Detecting Peer-to-Peer Activity by Signaling Packet Counting[*]

Chen-Chi Wu[†], Kuan-Ta Chen[‡], Yu-Chun Chang[†], and Chin-Laung Lei[†]

[†]Department of Electrical Engineering, National Taiwan University
[‡]Institute of Information Science, Academia Sinica
{bipa,congo}@fractal.ee.ntu.edu.tw, ktchen@iis.sinica.edu.tw, lei@cc.ee.ntu.edu.tw

1. MOTIVATION

Peer-to-peer traffic now constitutes a substantial proportion of Internet traffic. However, managing such traffic is a major challenge for network administrators, because peer-to-peer applications tend to use *dynamic port numbers* and *proprietary protocols*. There has been a great deal of research in the area of peer-to-peer traffic detection, and a number of approaches have been proposed; among them, the application-layer approach and the transport-layer approach are considered the most promising. In the following, we briefly discuss the strengths and weaknesses of both approaches.

The application-layer approach [3] identifies a protocol-specific signature in the packet payload. This approach achieves high detection accuracy because false positives do not occur if the signature is sufficiently unique; however, it cannot identify applications with unknown signatures and cannot be used for encrypted traffic. As a result, a peer-to-peer application might easily avoid detection. Furthermore, examining user payloads incurs a high computation overhead and raises privacy concerns. On the other hand, the transport-layer approach [1, 2] is based on the observation that a host running peer-to-peer applications usually interacts with a large number of other hosts. This approach is more lightweight, as no payload examination is required, and it is more robust against countermeasures of applications. However, as most peer-to-peer applications share the same property, i.e., communication with many hosts, it is hard to recognize *a particular peer-to-peer application* based on the network-level information only.

In this poster, we propose an approach that combines the advantages of payload-based and transport-layer approaches, and avoids their disadvantages. Specifically, our approach can *recognize particular peer-to-peer applications running on the monitored host without checking packet payloads*. The key to our approach is recognizing *the signaling behavior* of a peer-to-peer application. Although a peer-to-peer application can easily change its port number, payload, and even message format, the signaling patterns between peers are more fundamental and unlikely to change. For example, a BitTorrent client needs to regularly exchange the file bitmap containing the block status of its files with neighboring peers. This signaling behavior is essential for the maintenance of the BitTorrent network; and changing it would lead to state inconsistency and software incompatibility problems. Therefore, it would be difficult for a peer-to-peer application to change its signaling behavior without affecting its normal operations.

Our methodology is unique in three respects:

- It does not require any application-layer information, and only traffic associated with the monitored host is required, i.e., *a global view of the network is not necessary.*
- It recognizes particular applications running on the monitored host, so it does not treat all peer-to-peer traffic in the same way.
- It recognizes peer-to-peer applications based on their *unique signaling behavior*; thus, it is unlikely that an application will evade recognition.

2. METHODOLOGY

The signaling behavior of each peer-to-peer application is regulated by its underlying peer-to-peer protocol; therefore, each application possesses a distinguishing characteristic. Based on this concept, we keep track of all the signaling packets sent from and received by a monitored host. By signaling packets, we mean the small packets exchanged between peers to maintain network connectivity and other peers' states, not packets used to transfer files. Because we cannot distinguish signaling packets from file or media transfer packets, we simply assume that packets smaller than a certain threshold, e.g., 100 bytes, are signaling packets.

Given a stream of signaling packets from a host, we characterize its signaling behavior on two levels: the host level and the message level.

- **Host level.** A host regularly exchanges information with other hosts that are known to it, and also with new contacts, i.e., previously unknown hosts. Based on the number of new or old hosts it communicates with, we characterize the signaling behavior at the host level through a number of features, such as the ratio of new hosts contacted within a certain period.

[*]This work was supported in part by Taiwan Information Security Center (TWISC), National Science Council of the Republic of China under the grants NSC 96-2219-E-001-001, NSC 96-2219-E-011-008, and NSC 96-2628-E-001-027-MY3.

SIGCOMM'08, August 17–22, 2008, Seattle, Washington, USA.

Table 1: Features used to characterize the signaling behavior of peer-to-peer applications

Host level
Ratio of new / old hosts (mean, sd[†])
Growth rate of new / old hosts (mean, sd)
Correlation coefficient between the number of new and old hosts
Message level
Ratio of new / old packets (mean, sd)
Growth rate of new / old packets (mean, sd)
Correlation coefficient between the number of new and old packets
Alternate rate of new and old packets (mean, sd)

[†] Standard deviation

- **Message level.** For a monitored host, we denote a signaling packet exchanged with a new host as a new packet; otherwise, it is regarded as an old packet. Based on the number of new or old signaling packets, we define a number of features to represent the signaling behavior at the message level, e.g., the ratio of new packets and the alternate rate of new and old packets.

In our approach, we monitor hosts for a certain period of time and then apply the derived features to recognize the application running on the hosts. For each monitored host, we count the number of hosts contacted and the number of packets sent and received every minute during the monitor time. A host is considered old if it has been observed sending/receiving packets within the previous 5 minutes; otherwise, it is considered a new host. Based on the statistics of the host and message levels, we derive features at the end of the monitor time, as shown in Table 1. Each feature comprises a pair of values computed from traffic transmitted in the incoming and outgoing directions. Furthermore, we compute the correlations between the features in both directions.

The simple scenario in Fig. 1 shows how we derive the features. For the incoming traffic, five hosts communicate with the monitored host in the 6th minute. Since hosts B and D appeared in the 4th and 5th minutes respectively, they are labeled old hosts. On the other hand, because hosts A, G, and H were not observed in the previous 5 minutes, they are considered new hosts. Therefore, the ratio of new hosts and old hosts in the 6th minute is 3/5 and 2/5 respectively. At the end of the monitor time, we derive the mean ratio of new and old hosts based on the values we computed for each minute.

Our scheme comprises two phases: a training phase and a recognition phase. In the training phase, we derive features from each training stream of signaling packets, and then apply a support vector machine (SVM) to train the classifier. In the recognition phase, given a stream of signaling packets, we extract their features and determine the type of peer-to-peer application with the trained classifier.

3. PRELIMINARY RESULTS

We demonstrate the recognition accuracy of our scheme through 10-fold cross validation, using traces captured on hosts running BitTorrent, eMule, or Skype. In Fig. 2(a), we plot the effect of the length of the monitor time on the recognition accuracy. Our scheme correctly recognized 98% of the traffic traces within 6 minutes and 99% within 15 minutes. Therefore, we achieve a highly satisfactory performance within a short monitor time, and we can increase the length of monitor time if a nearly perfect recognition accuracy is necessary. In Fig. 2(b), we plot the ROC curves for each application. The ROC curve of a certain application depicts its performance while treating that application as the positive class and all other applications as the negative class. A classifier with a high performance has a ROC curve that passes through the upper left corner of the plot, i.e.,the true positive rate is close to 1 with a small false positive rate. Therefore, these curves evidence that our proposed approach can recognize each peer-to-peer application with a high true positive rate and an extremely low false positive rate. Overall, the results demonstrate that our scheme can recognize peer-to-peer applications running on monitored hosts with a high degree of accuracy.

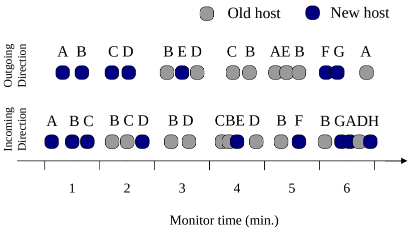
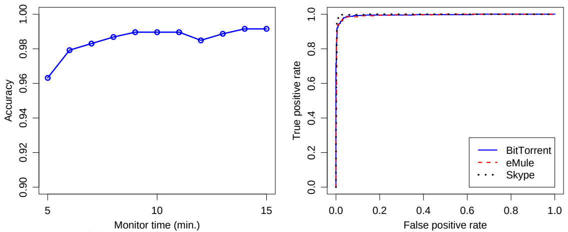


Figure 1: A simple scenario of signaling behavior in the host level



(a) Recognition accuracy

(b) ROC curves

Figure 2: (a) The influence of detection time on the accuracy, and (b) ROC curves for each applications

4. CONCLUSION AND FUTURE WORK

In this poster, we propose a scheme for recognizing peer-to-peer applications running on monitored hosts based on their signaling behavior. By analyzing such behavior at the host and message levels, we show that 99% of traffic traces can be correctly recognized within 15 minutes, which is a promising result. In our future work, we will consider more peer-to-peer applications, and improve our scheme's ability to correctly recognize a host running two or more applications simultaneously.

5. REFERENCES

[1] F. Constantinou and P. Mavrommatis. Identifying known and unknown peer-to-peer traffic. In *NCA '06: Proceedings of the Fifth IEEE International Symposium on Network Computing and Applications*, pages 93–102, Cambridge, MA, USA, 2006.

[2] T. Karagiannis, A. Broido, M. Faloutsos, and K. claffy. Transport layer identification of p2p traffic. In *IMC '04: Proceedings of the 4th ACM SIGCOMM conference on Internet measurement*, pages 121–134, Taormina, Sicily, Italy, 2004.

[3] S. Sen, O. Spatscheck, and D. Wang. Accurate, scalable in-network identification of p2p traffic using application signatures. In *WWW '04: Proceedings of the 13th international conference on World Wide Web*, pages 512–521, New York, NY, USA, 2004.