Survey of the State of P2P File Sharing Applications

Keita Ooi†*, Satoshi Kamei, and Tatsuya Mori*

Abstract

Recent improvements in Internet access have been accompanied by a dramatic spread of peer-to-peer (P2P) applications. P2P file sharing is affecting telecommunications carriers, copyright holders, and other areas of society. Since P2P applications do not employ servers to transfer files, it is difficult to gather traffic information on a large scale. This article presents i) techniques for measuring P2P file-sharing application traffic by collecting packets using a modified application and ii) measurement results for a Gnutella network. It also examines the current scope of P2P file sharing and the types of files that are shared.

1. What is P2P file sharing?

The origins of P2P file sharing go back to the Napster service, which has become synonymous with P2P. A computer running a P2P file-sharing application participates in a P2P file-sharing network as a "peer" of other computers running the same application. A P2P file-sharing application has two key functions. The first searches for a certain file on the hard disk of a peer via the network, and the second transfers a file discovered by that search back to the requesting computer either directly or via another peer.

2. Purpose of survey

The increasing popularity of P2P file-sharing applications has led to explosive growth in Internet traffic. Some telecommunications carriers have already responded to this increase by stipulating in user agreements that users who generate excessive amounts of traffic may have their contracts cancelled. Thus, there is a need to evaluate the effects of such traffic on telecommunications carriers and to predict its future effects. To this end, it is essential to survey actual traffic generated by P2P file sharing and to understand the nature of file sharing itself.

3. Various P2P file-sharing applications

Napster required a central server for file searching, which means that the traffic generated by searching had a pattern similar to that of ordinary Web traffic. Most Napster-generated traffic, however, consisted of file transfers made directly between peer computers without passing through the central server. This type of traffic generates a pattern quite different from that of Web traffic. While Napster the company discontinued its P2P service, compatible servers and clients using the same protocol as the Napster application continue to be used elsewhere.

For example, the WinMX application, which features a compatible-client function, has found widespread use in Japan. Of all the P2P traffic that now exists on Japanese networks, WinMX is believed to generate the most. In addition, Gnutella, which came after Napster, employs an architecture that does not require a central server. It achieves autonomous-distributed communications among peer computers for searching as well as for file transfer. The absence of a central server makes it difficult for a third party to deny or control services.

The above P2P file-sharing applications have been followed by others, and at present, the one responsible for the largest P2P network in the world is KaZaA. These new applications, however, do not

† NTT Information Sharing Platform Laboratories
Musashino-shi, 180-8585 Japan
E-mail: ooi.keita@lab.ntt.co.jp

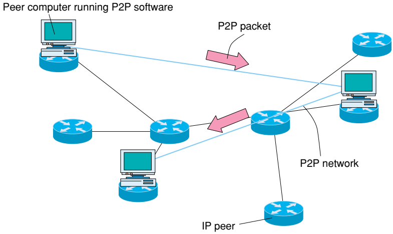

Fig. 1. P2P network.

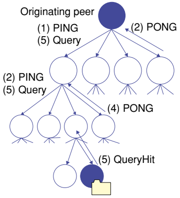

Fig. 2. Gnutella operation flow.

support multilingual file names, and in Japan, none of them has become as popular as WinMX.

In Japan, there is also an application called Winny that was created on the basis of Freenet, a P2P file-sharing application having a high degree of anonymity. Winny can generate sudden increases in traffic, and there are reports claming that it is generating more traffic than WinMX at some measurement locations within carriers' networks.

4. Measurement on the application layer

As shown in **Fig. 1**, peer computers running a P2P application form a logical P2P network that overlays the IP network on a layer independent of the physical network. It is therefore difficult to obtain information like the scale of users and the state of file sharing only from measurements made on the IP layer. In other words, measurements on the IP layer must be combined with those on the application layer to obtain an overall picture. Furthermore, when there are applications whose protocol specifications have not been released, it is even more difficult to determine usage patterns by measurements on the network layer. In addition, the type of information that can be collected differs from one application to another, which means that special measurement techniques must be developed for each application. In the following sections, we describe a traffic measurement technique for a logical network taking the Gnutella file sharing application as an example, and we present measurement results.

4.1 Measurement technique

To make measurements, we modified a servant to collect control packets as shown in **Fig. 2**. In Gnutella, a P2P network is constructed through the autonomous operation of peer computers (called "servants" in Gnutella). All control packets for checking presence, searching, and maintaining the network are exchanged on this network. There are five types of control packets as described below.

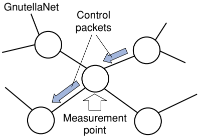

Fig. 3. Gnutella measurement concept.

- PING: A packet issued to discover connected servants.
- PONG: A packet issued by a servant in response to a PING packet. It includes an address and available capacity.
- Query: A packet issued to request a search. It includes a search character string. The number of times that this packet can be transferred from one servant to another is specified beforehand.
- QueryHit: A packet issued in response to a Query packet by a servant that possesses the file requested. It includes the file's size and URL.
- PUSH: A packet issued to request the sending of a file when the sending side is behind a firewall.

These packets are sent and received as shown in **Fig. 3**. Operations at each step of this process are described below.

1. The originating servant sends a PING to other servants.
2. Servants receiving a PING return a PONG and then send a PING containing the identifier of the originating servant to other servants.
3. Servants decrement a number-of-transfers variable every time a PING passes through.
4. A PONG arrives at the servant where the PING originated traveling in reverse order via the servants through which that PING passed.
5. In a search, the originating servant sends out a Query, and the servant that has the file in question returns a QueryHit to the originating servant.

4.2 Measurement conditions

File transfers that are performed directly between two servants are not targeted for measurement here. It must therefore be kept in mind that the results presented below are limited to information about files targeted for retrieval. In addition, by simultaneously measuring traffic on the network layer formed by operating servants using a traffic measurement technique for a logical network, we also collected information about the P2P logical network and measured the scale of that network. This technique can be applied to P2P applications in general.

Table 1. Measurement results for select applications.

Type	Unique IP addresses	Unique files	Total file capacity	Average file capacity
Gnutella	30,052	3,883,752	1.3 PB	330 MB
WinMX	350	165,933	25 TB	150 MB

P: 10^{15} T: 10^{12}

4.3 Measurement results

(1) Network scale

Table 1 shows the number of unique IP addresses, number of unique files, total file capacity, and average file capacity for Gnutella and WinMX as data reflecting network scale. The results shown for Gnutella were obtained over a 68-hour period on a weekend in the first half of 2003. Those for WinMX were obtained by sample measurements for comparison. About 30,000 IP addresses were collected for Gnutella making it possible to estimate the network size. Note that even if more IP addresses were to be collected by making measurements over a longer time period, the number of IP addresses does not simply equate to the number of users. In other words, the effective period of IP addresses must be taken into account.

(2) File size

We compared the sizes of files shared on Gnutella, WinMX, and the Web. The resulting file-size histograms are shown in **Fig. 4**, where both axes are log scales. Gnutella files were considerably larger overall than Web files. **Figure 5** shows the distribution of shared-file extensions searched for on the Gnutella network. The mp3 audio-file extension was most popular followed by avi and mpg moving-picture file extensions that generally correspond to large files.

(3) Distribution of number of references

Figure 6 shows the distribution of the number of references in Gnutella on a log-log scale. The x-axis represents the number of file references (n) and the y-axis represents the

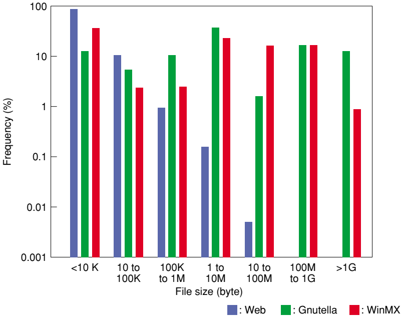

Fig. 4. File size distribution.

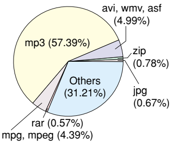

Fig. 5. Distribution of file extensions in Gnutella search words.

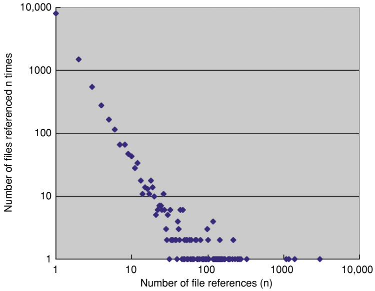

Fig. 6. Distribution of number of shared files in Gnutella.

number of files referenced n times. The linearity exhibited by the plots in this figure is called Lotoka's law (or Zipf's second law). In general, when recording what individual users select, the resulting shape (Lotoka-type distribution) conforms to this law. For a log-data Lotoka-type distribution, it is known, for example, that if the total number of files is small compared with the total number of references, then linearity will not be maintained in the area corresponding to a small number of file references (upper left area). That is to say, the plot will form a curve in the downward direction [1].

The log-data Lotoka-type distribution also indicates that cooperative filtering systems are effective [2]. And this suggests the possibility of applying P2P file-sharing log to fields such as market analysis and marketing.

5. Conclusion

At NTT Information Sharing Platform Laboratories, researchers are working on analysis systems that can utilize the historical properties in Lotoka distributions (as mentioned in this article) and ones that can analyze and use histories on an even larger scale. There is also a need for ongoing surveys to evaluate and predict the various effects of user scale, which tends to increase in a P2P file-sharing application. The main approach adopted by systems that aim to identify P2P traffic is to store the features of each current P2P application in a pattern file (as in Ellacoya and P-Cube). The drawback of these systems, however, is that new pattern files must be prepared whenever new or upgraded P2P applications appear, which increases the possibility of erroneous results. The need to monitor the contents (payload) of transmissions also makes it difficult to deal with large-scale systems.

At NTT Service Integration Laboratories, researchers are developing traffic separation systems for identifying P2P traffic using the logical-network traffic measurement technology introduced in this article and information on peer activity collected by that technology [3]. These systems will make it easier to observe the effects of P2P traffic on the network and to perform independent control on separate types of traffic.

References

[1] K. Muranaka, M. Matsuda, M. Aida, T. Motohashi, and M. Sato, "Analysis of Internet Access Patterns in Finite Address Space," IEICE Technical Report, IN2001-56, 2001 (in Japanese).

[2] T. Motohashi, M. Sato, and A. Kanai, "Advertising and Marketing Technology for Portal Services," NTT Technical Journal, Vol. 14, No. 1, pp. 85-87, 2002 (in Japanese).

[3] T. Mori, S. Kamei, and K. Ooi, "Measurement Analysis and Characteristics Evaluation of P2P Traffic," IEICE Society Conference, SB-3-1, Niigata, Japan, Sep. 2003 (in Japanese).

Keita Ooi

NTT Information Sharing Platform Laboratories.

He received the B.S. and M.S. degrees in mathematical science from Kyoto University, Kyoto in 1996 and 1999, respectively. In 1999, he joined NTT Information Sharing Platform Laboratories, Tokyo, Japan. His current interest is collaborative filtering.

Satoshi Kamei

NTT Service Integration Laboratories.

He received the B.S. and M.S. degrees in applied mathematics and physics from Kyoto University, Kyoto in 1997 and 1999, respectively. In 1999, he joined NTT Service Integration Laboratories. His current interests are measuring and analyzing overlay network traffic.

Tatsuya Mori

NTT Service Integration Laboratories.

He received the B.S. and M.S. degrees in applied physics from Waseda University, Tokyo in 1997 and 1999, respectively. Since joining NTT in 1999, he has been researching computer network measurement and analysis.