An Efficient Flow Analysis Scheme for Identifying Various Applications in IP-based Networks

Yong-Hyuk Moon†, Jae-Hoon Nah†

†Electronics and Telecommunications Research Institute, Daejeon 305-700, Korea

Abstract

The purpose of this paper proposes a network flow identification methodology considering network traffic dynamics caused by various types of recent network-based applications. In particular, P2P applications that generate encrypted packets are rarely visible to previous schemes, which have been proposed to handle packet filtering and control. Therefore, as a new strategy, we propose an efficient flow analysis scheme using a distributed monitoring agent for successful P2P and non-P2P traffic identification and management with statistical and application-level heuristic algorithm.

Key words:
Network application identification, flow detection, flow analysis, network security

1. Introduction

In recent years, Peer-to-Peer (P2P) [1] [2], an emerging technique to coordinate distributed resources with a decentralized manner in large scale networks, has obtained an immense popularity and has been rapidly evolving widely and deployed in various types of network-based applications, such as file sharing, voice of IP (VoIP), audio/video streaming, Internet games, and so forth. Despite its attractive characteristics such as efficient search mechanisms, scalable structure, and resource coordination, this emerging technology suffers from a degraded level of security and management service because it generally consists of a tremendous number of voluntary peers, tends to share security requirements with common distributed networks, and begins to use encryption. Accordingly, P2P inevitably involves the inherent weaknesses and complexities with respect to security and management.

In particular, unknown types of traffic, suspected to be generated by P2P applications, currently cause a lot of security vulnerabilities, such as hidden intrusion, malicious code propagation, and generation of huge traffic volume, approximately 70–80% of Internet traffic volume. This problem also induces the secondary attacks like denial of services (DoS). For example, a promising IPTV service required strict QoS or QoE has a large potential for revealing its security holes due to unexpected occurrences of incoming/outgoing traffics generated by unrecognized network applications; therefore, a home gateway or STB, which are necessary hardware platform for IPTV might experience a functional failure or might be maliciously used for leakage of personal information. Unfortunately, P2P technique makes this matter more serious because it has been born of nature which enables anonymous, dynamic, and open communications. Despite this situation, previous approaches for network measurement have not adequately dealt with P2P traffic well, since the recent P2P applications, such as Skype [3] [4] and KaZaA [5] tend to use proprietary application layer protocols and its closed specifications frequently are changed to the new version within a short development cycle [x]. As a result, such P2P traffic has not been easily visible to network security devices, such as firewalls or intrusion detection systems (IDSs); thus, sometimes P2P seems to be on decline. As a matter of fact, however, it is hidden.

Although a vast literature has been dedicated to the issue on P2P traffic identification and management, most methodologies have not satisfied critical requirements: low complexity, processing-overhead reduction, and real-time classification. Therefore, a close investigation on P2P traffic identification is significantly needed with deliberations on a peculiarity of P2P applications. For this we classify leading studies on methodologies for P2P detection into three classes as follows.

A. Signature-based Packet Inspection

Signature-based packet inspection [6] requires reverse protocol engineering as off-line work to extract a signature. Obviously it is too costly to make a signature for each P2P application, even though it guarantees a high detection rate for particular P2P traffic. In addition, looking into all payloads poses privacy and security concerns and induces the heavy processing overload as well because of packet capture and comparison in the central device.

B. Blind Techniques using Network Behaviors

Unlike the scheme above, the blind technique concentrates on how to identify P2P traffic using network behavior
without knowing any information extracted from the payload. Three different types of network behavior-based methods, such as host-level, flow-level, and signal-level have been discussed [7] [8]. The blind technique is expected to be useful for handling previously unknown or encrypted P2P packets because it obviates the need for packet inspection. However there is not much promise of high detection accuracy if the system has insufficient knowledge of the traffic.

C. Application-level Heuristic Schemes

In recent years, most application-level heuristics use an approach based on flow classification [9], machine learning [10], or support vector machine [11] with application specific features discovered by an experimental observation. Compared with packet inspection, some heuristics using statistic have shown great feasibility. It is convenient to apply them to large-scale networks and to handle unknown type of traffic. However most are still highly diverse and strongly dependent on the structure and condition of the networks and are sometimes too theoretically complicated.

As described above, we summarized this issue in Fig. 1 which shows the actual circumstances regarding to traffic identification against P2P dynamics. Key challenges still remain to be explored as shown in Fig. 1: allocation of random port number, port shifting (hopping), camouflage as http for NAT traverse, and encrypted packet. They aggravate the difficulties applying existing mechanisms. Thus, identification of new P2P activities always falls behind the innovation of P2P protocols. This fact inspires us to focus on designing the new identification scheme for P2P traffic.

Our aim of the paper is to identify the real-time P2P traffic, even encrypted, among various types of network-based applications with a high level of detection accuracy. In this paper, therefore, we propose a new flow analysis scheme for an efficient management of P2P dynamics by combining the flow-level packet analysis with the application-level flow classification in a hybrid manner.

2. Flow Statistic Extracting using DMA in IP-based Networks

To obtain data regarding network flows, a network device need to monitor and aggregate the principal information of real-time traffic traversing across the P2P networks, prior to analyzing and then classifying the P2P traffic. However, a conventional monitoring model depending on a central network device is limited to reflect various aspects of P2P applications working in many different service domains: file-sharing, VoIP, streaming, gaming, conferencing and so on. Hence, we located the distributed monitoring agents (DMA) built on a high-capable peer into each service domain for supporting the scalable network observation under dynamic circumstances. While individual DMA acted like a normal peer that joins in specific P2P networks, it tried to collect the network flow statistics. Subsequently, DMA extracted particular values as main parameters which will be used for distinguishing network flows. The aggregated data in one agent was periodically synchronized with others for creating global views on P2P traffic with respect to statistic characteristics. Multiple agent-based network monitoring models were well discussed in our previous work [15]. However, it is necessary to consider how many peers are chosen or how to evaluate who has enough qualifications working as DMA. In particular, the issue on deciding an optimal ratio between DMAs and normal peers [12] is out of scope in this paper; instead, we assume that every P2P community or group has only one DMA. This simple assumption guarantees a stable acquisition of status information in both of P2P and non-P2P networks. With this consideration, thus selecting one representative DMA located in a single P2P group can be achieved by estimating peers’ behavior and capacity; using the Eq. (1), the most capable peer was selected as DMA. Accordingly, as a preliminary step we propose the mutual recommendation-based distributed election algorithm for DMA composed of two factors such as transaction behavior and system capacity as followings:

\[
T_i^f = w_{\text{beh}} f(\text{Behavior_{success}}) \cdot w_{\text{cap}} f(\text{Capacity_{available}})
\]

\[
T_i^f = W_i \left( \frac{b_i - b_f^c}{b_f^c + b_f^c} \right) \cdot w_c \left( \frac{c_i}{c_{\text{total}}} \right)
\]

(1)

Figure 1. Late Progress in P2P Identification Methodologies
Where $T$ denotes the degree of peer $j$'s trust evaluated by peer $i$ and it consists of two ratio functions ($f_i$): first $f_i$ with behavior represents a success or failure ratio of transaction such as file request query, download/upload rate and correctness of result, second $f_i$ with capability means an allocation ratio of bandwidth, processing power, and memory to peer $j$ for communicating with peer $i$. Weighted factor $w_{ob}$ is defined to $\frac{(b_i' + b_j')}{n}$ where $n$ is the minimum number of transactions for the reliable trustworthiness estimation. And $w_{sc}$ could be determined by variation of resource provision level in terms of consistency.

In this proposed scheme, we assumed that the DMA-based P2P traffic monitoring model follows the group peer-based P2P network architecture which has been discussed in our previous study [13] and Beverly Yang's research paper [14].

3. Network Flow Analysis Scheme

In the discussion of identifying tricky P2P traffic, we proposed a flow statistic analysis-based scheme as shown in Fig. 2 with considerations of traffic dynamics originated from different types of P2P applications. Fig. 2 shows a conceptual structure of this scheme with respect to the entire procedure: consisting of three steps such as flow statistic sampling, flow analysis, and flow classification. Each step is carefully explained in the next sub-sections, respectively.

![Figure 2. Overall Procedure of 3-Step Flow Statistic Analysis](image)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Distributed Peer Election working on DMA</td>
<td>Aggregate P2P Flow Statistics by DMA</td>
<td>Direction-based FAST Flow Grouping (FFG)</td>
</tr>
<tr>
<td></td>
<td>Traffic Characteristic Monitoring (TCPM)</td>
<td>Application-Level Flow Statistic Reference (AFSR)</td>
<td>Flow Groups</td>
</tr>
<tr>
<td></td>
<td>Extract Flow Statistics from each Flow Group</td>
<td>Match a Traffic Group in a Specific Application</td>
<td>Application Groups</td>
</tr>
</tbody>
</table>

3.1 Network Statistic-based Flow Analysis

To perform the second (flow analysis) step through the network statistic-based flow analysis (NSFA) system producing two statistical references as shown in Fig. 2, we declared important flow statistic parameters as the indicator of P2P trafficked characteristic in table 1. A network traffic analysis system used the aggregated flow statistics gathered by the DMA model in order to periodically generate or update traffic characteristic positioning map (TCPM) and application-level flow statistic reference (AFSR); TCPM is the first statistical indicator to show characteristics of individual flow by using the average packet length and packet size variation and as the second reference to distinguish each flow, AFSR was created by using the packet size distribution and inter-arrival time as a main parameter. In our scheme, these two references were mainly used to group flows to one particular network-based application.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Packet Size</td>
<td>Distribution</td>
</tr>
<tr>
<td>Average length</td>
<td>Bytes</td>
</tr>
<tr>
<td>Median value</td>
<td>Bytes</td>
</tr>
<tr>
<td>Variance</td>
<td>Bytes</td>
</tr>
<tr>
<td>Inter-arrival time</td>
<td>msec</td>
</tr>
<tr>
<td>Average arrival time</td>
<td>msec</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>msec</td>
</tr>
<tr>
<td>IP address</td>
<td>The number of destination IP addresses count</td>
</tr>
</tbody>
</table>

The distribution includes lowest and highest values of the specific parameters. The main network parameters associated with flow statistic analysis have been addressed in [4] and [8].

3.2 Network Statistic-based Flow Classification

Most network flows are sent by a unidirectional packet first; however, they have always reverse flow, so that a bidirectional flow can be grouped to the same flow group. Accordingly, using the direction-based fast flow grouping (DFFG) method [5] [8] based on this property of network flow, we classified the real-time P2P network flows in the third (flow classification) step; DFFG compared a packet with the others using not entire payload information but only five columns such as source IP, destination IP, source port, destination port, and protocol type which were simply extracted from each packet. Namely, the flow analysis system has some flow groups. Since some flows belonging to a different flow group might be generated by the same application, even though different values of five columns between two flow groups are found. In order to remove this type of error, therefore, a mergence process of individual flow group (MPFG) should be executed continuously using the TCPM created in the second step. Through this process, then application
groups were obtained. In particular, therefore, MPFG can be considered as the essential process to lessen instability in P2P traffic identification caused by some particular features of P2P summarized as simultaneous use of TCP and UDP, a number of connections with different destinations, randomly selected or camouflage port number, and so on.

Fortunately, for example, if traffic control system has knowledge about a frequently used service port number by specific P2P application [5], flow groups in the third step can be quickly matched to an exact P2P application without MPFG process. However, it is rarely expected that there are representative port numbers in recent P2P applications.

3.3 Application-level Flow Identification

In order to guarantee highly accurate detection of traffic, now we need to reflect on the application-level P2P traffic identification in which each application group is correspondent to an exact network application. For this, the flow analysis system extracted special traits (e.g., packet size distribution and inter-arrival time distribution) from flow statistics of application groups and then compared them with the AFSR for matching an individual application group to a particular application.

At the end of this process, the detection accuracy can be arithmetically estimated by Eq. (2) [15]; in other words to calculate an accuracy of identifying a particular P2P application were performed at every fixed interval time.

This probabilistic measurement could provide a comprehensive glance of identifying P2P dynamics among other flows of network-based applications.

\[
P_x(t) = \frac{1}{N} \sum_{i=1}^{N} r(x|t-i\lambda)
\]

\[
r(x|t) = \begin{cases} 
1, & \text{AFSR of } x \text{ is equal to the flow's} \\
0, & \text{otherwise}
\end{cases}
\]

Where \( \lambda \) is meant to a measurement interval time, and \( N \) is defined to a number of measurement trials. \( r(x|t) \) is a function indicating whether AFSR instance \( x \) is equal to the real-time flows in application groups at time \( t \) or not.

So far, we explained how the proposed scheme for efficient network flow analysis can identify the dynamic P2P network flows. This algorithm described above can be summarized as follows.

**Algorithm:**

**STEP 1 Begin::**
1: While Do
2: If DMA does not exist
3: If \( T \geq a \)
4: Elect DMA peers by using Eq. (1) for a P2P community.
5: Else select another peer and repeat it.
6: Else
7: Extract statistics (refer to table 1) from aggregated P2P flows.
8: Share this information among DMAs.
9: While End

**STEP End::**

**STEP 2 Begin::**
10: Generate TCPM using average packet length and average packet size variation
11: Generate AFSR using inter-arrival time and packet size distribution

**STEP 2 End::**

**STEP 3 Begin::**
12: Joining in P2P community and exchanging traffic
13: Extracting five values (src ip, dest ip, src port, dest port, protocol type) from each packet received
14: Classifying each flow to a particular flow group using TCPM
15: Merging the same flow groups to one application group
16: Matching each application group to a specific application using AFSR

**STEP 3 End::**

The term \( a \) is a threshold value of trust for electing DMA. This algorithm describes whole steps in the linear order of sequence.

4. Analysis of Experimental Results

In this section, we offered experimental results with 4 aspects such as packet size distribution, traffic characteristic positioning, application-level reference, and detection accuracy. To consider various specialties and structures of network applications, 8 different types of widely-used or general-type programs were selected for identifying P2P flow dynamics; each type of application can be seen in Fig. 3, 4, and 5. In addition, for reducing the unexpected fluctuations and eliminating sub-network’s dependency, an experiment which gathered flows and then extracted statistical parameters by using the ethereal-like packet capturing program built in DMAs were designed and performed in separate locations such as ICU (Information and Communication University) campus and ETRI for 24 hours, respectively. Additionally, statistical flow analysis and estimating detection accuracy steps were executed by a simulation manner.

4.1 Flow Aggregation - Packet Size Distribution

Generally, data communications are established by exchanging control and data packets between source and destination node, thus examining packets’ attributes can be the first step for analyzing the network flows. Among the
statistic parameters extracted by the DMAs, the packet size distribution offered a good view to understand a possible and valid range of transferring packet’s size. Fig. 3 shows the overall application’s behavior in terms of packet size distribution.

In this result, AlFTP with passive mode and Kong (Internet Audio Streaming) had few number of less than 64 bit packets, since these applications are tended to exchange control packets for the initial connection establishment within short time. Moreover, Skype’s packets (P2P VoIP) and NateOn (Instance Messaging) have a small number of the largest packet compared with the other applications because this kind of network applications frequently exchanges small size of messages for text or voice communication in most cases. On the other hand, Joost (P2P Video Streaming) and AlFTP with active modes showed that approximately 70% had over than 1023 bits; because of its longer length of downloading data fragment than VoIP and instance messaging. In particular, it was found that BitTorrent (P2P File-Sharing) and GomTV (Internet Video Streaming) showed a similar pattern of the packet distribution during communication.

Obviously, this result can be understood as a pattern of particular packet’s occurrence or a ration of control to data packets. Hence, application types might be roughly determined by the frequency of occurrence of control and data since network-based applications usually are designed and followed to send or receive relatively long payloads, after finishing the session (initial communication step) establishment with quickly exchanging short type control messages in tens of milliseconds.

Although this result could provide rough clues which might include some errors or inaccuracy, the packet distribution can be changeable under particular network conditions or application-level protocol specifications. Therefore, more statistically meaningful reference is required to identify which the real-time flows probably is matched to the application group.

4.2 Flow Analysis – TCPM

As an index for deliberate classification of flow groups by specific traffic feature, TCPM generated from experiments with DMA and NSFA can be used to map each flow group to six recognizable application groups including file-sharing, VoIP, audio/video streaming, instant messaging, and ftp. First, from each flow group classified by DFFG, NSFA system investigates the average packet length and packet size variation. Next, with these statistic values, it tried to locate a flow group at around one point in Figure 3. Finally, individual flow group can be merged to one feature among six traffic classes.

Accordingly, each application of this result could be classified into four classes by a graphical position with respect to a variation and an average length of packet: 1) high variation - short length (i.e., P2P instance messaging), 2) low variation – short length (i.e., P2P VoIP), 3) high variation – medium/long length (i.e., P2P video streaming, P2P file-sharing, and non-P2P audio streaming), 4) low variation – long length (i.e., FTP and non-P2P video streaming). Since this showed more specific statistical aspects of each application compared with the result of packet size distribution discussed above, MPFG process could well recognize the flow features of network-based applications.

4.3 Flow Classification - ASFR

As a good illustration of ASFR(s) representing eight network applications was shown in Fig. 4. Among the sample applications below, we found that there were considerably distinguishable results, which can be thought of as a fingerprint of network application. This analysis
result presented how NSFA system can guarantee application-level network flow identification; namely, individual flow group involved in a specific application group will be matched to a corresponding application through comparing pattern and range of flow group’s statistics with well-made ASFR as a reference. We also expected that this approach is possibly extensible to the general P2P and non-P2P IP-based network applications because the proposed NSFA process showed that it guaranteed to extract the unique traffic features for a various type of network applications.

Figure 5. ASFR Example for Eight Network Applications

Statistical characteristics of P2P and non-P2P flows were significantly distinguishable as shown in Fig. 6; this result described more particular and refined feature of each application with respect to the packet size distribution and inter-arrival time.

4.4 Detection Accuracy Rate

Using three findings as can be seen in Fig. 3, 4, and 5, we tried to identify P2P and non-P2P network applications for demonstrating the performance efficiency of the flow statistic based analysis scheme in an aspect of detection rate. Concretely, analysis steps in this simulation were performed by the sequence of order by the packet size distribution, MPFG using TCPM, and the application-level flow identification using ASFR, after executing DFFG step. Fig. 6 compared the detection accuracy of executing each flow identification process in the proposed scheme discussed in the section 3; comparing real-time flows with reference such as packet size distribution, TCPM, and ASFR were achieved by simple value range and geometrical boundary matching as an off-line process.

As the first step, DFFG classified flows among real-time traffic into hundreds or thousands of flow groups for instant time by investigating who were talking to whom; these data regarding the classified flow groups were accumulated a single record data which might be central. Secondly, we examined packet size distribution of each flow groups with simple concept that a particular network application’s packets tends to have somewhat fixed percentage of a particular packet size; however for reducing matching errors of this rough detection process, we did not strictly apply the rule of packet size distribution to each flow. Thus, MPEG process was performed using TCPM for unrecognized and classified flow groups in order to generate more refined group, namely application groups; for almost application types as shown in table 1, higher and more reliable detection rate were found than the first process (approximately over than 80%). Finally, we compared application groups with ASFR in a similar manner; the best identification accuracy was particularly achievable in this process. Furthermore, a successful detection rate was evaluated by Eq. (2).

<table>
<thead>
<tr>
<th>Application Types</th>
<th>Packet Size Dist.</th>
<th>MPEG</th>
<th>ASFR</th>
</tr>
</thead>
<tbody>
<tr>
<td>P2P VoIP</td>
<td>62%</td>
<td>88%</td>
<td>97%</td>
</tr>
<tr>
<td>Non-P2P VoIP</td>
<td>22%</td>
<td>86%</td>
<td>94%</td>
</tr>
<tr>
<td>P2P Video Streaming</td>
<td>23%</td>
<td>72%</td>
<td>95%</td>
</tr>
<tr>
<td>Non-P2P Video Streaming</td>
<td>33%</td>
<td>85%</td>
<td>92%</td>
</tr>
<tr>
<td>P2P File-Sharing</td>
<td>30%</td>
<td>83%</td>
<td>97%</td>
</tr>
<tr>
<td>P2P Instance Messaging</td>
<td>46%</td>
<td>90%</td>
<td>97%</td>
</tr>
<tr>
<td>FTP (Active or Passive)</td>
<td>29%</td>
<td>82%</td>
<td>94%</td>
</tr>
<tr>
<td>Non-P2P Audio Streaming</td>
<td>43%</td>
<td>73%</td>
<td>90%</td>
</tr>
</tbody>
</table>

Flow data gathered from Skype, Window Live Messenger, BitTorrent, Joost, RTSP-based Internet audio streaming service, Yahoo Messenger, and FTP were representatively used for each application type.

Consequently, the cooperative processes such as DFFG, packet size distribution, MPFG, and NSFA offered us great possibilities to recognize which flow will belong to a corresponding application, even though flow (or traffic) is hardly visible to a network security device due to the packet encryption, camouflage, closed protocols, or rapid version changes.
5. Discussion and Further Study

In this paper, we have proposed the DMA-based adaptive traffic monitoring algorithm against the fluctuations in networks, in order to collect globally meaningful statistic parameters in large scale networks. Despite P2P dynamics such as unexpected traffic behaviors, proprietary application protocols, rapid development cycle, and so on, meaningful statistics have been extracted from each peer using the DMA concept, and the proposed 3-phase flow analysis scheme, which consists of TCPM, ASFR, and NSFA has successfully identified some types of network flows from sample applications, for example P2P/Internet VoIP, Instance messaging, P2P file-sharing, and P2P/Internet streaming, with respect to the accuracy of detection ratio (approximately more than 95% success).

Although this flow analysis scheme shows a good detection ratio for some network-based applications, feasibility and practicality has not been clearly proved yet for a real deployment into network systems such as a router, switch, or gateway in our experiments. However, it is certain that our heuristics can show highly stable results and cannot be strongly dependent on a particular type of networking structures or conditions. For designing and implementing a promising methodology to identify various network traffic, challengeable issues such as an overhead of real-time processing, restriction of extracting statistic parameters, and self-verification operation should be more carefully examined in further study.

Acknowledgement

This work was supported by the IT R&D program of MKE/IITA [2008-S-001-01, Development of Open-IPTV (IPTV2.0) Technologies for Wired and Wireless Networks].

References


Yong-Hyuk Moon received BS degree in Computer Engineering from Dankook University, Seoul, Korea in 2003. And He received MS degree in Information and Communications University (ICU), Daejeon, Korea in 2006. Currently he is with Division of Information Security in Electronics and Telecommunications Research Institute (ETRI), Daejeon, Korea. His research topics are related to IPTV Security, Grid Computing, P2P networking architecture, distributed computing security, and various networked computing issues.

Jae-Hoon Nah received MS degree in Computer Engineering from Chung-Ang University in 1987. He received the Ph.D. degree in Electronic and Information Engineering from Hankuk University of Foreign Studies in 2005. He is a principal research engineer and a team leader in Division of Information Security in Electronics and Telecommunications Research Institute (ETRI), Daejeon, Korea. His research interest includes IPTV security, distributed network security, peer-to-peer network, and overlay multicasting.