A Measurement-based Study on the Correlations of Inter-domain Internet Application Flows[☆]

Xiaofei Wu[a,*], Xin Wang[b], Ke Yu[a], Frank Y. Li[c]

[a]*School of Information and Telecommunication Engineering, Beijing University of Posts and Telecommunications, Beijing, China, 100876*

[b]*Department of Electrical and Computer Engineering, Stony Brook University, Stony Brook, New York, USA*

[c]*Department of Information and Communication Technology, University of Agder (UiA), Grimstad, Norway*

Abstract

Internet traffic characterization has a profound impact on network engineering and traffic identification. Existing studies are often carried out on a per-flow basis, focusing on the properties of individual flows. In this paper, we study the interaction of Internet traffic flows and network features from a complex network perspective, focusing on six types of applications: P2P file sharing, P2P stream, HTTP, instant messaging, online games and abnormal traffic. With large-volume traffic flow records collected through proprietary line-speed hardware-based monitors, we construct flow graphs of these different application types. Based on the flow graphs, we calculate the correlation coefficients on various properties for individual or multiple applications. Our studies on associativity among degree and strength of individual hosts and connected nodes reveal distinct correlative behavior of different types of applications. Especially, the correlations of P2P applications are observed to be much stronger than those of the other applications. We also investigate the correlations between different types of applications, and observe that HTTP has remarkably different correlations from those of the two P2P applications due to the fact that multiple application types rely on HTTP. Finally, we study the dynamics of correlations for a period of 24 hours and reveal a few interesting trends. We believe that our work which focuses on the assorta-

[*]Corresponding author

Email addresses: wuxf@bupt.edu.cn (Xiaofei Wu), xwang@ece.sunysb.edu (Xin Wang), yuke@bupt.edu.cn (Ke Yu), frank.li@uia.no (Frank Y. Li)

tivities of Internet applications provides insightful understanding on Internet traffic classification of up-to-date applications and will be helpful for Internet traffic classification and engineering.

Keywords: Internet traffic monitoring, Application flow, Assortativity, Complex Network

1. Introduction

Study on the characteristics of Internet traffic is important for understanding the activities and behavior of the Internet, and it is also essential for Internet service providers (ISPs) to better manage the operation of the Internet [1]. Moreover, understanding the characteristics of applications and traffic flows is helpful for many network operations, including planning and provisioning of the network infrastructure, traffic engineering and performance optimization, guaranteeing of service quality and protection of network from fraud. Indeed, the characteristics of Internet traffic have been studied extensively, and properties such as heavy-tail distribution, self-similarity and fractal behavior have long been understood [2].

Many recent studies on Internet traffic are carried out at the flow level. An Internet flow is formed with a series of packets exchanged between two hosts, identified by the well-known five tuples: source IP address, destination IP address, source port, destination port, and protocol type. Usually, a flow tracks the information exchanged for a complete Internet interaction, and the study of flow characteristics reveals how the Internet is accessed. Therefore, investigating Internet flows, both at the aggregate level and at the individual level, can provide insight on the features of Internet traffic. Flow records are widely used to examine the characteristics of Internet traffic and applications in the literatures, e.g., the studies for inter-domain traffic [3], peer-to-peer (P2P) applications [4], security reasons [5] and entertainment [6].

So far, most of the existing studies on flows were based on the observations of single flows, where millions of flow records were examined one by one and the properties of flows were retrieved. The properties studied include the number of packets and bytes in a flow, flow duration, distributions of packet size and interval, etc. Based on the information retrieved from flows, such as destination port and protocol type, flows can be classified into different applications and the statistics of applications are obtained from the observation of a large number of flows. While this method is able to reveal the characteristics of an application flow, it does not fully utilize the flow records for more thorough understanding of the features of Internet traffic. In fact, flows do not exist independently but correlate with each other via network elements, e.g., hosts and connections. Therefore, traffic is also observed to present complex network characteristics in Internet [7]. For instance, in P2P file sharing, flows between two hosts do not only carry file contents of the two hosts, but also include information of other participants. Therefore, flows sharing the same content are related and properties such as flow duration are affected by the composition of concurrent flows. The Internet flows form dynamic overlays on top of the physical and logical network structures, and these dynamic overlays manifest the vivid prospect of Internet interactions, therefore are very important for us to better comprehend the nature of Internet applications. Instead of constraining our studies to individual flow traffic, in this paper, we investigate the important properties of the overlaying flow infrastructure from a complex network point of view based on aggregated and correlated flows.

The main objective of this study is to investigate the degree and strength correlations within and across the Internet applications for inter-domain traffic. The concept of complex networks and graph [8] is adopted in this study. We construct flow graphs from the Internet traffic flow records we collected from operational networks and investigate their properties. In our previous paper [9], we have investigated the important characteristics of the flow graphs such as the distributions of node degree and the strength of different applications. In this paper, we further examine the flow graphs and analyze the correlations between different properties of network elements, both within one type of application and across different types of applications. More specifically, we study the mixing pattern and correlations, which are the important properties of complex networks [10]. To the best of our knowledge, this is the first effort to study the mixing pattern of different Internet applications based on a large number of inter-domain flow records and new traffic types such as online game and abnormal traffic. Besides the pattern mixing of individual application, we also study the assortativities between different applications, for example, between HTTP and P2P streams. The flow records were collected by a proprietary line-speed hardware-based monitor with a capturer and a classifier to track the traffic of a 10 Gbps trunk link between an access network and the backbone. The records consist of detailed traffic information with application classifications. Taking the advantage of this application classification ability, we are able to analyze the properties of multiple types of applications. In addition to the widely studied application types, i.e., HTTP and P2P file sharing (P2PF), we also analyze four other types of applications: instant messaging (IM), online games (OG), P2P streams (P2PS), and abnormal traffic (AT). The AT traffic includes virus and attacks detected. The mixing patterns and assortativities of specific Internet applications revealed by our work would help better predict traffic trends, which will in turn serve as a guideline for better Internet service provisioning and traffic engineering.

The main contributions of this study are as follows:

1. We examine flow records for six different types of traffic, namely P2PF, HTTP, IM, OG, P2PS and AT. We exhibit the profiles of nodes, connections and traffic volume of these applications, as well as their variations of traffic parameters within a 24-hour period. The analysis is performed based on a huge volume of data captured in a two-day period by a proprietary hardware monitor located between an access network and the backbone network. From the traffic analysis, we find that many hosts run more than one application simultaneously, and on an average each host runs 1.43 application.
2. We construct graphs from the flow records and examine the assortativities among node degree and strength of a specific application. More specifically, we calculate four types of correlation coefficients for a node and six types of coefficients for a connected node pair, and identify different assortative behavior for different type of applications.
3. We investigate the correlations across different types of applications by examining correlations between the properties of nodes participating in multiple applications. Based on the calculation of four types of correlation coefficients related to degree and strength, we conclude that the correlations between different pair of applications are very different.
4. We present the correlation coefficients in a 24-hour period to explore the time variation of application relationship, from which we observe different trends of different types of applications.

The rest of the paper is organized as follows. After a brief review of the related work in Section 2, our data sources and the flow graph construction procedures are introduced in Section 3. In Section 4, The methodology of correlation analysis is expatiated. Then our results are presented and analyzed in Section 5. Finally, the paper is concluded in Section 6.

2. RELATED WORK

Internet traffic has been studied extensively to reveal its characteristics. In general, data traffic is well known to be self-similar and fractal [2][11]. The traffic of a residential access network was observed [12], the volumes and the ratios of different applications such as web browsing, P2P file sharing, streaming, messaging and online games were presented. As P2P applications are taking a significant share of the current Internet traffic, they have drawn a lot of attention. In [4], two P2P file sharing applications were studied and the results show that the inter-arrival time and the connection duration vary significantly, and thereafter a novel measurement method named CTI was introduced. The traffic of online game was investigated both through experiments and simulations in [13], where the inter-departure and inter-arrival times of the packets were reported. Besides these normal applications over the Internet, virus and attacks also constitute an important part of the Internet traffic, and are a major concern for network security. Based on packets captured by monitors deployed in the network, the scanning pattern of distributed denial of service (DDOS) as well as the packet numbers and ports were presented in [14]. A more comprehensive study has been made in [5], which provides profiles of abnormal traffic corresponding to IP addresses and ports using data mining and significant cluster extraction. In [15], the authors conducted insight investigations on residential Internet traffic and many properties of digital subscriber line (DSL)-level sessions and TCP connections were studied.

Literature research has observed the existence of complex network in many Internet-related infrastructure and applications. The topology of physical connection of routers [16], routing connection of autonomous system (AS) [17] and the World Wide Web as a network of web pages and hyperlinks [18] all show power law degree distribution, which is a fundamental property of complex networks. Traffic Dispersion Graphs (TDGs) constructed from traffic flows are applied to model social behaviors, degrees and other properties of TDG as analyzed in [7]. TDGs formed by P2P traffic are studied thoroughly to develop a P2P traffic classification framework [19]. Another kind of traffic graph, i.e. Traffic Activity Graphs (TAGs) are revealed to possess interesting and meaningful block structures [20]. Graph features of various application traffic are shown to be helpful for Internet traffic classification [21]. Meanwhile, the Internet traffic studies continuously attract research attentions. The flow records of aggregated traffic and HTTP traffic captured in Internet2 were analyzed in [22]. The strength of edges was shown to have power law behavior over several orders. Based on the power law property observed, traffic can be reconstructed from partially measured network flows to facilitate research [23]. The correlation and pattern mixing of complex networks were studied in depth in [10], where the authors presented parameters, models and simulations of several types of networks, including social, technological and biological ones. The correlations of the properties of several network models were presented in [24]. Network traffic studies rely heavily on the accuracy of flow capturing and identifications. [25] shows that the accuracy of current port-based and DPI methods varies greatly with respect to different applications. The correlation and pattern mixing of complex networks were studied in depth in [10], where the authors presented parameters, models and simulations of several types of actual networks, including social, technological and biological ones. The correlations of the properties of several network models were presented in [24].

Existing results on the correlation of the networks are all based on static scenarios with the studies focusing on the topology of the Internet and Web. In contrast, the application flows which we study in this paper reflect the dynamic *user behavior* of the Internet. Compared with existing results, our work enriches the field of inter-domain traffic characterizations in two aspects. First, we analyze flow records of six major types of Internet applications; second, we investigate detailed mixing patterns and assortativities within one type and between different types of applications, and our studies reveal the existence of correlations between different properties.

In this paper, we focus on the interaction properties of diverse types of application flows and analyze the results from the perspective of complex networks. In our previous paper [9], we presented some general results of complex network analysis such as the incoming and outgoing degree and strength of nodes, as well as the process of application network growth. The focus of this paper is on the study of new properties, i.e., the mixing patterns and assortativities of applications. These new properties are investigated thoroughly through many types of correlation coefficients, both within one type of application and between different types of applications. In addition to HTTP, IM and P2PF studied in [9], three new types applications are investigated in this paper, namely P2PS, OG and AT.

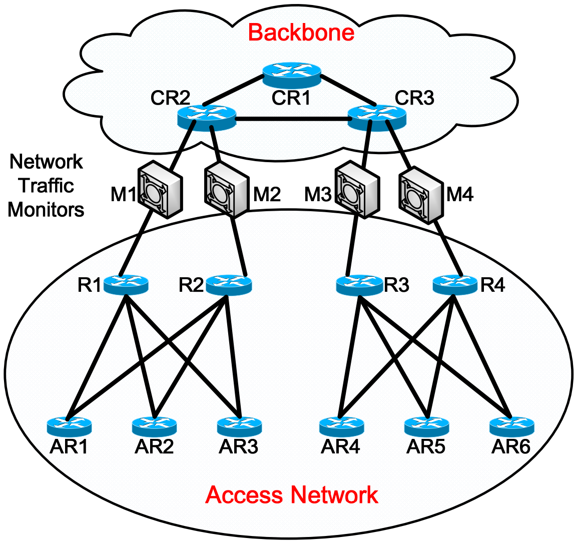


Figure 1: Network Traffic Monitor and Network Topology.

3. DATA COLLECTION AND FLOW DATABASE CONSTRUCTION

In this section, we provide information on data flow collection and database construction before performing our data analyses in the next section.

As mentioned earlier, the flow data were collected by placing high-performance network traffic monitors on the trunks between an access network (AN) and the backbone network. However, the AN itself is large in scale. It covers a part of a province in the Mid-West of China, which consists of millions of end users. The conceptual diagram of the network architecture is illustrated in Fig. 1.

Each trunk monitored has the capacity of 10 Gbps. The monitor captures every packet passing through the trunk in both directions, and reports the results to a server periodically. Each flow record contains information about a single network flow, which is defined as one or more packets sent from a source host and port, to a destination host and port, using a particular protocol. The traffic are continuously monitored during a two-day observation period, and the flow data are fully captured by the monitors[1]. The records are then further classified into applications such as Web, FTP, Email, VoIP, Video

[1] In our study, hosts include both clients and servers.

Stream, P2P, etc.

The flow database used in this paper was constructed based on data traffic collection for a period of two days, starting from 8 pm, 12 October 2010 to 8 pm, 14 October 2010. The traffic was captured in an unsampled manner, i.e., on a per-packet basis where every packet is tracked and classified into a flow and every flow was recorded. In total, more than 3.5 billion flow records were collected with detailed entries including time stamp, source and destination IP addresses and ports, total number of packets and bytes in the flow, and application type of each flow record. In the cases of studying the dynamics of flow and traffic characteristics within a 24-hour period, for more stable results, the traffic properties from the same period of different days are averaged.

The monitors which provide the traffic identification results are commercial products which are widely deployed in the carriers' networks. To classify flows into different types of applications, the proprietary monitors apply various parameters and methods, including application signature, special packet sequence and some deep packet inspection (DPI) methods together with port-based method. For example, BitTorrent, eMule, Poco, DirectConnect, Gnutella, KaZaa, Thunder, Kugoo, Maze, Winny, Share, PerfectDark, Vagaa are classified as P2P download. Over 200 specific applications are identified and classified into 15 types of applications. The results are confirmed by the carrier and used in daily network management. We take the results as the ground truth in our study. Of the 15 types of applications, this paper focuses on P2PF, HTTP, IM, OG, P2PS and AT, which are the most typical traffic types and consist of more than 70% traffic volume of the total traffic load[2] Although sophisticated methods have been used to identify the traffic, there is still about 25% of flows that can not be recognized. These flows are difficult to be classified by investigating individual flows, but may be easier to identify when considering correlations among flows. We hope our results in this paper can provide some guidance to the development of more efficient methods to identify the remaining traffic in the future.

Furthermore, the flow records are applied to construct graphs, called *flow graphs*, in the following way. Each IP address in the records represents a node; each flow record between two hosts forms an edge; multiple records

[2] The others are FTP, Email, eBusiness, VoIP, RouteNMS, UserDefine, DirectvideoStream, GenericTCP, GenericUDP.

between the same pair of nodes are considered to belonging to the same edge. The bytes over an edge are summed up and the total rate is a weight property of the edge, which is called *strength* of the edge. The records of different applications are used to construct different flow graphs. Since the flows are directional, the flow graphs generated are directed as well. We also aggregate all the flow records to construct an overall graph as a reference.

It is worth mentioning that the flow records we used are far from exhibiting all the traffic in the network. Instead, only inter-domain traffic flows, i.e., the flows between AN and the backbone network, were captured, while the intra-domain flows within AN or backbone network are not visible to the monitors. Therefore, the traffic details within the two domains are not investigated in this study. This can pose bias on our results, since these records may be in favor of global traffic like HTTP, but less favorable for local applications, such as P2P and IM. However, by focusing on the flows crossing the border, we can understand more precisely the interactions between different network segments. In real-life networks, there are often interactions between autonomous systems, and the interfaces under monitoring are often the places where billing and traffic engineering are applied. They are also the traffic bottlenecks in many cases. Therefore the analysis results in this paper will be valuable for making network management and traffic engineering decisions.

4. Description of Correlation Analyzing Methods

Assortative mixing, or assortativity, is an important feature of complex networks besides the well-known power law distribution and the small world effect [10]. Assortative mixing represents the correlations between the properties of network elements, i.e., nodes and edges, and may reveal profound structural and dynamic properties of the network. Assortative mixing of the flow graph can help answer questions such as: "Does a node which has a larger number of incoming connections tends also to have a larger number of outgoing connections?", or "Do the two connected nodes have similar amount of traffic?". Therefore, it can provide better understanding of interactions among applications. Assortative mixing is measured differently for different types of networks and properties. For a network with nodes of several types, the fraction of edges that connect a node of type p to a node of type q, noted as m_{pg}, shows extent of mixing between the two types. The assortative mixing can be measured as an *assortative coefficient* r_d as [10]:

$$r_d = \frac{\sum_p m_{pp} - \sum_p a_p b_p}{1 - \sum_p a_p b_p}, \tag{1}$$

where

$$a_p = \sum_q m_{pq}, b_q = \sum_p m_{pq}. \tag{2}$$

For the network element with a scalar property, such as the incoming degree of a node, the assortative coefficient r_s is calculated as follows. Let m_{xy} be the fraction of connections between a network element with the property value x and the element with the property value y, and

$$a_x = \sum_y m_{xy}, b_y = \sum_x m_{xy}. \tag{3}$$

Then r_s is calculated as the standard Pearson correlation coefficient:

$$r_s = \frac{E(xy) - E(x)E(y)}{\sigma_x \sigma_y} = \frac{\sum_{xy} xy(m_{xy} - a_x b_y)}{\sigma_a \sigma_b} \tag{4}$$

where $E()$ and σ are the expectation and the standard deviation of the corresponding variable respectively.

The calculated assortative coefficient is +1 in the case of a perfect positive (increasing) linear correlation between the values of the connected network elements, -1 in the case of a perfect negative (decreasing) linear relationship (anti-correlation), and certain value between -1 and +1 in all other cases. When r approaches zero, it indicates that there is a looser relationship (close to uncorrelated) between the corresponding properties. The closer the coefficient is to either -1 or +1, the stronger the correlation between the studied properties is.

To estimate the statistical error of value r, the Jackknife method [10] can be used to evaluate the expected standard deviation of r as:

$$\sigma_r^2 = \sum_{i=1}^{M} (r_i - r)^2 \tag{5}$$

where M is the number of elements under inspection, r_i is the value of r with the ith element removed from the calculation.

In our flow graphs, the properties of nodes include: d_{in}, the incoming degree of a node, d_{out}, the outgoing degree, s_{in}, the incoming strength and s_{out}, the outgoing strength of the same node respectively. The properties of edges include: s_e, the strength of an edge. By definition, s_e of an edge (i, j) between node i and j, is obtained by:

$$s_e(i,j) = (bytes\ from\ node\ i\ to\ node\ j) \quad i,j \in [0, N-1] \tag{6}$$

and

$$s_{in}(i) = \sum_{j=0}^{N-1} s_e(j,i) \tag{7}$$

$$s_{out}(i) = \sum_{j=0}^{N-1} s_e(i,j), \tag{8}$$

where N is the number of nodes in the graph.

Among these properties, we study two types of assortative mixing. The first type, referred to as *individual node assortativity* (INA), describes the property correlation of the same node, e.g., the assortativity between the incoming degree and outgoing degree of the same node. The second one, referred to as *connected node assortativity* (CNA), describes the related properties between connected nodes, e.g. the incoming degrees of two connected nodes.

The INA coefficient between incoming degree and outgoing degree, denoted as r_{dido}, are evaluated with a *sample calculation* as:

$$r_{dido} = \frac{N \sum_n d_i(n) d_o(n) - \sum_n d_i(n) \sum_n d_i(n)}{\sqrt{N \sum d_i^2(n) - (\sum d_i(n))^2} \sqrt{N \sum d_o^2(n) - (\sum d_o(n))^2}} \tag{9}$$

where $d_i(n)$ and $d_o(n)$ are the incoming and outgoing degrees of node n respectively, and N is the number of nodes in the graph. The other INA coefficients are calculated in the similar way. The CNA coefficients are calculated with properties of two connected nodes. For instance, re_{didi} is the CNA coefficient between the incoming degrees of connected nodes.

5. TRAFFIC ANALYSES

In this section, we present the results derived from the flow database and analyze the records. We first present the basic traffic characteristics

Table 1: Statistics of Applications.

	P2PF	P2PS	HTTP	IM	OG	AT
maximun nodes	1,647,539	2,358,726	372,356	53,729	32,058	103,279
average nodes	990,248	1,205,864	207,672	33,950	20,899	66,538
minimun nodes	320,294	268,983	48,425	6,398	7,202	20,809
maximun edges	2,199,836	3,531,262	832,553	61,766	27,928	77,725
average edges	1,389,872	1,807,927	554,083	39,644	17,357	56,023
minimum edges	397,824	369,763	102,211	6,258	5,725	17,487
maximun traffic	1,147,749	959,782	365,998	5,772	12,135	8,810
average traffic	705,376	457,311	229,803	3,012	6,104	3,267
minimum traffic	302,790	100,025	57,407	344	2,792	2,192

of different types of applications, and then study the mixing patterns and assortativities of traffic inside the same type of application as well as across different types of applications respectively. Lastly we show the correlation dynamics within 24 hours in different scenarios.

5.1. Traffic Characteristics of Diverse Applications

The duration of a flow varies depending on application type and user demand. While browsing a simple Web page takes only a few seconds, a P2P download may last for several hours. As a compromise, the flow records are inspected on an hourly basis in our study. As traffic varies during a day, we examine also the traffic dynamics on hourly basis. The basic parameters of the six types of applications are summarized in Tab. 1 and the variations are presented in Fig. 2 and Fig. 3.

In Tab. 1 and Figs. 2 and 3, n and e represent the number of nodes and connections respectively, t is the traffic volume in the unit of megabyte . Tab. 1 shows the maximum, average and minimum volume of node, edge and traffic. As we can see, P2PS, P2PF and HTTP are the three largest application volumes, accounting for more than 90% of the traffic classified. In Fig. 2, the label on the left side of Y axis shows the volume of nodes and connections, while the right label represents the traffic volumes. The volume of nodes is the summation of nodes involved in the flow records in each hour, so are the volumes of the connections and traffic. Parameters of P2PF traffic (Fig. 2a vary with the time of the day, with the lowest value appearing around 3∼5 am in the morning when most people are sleeping. The other normal traffic, such as P2PS (Fig. 2b), HTTP (Fig.2c), IM (Fig. 3a) and OG (Fig. 3b) have similar trends. However, although the numbers of hosts and connections for AT in Fig. 3c are similar to the other applications, the behavior and the traffic intensity of AT have different trends. The AT generates a large volume of traffic during the mid-night but traffic load remains low in the day time except for the peak in the noon. This is because that AT traffic is mostly likely generated by computer programs automatically in a scheduled manner.

The nodes and traffic volumes from different applications in percentage are shown as histogram in Fig. 4. The percentages are calculated based on the total traffic monitored, including the unknown traffic. We can observe that P2PS and P2PF account for about 60% of the total traffic, but over 25% of the traffic cannot be identified. The percentage of nodes corresponding to unidentified applications is even higher. The possible reason may be that shorter connections which contain a smaller amount of data are more difficult to be identified. The results in Fig. 4 are derived from the traffic records in the first hour of our collected data. We have also made an investigation on the data of the rest of the hours, and the percentage relationship of different application types is similar to the first hour, although the total number of nodes and amount of traffic vary over time. While the percentage of different types of traffic in Fig. 4 sums up to 100%, the summation of the number of nodes is over 100%. This is because many hosts are involved in more than one type of applications at the same time and are counted multiple times. *We discover from the records that the average number of applications per host varies between 1.4 and 1.5 within the day, with an average ratio at about 1.43 (Fig. 5a).* We further count the number of hosts involved in different number of applications simultaneously, and present our results in Tab. 2, where n_a is the number of hosts that have $\#_a$ number of applications. It is shown that over 25% of hosts participate in more than one application at the same time. In order to find out whether a host with a higher number of connections participates in more applications than a host with a fewer number of connections does, we depict the relationship between the average number of applications and the number of node connections in Fig. 5b. The average number of applications involved by the nodes with lower than 20 connections do increase slowly as the connection number becomes larger. However, for the hosts with a large number of connections (e.g., over 100), many of them are involved in three applications simultaneously, and there is a big variation in the number of applications involved by hosts with a large number of connections. This result indicates that the majority of users involve in a limited number of applications simultaneously, and some users may establish a few connections for the same type of application.

Table 2: Number of Nodes vs. the Number of Node Applications.

$\#_a$	1	2	3	4	5	6	7	8	9	10
n_a	54,500,625	12,887,185	2,802,451	1,389,541	158,377	371,912	438,233	316,697	170,577	80,360
%	74.5	17.6	3.8	1.9	0.2	0.5	0.6	0.4	0.2	0.1

Table 3: Individual Node Assortativity of Applications.

		d_id_o	d_is_i	d_os_o	s_is_o
P2PF	r	0.819	0.277	0.463	0.333
	σ	0.055	0.028	0.029	0.026
P2PS	r	0.900	0.694	0.503	0.523
	σ	0.007	0.022	0.041	0.090
HTTP	r	0.431	0.174	0.397	0.033
	σ	0.038	0.011	0.028	0.004
IM	r	0.293	0.126	0.212	0.019
	σ	0.132	0.023	0.036	0.012
OG	r	0.040	0.125	0.292	0.023
	σ	0.202	0.023	0.060	0.010
AT	r	0.260	0.001	0.006	0.000
	σ	0.038	0.001	0.001	0.000
Overall	r	0.840	0.620	0.524	0.398
	σ	0.019	0.070	0.040	0.055
Unknown	r	0.891	0.392	0.368	0.192
	σ	0.028	0.075	0.055	0.047

5.2. Mixing Pattern and Assortativities within One Type of Application

We define and calculate four INA coefficients for each application using 9, namely the INA coefficient of the incoming and outgoing degree (r_{dido}), the incoming degree and incoming strength (r_{disi}), the outgoing degree and outgoing strength (r_{doso}) and the incoming and outgoing strength (r_{siso}). The results are presented in Tab. 3 together with the estimated σ using (5).

In addition to the coefficients of the six specific types of applications, we also include those of the aggregated traffic (overall) and unidentified traffic (unknown) for reference. For r_{dido}, the large values of the P2P applications clearly show strong correlations between incoming and outgoing connections. This is in accordance with the P2P feature, i.e., a node with a larger number of outgoing connections is easier to get more incoming connections. However, the coefficients concerning the strength of P2PF are smaller than those of P2PS. The reason might be that for P2P file sharing, especially for files that are not very popular, seed nodes which do not receive data contribute to a larger share of the traffic volume. In the P2P stream case, however, users share data when viewing the same media at the same time, thus the coefficients are relatively higher. The HTTP and IM applications have middle ranges of correlations, with lower values between the incoming and outgoing

Table 4: Connected Node Assortativity.

Application	re_{didi}	re_{dodo}	re_{sisi}	re_{soso}	re_{dodi}	re_{sosi}
P2PF	-0.037	-0.030	-0.044	-0.041	-0.037	-0.040
P2PS	-0.063	-0.073	-0.014	-0.025	-0.069	-0.018
HTTP	-0.112	-0.193	0.004	-0.104	-0.213	-0.028
IM	-0.061	-0.108	-0.041	-0.088	-0.175	0.005
OG	-0.051	-0.102	0.001	-0.041	-0.094	-0.002
AT	-0.076	-0.057	0.014	0.000	-0.096	0.499
Overall	-0.047	-0.047	-0.034	-0.025	-0.050	-0.029
Unknown	-0.031	-0.029	0.032	-0.019	-0.031	-0.022

strength, indicating that the hosts of HTTP and IM do not usually transmit or receive data at the same time. The small value of r_{dido} and r_{siso} of OG implies that there is very low correlation between incoming and outgoing connections or traffic volumes. The incoming traffic for online games may involve a large amount of data from other players while the outgoing traffic may consist mainly of small-size game control commands.. However, its values are larger for r_{disi} and r_{doso}, meaning that nodes with more connections tend to generate higher traffic. For AT, there is no correlation concerning the traffic volume, but the value of r_{dido} does show some correlation. As we know, most virus infection and attacks are completed by malware exchanging messages with remote attackers, so there are certain degree of correlation between the incoming and outgoing connections. The σ values are generally one order smaller than the corresponding r values when the r values are not too small, indicating that our statistics have reasonable precision.

The CNA coefficients are presented in Tab. 4. Similar to the INA, we use the subscripts to indicate the properties investigated, and use re for CNA instead of r for INA. Six CNA coefficients are studied, namely re_{didi} for incoming degrees, re_{dodo} for outgoing degrees, re_{sisi} for incoming strength, re_{soso} for outgoing strength, re_{dodi} for outgoing degree of source node and incoming degree of destination, and re_{sosi} for outgoing strength of source and incoming strength of destination. Most values in Tab. 4 are smaller than zero, which indicates a negative assortativity, and means that nodes having larger degree or strength tend to connect to nodes with smaller corresponding values. However, as most values are close to 0, it indicates the correlations are pretty weak. The assortative correlations of HTTP and IM, on the other hand, are noticeable.

To further investigate the relationship, besides providing the assortativity on the same properties, we calculate two new coefficients for connected nodes: one is between the outgoing degree of the source node and the incoming degree of the destination node (re_{dodi}), and the other is between their corresponding strength (re_{sosi}). Since an edge creates one incoming degree and one outgoing degree on the two end nodes of the edge, we expect some positive correlations between them. However, it turns out not to be the case. We observe stronger negative assortativities for HTTP and IM, and the two new coefficients are even larger than the other coefficients of the two applications. This implies that connections are more likely to go from high output nodes to low input nodes or vice versa. In the mean time, other applications show almost no such correlations. This may be because that HTTP applications are generally carried out between the servers which have larger number of connections and clients which have smaller number of connections. For IM, besides servers for registration etc., there may exist "hub" persons who maintain a lot of contacts with others at the same time, which is a phenomena often observed in social networks. Most re_{sosi} values are small except for AT, showing a strong positive assortativity. We inspect the flow records, and find that heavy traffic generated by a few connections takes a significant share of the total traffic load among AT nodes, hence strong correlation exists. However, the σ of re_{sosi} is 0.312, which is also quite large, indicating that the distribution is highly skewed. This implies that the traffic of AT is exchanged mostly among some particular hosts, while the others do not transmit or receive much traffic. All other σ values of the coefficients are small, around or less than 0.001, so they are not listed in Tab. 4.

5.3. Assortativities across different types of applications

As observed from the analysis in Section 5.1, about 25% of the hosts run multiple applications. This is because that one may browse the web while downloading files in the background, or exchange messages with friends while being attacked unperceptively. It is therefore interesting to find out which kinds of applications are usually used together by end users and which are not. In this subsection, we investigate the correlations between applications running together by the same host. The same method mentioned in the previous subsection is adopted, but the correlation coefficients of the properties across different types of applications are calculated instead. From the flow records, we identify nodes associated with multiple applications, e.g., IP addresses with both HTTP and P2PS records. For these common nodes among applications, the correlation coefficients of degree and strength are calculated, and the results are illustrated in Tables 6 and 7.

Table 5: Number of Common Nodes between Applications.

	P2PF	P2PS	HTTP	IM	OG	AT
P2PF	1,318,392	268,335	147,745	24,683	13,539	44,479
P2PS		2,209,628	143,085	29,434	12,190	50,162
HTTP			342,108	32,364	16,933	25,582
IM				51,885	4,052	6,885
OG					30,262	4,013
AT						93,151

Table 6: Correlation Coefficients of Degrees across Applications.

		P2PS	HTTP	IM	OG	AT
P2PF	r_{di}	0.717	0.186	0.228	0.238	0.486
	σ_{di}	0.027	0.023	0.099	0.044	0.049
	r_{do}	0.619	0.172	0.096	0.022	0.140
	σ_{do}	0.029	0.020	0.039	0.203	0.013
P2PS	r_{di}		0.708	0.625	0.577	0.663
	σ_{di}		0.040	0.035	0.021	0.041
	r_{do}		0.570	0.099	0.545	0.226
	σ_{do}		0.058	0.020	0.018	0.014
HTTP	r_{di}			0.105	0.678	0.391
	σ_{di}			0.017	0.049	0.076
	r_{do}			0.068	0.078	0.128
	σ_{do}			0.008	0.499	0.023
IM	r_{di}				0.076	0.178
	σ_{di}				0.036	0.095
	r_{do}				0.002	0.128
	σ_{do}				0.007	0.012
OG	r_{di}					0.326
	σ_{di}					0.061
	r_{do}					0.145
	σ_{do}					0.025

The number of common nodes between applications is shown in Tab. 5 first. For reference purpose, the number of nodes of a specific type of application is shown along the diagonal line of the table corresponding to another type of application, e.g., P2PF has 1318392 nodes in its flow graph. As illustrated, although the number of nodes for HTTP is significantly smaller than those of P2P applications, the common nodes between HTTP and IM/OG are more than those between P2P applications and IM/OG. This result indicates that people tend to perform other activities while browsing the web or vice versa, but not when they are downloading files or involved in an application with streams such as watching movie. IM/OG involves a lot of client-to-server communication as HTTP does, and many IM/OG messages are carried by HTTP protocol. This may also be the reason they have many common nodes.

As shown in Tab. 6, the degree correlations between the two P2P applica-

Table 7: Correlation Coefficients of Strength across Applications.

		P2PS	HTTP	IM	OG	AT
P2PF	r_{si}	0.410	0.378	0.130	0.150	0.017
	σ_{si}	0.024	0.027	0.029	0.043	0.068
	r_{so}	0.504	0.021	0.039	0.030	0.024
	σ_{so}	0.034	0.003	0.014	0.011	0.003
P2PS	r_{si}		0.396	0.096	0.156	0.228
	σ_{si}		0.120	0.036	0.040	0.083
	r_{so}		0.083	0.066	0.076	0.014
	σ_{so}		0.012	0.018	0.021	0.044
HTTP	r_{si}			0.139	0.164	0.097
	σ_{si}			0.028	0.038	0.082
	r_{so}			0.023	0.108	0.003
	σ_{so}			0.006	0.022	0.001
IM	r_{si}				0.040	0.015
	σ_{si}				0.014	0.123
	r_{so}				-0.002	0.128
	σ_{so}				0.034	0.027
OG	r_{si}					0.126
	σ_{si}					0.024
	r_{so}					0.037
	σ_{so}					0.005

tions are quite strong. However, there is a clear difference between their respective correlations with HTTP. The correlations between P2PS and HTTP are apparently stronger than those between P2PF and HTTP, which may be because that the index and the directory of streaming media are often carried by the HTTP protocol, and many media links are embedded in web pages. Moreover, P2PS has strong correlations with all other applications, while the correlations between P2PF and others are not that significant. This is perhaps because people tend to start P2P downloading when they do not use their computers, e.g., when watching stream media. AT traffic is observed to have quite large correlations with almost every application, showing that abnormal traffic affects all applications. The correlations related to r_{di} of AT are consistently stronger than those of r_{do}, probably because some connections are dropped by anti-virus software.

While degree correlation indicate the relationship between connection numbers of hosts, strength coefficient present the relationship between traffic volumes. In Tab. 7, the incoming strength coefficient (r_{si}) and the outgoing strength coefficients (r_{so}) are shown. The two P2P applications show strong correlations with their incoming and outgoing volumes, which is in conformance with the degree correlations. Both types of P2P applications carry mostly the same type of contents, i.e., videos and multimedia. So it is not surprising that people who like such contents tend to use both of them.

However, there are sharp differences between r_{si} and r_{so} for the two P2P applications and HTTP. The incoming strength of P2PF and HTTP has a high correlation with a value 0.378, while the outgoing strength has a very low correlation at 0.021. Similar relationship is also observed between P2PS and HTTP. A possible reason is that most hosts receive the traffic from the network using browsers and viewers, so the incoming traffic presents stronger correlations. However, the outgoing traffic flows are generated by different application servers, so the correlations between the outgoing traffic from different servers are low. In general, the correlation values of strength are smaller than those of degree, implying that the traffic from different applications is more random at hosts since hosts usually do not involve in transmissions of large amount of data for different applications at the same time.

5.4. Correlation Variation within a Day

Finally, we further investigate the variation of correlation coefficients over a 24-hour period. We average the 24 hour results over 2 days, and the four individual node assortativities shown in Tab. 3 are depicted in Fig. 6 and Fig. 7, with σ plotted as error bars. As we may observe, most coefficients do not vary significantly, and the variations are generally within ±0.1. For the two P2P applications, the coefficients regarding the strength decrease during night, perhaps because more hosts finish downloading data and become seeds or simply stop transmitting and receiving any data. However, the coefficients of HTTP increase or become flat during night, with the exception of r_{doso}. The coefficients of AT are quite uniform, with an obvious r_{dido} and no other correlations just like we observe in the first hour. This is because that the abnormal traffic is likely generated randomly by machines and does not depend on the time of day. There are no clear patterns in the IM and OG applications and the large corresponding σ hinders us from making more meaningful observation. As the percentages of traffic corresponding to these two traffic types are lower in the total traffic, we will resort to more powerful machine to process a larger number of records in the future.

The variation of the correlation coefficients of the connected nodes is also shown. Most values are observed to be around 0, with the exception of the HTTP traffic, which is shown in Fig. 8. During late night, the values of the degree coefficients approach 0, meaning that the negative assortativities reduce during night. This is because fewer users will initiate HTTP applications in late night. The traffic strength correlation remains to be low and at the similar levels at different time of the day.

Furthermore, we inspect the correlation coefficients across applications over a 24-hour period. We focus on the three largest applications and the coefficients between them are plotted in Fig. 9. For the correlations between the two P2P applications, it is interesting to observe that during night, the strength correlations (r_{si} and r_{so}) decrease when the connection correlations (r_{di} and r_{do}) vary, while in the morning, all the four correlations increase and remain high for the rest of the day. The correlations between P2PF and HTTP exhibit a deviation for the strength and connections. The connection correlations increase and reach the peak at 2∼5 am. In contrast during the same time, the correlation of the incoming strength drops to a minimum. The correlation of incoming strength remains unchanged. The coefficients of the connections between P2PS and HTTP are larger than those between P2PF and HTTP, while the coefficients of the strength are almost the same.

6. CONCLUDING REMARKS

In this work, we have constructed flow graphs based on detailed Internet traffic flow records for different applications. Six types of applications, namely P2P file sharing, P2P stream, HTTP, instant messaging, online game and abnormal traffic are investigated. We first reported the profiles of the applications, showing that P2PF, P2PS and HTTP are the three largest ones in terms of both the host numbers and the traffic intensity. Then we revealed that the number of nodes and connections as well as the traffic volumes of applications vary during the day, being the lowest from midnight to early morning as expected. Over 25% hosts participate in more than one application, with the average number of applications per host as 1.43.

Furthermore, the correlations within one type and across different types of applications were investigated in various ways. For each application type, four kinds of correlation coefficients concerning the same hosts and six kinds regarding connected nodes were calculated. Different types of applications have distinct correlation behavior. The coefficients show that the properties of hosts generally are positively assortative, with the two P2P applications being the strongest, while the hosts connected have weak negative or no assortativities. The correlation between degree are more significant than those between strength.

We further studied the correlations between different types of applications. The incoming and outgoing degree and the strength of common nodes between different types of applications were examined, and the correlation coefficients were presented. According to our results, applications have very different relationships, and the coefficients vary significantly from almost 0 to over 0.7. A few other interesting facts were also explored. For instance, although the two P2P applications have strong correlations between themselves, they correlate with HTTP very differently.

Moreover we investigated the variations of correlations within a day. Different behavior was found with different types of correlation coefficients. The correlations of P2P applications decrease during night, while the correlations of HTTP remain the same or increase during the period. AT presents uniform correlations within a day. The correlations between the two types of P2P applications increase in the morning and remain high during the daytime, and the correlations between P2PF and HTTP exhibit a deviation between the strength and connections.

Our studies also shown a variety of behaviors regarding the correlations of applications. Some properties exhibit distinctive differences which are helpful for understanding the nature of Internet applications and interactions. As mentioned when reporting the traffic profiles, about 25% of traffic cannot be classified yet. In our future work, we will further investigate the relationship between the unknown traffic and the identified application types.

7. Acknowledgment

This work is supported by Project 61171098 of the National Natural Science Foundation of China and Project CNS 0628039 of the National Science Foundation of USA. It is also supported by the Fundamental Research Funds for the Central Universities of China, and is partially supported by the EU FP7 project FP7-PEOPLE-IRSES S2EuNet under Grant No. 247083 and the 111 Project of China under Grant No. B08004.

References

[1] A. Callado, C. Kamienski, G. Szabo, B. Gero, J. Kelner, S. Fernandes, and D. Sadok, A Survey on Internet Traffic Identification, IEEE Communications Surveys & Tutorials, vol. 11, no. 3, pp. 37-52, 3rd Quarter 2009.

[2] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson, On the Selfsimilar Nature of Ethernet Traffic, in Proc. of ACM SIGCOMM '93, pp. 183-193, Ithaca, NY, USA, Sept. 1993.

[3] C. Labovitz, S. Iekel-Johnson, D. McPherson, J. Oberheide, and F. Jahanian, Internet Inter-Domain Traffic, in Proc. of ACM SIGCOMM '10, pp. 75-86, New Delhi, India, Aug./Sept. 2010.

[4] R. Bolla, M. Canini, R. Rapuzzi, and M. Sciuto, Characterizing the Network Behavior of P2P Traffic, in Proc. of the 4th International Telecommunication Networking Workshop on QoS in Multiservice IP Networks, pp. 14-19, Venice, Italy, Feb. 2008.

[5] K. Xu, Z. L. Zhang, and S. Bhattacharyya, Internet Traffic Behavior Profiling for Network Security Monitoring, IEEE/ACM Transactions on Networking, vol. 16, no. 6, pp. 1241-1252, Dec. 2008.

[6] S. Ratti, B. Hariri, and S. Shirmohammadi, A Survey of First-Person Shooter Gaming Traffic on the Internet, IEEE Internet Computing, vol. 14, no. 5, pp. 60-69, Sep./Oct. 2010.

[7] M. Iliofotou, P. Pappu, M. Faloutsos, M. Mitzenmacher, S. Singh and G.Varghese, Network Monitoring Using Traffic Dispersion Graphs (TDGs), in Proc. of the 7th ACM SIGCOMM conference on Internet measurement (IMC'07), pp. 315-320, San Diego, USA, October 2007.

[8] M. E. J. Newman, The Structure and Function of Complex Networks, SIAM Review, vol. 45, no. 2, pp. 167-256, May 2003.

[9] X. F. Wu, K. Yu, and X. Wang, On the Growth of Internet Application Flows: A Complex Network Perspective, in Proc. of IEEE INFOCOM 2011, pp. 2082-2090, Shanghai, China, April 2011.

[10] M. E. J. Newman, Mixing Patterns in Networks, Phys. Rev. E, vol. 67, no. 2, pp. 1-13, Feb. 2003.

[11] D. Chakraborty1, A. Ashir, T. Suganuma, G. Mansfield Keeni, T. K. Roy, and N. Shiratori, Self-similar and Fractal Nature of Internet Traffic, International Journal of Network Management, vol. 14, no. 2, pp. 119-129, March/April 2004.

[12] M. Kihl, P. Odling, C. Lagerstedt, and A. Aurelius, Traffic Analysis and Characterization of Internet User Behavior, in Proc. of 2010 International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT), pp. 224-231, Moscow, Russia, Oct. 2010.

[13] Q. Zhou, C. J. Miller, and V. Bassilious, First Person Shooter Multiplayer Game Traffic Analysis, in Proc. of 11th IEEE Symposium on Object Oriented Real-Time Distributed Computing (ISORC'08), pp. 195-200, Orlando, FL, USA, May 2008.

[14] M. Kuruba, G. M. Keeni, Characteristics of Illegitimate Internet Traffic, in Proc. of 2008 IEEE Region 10 Conference (TENCON 2008), pp. 1-5, Hyderabad, India, Nov. 2008.

[15] G. Maier, A. Feldmann, V. Paxson, and M, Allman, On Dominant Characteristics of Residential Broadband Internet Traffic, in Proc. of the 9th ACM SIGCOMM conference on Internet measurement conference, pp. 90-102, Chicago, IL, USA, Nov. 2009

[16] M. Faloutsos, P. Faloutsos, and C. Faloutsos, On Power-law Relationships of the Internet Topology, in Proc. of ACM SIGCOMM '99, pp. 251-262, Cambridge, MA, USA, Aug. 1999.

[17] P. Mahadevan, D. Krioukov, M. Fomenkov, X. Dimitropoulos, K. Claffy, and A. Vahdat, The Internet AS-level Topology: Three Data Sources and One Definitive Metric, ACM SIGCOMM Computer Communication Review, vol. 36, no. 1, pp. 17-26, Jan. 2006.

[18] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener, Graph Structure in the Web, Computer Networks, vol. 33, no. 1-6, pp. 309-320, June 2000.

[19] M. Iliofotou, H. Kim, M. Faloutsos, M. Mitzenmacher, P. Pappu, and G. Varghese, Graption: A Graph-based P2P Traffic Classification Framework for the Internet Backbone, Computer Networks,vol. 55, no. 8, pp. 1909-1920, 2011.

[20] Y. Jin, N. Duffield, P. Haffner, S. Sen and Z. Zhang, Can't See Forest through the Trees? Understanding Mixed Network Traffic Graphs from

Application Class Distribution, in Proc. of 9th Workshop on Mining and Learning with Graphs (MLG2011), Aug. 2011, San Diego, USA.

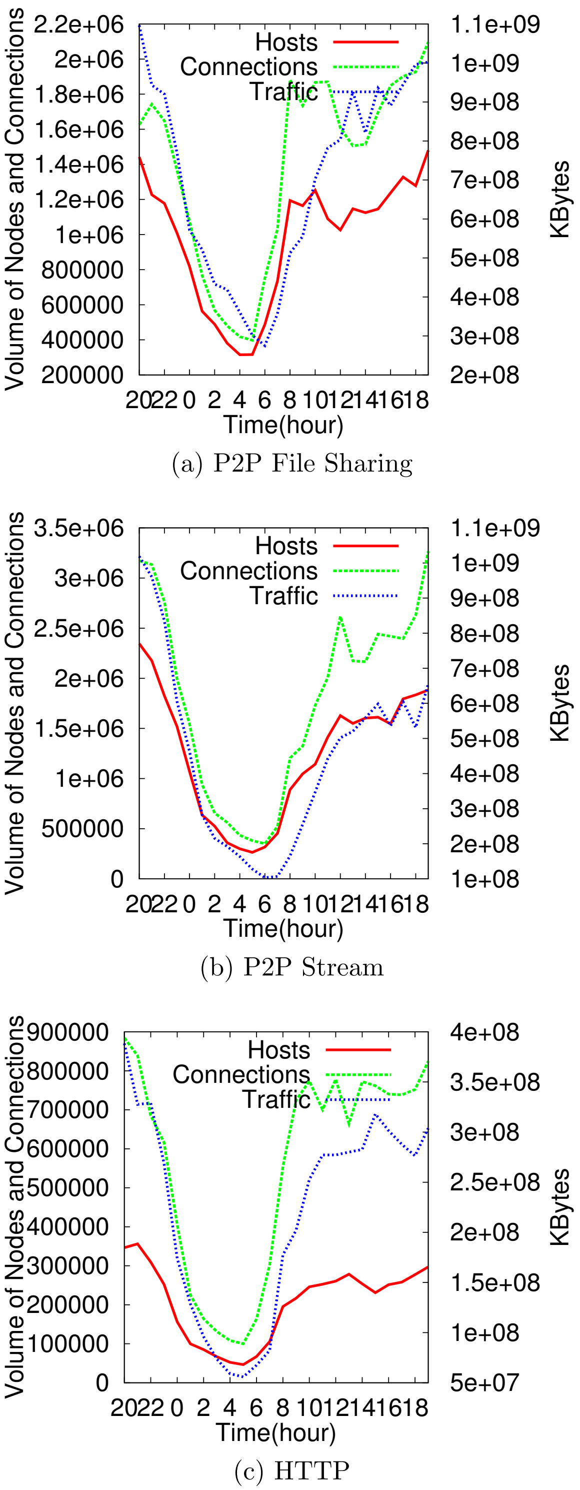
[21] A. Fahad, Z. Tari, I. Khalil, I. Habib, and H. Alnuweiri, Toward an Efficient and Scalable Feature Selection Approach for Internet Traffic Classification, Computer Networks, vol. 57, no. 9, pp. 2040-2057, June 2013.

[22] M. R. Meiss, F. Menczer, and A. Vespignani, Structural Analysis of Behavioral Networks from the Internet, Journal of Physics A: Mathematical and Theoretical, vol. 41, no. 22, June 2008.

[23] L. Nie, D. Jiang, and L. Guo, A Power Laws-based Reconstruction Approach to End-to-end Network Traffic, Journal of Network and Computer Applications, vol. 36, no. 2, pp. 898-907, March 2013.

[24] L. da F. Costa, F. A. Rodrigues, G. Travieso, and P. R. Villas Boas, Characterization of Complex Networks: A survey of Measurements, Advances in Physics, vol. 56, no. 1, pp. 167-242, Feb. 2007.

[25] M. Dusi, F. Gringoli, and L. Salgarelli, Quantifying the Accuracy of the Ground Truth Associated with Internet Traffic Traces, Computer Networks, vol. 55, no. 5, pp. 1158-1167, April 2011.

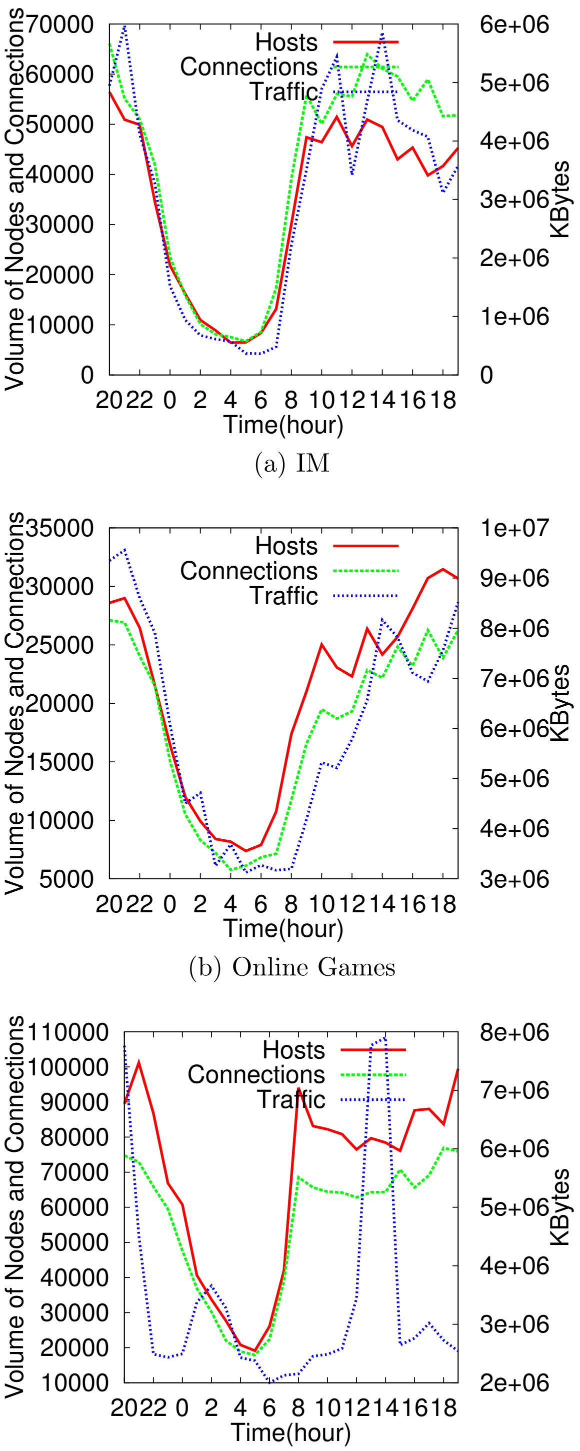


(a) P2P File Sharing

(b) P2P Stream

(c) HTTP

Figure 2: Hosts, Connections and Traffic over 24 hours.



(a) IM

(b) Online Games

(c) Abnormal Traffic

Figure 3: Hosts, Connections and Traffic over 24 hours(cont.).

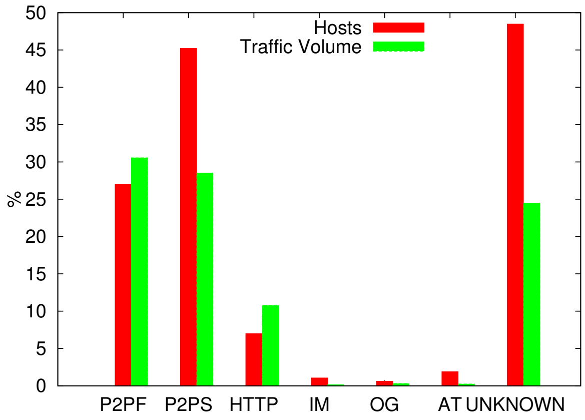
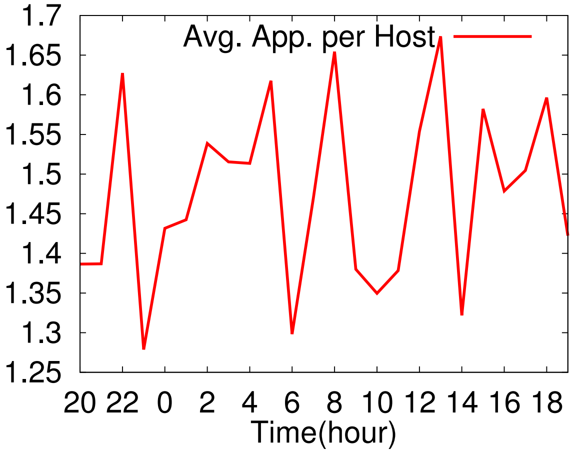
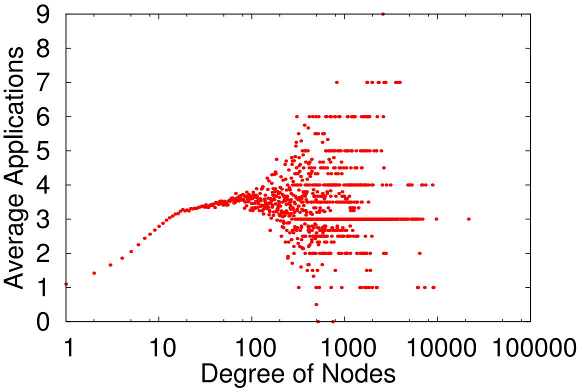


Figure 4: Nodes and Traffic Volumes of Applications.

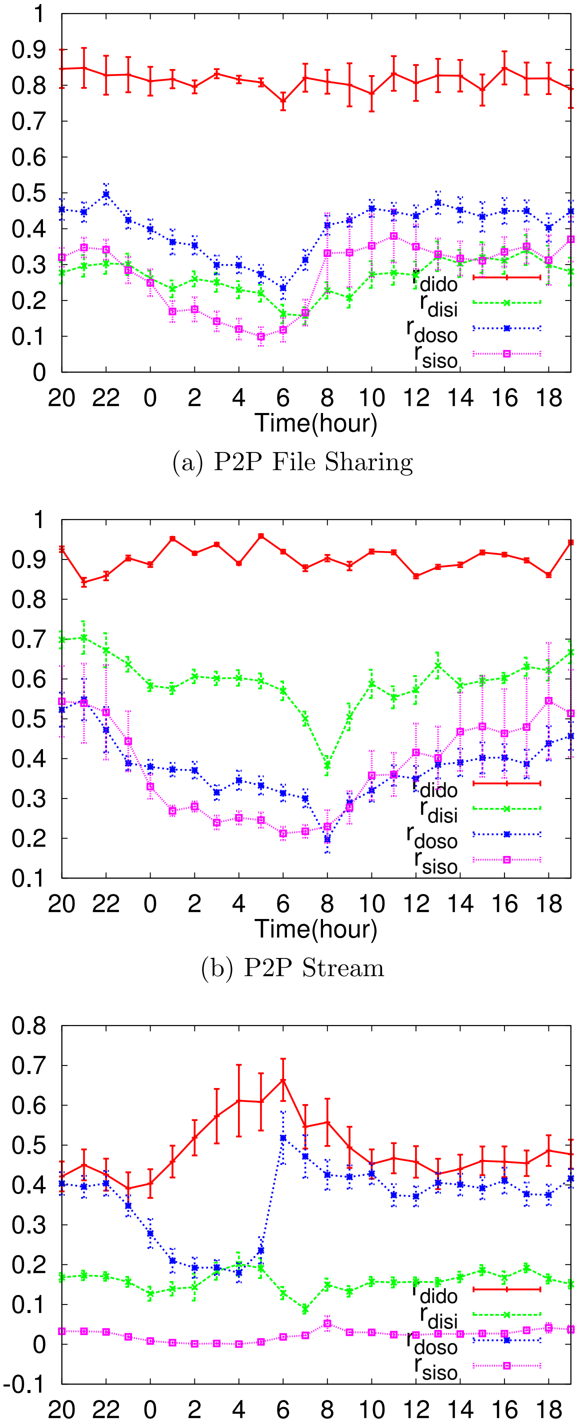


(a) Average Number of Applications per Host within a day



(b) Average Number of Applications

Figure 5: Average Number of Applications of Hosts.

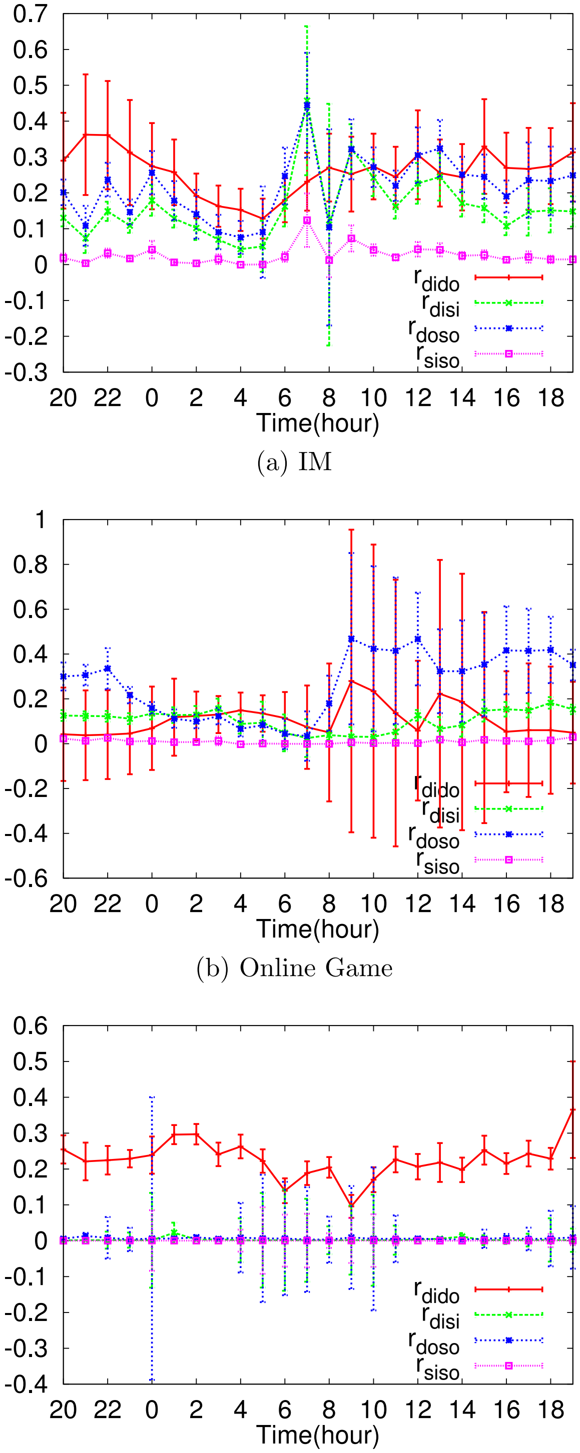


(a) P2P File Sharing

(b) P2P Stream

(c) HTTP

Figure 6: Correlation Coefficient within a Day.

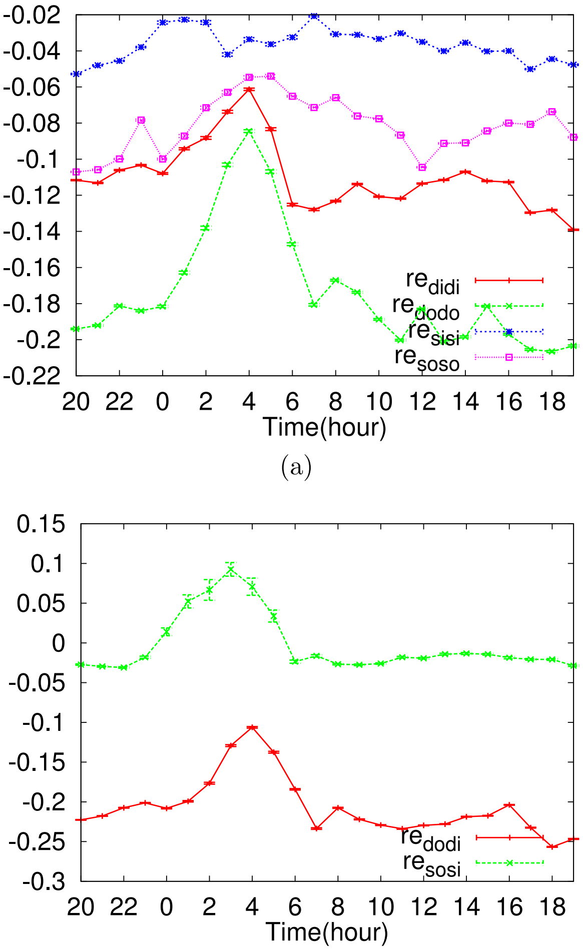


(a) IM

(b) Online Game

(c) Abnormal Traffic

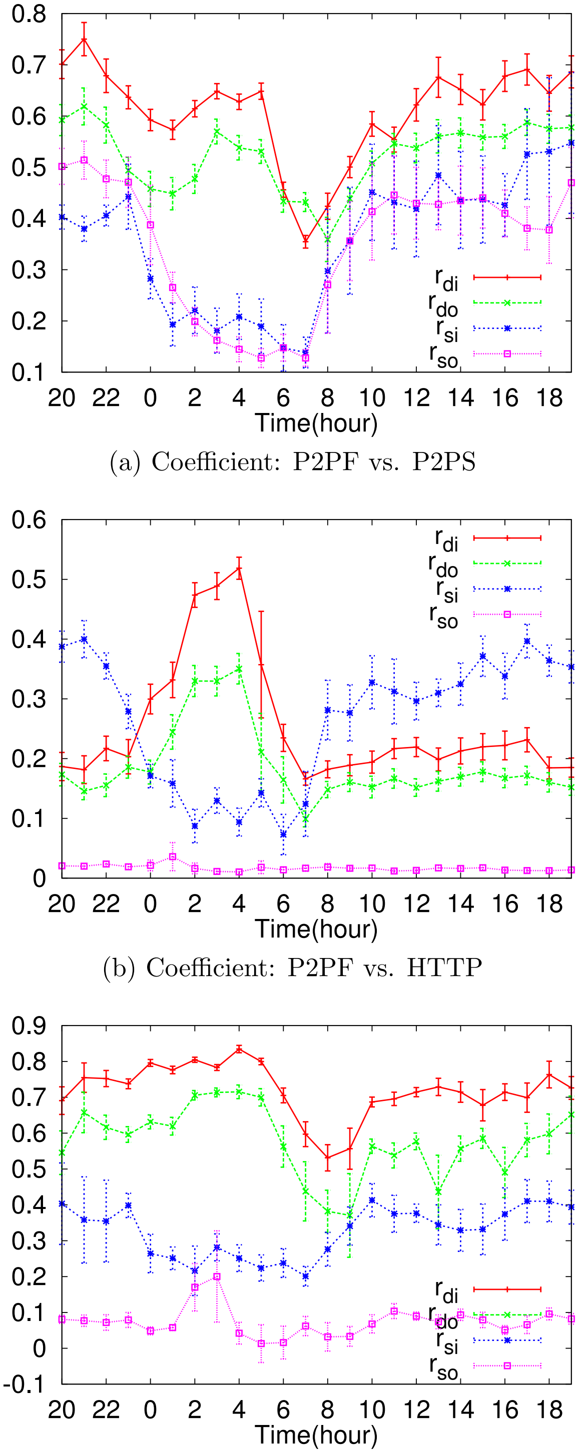
Figure 7: Correlation Coefficient within a Day(cont.).



(a)

(b)

Figure 8: Edge Correlation Coefficient of HTTP.



(a) Coefficient: P2PF vs. P2PS

(b) Coefficient: P2PF vs. HTTP

(c) Coefficient: P2PS vs. HTTP

Figure 9: Coefficients Between Applications.