NOTE: This is a version of an unedited manuscript that was accepted for publication. Please, cite as: J. L. García-Dorado, J. A. Hernández, J. Aracil, J. E. López de Vergara, F. J. Monserrat, E. Robles and T. P. de Miguel. "On the Dura-tion and Spatial Characteristics of Internet Traffic Measurement Experiments", IEEE COMMUNICATIONS MAGAZINE, Vol. 46, No. 11, pp. 148-155, Nov. 2008.

On the Duration and Spatial Characteristics of Internet Traffic Measurement Experiments

José Luis García-Dorado, José Alberto Hernández, Javier Aracil
and Jorge E. López de Vergara
Universidad Autónoma de Madrid (Spain).
Email: {jl.garcia, jose.hernandez, javier.aracil, jorge.lopez_vergara}@uam.es

Francisco J. Montserrat, Esther Robles and Tomás P. de Miguel.
RedIRIS - Spanish National Research and Education Network.
Email: {francisco.monserrat, esther.robles, tomas.demiguel}@rediris.es

April 4, 2008

Abstract

Often, Internet measurement-based studies have followed a three-step procedure: (1) Collection of network measurements; (2) measurement-based model inference; and, (3) generalization of the results obtained to other scenarios. Indeed, it has been a general belief that certain internetwork traffic statistics, such as the mostly used IP addresses and port numbers, show a similar behavior in networks with similar features, and the conclusions derived from the measurements of a given network could be extrapolated to a similar scenario.

This study makes no starting assumption concerning this issue and undertakes a "spatial" analysis of network measurements. The measurement set comprises a six-month trace collected by RedIRIS (the Spanish National Research and Education Network) at different monitored points across the country.

Our experiment shows that, although the frequency statistics of IP addresses and port numbers follow a Zipf distribution (as expected), the distributions' characteristic parameter values vary significantly in a spatial dimension, that is, across the individual university networks, even when the profile of the networks' user base are similar. In practical terms, this means that network designers, analysts, and operators should not assume that statistics for Internet site and applications usage for one network may accurately characterize other networks, even when those networks have similar user bases and environments. Furthermore, we show that experiment durations of approximately 30 days are necessary for the traffic processes to display stationarity. Hence, in order to obtain accurate statistics on traffic characteristics of large internetworks using state-of-the-art measurement techniques, long and spatially diverse experiments may be necessary.

Keywords: Network measurements; Spatial diversity; Temporal diversity; University networks; Internet Traffic Measurement Experiments.

1 Introduction

Network traffic measurements collected across the Internet provide very meaningful information for researchers, service providers and other members of the Internet community [1].

On the one side, network operators may benefit from such information in their goal of ensuring the appropriate quality of service (QoS) to their customers. Indeed, the ever-increasing user demands and wide variety of application requirements are forcing ISPs to develop network capacity plans very carefully, not only to maintain the quality of service provided, but also to reduce the need for investment. ISPs have not underestimated the benefits of traffic measurements, and have traditionally applied their potential to other related fields, namely the performance evaluation of networks, the detection of anomalies and denial of use attacks, and even the generation of the clients' invoices [2].

On the other side, the research community has also found essential the use of real network measurements to better understand Internet dynamics, and further apply this knowledge to the development of network models, with direct application to network operators' needs mentioned above [3].

However, the collection of representative traffic measurements is not a straightforward process. In the light of this, the authors in [4] provide a detailed explanation of the problems that can be found in the simulation of the Internet, some of which also arise in the process of measuring networks. Examples of such problems include the large size and the heterogeneous nature of the Internet, the ever-increasing number of new applications being introduced to the network, the fast and unpredictable way the Internet changes, the size and date of the sample collected and the handling of outliers in the measuring process. To overcome these limitations, the authors in [4] try to identify *invariants* in Internet behavior in order to reduce the complexity of its characterization. An invariant is defined as a facet of behavior that is empirically shown to persist for some time in a wide set of measured samples. Examples include the diurnal patterns of activity, the probability distributions which describe connection sizes and durations, the distribution of inter-arrival times between consecutive packets in aggregated Internet traffic and between network user sessions.

In this article, we pinpoint two additional difficulties: First, the *"spatial diversity"* of measurements, that is, whether the information arisen from measurements collected at diverse locations with similar features differs significantly or not; and secondly, the time required to capture stationarity, the *"temporal diversity"*, that is, the amount of measuring time needed to bring a sampled distribution which persists over time. Essentially, we try to answer the following two questions:

Can the conclusions derived from a measurement experiment in a given network be further applied to a similar network/scenario? And, how long should the measurement experiments last until stability in the metrics under study is reached?

Throughout this article, the term *similar networks* shall refer to networks which share certain common intrinsic features. In this light, the research community has generally accepted that the conclusions derived from a given network are valid for a scenario with similar characteristics, such as population size, bandwidth capacity and filtering policy. Therefore, measurements have been taken from links that are believed to be *sufficiently representative* of the Internet, typically university, residential or even smaller networks.

To answer the questions above, this work studies the distribution of the most popular IP addresses and port numbers (often bound to specific services/applications) in a set of university network access points nationwide. It is worth noticing that this study is not focused on the measurement

results themselves, which have been reported elsewhere, but instead on the *representativeness of network measurement experiments, in terms of spatial and temporal diversity*. Temporal diversity is related to the concept of "horizontal aggregation", as introduced in [5], whereby the authors study the necessary time-scale such that aggregated traffic follows a Gaussian distribution. However, in this paper we follow a rather different approach: the problem is not to estimate the time-scale to reach Gaussianity but to rather find the time horizon above which the distribution parameters remain stable. Such time horizon is typically in the range of days or weeks, a much coarser time-scale than the ones often considered in such horizontal aggregation studies (seconds or milliseconds). Other works have aimed at ranking the top traffic generators in a network scenario, often known as "heavy-hitters", and their persistence over time in such ranking [6]. Again, this is not the purpose of this article since we are only taking into account the traffic distribution for the most active ports and IP addresses, without making an explicit identification of them.

Concerning spatial diversity, this has received little attention from the research community. For instance, the authors in [7] make a comparison study of the inter- and intra-use of mainframes between seven Japanese regions in the late 1980s, but nonetheless the spatial diversity of the measurements was not analyzed. We believe that such lack of spatial diversity related studies is due to the difficulties in capturing traffic from a large number of distant networks and over large periods of time.

In fact, this work analyzes an extensive set of measurements (Netflow records) collected from a large number of university networks kindly donated by RedIRIS (the Spanish National Research and Education Network –NREN–). The following analysis is performed over the traffic flow records collected from April to September 2007, comprising a total of 13,000 million flows. RedIRIS spans more than 70 universities whose size, user population and organization is well documented in central repositories by the Spanish Ministry of Education for statistical purposes. Therefore, it is possible to group universities by similar features, for instance, number of users, bandwidth, traffic filters (e.g., restrictions on peer-to-peer (P2P) applications such as music file sharing), and proceed with the analysis to check *whether or not, university networks with similar intrinsic characteristics produce similar traffic patterns*.

The remainder of this work is organized as follows: Section 2 describes in detail the topology of the Spanish NREN and the measurement set under study. Section 3 presents the experiments performed and results obtained both in the time and space dimensions. Finally, section 4 concludes this work and summarizes the main findings obtained.

2 Measurements

The Spanish NREN serves more than 260 institutions, mainly universities and research centers, and comprises 18 Points of Presence across the country, as shown in figure 1. For the experiments, RedIRIS provided the traffic measurements at the access routers of a large number of universities interconnected by the Spanish NREN, typically of bandwidth ranging from 100 Mbps to 1 Gbps.

In what follows, we shall denote *downstream traffic* as the collection of flows that are sourced by a host located somewhere in the Internet and destined for a host located in the university network, and we shall denote *upstream traffic* as the converse, i.e., the collection of flows that are sourced by a host in the university network and destined for a host in the Internet (see figure 2). Note that with these definitions, inter-university traffic is neither downstream nor upstream traffic, indeed we did not include such traffic in our experiments.

2.1 Data collection infrastructure

In this section we describe the data collection infrastructure and data format. All the access routers feature flow monitoring (Netflow) capabilities. A flow is a sequence of packets that share the same source and destination IP addresses, port numbers and protocol. In this light, a flow summary includes: traffic volume in bytes and packets, port numbers, source and destination IP addresses, type of service, input and output interface indices (as per SNMP MIB), together with time-stamps for the flow beginning and end. For a thorough description of Netflow, the reader is referred to RFC 3955. The flow summaries are sent to a central repository, located at the Universidad Autónoma de Madrid (UAM) campus. The average input rate to the repository was 2 Mbps (flow summaries), over a six month period (April to September 2007).

Figure 1 shows the measurement system architecture. First, the *Flow-Tools* software package was used for data collection at the repository. Then, a number of statistics were obtained by the processing subsystem, which included total bandwidth consumption per university, peak-hour bandwidth requirements and most active IP addresses and port numbers. Finally, the Monitoring System provides a graphical interface, whereby such processed information can be accessed via web and properly visualized (this is the third stage).

2.2 University networks under study

The collected traffic sample comprised more than 70 universities, with different user base populations, access link capacities, filtering policies (P2P applications), proxies and Network Address Translation (NAT) capabilities. Clearly, such *intrinsic features* have an impact on the traffic pattern. For instance, if NAT services or proxies are available it is very possible to find that most traffic comes from a single IP address, but the truth is that a large number of traffic sources are sharing the same IP address. In the same way, NAT not only affects IP addresses but also port numbers, since every traffic source under the same IP address is given a different port number.

Consequently, we made a choice of universities with *similar features*, and compared the resulting most popular port and IP addresses distribution. In this light, we have carefully selected 9 universities out of the total set, for which the above intrinsic features are very much alike.

Firstly, regarding the filtering policy, we have chosen universities in which most non-educational traffic is allowed with no rate control except for well-know peer-to-peer applications. Additionally, it is worth noticing that the analyzed measurements comprise traffic to the Internet only, not between campuses. Thus, such inter-university traffic from supercomputing or grid facilities is explicitly not included. Furthermore, we also performed an inspection of the most active flows, in order to ensure that no outliers were present in the sample.

Secondly, concerning the use of NAT, we focus on most frequently accessed IP addresses and ports *on the Internet side*, i.e. destination IP addresses and port numbers of upstream flows from campus networks, and origin IP addresses and port numbers of downstream flows (see figure 2). Such measurements provide a more meaningful and representative portrait of the user behavior browsing Internet content, rather than pursuing a characterization of the Internet users that access hosts in the university campuses. Anyway, it is worth mentioning that the selected centers make negligible use of NAT and proxies, if any.

The population size of the universities under study ranges from $20,000$ to $40,000$ members with a similar proportion between subpopulations (strata), i.e. students, faculty and administration, thus

favoring the representativeness of the aggregated traffic (see table 1). Furthermore, this table shows the number of collected Netflow summaries for the selected universities, along with the number of active IP addresses in the peak traffic hour. The latter gives a hint of the population activity, to reinforce the fact that the sample is representative in terms of number of active users.

In addition to this, the access bandwidth capacity in all universities under study is exactly 1 Gbps and they are connected to the Internet through a single Exchange Point, located in Madrid.

We conclude that the selected universities are similar in terms of user base populations, access link capacity, filtering policies (P2P applications) and availability of proxies and Network Address Translation (NAT) services. It is finally worth remarking that the measurements were collected over the same time period, thus avoiding any contamination of the spatial diversity by temporal factors.

3 Experiments and results

The following presents a measurement analysis from the spatial diversity point of view, that is, whether or not equivalent universities share similar behaviour. It also shows the timescale for which the observed behaviour becomes stable, i.e. the sampling distribution does not significantly change as the sample size increases.

A typical invariant that can be observed from measurements of a university network concerns the IP addresses and port numbers most widely found in the traces. It is well-known [8] that, although the amount of possible destination IP addresses of flows and port numbers is huge, most users typically connect to the same sites and use the same services. Moreover, the amount of traffic either sourced or destined to the most popular IP addresses and port numbers follows a *Zipf distribution*. Zipf-like phenomena has been observed in the past in internetwork traffic traces [9], and often appears in other disciplines, such as economics, sociology and linguistics.

The Zipf cumulative distribution function is given by:

$$F(k) = \frac{\sum_{n=1}^{k} 1/n^s}{\sum_{n=1}^{N} 1/n^s}, \qquad k = 1, \ldots, N$$

where $s > 0$ characterizes the Zipf distribution, and N is the number of most popular IP addresses or port numbers included in the study, and k refers to their rank.

In our spatial analysis, we shall study the most popular (namely comprising most exchanged traffic in bytes) IP addresses and port numbers. Thus, we shall use $F(k)$ to represent the cumulative fraction of traffic (in bytes) over the total that are sent to the k-th most popular IP address or port number $k = 1, \ldots, N$ in the Internet.

For example, in Zipf distributions with $s = 1$, the most popular port number ($k = 1$) or IP address comprises as much as twice the traffic exchanged by the second ($k = 2$) most popular one, and thrice the traffic of the third ($k = 3$) popular one, and so on. For $s > 1$, the percentage of total traffic of the most popular one with respect to the others is even larger, and viceversa, i.e. if $s < 1$, such percentage is smaller. Hence, the s parameter is related to the tail decay of the Zipf distribution.

The purpose of the following experiments (spatial diversity) is to check whether or not university networks with similar intrinsic features, as discussed in the previous section, show the same behavior,

in terms of the s parameter of the Zipf law.

However, prior to any spatial analysis, it is first necessary to find a time-scale at which the parameters under study are stable. This is the purpose of next section.

3.1 Temporal diversity analysis

This section examines the temporal aspect of the measurement set over which we perform the spatial diversity analysis in the next sections. In other words, this section aims to check that the measurement set under study shows stationarity features, i.e. distributions that do not change with time. To do so, we evaluate the number of days worth of data required until the s parameter of the Zipf distribution for the most popular IP addresses and port numbers remains stable.

Figure 3 shows the most active destination IP addresses and port numbers of upstream flows for University U_1 (for 1-day and 1-month time slot), together with its most-likely Zipf distribution, obtained following the linear least squares regression technique described in [10]. The accuracy between the measured data and the theoretical Zipf fit can be visually checked in the figure. Note that only the fifteen ($N = 15$) most popular IP addresses and port numbers were taken into account in the estimation of the Zipf parameter s. We remark that similar behaviour was observed for $N = 8$ and $N = 20$, although such results have not been included for the sake of brevity.

This figure also shows that the Zipf model most accurately matches the measurements when 30 days worth of data is assumed (figure 3 right). Additionally, the estimated s values vary for different time-scales. Hence it is necessary to consider a large traffic sample until the s parameter becomes stable. Following this, figure 4 shows the estimated s value assuming several days of measurements. As shown, the s parameter estimate becomes smoother as we increase the trace length, bringing a stable value after 30 days of data. We consider a s estimate is stable if it varies less than 5% after five consecutive days.

It is also worth noticing that the s estimate after 30 days of data is different for all networks under consideration. This issue is analyzed in the next section.

3.2 Spatial diversity of most popular IP addresses and port numbers

Figure 5 shows the cumulative distribution function (CDF) of the fifteen most popular IP addresses (on the right) and port numbers (on the left) for all universities under study, both in the upstream (top) and downstream (bottom) direction from the Internet side. The numbers shown refer to the cumulative ratio of transferred bytes over the total in the trace. Following the results of the previous section, we have used 30 days worth of data in order to obtain a reliable estimate of the CDF.

Surprisingly, although the networks under study were carefully chosen with similar intrinsic features (large aggregation level, filtering policy, access bandwidth, proxies, NAT and population size and strata), the observed traffic profiles, as measured by the s parameter values, are different from one another. It is worth noticing that the population sizes of all networks under study are large enough (more than $20,000$ Internet users) such that the CDF are expected to converge to the same distribution.

In conclusion, the most popular IP addresses and port numbers of each university network follow a Zipf distribution, but the spatial analysis has shown that the particular s parameter is different

in each case (see figure 4). Hence, *measurements collected at one university are not generally valid to another, even if they have similar intrinsic features.*

4 Conclusions and future work

This work provides a new point of view in the study of network measurements: the spatial analysis. Essentially, the spatial analysis aims to check whether or not the conclusions derived from the analysis of a given set of measurements gathered from a particular network scenario are valid to another but similar network scenario. The answer to this question is negative. Although a number of invariants have been identified to persist across different scenarios, our findings show that, when measurements from networks with similar intrinsic features are compared, the distribution of the most popular port numbers and IP addresses differ from one network to another.

Additionally, the experiments have shown that the distribution of the most popular IP addresses and port numbers experience high variability, and only reach some stability when long periods of measurements are considered, typically in the range of weeks. However, it is important to remark that, given the heterogeneous nature of the Internet and the fast and unpredictable way it changes, the results do not remain valid for long periods of time, thus requiring continuous monitoring and measuring, as noted in [4].

This involves two important consequences: Firstly, the duration of internetwork experiments must last until the measurements under study become stable, which involve a much longer traffic trace than usually believed; and, secondly, single-link measurements do not suffice for a meaningful analysis, hence a spatially diverse measurement experiment must be carried out. As a result, the required measurement infrastructure must be designed accordingly, and may involve sophisticated and costly equipment, both in terms of storage capabilities and number of probes.

While this result is worthwhile to be reported itself, a number of interesting research directions appear from this conclusions, for instance, the search for an explanation on why spatial diversity occurs. As future work, the authors will focus their attention on the most active users, since they are highly variable and seem to be responsible for the majority of traffic generated in an internetwork.

Acknowledgments

This work has been partially funded by the Spanish Ministry of Education and Science under projects *DIOR* (TEC2006-03246), *CONSOLIDER i-MATH* (CSD2006-00032) and the F.P.I. fellowship program. The authors would also like to acknowledge the support of the IMDEA Mathematics Research Institute.

The authors are grateful to Dr. Michael O'Donnell for his careful proofreading of this article.

Finally, we would also like to thank the anonymous reviewers and liaison editor who helped us to improve the quality of the paper.

References

[1] N. Brownlee and K.C. Claffy, "Understanding internet traffic streams: Dragonflies and tortoises," *IEEE Commun. Mag.*, vol. 40, no. 10, pp. 110–117, Aug. 2002.

[2] C. Fraleigh, S. Moon, B. Lyles, C. Cotton, M. Khan, D. Moll, R. Rockell, T. Seely, and C. Diot, "Packet-level traffic measurements from the Sprint IP backbone," *IEEE Network*, vol. 17, no. 6, pp. 6–16, Aug. 2003.

[3] S. Floyd and E. Kohler, "Internet research needs better models," *ACM Comput. Comm. Rev.*, vol. 33, no. 1, pp. 29–34, 2003.

[4] S. Floyd and V. Paxson, "Difficulties in Simulating the Internet," *IEEE/ACM Trans. Netw.*, vol. 9, no. 4, pp. 392–403, Aug. 2001.

[5] J. Kilpi and I. Norros, "Testing the gaussian approximation of aggregate traffic," in *Proc. Internet Measurement Workshop*, 2002, pp. 49–61.

[6] J. Wallerich and A. Feldmann, "Capturing the variability of Internet flows across time," in *Proc. IEEE INFOCOM*, 2006, pp. 1–6.

[7] J. Murai, H. Kusumoto, S. Yamaguchi, and A. Kato, "Construction of internet for Japanese academic communities," in *Proc. ACM/IEEE conf. on Supercomputing*, 1989, pp. 737–746.

[8] A. Feldmann, A.G. Greenberg, C. Lund, N. Reingold, J. Rexford, and F. True, "Deriving traffic demands for operational IP networks: methodology and experience," *IEEE/ACM Trans. Netw.*, vol. 9, pp. 265 – 280, Jun. 2001.

[9] L.A. Adamic and B.A. Huberman, "Zipf's law in the Internet," *Glottometrics*, vol. 3, pp. 143–150, 2002.

[10] P.T. Nicholls, "Estimation of Zipf parameters," *J. American Society for Information Science*, vol. 38, no. 6, pp. 443–445, Nov. 1987.

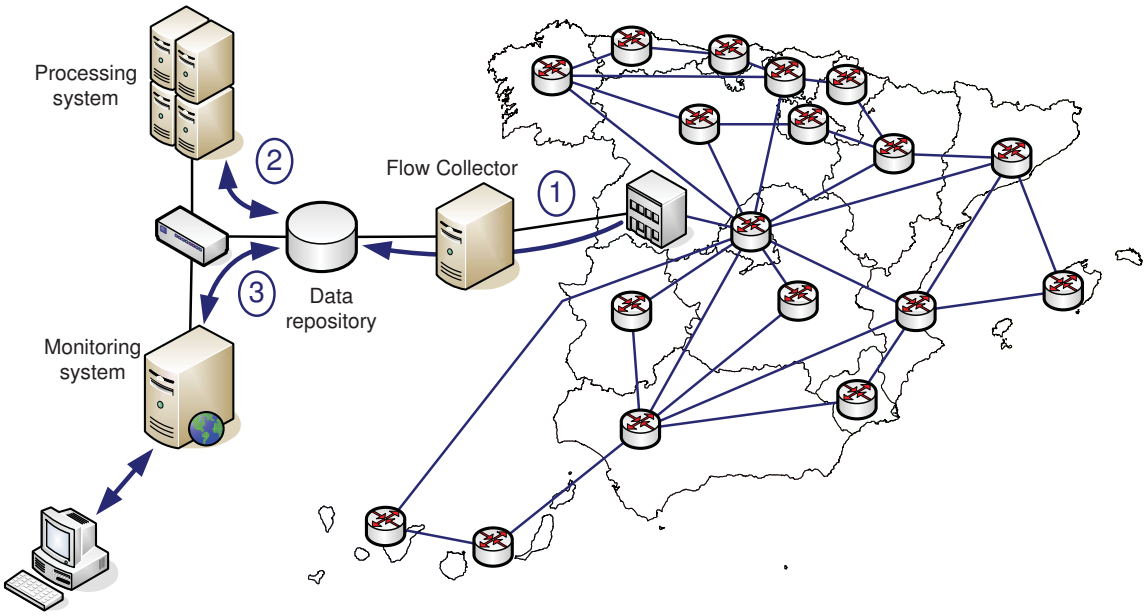

Figure 1: Measurement system architecture (on the left) and RedIRIS network topology (on the right).

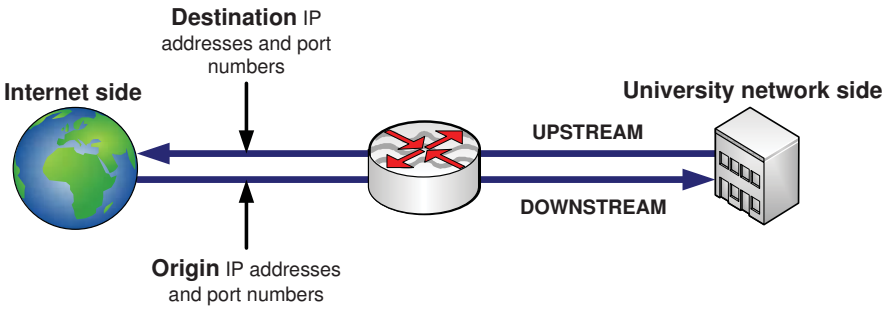

Figure 2: Analyzed data: Destination IP addresses and port numbers for upstream flows, and origin IP addresses and port numbers of downstream flows.

Universities	Population (ratio students / staff)	No. of Flows per day	Different IP addresses in the peak hour (From University/ to Internet)
U_1	40,000 (9.6)	1,400,000	4,000 / 23,000
U_2	29,000 (10.8)	1,300,000	4,000 / 22,800
U_3	20,000 (12)	2,000,000	3,200 / 30,000
U_4	31,500 (11)	5,870,000	5,000 / 90,000
U_5	30,500 (10.3)	3,000,000	4,300 / 66,000
U_6	36,000 (11.2)	4,000,000	5,600 / 66,000
U_7	33,500 (12.2)	3,500,000	4,500 / 58,000
U_8	26,500 (11.2)	2,400,000	6,500 / 30,000
U_9	28,000 (10.5)	2,500,000	2,000 / 30,000

Table 1: User-base population size, average number of flows collected per day and average IP addresses in the peak hour per day for all universities under study.

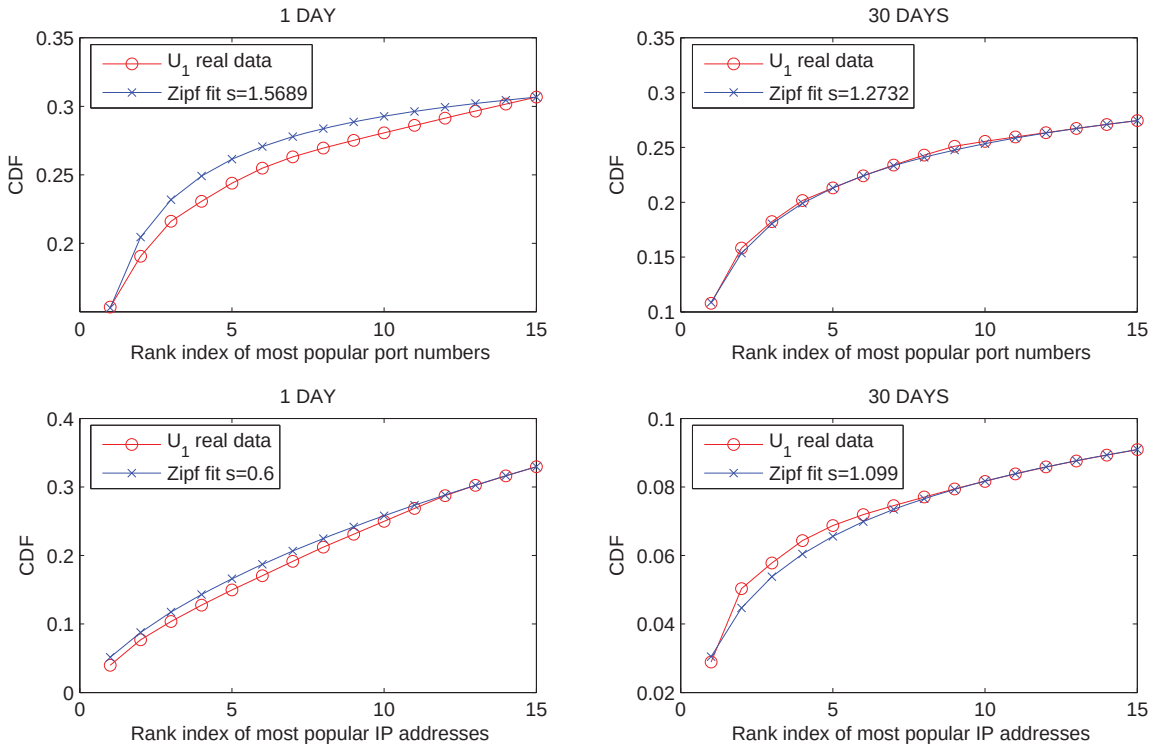

Figure 3: Cumulative distribution function of most popular port numbers and IP addresses (upstream) for U_1 and its Zipf distribution fit, assuming 1 day of data (left) and 30-days of data (right).

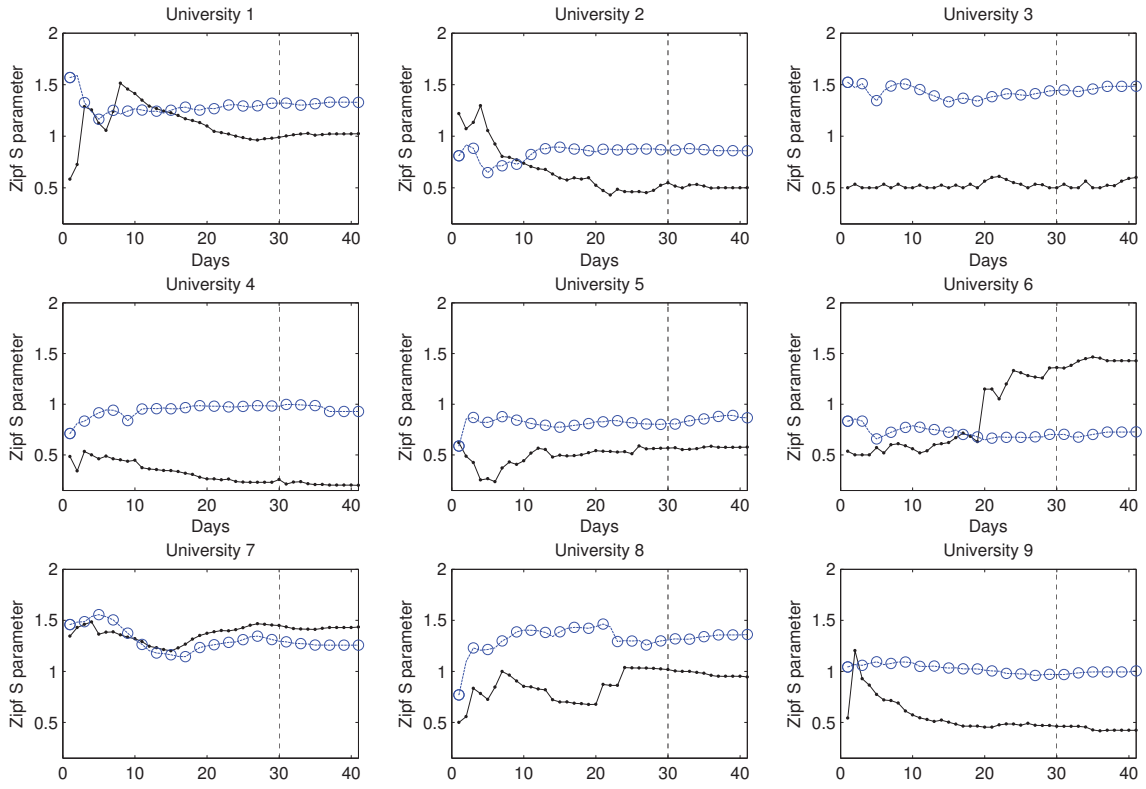

Figure 4: Most-likely Zipf distribution s value for the 15 most popular port numbers and IP addresses for all university networks (only upstream flow direction) for various time-scales of traffic statistics (from 1 day to 40 days of aggregated data).

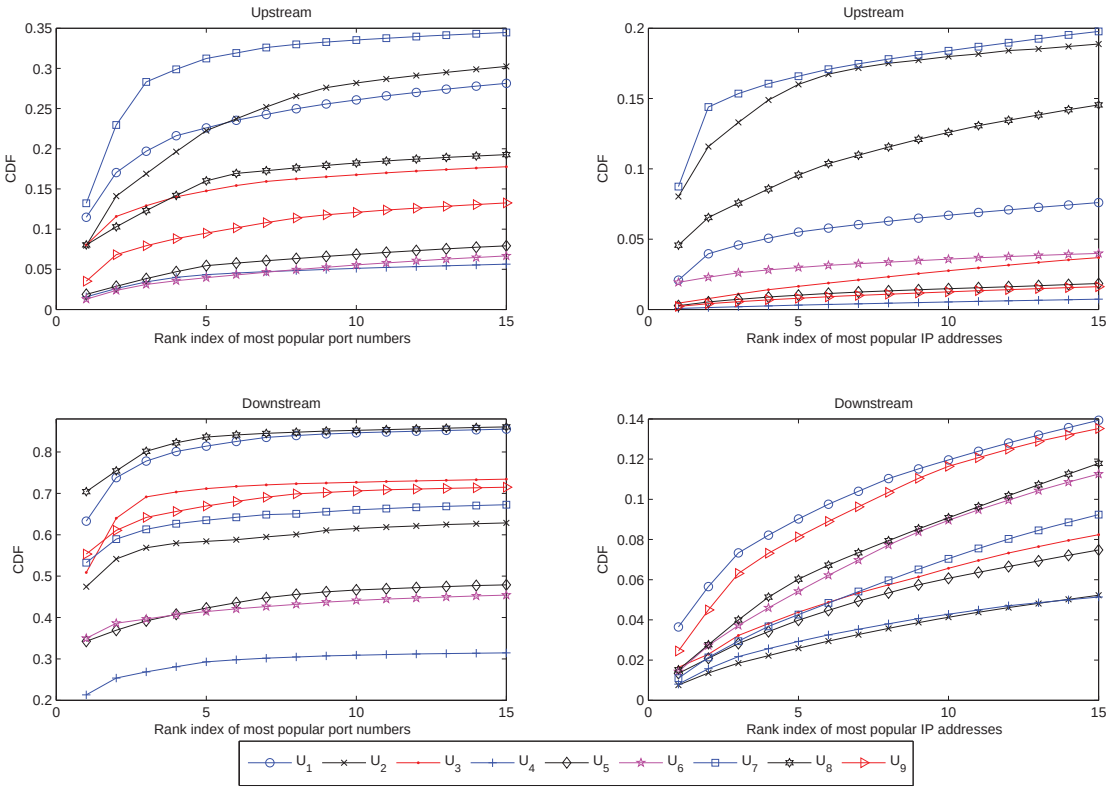

Figure 5: CDF of most popular IP addresses and port numbers for all universities under study.