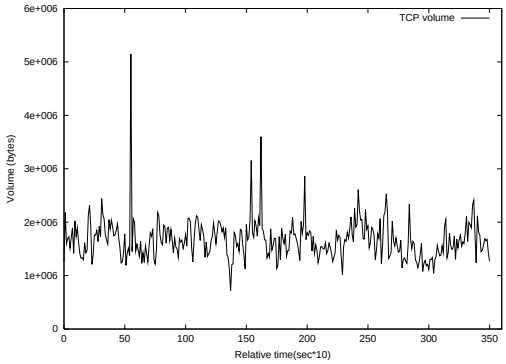

Figure 3. TCP volume

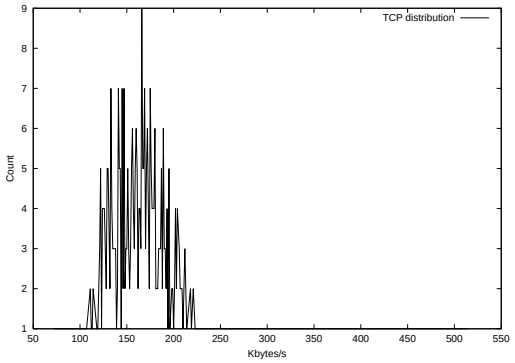

Figure 4. Distribution of TCP

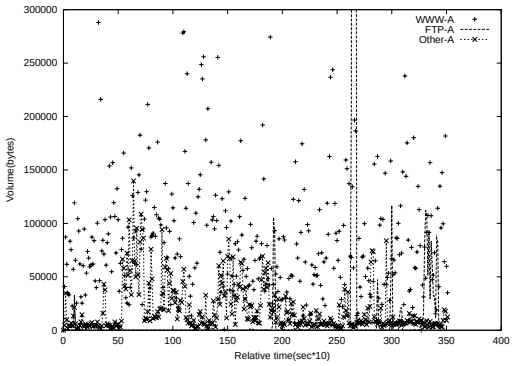

Figure 5. Activity for organization A

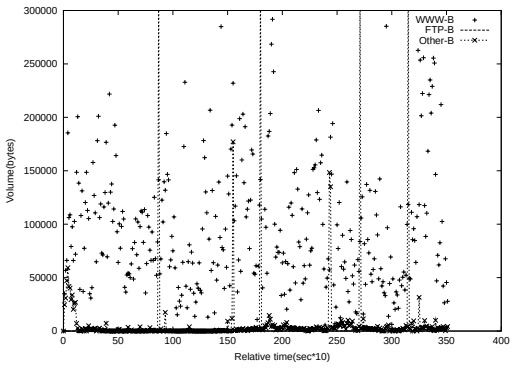

Figure 6. Activity for organization B

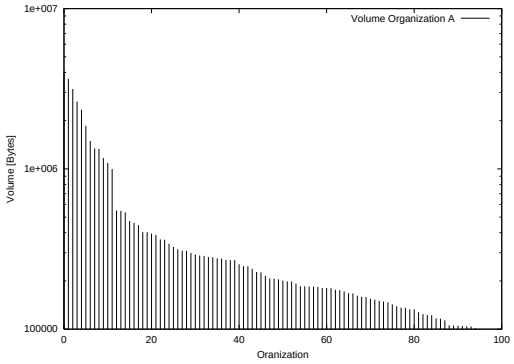

Figure 7. Interaction with organization A

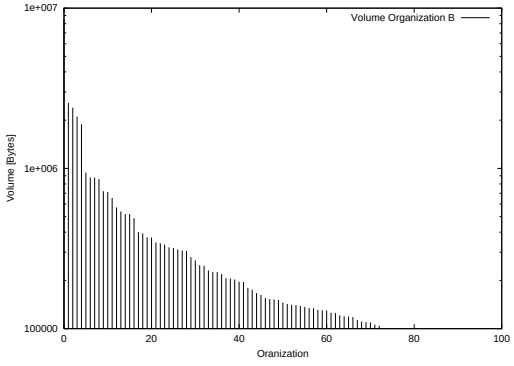

Figure 8. Interaction with organization B