2007－EVA－21　(2)
2007／6／22

ピュアP2Pアプリケーショントラヒック特性の評価

大坐畠 智 [†]　　川島 幸之助 [†]

[†] 東京農工大学 大学院 共生科学技術研究院 先端情報科学部門

概要　P2Pオーバレイネットワーク上では主として音楽，動画ファイルが交換されている．P2Pアプリケーションは，これまでのクライアント／サーバ型のアプリケーションと比較して非常に大きなトラヒックを生成しており，ネットワークへの膨大なトラヒックの源となっている．しかし，匿名性の高い通信方式を用いているピュア型のP2Pアプリケーショントラヒックの実態は，あまり良く知られていない．そこで，これを明らかにするため，まず日本で最も人気のあるP2Pファイル共有アプリケーションであるWinnyに対するトラヒック特定方式を開発した．提案する特定方式はピア間のトランスポート層でのクライアント／サーバ関係に着目して特定を行うものである．つぎに，提案方式を用いて特定した対象とするトラヒックの特性を明らかにする．

A Pure P2P Application Traffic Identification Method and Evaluations of Traffic Characteristics

Satoshi Ohzahata[†] and Konosuke Kawashima[†]

[†] Institute of Symbiotic Science and Technology, Tokyo University of Agriculture and Technology

Abstract　In P2P networks, it is mainly music and video files that are transferred, and it is known that the traffic volume is much larger than that of classical Client/Server applications. However, the nature of current P2P application traffic is not well known because of the anonymous communication architectures used. To solve this problem, we have developed an identification method for pure P2P application traffic, especially for Winny, the most popular pure P2P file sharing application in Japan. Our proposed method relies only on Client/Server relationships among the peers, without recourse to application header information. In addition to describing the method, we also give an evaluation of the characteristics of the identified traffic collected in an ISP.

1　Introduction

Internet applications are changing and the traffic volume continues to increase. A large proportion of the traffic volume is occupied by P2P traffic because huge files are shared via the overlay network, and this has a large impact on the network. The effect of P2P traffic has been estimated in order to construct networks and manage them appropriately. However, the status is not still well known because the architecture is constructed in an anonymous way and no administrator exists for the network. When we undertake research on Internet traffic, we have to identify the types of traffic. To meet this need, many traffic identification methods have been considered.

Classical P2P applications have their default service port: Gnutella (6346, 6347), Kazaa (1214), BitTorrent (6881–6889). However, current versions of the above P2P applications do not always use their default service port.

Signature matching techniques are effective when the applications exchange specific characters as part of the payload of packets [1], [2]. However, in these signature matching methods, the application signatures need to be updated with changes in the application protocol and every packet needs to be analyzed.

In modern P2P applications, such as Winny [3], Share [4] and BitTorrent, the signature matching method cannot be applied easily. Since the communications are encrypted, the signatures cannot be extracted from them. An effective way of overcoming this problem is the use of identification methods in the transport layer [5], [6]. These methods enable P2P traffic to be identified using only header information at layers lower than the transport layer. However, with these methods the accuracy of identification for each application still has significant room

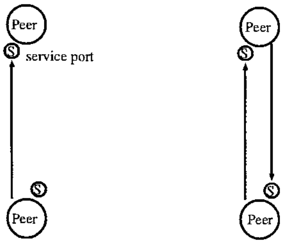

Figure 1: Communication relations in the Transport layer; (a) One-way Client/Server model (on left). (b) Two-way Client/Server model (on right).

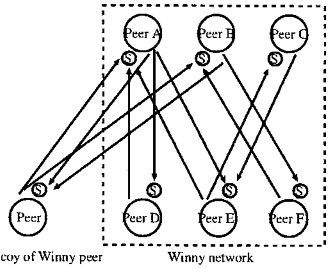

Figure 2: Collection of IP and service port number of Winny peers by the Decoy peer.

for improvement.

To address the above problems, we previously proposed an improved service port number identification method specifically designed for pure P2P application traffic [7]. In this paper, we propose a further improved service port identification method, which uses the model of the access events made among peers. Then, we applied the proposed method to a popular pure P2P application Winny. We also give characteristic evaluations for the identified traffic measured in ISP network.

2 Models of Access between Peers in Pure P2P Networks

In this section, we discuss the Client/Server model applying between a pair of peers in a P2P network. By expanding the discussion, we explain our proposed service port identification method in the next section.

Figure 1 shows examples of Client/Server models between peers. The service port is depicted by the circles beside each peer. Each arrow corresponds to a connection made by one peer to another, the access being from the ephemeral port of one peer to the service port of the other.

Figure 1 (a) is a One-way Client/Server model, in which only a one-way connection is established between the peers. In this model, one peer plays only the role of the server and the other peer is only a client for communication between the two in the Transport layer, even if both peers have their own service port. A pure P2P network can be composed of only a One-way Client/Server model because a TCP connection provides bi-directional communication. If one TCP connection is established between the two peers, then, they can communicate with each other. At the initiation of connection, a peer has to give attention to the Client/Server direction at the Transport layer level. However, the peers need not be concerned about the direction/relation at the Transport layer level after the connection has been established, and the role of server and client has been configured in the application layer level network.

Figure 1 (b) is a Two-way Client/Server model in which two Client/Server connections are established between the peers. In this model, both peers play the roles of both server and client in the connection between the two peers. In general, combinations of the two models are often found in a pure P2P network. In many pure P2P networks, the file search network uses the One-way Client/Server model and the Two-way Client/Server model is sometimes constructed when file transfer connection is established between the peers which have previously been connected directly through the file search network.

The two-way Client/Server model is introduced to maintain network stability by confirming the service ports to each other or directly communicating between the peers. In a Winny network, the two-way Client/Server model is adopted and if the service port is accessed, the peer will communicate back to confirm existence of the service port of other peer. This is because the connection controls are different for NATed peers, which do not establish a special service port for the other peers, and for the ordinary peers which do use a service port. Then, two connections are established between the two peers, in reverse directions.

3 Proposed Traffic Identification Method for Winny

The basic idea of our proposed identification method is to set up a decoy peer, which joins a P2P network and collects all pairs of IP addresses and service ports of the other peers (Figure 2). However, collecting these for all peers is difficult because of the restricted search capacity of the decoy peers (it finds only Peers A and B). Therefore, we need to find the missing peers (Peers C–F) by finding the

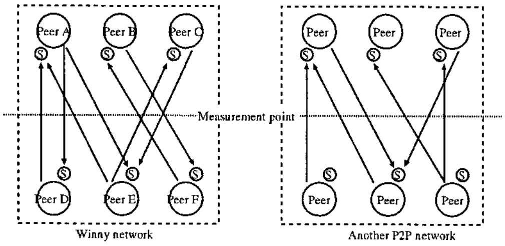

Figure 3: Simple P2P network model.

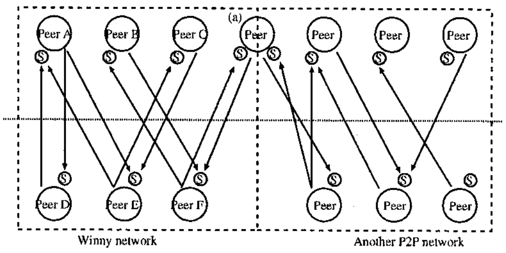

Figure 4: Complex network model 1.

Client/Server relationships between the peers as described in the previous section. We will use three network models in the following to explain our identification method.

3.1 Identification method 1

Figure 3 shows a simple P2P network model. There are two pure P2P application networks. In this model, each IP address (node) only acts as a peer for one application and so each node prepares only one service port for that application. Some P2P applications may use more than one service port. However, the fundamental discussion in the following description is not affected by this and it is easy to expand the discussion to cover this case. In Figure 3, each P2P network serves only one application, and is composed of a set of peers which use that application. Since one node runs one P2P application, the accesses among the peers are closed within a network for each application and so a Winny peer will never access a peer in another P2P network.

We assume that the service ports and IP addresses of Peers A and B in the P2P network have been identified by the decoy peer, which joins the network and collects the IP addresses and service ports of the other peers. These identified peers access other peers and it is possible to be sure that these peers are running the same P2P application. If each peer in this network is only using one service port, and the Two-way Client/Server access model has been established, these peers are using the same P2P application (Peer D, E and F). Thus we can learn the service ports of the application. By repeating this procedure, we can eventually find new peers (e.g. Peer C). For example, once we know one Winny service port, we can identify Winny peers one after another by using the Two-way Client/Server access model because the peers always access each other within the same network.

We also show a measurement point in Figure 3. Generally, a measurement point is installed between a backbone network and a stub network. At the measurement point, we can measure the access relations of peers communicating between these networks.

3.2 Identification method 2

Figure 4 shows Complex network model 1. In general, some P2P users run several P2P applications at one node and so the node has two or more service ports for different applications (as with Peer (a)). In the Complex network model 1, these peers are included in upper or lower side network. Each application peer only accesses the network it belongs to, but, if Peer (a) is not managed correctly, it may seem as if all these peers are part of a single P2P network connected via Peer (a) although the network is actually formed of two P2P networks. This means that Peer (a) may cause the possibility of a False Positive (FP). So we have to distinguish between the two P2P networks by means of the following procedures.

The simplest solution is to exclude Peer (a) from use in the identification method and then the network can be divided into two networks and we can apply Identification Method 1 to this situation. In using this method, the service ports of nodes which have a number of service ports are excluded. An identification method of these nodes is discussed in the next subsection.

In Identification method 2, we consider the access relations between Peer (a) and the other peers. The Winny service port of Peer (a) is only accessed by Winny peers and the Winny peer in Peer (a) also communicates back to Winny peers (Peer F). If we focus on "communicate and communicate back" relations, Identification method 1 can be applied because these access relations are also closed for each P2P network. In the case that Peer (a) accesses a peer which has only one service port, this procedure works well. Then, we can also apply Identification Method 1 to this situation by focusing on Two-way Client/Server access relations.

3.3 Identification method 3

Figure 5 shows Complex network model 2. This model is more realistic than Complex network model 1. There are a number of nodes that have several

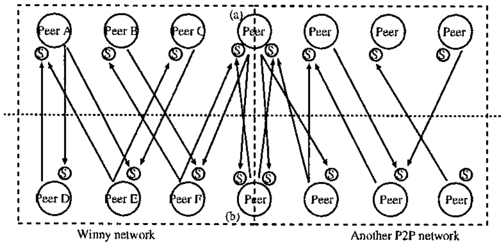

Figure 5: Complex network model 2.

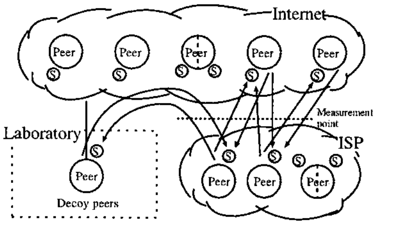

Figure 6: Measurement point in ISP.

peers, using different service ports at the same node. In the Complex network model 2, these peers are included in upper and lower side network.

In Complex network model 1, Two-way Client/Server access relations are closed within each P2P network. However, when peers which have more than two service ports, access each other, the "communicate and communicate back" relation is not closed in each P2P network, as a result of the relations between Peer (a) and Peer (b). In this case, we cannot decide which peer is Winny and which is not, and this leads to a high possibility of FP. We have to eliminate this situation in order to apply our method to it.

An effective solution is to modify the situation so that it reflects Figure 4. If either those upper or lower peers which have more than two service ports are not removed in the procedures, Two-way Client/Server access relations are closed within one P2P network as in Complex network model 1. This is because, in this situation, Peer (a) or Peer (b) only accesses other peers which have one service port. Consequently the Two-way Client/Server access relation is still effective even in complex network model 2.

3.4 Proposed identification procedures

In this subsection we summarize the discussion of Sections 3.1–3.3 and propose a traffic identification method for pure P2P applications.

Step 1: We prepare decoy peers for the P2P applications for which we want to identify the traffic (for example, Winny). The decoy peers join the P2P network and directly collect the IP address and service port number of the peers. Therefore, at least one service port for Winny in Figure 3 has to be identified by this step. It is difficult to identify the IP address and service port number of all peers. Therefore, we identify them by the accesses relations existing between the peers. We can start from those peers which are identified by the decoy peer, and then take the following steps.

Step 2: Then, we consider the situation of Identification Method 3 because we have to find nodes which have more than two service ports at the one node. To keep a low value of FP, the two-way Client/Server model (Figure 1 b) is required, in which a service port is identified by the attempts to access it. In Step 2, we identify the entire P2P network, which includes Winny and many other kinds of P2P network.

Step 3: With the results of Step 2, we establish a situation in which Identification method 2 can be used. Then, we remove one side (upper or lower) of the peers which have more than two service ports, and all access relations which include these peers are not used. These peers were identified in the Step 2. By this method, we can identify one set of peers (on one side), which have more than two service ports (see Figure 4). These procedures must then be repeated for the other side with the same way.

4 Application to an ISP Traffic

4.1 Basic flow statistics in the ISP

In this section we describe how we applied our proposed identification method for ISP traffic and evaluated the traffic characteristics. Figure 6 shows measurement setup. We installed two measurement points, one in our laboratory and the other in an ISP's network. In our laboratory, we set-up a PC which ran 20 Winny peers, and the decoy peers collected the service ports and IP addresses of Winny peers. In the ISP, the traffic of a stub network was measured for a research use, the link speed being 1 Gbps in the each direction. The ISP traffic log includes communications between the ISP and the Internet. This traffic log also included accesses between the decoy peer and Winny peers in the ISP network. Then, we were able to apply our method because the decoy peers accessed the service port of Winny in the ISP and the service ports identified are accessed by Winny peers outside of the ISP. We analyzed 5 hours of the ISP traffic log from 14:00–19:00 on June 8, 2004, and the log analysis period was also

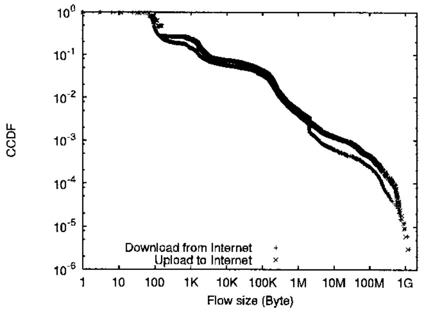

Figure 7: Flow size.

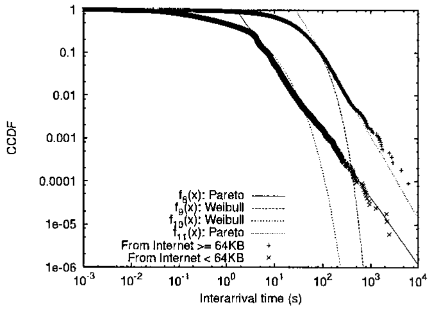

Figure 8: Flow interarrival time at a service port of Winny for file search flow and file sharing flow .

5 hours. The traffic log includes 1,807 IP addresses in the ISP network, and 128,129 IP addresses in the Internet. We identified 72 Winny peers in the ISP and these peers had communications with 48,049 of Winny peers outside the ISP. Thus, we can estimate that 72/1807 = 0.0398 of the ISP users was a Winny user.

4.2 Flow size

Figure 7 shows the complementary cumulative distribution function (CCDF) of the Winny flow size for the flow directions of downloading from and uploading to the Internet. It will be seen that on the curves in the graph, there are 3 discernible steps, at 100B, 1KB and 64KB. 75% of the flows are smaller than 150 B. Winny is a pure P2P application and each peer has to maintain contact with adjacent peers (maximum 600 peers) to maintain the network. These flows are used to confirm the existence of the adjacent peers and exchange "cluster words." The cluster words are set in each peer to form clusters of peers with the same interest, and are used when the peers establish the file search link. With regard to the second step, the flows with sizes ranging from 1 KB to 64 KB are mainly used for file searching because shared files are transferred by in units of 64 KB in the network. Then 95.4 % of flows smaller than 64 KB are not used for file transfer. The third step occurs around 64 KB and the 4.6 % of the flows greater than this are used for file transfer. The sum of the flows over 64 KB was 95,717,777,255 B and formed about 99 % of the Winny traffic. This shows that a small number of flows produce huge volume of traffic. In spite of the fact that it is mainly DVD/CD iso, avi and mpeg files that are shared in a Winny network, the distributions do not show a long tail. Since Winny employs multiple downloading from peers for a single file, and includes a download resume function, one file is downloaded by multiple flows and so the size of each flow does not become excessively large.

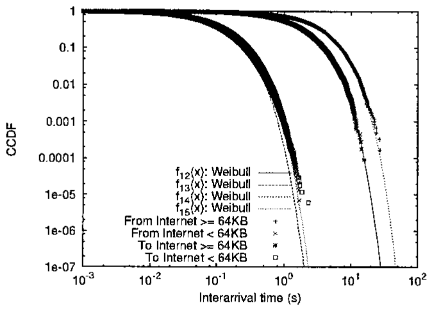

Figure 9: Flow interarrival time at the measurement point for file search flow and file sharing flow .

4.3 Characteristics for file search flow and file sharing flow

A Winny network is composed of an adjacent peer search network, a file search network and a data sharing network. Traffic for these networks will have different characteristics and we analyze them in this subsection. Since Winny uses the same service port for each of these networks, it is difficult to distinguish these flows. However, since Winny transfers a file in units of 64KB during file exchange, we can define flows smaller than 64 KB as file search flows (includes adjacent peer search flows and file search flows), and flows larger than 64 KB as file sharing. The basic statistics obtained are given in Table 1.

Figure 8 shows the flow interarrival time at a service port of Winny for a file search flow and file sharing flow. This graph shows that accesses to the Winny service port in the ISP are described. The average interarrival time for the file search flow is 3.16 seconds and for the file sharing flow is 42.6 seconds. To maintain the file search network, the file search flow is generated more often than the data flows. Both graph have long tail distributions. For file search flow, the connection is maintained by the

Table 1: Winny flow statistics for file search and data sharing, (14:00–19:00 on June 8 2004).

	Flow size < 64KB	Flow size >= 64KB
Number of flows (from Internet)	314,032	13,654
Number of flows (to Internet)	310,954	16,732
Total flow size in Bytes (from Internet)	272,285,653	34,541,836,171
Total flow size in Bytes (to Internet)	402,165,683	61,175,941,084
Av. flow size in Bytes (From Internet)	867.1 (4,730)	2,529,796 (32,031,000)
Av. flow size in Bytes (to Internet)	1,293 (5,792)	3,656,224 (36,918,534)
Av. flow duration in seconds (from Internet)	6.32 (76.3)	101.8 (464.9)
Av. flow duration in seconds (to Internet)	5.50 (73.4)	90.8 (399.4)
Av. flow interarrival time at a peer (s)	3.16 (18.1)	42.6 (137.4)
Av. flow interarrival time at measurement point from Internet (s)	0.108 (0.178)	1.66 (2.40)
Av. flow interarrival time at measurement point to Internet (s)	0.127 (0.203)	2.04 (2.91)
Av. flow rate in Bytes/second (from Internet)	248.9 (968.1)	18,124 (68,081)
Av. flow rate in Bytes/second (to Internet)	335.7 (958.7)	22,295 (78,886)

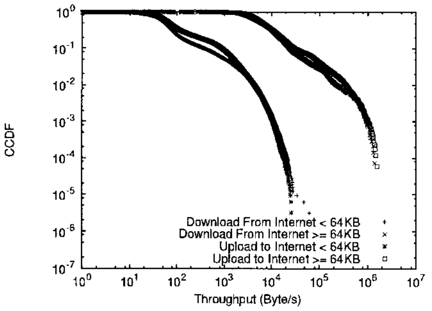

Figure 10: Flow rate for file search flow and file sharing flow.

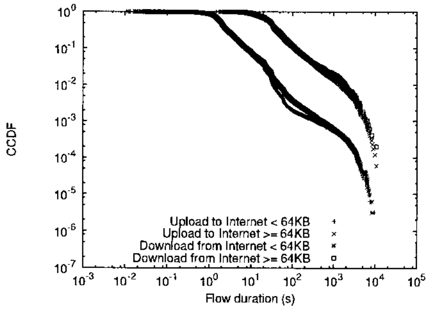

Figure 11: Flow duration for file search flow and file sharing flow.

keep alive mode of TCP and the clustering for the file search network. Since the number of connections for the file search network is restricted in Winny, the interarrival time becomes long. For the file sharing flow, the interarrival time depends on the number of attempts of file downloading. In the case that the file is not found in the network or all the files searched for are downloaded, the interarrival time becomes long.

Figure 9 shows the flow interarrival time at the measurement point for file search flows and file sharing flows. For flows smaller than 64 KB, there is almost no difference between the directions. However, for flows larger than 64 KB, flows from the Internet arrive much more frequently at the measurement point than outgoing flows to the Internet. Flows for maintaining the file search network between the ISP and the Internet are produced with almost the same interval, but the flows for file sharing to the Internet are more generated than flows for file sharing from the Internet. This is the main reason that the number of bytes transferred from the ISP is larger than that of bytes transferred to the ISP.

Flow rates for file search flows and file sharing flows are shown in Figure 10. These shapes are in the same size range and are almost the same for both directions. There is a step in each graph at a point less than 64 KB. The flow rate is very slow; 90 % of flows are slower than 100 B/s because these flows are maintained by the keep alive mode of TCP without sending file search requests. Then, a small number of flows are used actively for the file search and almost all flows are used to maintain the file search network in the Winny network. The file sharing flows are much faster than the file searching flows and the distribution is long tailed with the limit of the access link speed.

Flow duration for file search flow and file sharing flow are described in Figure 11. For file search flow, over 99 % of flow duration is shorter than 10 seconds, because that Winny peer disconnects the connection after confirming establishment of the Two-way client/server access in almost cases. However, the other flow is maintained for long period with the

Table 2: Peer level traffic statistics (14:00–19:00 on June 8 2004).

	Peers inside the ISP	Peers outside the ISP
Number of peers	72	48,049
Total downloaded/uploaded data in Bytes	96,392,228,591	96,392,228,591
Total downloaded data in Bytes	34,814,121,824	61,578,106,767
Total uploaded data in Bytes	61,578,106,767	34,814,121,824
Average transferred/received data in Bytes	1,338,780,952 (2,379,565,463)	2,006,123(28,725,585)
Average downloaded data in Bytes	483,529,469 (890,227,223)	1,281,568 (21,983,104)
Average uploaded data in Bytes	855,251,482 (1,894,652,200)	724,554 (18,013,326)

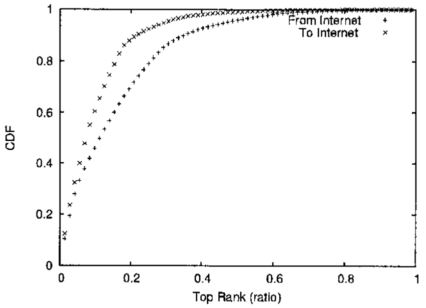

Figure 12: Traffic volume of Winny peer (Top-ranking IPs inside ISP).

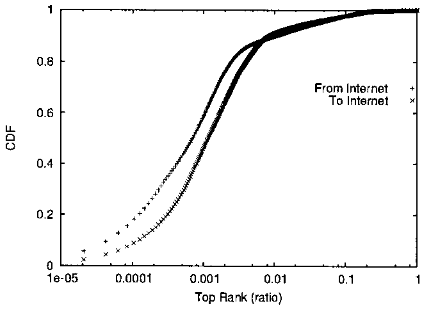

Figure 13: Traffic volume of Winny peer (Top-ranking IPs outside ISP).

keep alive mode because the peers have the same interest for file sharing. For file download flow, it depends on a file size of downloading and a large file take long period to download.

5 Peer Level Traffic Statistics

In this section we analyze the Winny traffic for each peer (IP address), and the Winny peer level traffic characteristics are given in the following.

Table 2 shows the statistics for traffic volume of Winny peers. 72 of Winny peers in the ISP downloaded 34.8 GB of data and uploaded 61.6 GB of data in 5 hours. Each Winny peer in the ISP downloaded 483.5 MB of data and uploaded 855.3 MB of data on average. The uploaded data is almost twice as large as the downloaded data. Each Winny peer in the ISP downloaded 724.6 KB of data from a Winny peer outside of ISP and uploaded 1.28 MB of data to a Winny peer outside of ISP on average. The details are discussed with reference to Figures 12–15.

Figure 12 and 13 show the CDF for the transferred and received data of Winny peers inside and outside the ISP, respectively. These CDFs were obtained by first ordering the IP in order of traffic volume. Figure 12 show that 40 % of Winny peers in the ISP download 90 % of the data volume, and 20 % of Winny peers in ISP upload 90 % of data volume. Some of the Winny users dominate the file uploading and downloading traffic, and the number of dominant file uploading users is smaller than that of dominant file downloading users. This means that the number of content uploaders was smaller than the number of content downloaders at this time.

Figure 13 show that in the case of Winny peers in the ISP downloading/uploading data from/to Winny peers outside of ISP, 1% of peers outside the ISP accounted for 90 % of the volume. A Winny peer accesses many Winny peers outside of ISP to maintain the file search network but file sharing are done with a very restricted number of peers and so the traffic volume is concentrated in this small number of peers. Since the Winny peers inside the ISP upload a larger volume of traffic than the download traffic, there are many more outside ISP peers receive traffic than those of transfer traffic.

Figure 14 and 15 show that the relationship between the download and upload traffic for Winny peers inside and outside the ISP. In Figure 14, peers which download a large volume of data also upload a large volume of data. In Winny, to avoid free loaders, an "encrypted file cache" mechanism is implemented and the downloaded files are uploaded without user's intervention. In addition, a file has to be shared via an intermediate peer in many cases, and

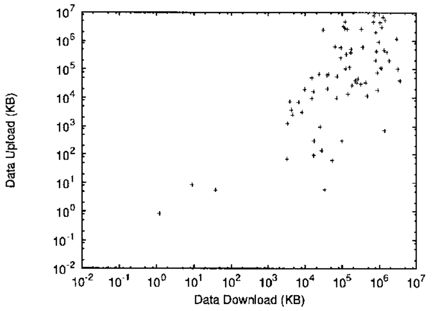

Figure 14: Relation between download and upload Winny traffic volume for IP inside ISP

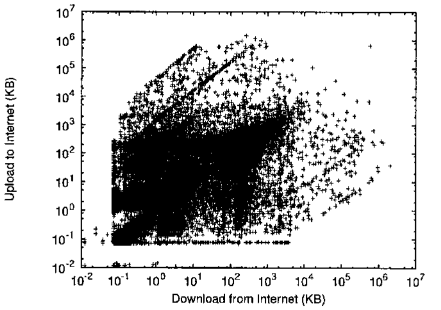

Figure 15: Relation between download and upload Winny traffic volume in IP for outside ISP.

the encrypted file is also cached in the intermediate peer. Then, when a file is shared, two caches are produced in the network. Furthermore, if the user sets some keywords, the peer always collects the files which match the keywords. These are reasons why Winny produces such a huge volume of traffic.

In Figure 15, there are a few linearly-increasing lines for large download/upload traffic. This means that a file is shared between peers which, in many cases, are not directly connected by the file search network. These peers are only used for file uploading or downloading without file searching, and then points of transferred data size and the corresponding acknowledgement data size are plotted in the graph. In a Winny network, there is no accurate incentive mechanism and file uploading peers do not always download files from the peer.

6 Conclusions

We have proposed a traffic identification method for the pure P2P application, Winny, and applied our method to ISP traffic and give an evaluation of the identified Winny traffic. The characteristics of the Winny flows is symmetric for the outgoing flows to the Internet and the incoming flows. A small number of peers is responsible for producing almost all the traffic volume.

We have only shown that our identification method is effective for Winny but the modified method will work well for the other P2P applications, even for P2P applications which will appear in the future. This is because our method depends on the basic relationships involved in client/server computing in Internet applications. In the case that the communication is encrypted, our proposed method will give some clues to help determine the traffic status. Our analysis only uses the traffic logs in the transport layer and network layer. With a combination of IP layer, transport layer and application layer traffic log which includes controls of Winny application, we will further understand the characteristics of the traffic.

Acknowledgment

This research is partly supported by Grants-in-Aid for Scientific Research (KAKENHI), No. 18500047.

References

[1] S. Sen O. Spatscheck and D. Wang, "Accurate, Scalable In-Network Identification of P2P Traffic Using Application Signatures," *Proc. ACM WWW'04,* 2004.

[2] T. Karagiannis, A. Broido, N. Brownlee, k. claffy and M. Faloutsos, "Is P2P dying or just hiding?," *Proc. IEEE Globecom 2004,* pp. 1532–1538, 2004.

[3] Isamu Kaneko, *"The Technology of Winny,"* ASCII, 2005 (in Japanese).

[4] Share, "http://en.wikipedia.org/wiki/Share_%28p2p%29."

[5] T. Karagiannis, A. Broido, M. Faloutsos and k. claffy, "Transport Layer Identification of P2P Traffic," *Proc. ACM IMC '04,* pp. 121–134, 2004.

[6] F. Constantinou and P. Mavrommatis, "Identifying Known and Unknown Peer-to-Peer Traffic," *Proc. 5th IEEE International Symposium on Network Computing and Applications,* pp. 93–102, 2006.

[7] S. Ohzahata, Y. Hagiwara, M. Terada and K. Kawashima, "A Traffic Identification Method and Evaluations for a Pure P2P Application," *Proc. PAM 2005,* pp. 55–68, 2005.