Flow-concurrence and bandwidth ratio on the Internet

José Luis García-Dorado and Javier Aracil

High Performance Computing and Networking research group,
Universidad Autónoma de Madrid, Spain

Abstract

The relevance of flow-based monitoring in tasks such as the detection of anomalies and denial of use attacks, traffic reporting, performance evaluation, software routing among others motivates the study of the Internet traffic in terms of flows. However, by the time a network manager or practitioner start any of these tasks, they face the challenge of dimensioning probes' capacities, which depends on the number of concurrent flows. Unfortunately, while the bandwidth of a network or link, both in operation and in future deployments, can be known, or at least estimated in advance, nothing is known about the load in terms of concurrent flows. We aimed at filling this gap by studying the concurrence of flows with respect to the bandwidth normalized by factors such as protocol shares, timeouts, applications, time and years among others. As a result, we provide the research community with several models, based on lognormal distributions, and parameter estimates in such a way that any player in the Internet arena can estimate the expectable number of concurrent flows in their own infrastructure. Moreover, such results emerge from a diverse set of network traces so making the extrapolation of conclusions viable.

Keywords: Flow concurrence; Flow-Bandwidth ratio; Netflow; Capacity planning; Network monitoring.

1 Introduction

Flow-based monitoring has become a vital tool for numerous management tasks that operators and service providers carry out. The examples span a number

1

of fields: monitoring [1, 2, 3], performance evaluation of networks [4], traffic engineering [5], the detection of anomalies and denial of use attacks [6, 7], traffic classification [8, 9, 10] and even the generation of clients' invoices [11]. Moreover, the research community has also exploited flow-based records as a powerful tool to measure the Internet in an attempt to further expand the knowledge of its dynamics [12, 13, 14]. Leveraging on network flows to carry out all these tasks comes with a decrease in the traffic volume to analyze with respect to packet traces, and with an increase of the available information with respect to more aggregated measurements as those from SNMP [15]. Therefore, the use of network flows has turned out to be an interesting trade-off between both edges, packet traces and SNMP time series [16], which explains its tremendous success. Moreover, with the advent of the IPv6 protocol, which specifically includes a flow label in its header, the relevance of handling traffic at flow level may even increase. Finally, programmable routers brought by software-defined networks are expected to work at flow level, with open APIs that allow providing differentiated Quality-of-Service (QoS) on a per-flow basis [17].

A network-flow monitoring system, regardless its latest finality, has as its first task to construct network flows following Netflow protocol as Cisco designed it decades ago [18], which was later standardized by IPFIX [19]. A network flow (or, in short and hereafter, a flow) record comprises fields: <IP_source (32 bits), IP_destination (32 bits), source_port (16 bits), destination_port (16 bits), protocol (8 bits)>. Netflow uses such fields, typically named as the flow 5-tuple, as keys to group packets into flows in the table of active flows. In addition to these fields, flows may also comprise optional pieces of information such as timestamps, next hop, subnet masks, TCP flags, a fraction of payload, statistics of packet sizes to cite some of them [20].

Flow construction consists of the extraction of the 5-tuple of each arriving packet, then, such 5-tuple is used as a key to search in the table of active flows and, finally, if an active flow is found it is updated and, otherwise, a new one is created. In general, flow records are exported by connection termination flags in TCP or, in general, when no packet arrives for a given active flow during a configurable idle timeout. Other less common reasons are router resource exhaustion, which denotes a wrong configuration, or flows longer than a given interval, typically 30 minutes. However, often flow construction, given its per-packet basis operation, is incompatible with other routers' duties, among which routing stands out. In such a case, typically, routers sacrifice the accuracy of flow construction applying packet sampling, e.g., simple random sample or applying smarter methodologies in an attempt to cut burden and keep precision [21].

With all this in mind, it turns out that the correct dimensioning of flow-based systems depends on the concurrence of the number of flows. That is, how many flows per unit of time the system needs to handle, instead of traffic volumes aggregates, such as bandwidth in Mb/s or Gb/s or packet rates, metrics that are easier to estimate. As it turns out, the capacity planning of the flow table is of utmost importance, as it is visited once per incoming packet and at line rate. As networks are evolving towards 100 Gb/s per link and beyond, the read/write latency required is very small, which calls for ultra-fast-access memories. The latter are scarce and limited in size as they normally reside on-chip with the processor.

From our experience, any time a network scenario must be monitored, from academic to commercial networks, network managers possess knowledge about

their network traffic volume aggregates. For example, the average bandwidth used in a link at a given time is about 1.2 Gb/s, typically extracted from SNMP-based tools. And also, some hints about the traffic the network carries. For example, the fraction of UDP traffic, or the high/low popularity of P2P applications in the network.

In light of this, this paper proposes to exploit the relationship between bandwidth (the available information) and the flow concurrence (the target information). In other words, to study the metric number of concurrent flow per unit of bandwidth (e.g., Mb/s). This way, network managers may estimate load in terms of flows by means of already available SNMP measurements or bandwidth measurements from other networks. Intuitively, such metric would depend on the particularities of each network. For example, the application mix, protocol shares, timeouts, applications, time, years among other factors. Consequently, we have carried out a trace-driven analysis of the flow concurrence as general as possible on the Internet by crawling for public traces and achieving, eventually, a varied and rich set of traces. We have been able to analyze traces that span different classes of users, dates, volumes and countries. However, during the study of the traces, we have re-learned how far real traffic is from ideal behaviors, as several unexpected issues emerged. In particular, we have found that most of the TCP flows are not closed as taught in books or explained in RFCs. For example, some of them suddenly stop transmitting and flow is not exported until timeouts expire, which entails a waste of memory as flow records have to be in cache waiting for a packet that never arrives. Similarly, some TCP flows are closed and, then, re-opened several times, which makes flow cache export and re-open flow records several times.

Despite this, we have found some rules and propose several models, based on lognormal distributions, to relate the load in flows and the bandwidth. Initial results suggested that one Mb/s of traffic may carry as much as 1000 flows in average with 95th percentile over 10,000. However, after considering other factors, these figures can be reduced by an order of magnitude. Note that in a Multi-Gb/s network, this can imply overestimates of millions of flows. In particular, we have found that factors such as the timeout and the layer-4 protocol may explain as much as 75% of the diversity of the phenomenon. Finally, the application of the models with the parameter estimates in both the set of networks under study and some evaluation traces has proven the utility of our approach. Therefore providing both network managers and practitioners with a robust tool in their task of planning capacity for flow-based systems.

2   Problem statement and datasets

Let us define the flow concurrence as the number of active flows given a unit of time. Then, let the traffic volume be the number of bits transmitted during the same unit of time. Finally, the ratio between the number of active flows and the traffic volume gives the metric we are interested in. Let us refer to it as the flow-bandwidth ratio, or, in short, as flow ratio. In general, we measure the bandwidth in Mb/s, the number of active flows in flows per second, and, as a result, the flow-bandwidth ratio is measured in flows per Mb in a granularity of one second.

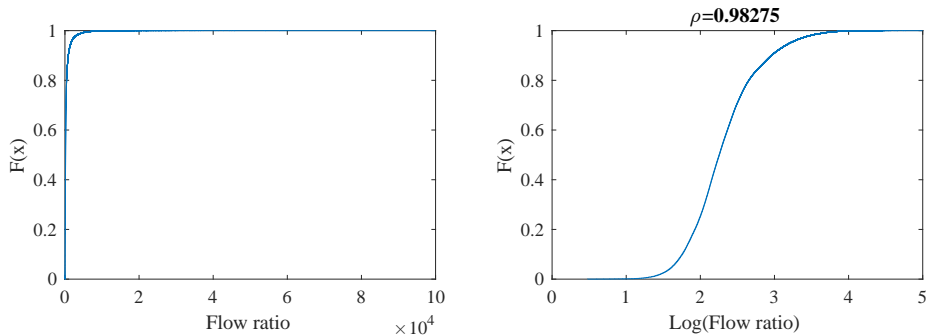This way, a network manager, simply, by multiplying their estimated band-

Figure 1: Flow ratio distribution in the full set of traces

width and the flow ratio obtains the estimation of the concurrent number of flows. Then, our aim is delving into the dynamics of this metric in diverse scenarios, assessing the impact of factors in it, modeling and reporting parameter estimates and, finally, providing feedback to the Internet community.

This search for diverse Internet scenarios calls for the capture of traffic in diverse locations for long periods of time. To this end, the use of public traces' repositories emerges as a natural approach. Unfortunately, the Internet community faces some problems to share captures freely, not from a technical point of view, but closer to privacy and security concerns as well as legal issues [22]. Even more, things are even more complicated when traces include application payloads. Note that this sort of traces may result of great interest to relate flow ratios and application popularity. Fortunately, the research community has circumvented these problems, and institutions such as AMPATH [23], HPCN [24], Caida [25], University of Twente (T) [26], University of Southern California (USC) [27] and Wide [28] have released a rich set of traces. Table 1 summarizes the most significant aspects of the traces we have had available. They span thirteen years of measurements in nine networks, at different times, involving several countries, with different network capacities among other characteristics.

3 Flow-Bandwidth ratio

As a first result, Figure 1 shows the distribution of the flow ratio metric for the full set of traces both in absolute terms and after a logarithmic transformation. Several observations arise.

- The distribution shows a long tail, with some infrequent samples over 200,000 flows per Mb, however the most significant fraction of the probability mass is below some thousands flows per Mb.

- A logarithmic transformation permits an easier visualization. Interestingly, 90% of the probability resides in the interval 100-10,000 flows per Mb, 70% between 100 and 1000, with a median and mean below 1000. This can be considered as a first empirical implication. For example, given a GbE network, the number of concurrent flows tends to range between 100 thousand and 10 million concurrent flows, although typical values would

4

Table 1: Summary of traces analyzed

Network	Total capture time	Measurements Distribution	Year	Date	Type	Typical bandwidth	Capacity	Location
AMPATH	2 days	Continuous	2007	9JAN	Academic	10 Mb/s	OC-12	USA
International Bank	15 days	Continuous	2013	22JAN	Branches traffic	30 Mb/s	1 Gb/s	LATAN
Equinix Chicago Link	10 hours	1 hour per day	2008	19MAR 30APR 15MAY 19JUN 17JUL 21AUG 18SEP 16OCT 20NOV 18DIC	Internet backbone links	5 Gb/s	OC-192	USA
Equinix Chicago Link	12 hours	1 hour per day	2009	15JAN 19FEB 31MAR 16APR 21MAY 18JUN 16JUL 20AUG 17SEP 15OCT 19NOV 17DIC	Internet backbone links	6 Gb/s	OC-192	USA
Equinix Chicago Link	7 hours	1 hour per day	2010	21JAN 25FEB 25MAR 14APR 19AUG 16SEP 29NOV	Internet backbone links	6 Gb/s	OC-192	USA
Equinix Chicago Link	6 hours	1 hour per day	2011	17FEB 24MAR 13APR 19MAY 21JUL 25AUG	Internet backbone links	3 Gb/s	OC-192	USA
Equinix Chicago Link	9 hours	1 hour per day	2013	25APR 29MAY 20JUN 18JUL 15AUG 19SEP 24OCT 21NOV 19DIC	Internet backbone links	5.5 Gb/s	OC-192	USA
Equinix Chicago Link	10 hours	1 hour per day	2014	16JAN 20FEB 20MAR 25APR 19MAY 19JUN 17JUL 21AUG 18SEP 18DIC	Internet backbone links	5.5 Gb/s	OC-192	USA
Equinix New York Link	6 hours	1 hour per day	2018	15MAR 19APR 17MAY 21JUN 19JUL 16AGO	Internet backbone links	4 Gb/s	OC-192	USA-Brazil
Equinix San Jose Link	6 hours	1 hour per day	2008	17JUL 21AUG 18SEP 16OCT 20NOV 18DIC	Internet backbone links	3 Gb/s	OC-192	USA
Equinix San Jose Link	12 hours	1 hour per day	2009	15JAN 19FEB 31MAR 16APR 21MAY 18JUN 16JUL 20AUG 17SEP 15OCT 19NOV 17DIC	Internet backbone links	5 Gb/s	OC-192	USA
Equinix San Jose Link	11 hours	1 hour per day	2010	21JAN 25FEB 25MAR 14APR 17JUN 15JUL 19AUG 16SEP 20NOV 18OCT 17DIC	Internet backbone links	5.5 Gb/s	OC-192	USA
Equinix San Jose Link	11 hours	1 hour per day	2011	20JAN 17FEB 24MAR 13APR 19MAY 21JUL 25AUG 15SEP 20OCT 17NOV 22DIC	Internet backbone links	5.5 Gb/s	OC-192	USA
Equinix San Jose Link	12 hours	1 hour per day	2012	19JAN 16FEB 15MAR 18APR 15MAY 21JUN 19JUL 16AUG 20SEP 18OCT 15NOV 20DIC	Internet backbone links	5 Gb/s	OC-192	USA
Equinix San Jose Link	12 hours	1 hour per day	2013	17JAN 21FEB 21MAR 25APR 29MAY 20JUN 18JUL 15AUG 19SEP 24OCT 21NOV 19DIC	Internet backbone links	5.5 Gb/s	OC-192	USA
Equinix San Jose Link	9 hours	1 hour per day	2014	16JAN 20FEB 20MAR 25APR 16MAY 19JUN 17JUL 21AUG 18SEP	Internet backbone links	3 Gb/s	OC-192	USA
T1	4 hours	15 mins. per day	2002	23MAY-29MAY 23JUN-26JUN	University dorms	160 Mb/s	300 Mb/s	Netherlands
T5	50 hours	15 mins. 3 times a day	2003-2004	5DIC-9FEB	Web hosting	10 Mb/s	100 Mb/s	Netherlands
USC	3 days	Continuous	2008	18MAR	Academic	300 Mb/s	1 Gb/s	USA
Wide	2 days	Continuous	2007	12APR	Academic	100 Mb/s	1 Gb/s	Japan
Wide	3 days	Continuous	2008	10MAR	Academic	100 Mb/s	1 Gb/s	Japan
Wide	3 days	Continuous	2010	13APR	Academic	200 Mb/s	1 Gb/s	Japan
Wide	63 hours	Continuous	2012	30MAR	Academic	250 Mb/s	1 Gb/s	Japan
Wide	3 days	Continuous	2013	25JUN	Academic	250 Mb/s	1 Gb/s	Japan
Wide	1 day	Continuous	2014	10FEB	Academic	500 Mb/s	1 Gb/s	Japan
Wide	2 days	Continuous	2017	12APR	Academic	600 Mb/s	1 Gb/s	Japan
Wide	2 days	Continuous	2018	9MAY	Academic	700 Mb/s	1 Gb/s	Japan

be smaller than a million concurrent flows. It turns out that such interval width, while coherent to the Internet diversity and heterogeneity [29], gives a limited applicability to real scenarios.

- After the logarithmic transformation, we assess the goodness-of-fit to a Gaussian distribution by means of the correlation test [30], resulting in a determination coefficient of 0.99 over the typical thresholds of 0.9 or 0.95. This suggests the modeling of the phenomenon by a fairly Gaussian process, which makes any analysis more tractable.

This way, the aim from here on is to reduce the width of the estimated intervals for the flow ratio. To do so, we apply a factorial approach where the metric under study is analyzed grouped into characteristics (or more formally, factors). For example, by grouping by level-4 protocols (i.e., TCP, UDP or other IP traffic) the dispersion in the flow ratio becomes lower inside each possible group (e.g., TCP). In practical terms, this means that when it comes the time to estimate the flow ratio, instead of observing the values of Figure 1, network managers aware of protocol shares in their networks may infer the flow ratio according to the popularity of each protocol. And the same applies to other factors such as the time when measurements were gathered, timeout used, traffic volumes, among others, reviewed in the next section.

3.1 Factor analysis

We entrusted ANOVA with the factor analysis of the data [31]. ANOVA is a centenary methodology typically applied in the social science area, but that recently has attracted the interest of the Internet community. To cite some examples, ANOVA was used in [32] as a mechanism to compare the performance of deployments in the public Cloud in terms of bandwidth and downtime occurrence. The authors in [33] applied ANOVA to model the internal structure of institutional websites. ANOVA was also useful to assess if Twitter activity and patterns, specifically twitters published in computer science conferences, have changed over time [34].

In particular, ANOVA takes the observed variance of a given response variable and describes it in terms of explanatory factors (or groups), specifically as an addition of terms that accounts for the effect of such factors (or their interaction). In this way, ANOVA determines whether a set of factors has statistical importance in explaining the response variable (and if so, to what extent) and provides an estimate for the terms. More intuitively, ANOVA compares the mean of a set of observations before and after they are grouped into factors and interactions of factors. If the difference is relatively large, then the factor (or interaction) is considered significant, otherwise, it is irrelevant. Considering the relevant factors, ANOVA poses a model where any observation is a result of the addition of a constant plus a set of terms accounting for each possible value each factor can take (named, levels) plus the effect of the interactions of factors and, finally, plus an experimental error (or unexplained variance) [32].

In particular, we are considering factors and interactions with the following levels:

- (Idle) timeout: That is, the maximum time a flow remains in the active-flow table without receiving a new packet. We have considered three

6

possible levels: 15, 60 and 120 seconds. Note that timeout is 15 seconds by default in most of the router vendors. However, some researchers have shown that this should be larger to attract human interactions as shown in [35], which estimates the optimal timeout in 120 seconds, and other studies that have assumed 60 seconds [36].

- Maximum flow-active time: Maximum time a flow can remain in the active-flow table, this value makes sense to avoid saturating routers with long tables. In our case, capacity was planned to suffice for all the analyzed traces, so a maximum was not fixed.

- (Layer-4) Protocol: TCP, UDP or other IP traffic.

- Day/Night: Daytime is considered to spans between 6 a.m. to 6 p.m. in local time.

- Time: Each of the 24 hours that a day spans is a level.

- Weekday/Weekend: Monday to Friday compared to Saturday-Sunday.

- Day of the week: Sunday to Saturday.

- Month: January to December.

- Year: 2002 to 2018.

- Day of the month: 1 to 31.

- Bandwidth: Average daily bandwidth. Defined as five levels, very low intensity (less than 50 Mb/s), low (50-200 Mb/s), middle (200-1000 Mb/s), high (1-3 Gb/s) and very high (more than 3 Gb/s).

- Network: The specific network where measurements were gathered. Note that this factor specifically accounts for the particularities of a given network, not the generalities we are searching.

- Significant 2-way interactions of these factors.

We note that our objective is to provide network managers with mechanisms to estimate flow ratio, not to model the specific network behaviors from which data was gathered. This way, we aim to extract the commonalities of the flow ratio metric across different networks in diverse scenarios to find general patterns. This implies that any factor related to a particular network is of less interest than a more general one or simpler factors. For example, a network that, for any reason, presents a flow ratio larger than usual at 9 p.m. provides no information for network managers on the Internet. This phenomenon can be likely explained for a particular heavy-hitter user of such network or other uncharacterized artifacts [37] that can obscure general conclusions. As another example, the USC ANT project has probed the entire IPv4 space of Wide [38] for years generating an artificial amount of ICMP traffic. Such traffic lacks interest for the general characterization of Internet behavior. This is way, a good practice is not to consider measurements during outages, very low-activity and abnormal-operation periods.

On the other hand, variance explained by protocols in a rich set of networks are pieces of behavior that can be extrapolated to other scenarios. This way,

Table 2: ANOVA table

Dependent variable: Flow ratio

Factor	DF	Sum of Squares	% Variance	Mean Square	F	p-value
Timeout	2	109577	13.07	54789	1107428	0.00
Protocol	2	521169	62.18	260585	5267120	0.00
Day/Night	1	231	0.03	231	4675	0.00
Time	23	1920	0.23	83	1688	0.00
Weekday/Weekend	1	700	0.08	700	14156	0.00
Day of the week	5	2597	0.31	519	10499	0.00
Month	11	5040	0.60	458	9261	0.00
Year	12	15434	1.84	1286	25997	0.00
Day of the month	30	4168	0.50	139	2808	0.00
Bandwidth	4	20117	2.40	5029	101654	0.00
Network	6	10848	1.29	1808	36546	0.00
Protocol:Timeout	4	3947	0.47	987	19943	0.00
Protocol:Weekday/Weekend	2	962	0.11	481	9720	0.00
Protocol:Time	46	6517	0.78	142	2864	0.00
Protocol:Month	22	10147	1.21	461	9323	0.00
Protocol:Year	24	22638	2.70	943	19066	0.00
Protocol:Network	8	3661	0.44	458	9249	0.00
Network:Timeout	16	705	0.08	44	890	0.00
Network:Day/Night	4	2438	0.29	609	12319	0.00
Network:Time	85	5941	0.71	70	1413	0.00
Network:Weekday/Weekend	3	701	0.08	234	4723	0.00
Network:Day of the week	22	1046	0.12	48	961	0.00
Network:Month	18	774	0.09	43	869	0.00
Network:Year	6	348	0.04	58	1174	0.00
Network:Day of the month	39	611	0.07	16	317	0.00
Network:Bandwidth	5	19379	2.31	3876	78342	0.00
Residuals	1343868	66486	7.93			
Total		838104	100.00			

Adjusted R^2=0.91

how UDP's flow ratio behaves in contrast to TCP traffic can be used for other managers, simply, by rescaling by the proportion of UDP/TCP traffic of their networks.

In this light, we propose to follow an ANOVA iterative approach whereby the most general factors are first used to explain variance, and the most specific ones are considered progressively (often named as ANOVA Type I). In practical terms, such top-down approach means that well-known characteristics as transport-layer protocols, configurable timeouts, or the time and day when measurements were gathered are preferred to two-ways factors (e.g., Protocol:Month) or those related to particularities of a given network (e.g., Network or interaction of Network with other factors).

Table 2 shows the results after applying ANOVA in the data. Several observations arise.

- All groups pass the correlation test [30] for assuming a Gaussian distribution, after the logarithmic transformation, with coefficient values over

0.95.

- All of the considered factors and a number of their interactions are quantitatively significant, i.e., p-values are less than 0.05 as the last column of the table shows.

- The fourth column of the table shows the normalization of the variance that each factor explains (Sum of Squares) by the total variance of the phenomenon. In other words, the quantitative relevance of each factor. The results show that Timeout and Protocol are able to explain more than 75% of the variance, which may represent a reasonable error-bound for the modeling of the phenomenon. After them, only Bandwidth accounts for a certain percentage.

- Figure 2 shows the number of concurrent flows per protocol, timeout and certain year, in logarithmic scale in the Y-axis. In particular, the effect of Timeout translates into smaller ratios for smaller timeouts. For example, a timeout of 15 seconds may give 5 times less concurrent flow per Mb than larger timeouts. Regarding protocols, TCP tends to present more than 10 times less concurrent flows in the same bandwidth than UDP and even less for other protocols.

- Factors such as day/night and weekend/weekday have turned out to be of small significance. Note that this does not mean that the traffic volume during weekdays or during daytime is equal to the traffic during weekends or at night, but that variances of the flow ratio can only be marginally explained by these factors once applied the previous ones. For completeness, we have found that weekends tend to have larger ratios.

- The factor Year and, especially, 2-way-factor Protocol:Year show some relevance. By visually inspecting the flow ratio variations over time (the last boxplot of Figure 2) a clear tendency does not appear. The boxplot suggests that during the last years under study, i.e., 2008-2018, the ratio has behaved as a stationary process.

  Anyhow, in practical terms, a network manager should be more interested in the most recent figure (e.g., 2014-2018), not in the historical values of the ratio.

- The factor Bandwidth accounts for about 2.4% of the variance, and Network:Bandwidth an additional 2.31%. After a more detailed inspection, we found a pattern whereby those moments with less aggregate traffic showed a higher flow ratio. This makes sense as larger traffic aggregates make bandwidth time series smoother and less sensitive to spurious behavior, i.e., a heavy-hitter user whose particular traffic patterns exerts a significant impact on the aggregated metrics. However, the Network:Bandwidth factor points in another direction. It means that the bandwidth may affect differentially to the set of networks.

- The rest of the factors show results, quantitatively, little significant, or, intuitively, little useful for network management. Interestingly, the factors related to a particular network give explanation to only 1.29% of total variance. This marginal amount of percentage points suggests that,
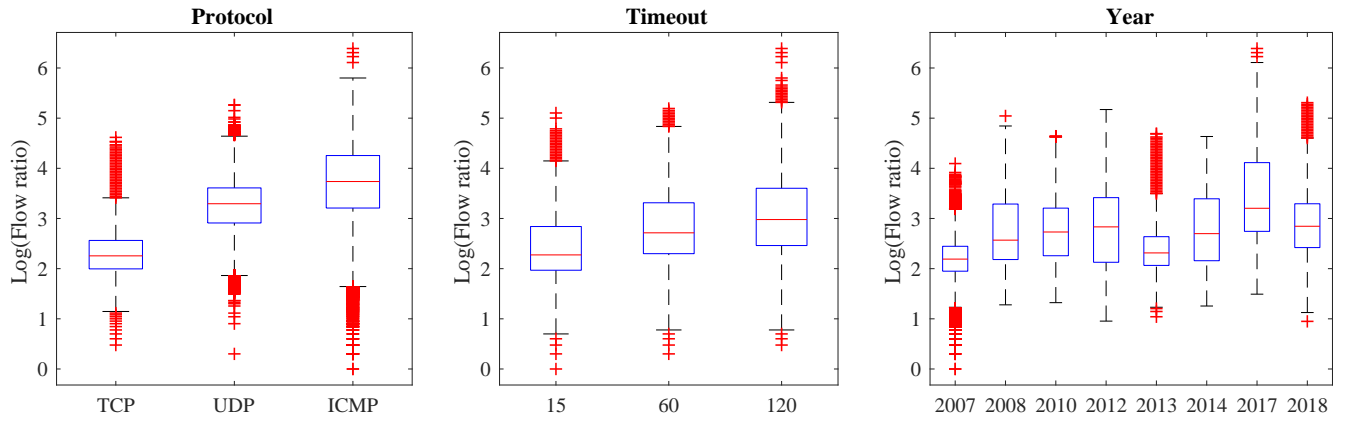
Figure 2: Boxplots for flow ratio per factors Protocol, Timeout and Year
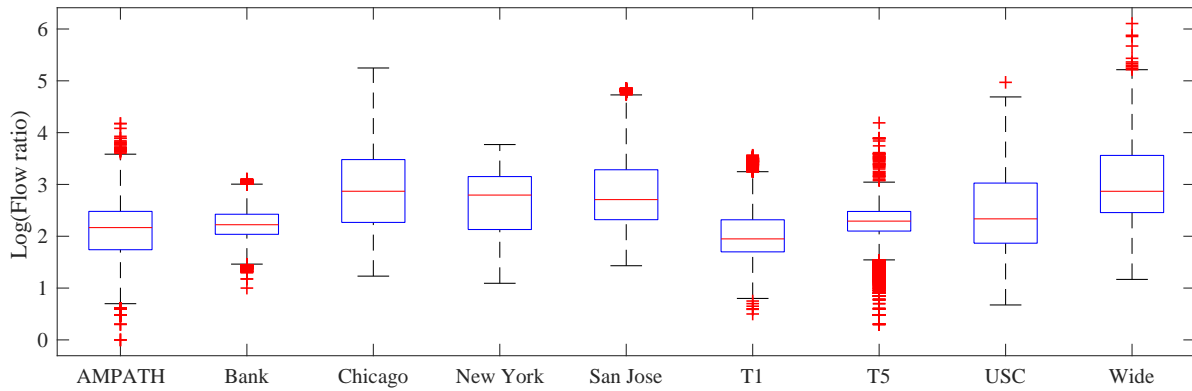

Figure 3: Boxplots for flow ratio per network under study

once other factors were applied, the set of networks behave fairly homogeneously. Then, the characterization of the phenomenon as a general process makes sense. For the sake of completeness, Figure 3 illustrates the differences between networks. Equinix's links tend to have a larger flow ratio than the other networks but with no significant difference with respect to Wide for example.

- Finally, the factor Residuals represents the unexplained variance. In particular, it accounts for less than 8%.

All of this leads us to propose to take into consideration the factors Timeout and Protocol to estimate flow ratios as a good trade-off between simplicity and coverage (i.e., explained variance over 75% in a strong model with R^2 over 0.9). The final aim is to provide network managers, who know the timeout in use on their networks as well as some estimation of the protocol share, a rule such that flow ratio can be estimated.

10

4 Gaussian modeling

Given that the previous section shows Gaussian patterns in the flow ratio, we hypothesize that the flow ratio $\mathbf{R}$ can be modeled as a Gaussian multivariate random variable, i.e.:

$$\mathbf{R} = (R_1, \ldots, R_N) \sim \mathbf{N}(\mu, \mathbf{\Sigma})$$

being each $R_i, i = 1, \ldots, N$ the contribution of each factor $1, \ldots, N$ to the overall flow ratio $\sum_{i=1}^{N} X_i$, whereby μ is the vector of means and $\mathbf{\Sigma}$ is the covariance matrix $[\sigma_{ij}], i = 1 \ldots N, j = 1, \ldots, N$. To do so, the dependence of the flow ratio with factors must be assessed.

4.1 Testing the null hypothesis of independence between factors

In order to derive $\mathbf{\Sigma}$, we pose the null hypothesis of independence between factors. Let $\hat{\rho_{ij}}$ be the estimated correlation coefficient between factor i and j, namely $\hat{\rho_{ij}} = s_{ij}/\sqrt{s_{ii}s_{jj}}$, being s_{ij} the quasivariance of factor i and j. As it is well-known [39], the random variable

$$\zeta = \frac{1}{2}log\left(\frac{1 + \hat{\rho_{ij}}}{1 - \hat{\rho_{ij}}}\right) \tag{1}$$

follows a Fishers' Zeta distribution and if

$$z = \frac{1}{2}log\left(\frac{1 + \rho_{ij}}{1 - \rho_{ij}}\right) \tag{2}$$

then $\sqrt{M - 1}(\zeta - z)$ converges in distribution to the standard Gaussian distribution, being M the sample size. Thus, for large values of M, ζ can be approximated by a Gaussian random variable with mean

$$z + \frac{\rho_{ij}}{2(M - 1)} \tag{3}$$

and variance $\frac{1}{M-3}$.

As a result, testing the null hypothesis $\rho_{ij} = \rho_0$ at the significance level α can be performed with the following critical region

$$\sqrt{M - 3}\left|z - z_0 - \frac{\rho_0}{2(M - 1)}\right| > \lambda_{\frac{\alpha}{2}} \tag{4}$$

where $\lambda_{\frac{\alpha}{2}}$ is the $1 - \frac{\alpha}{2}$ percentile of the standard Gaussian distribution.

In our case, $\rho_0 = 0$ and Equation 4 yields

$$|z| > \frac{\lambda_{\frac{\alpha}{2}}}{\sqrt{M - 3}} \tag{5}$$

as the critical region to reject the null hypothesis.

After applying the test, it turns out that the hypothesis of independence between the factors Timeout and Protocol can be rejected at $\alpha = 0.95$. That means that there is evidence of dependence between both factors. In other words, that by changing timeouts the flow ratio varies differently per protocol.

The reason for this is that TCP flows are exported by not only timeouts but also when TCP FIN flag is found, conversely UDP and other IP traffic such ICMP are only exported by timeouts. By increasing the timeout, UDP and ICMP flows become longer, but the same does not apply, or at least to the same extent, to TCP. This suggests handling each timeout value separately. Note that, typically, network managers set the timeout value and it is used for all traffic for an indeterminate time. Although, changes are possible in adaptive sampling [40].

Then, we apply the test to the factor Protocol assuming timeouts fixed to a given value. In this case, no evidence of dependence was found in all the networks under study with the exception of USC trace. This means that, for example, changes in the flow ratio in TCP does not imply a change in the UDP ratio counterpart. The reason for this is that both ratios change according to the variations of the applications running over them, but not necessarily at the same time. For example, a network whose users are now allowed running P2P applications while such applications were banned previously. Likely, the use of P2P applications may change the typical ratio of TCP traffic but unlikely there would be changes in the ICMP traffic.

4.2 Implications

Given the results of the previous section, we propose to model flow ratio (FR) independently per timeout value, so given a timeout, each protocol behavior is modeled as:

$\mathbf{FR}(timeout, fractionTCP, fractionUDP, fractionOthers) =$

$fractionTCP \cdot \mathbf{N}(\mu_{tcp}^{timeout}, \sigma_{tcp}^{2timeout}) + fractionUDP \cdot \mathbf{N}(\mu_{udp}^{timeout}, \sigma_{udp}^{2timeout}) +$

$fractionOthers \cdot \mathbf{N}(\mu_{others}^{timeout}, \sigma_{others}^{2timeout})$

$$(6)$$

And given the evidence of independence of the flow ratio variance between protocols, the final flow ratio estimate turns out to be the weighted sum of three independent Gaussian random variables. That is:

$\mathbf{FR}(timeout, fractionTCP, fractionUDP, fractionOthers) =$

$\mathbf{N}\Big((fractionTCP \cdot \mu_{tcp} + fractionUDP \cdot \mu_{udp} + fractionICMP \cdot \mu_{icmp})^{timeout},$

$(fractionTCP^2 \cdot \sigma^2{}_{tcp} + fractionTCP^2 \cdot \sigma^2{}_{udp} + fractionTCP^2 \cdot \sigma^2{}_{icmp})^{timeout}\Big)$

$$(7)$$

Table 3 depicts the parameter estimates according to the factors Timeout and Protocols, and all traffic (* represents the three protocols and the three timeouts).

By comparing means for a given timeout, e.g., the 15-second row of the table, it becomes apparent that TCP traffic exhibits less concurrency by an order of magnitude for same traffic volume than UDP and even more for other traffic. It is also worth remarking that the variance of TCP and UDP are clearly smaller than the variance considering all the traffic (for a 15-second timeout, 0.11 and 0.08 in contrast to 0.46). That implies a significant reduction of the variability of the phenomenon, even more, when it is expected that TCP and somehow UDP dominate most of the traffic on the Internet.

The table also highlights the importance of considering the timeout. Inspecting the table by columns illustrates how estimates increase significantly. Finally, the table includes the estimates for 95th percentiles. It may represent a conservative bound for those network managers more concerned with extreme values.

Table 3: Parameter estimates according to the factors Timeout and Protocol

Protocols	TCP			UDP			Others			*		
Timeout	μ	σ^2	95th	μ	σ^2	95th	μ	σ^2	95th	μ	σ^2	95th
15	2.00	0.11	2.50	2.82	0.08	3.34	3.29	0.52	4.68	2.44	0.46	3.77
60	2.34	0.15	2.98	3.34	0.09	3.84	3.75	0.55	5.22	2.83	0.53	3.98
120	2.53	0.19	3.39	3.61	0.09	4.15	4.05	0.60	5.51	3.07	0.62	4.26
*	2.30	0.19	3.05	3.26	0.19	4.06	3.70	0.64	5.25	2.78	0.60	4.12

4.3 Numerical examples

Let us now evaluate how the modeling can help network managers using the set of traces under study as well as other measurements as a validation set, i.e., traces not used in the modeling process.

According to the general results of Figure 1, the worst-case scenario for a network should be more than 10,000 flows per Mb (i.e., flow ratio in the tail of distribution) with an expected average of some 600 flows per Mb. This applied to Wide network, assuming a typical bandwidth between 100 and 500 Mb/s, gives an estimate of between 60,000 and 300,000 active flows in average, and worst case between 1,000,000 and 5,000,000 flows.

Let us apply the factor analysis, for example for a timeout of 60 seconds, and with a typical protocol share of 92% TCP, 6% UDP and 2% ICMP for this network [41]. With this, and applying Equation 7 with the parameters of Table 3, flow ratio varies as a Gaussian distribution with mean 2.43 and variance 0.13. This yields an estimated average in the number of active flows between roughly 27,000 and 135,000 with a worst case in between 132,000 and 650,000 flows for bandwidths of 100 and 500 Mb/s respectively. We remark that the width of the intervals has changed dramatically. It is also worth remarking that the reduction of variation of the flow ratio phenomenon is even more significant in high-speed networks. In the case of Equinix, with aggregates over 5 Gb/s, a non-factorial approach may give an estimation of active flows between 3 and 64 million, whereas after applying factor estimation are between 1 and 6 million flows.

In general, the reduction can be illustrated by changes in the coefficient of variation (CV). The coefficient of variation for non-factorial approach is 0.26. Then, these figures reduce to 0.15 and 0.14 to Wide and Equinix networks respectively, more than 40%. Similarly, factorial approach in the rest of the networks gives CVs in the range from 0.14 to 0.16, as the protocol distributions are similar (although differences exist, e.g., the Bank network does not include ICMP traffic and UDP was marginal whereas Equinix networks' UDP traffic achieves typically figures over 5%).

Regarding the error that the chosen model is making, ANOVA stated that its explained variance was roughly 75% on the set under study; hence, we can

expect deviations in such a range of +/- 25%. In general, most of this error comes from the two Equinix networks. As pointed in Section 3, they showed higher ratios than the other networks. Note that the ANOVA table has already predicted that the factor Network has some impact. However, in our approach, we did not include it to make results general for other Internet users and not only for the managers of these networks. Coherently the modeling tends to underestimate these networks. Actually, the authors in [42] suggested that ISPs might present a higher number of concurrent flows due to the sort of traffic they carry in contrast to domestic or academic networks.

To illustrate this, Figure 4 shows the mean and 5/95th percentiles estimates after applying the modeling overlapped with the real number of concurrent flows measured at 5-minute grain in a day for Bank network, Wide and Equinix's San Jose link. Certainly, in all of the cases, the measurements fall between the 5/95th estimates intervals, and the average is a good predictor. Despite this, we found cases such as the day chosen for San Jose link (5th of May, 2012). This day has been chosen as the estimates presented the worst results for the full measurement campaign. However, note that we are searching for the typical behavior users may expect in their networks, not the exceptional behavior of one day like this one.
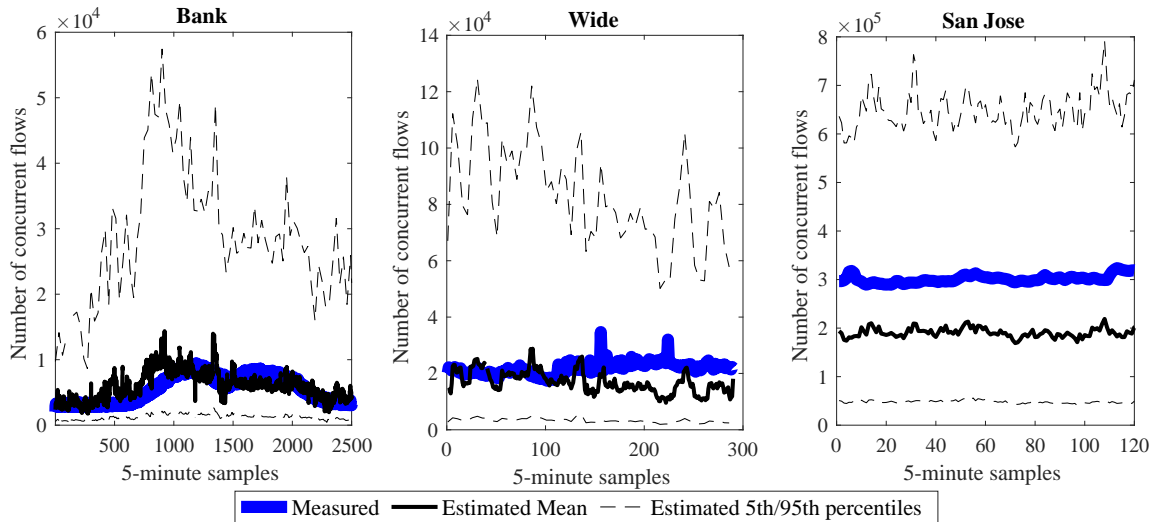

Figure 4: Examples of estimation (Mean and 5th/95th percentiles) for one day in Bank network, Wide, and Equinix's San Jose link in contrast to the measurements actually gathered

Let us now focus on the performance in other traces. In particular, we have used traces from the network of an ISP in Spain (ISP), from a research laboratory (RL) and a residential link given service to some thousands of users (RE). The results showed larger average errors for ISP trace up to 30%, lower ones for RL traces (10%), and equivalent errors for RE link (20%).

We believe these errors are reasonable for network managers that search for a guide not only over the non-factorial approach but also over the current lack of any foreknowledge about how many flows they can expect in traffic aggregates.

14

Table 4: ANOVA table including Application factor for Wide's measurements
Dependent variable: Flow ratio

Factor	DF	Sum of Squares	% Variance	Mean Square	F	p-value
Timeout	2	20938	10.20	10469	74317	0.00
Protocol	2	120479	58.72	60239	444526	0.00
Applications	5	10084	4.91	2017	5412	0.00
Residuals	224964	53662	26.16			
Total		205164	100			

Adjusted $\bar{R}^2$=0.829

5 Flow ratio with application mix

It becomes apparent that changes in the flow ratio may come from the different applications that clients use. For example, banning P2P applications in an academic network have a direct impact on the traffic of such a network. This way, network managers can have some knowledge of the popularity in term of applications of their users. Let us study how much variance this can help to explain.

Unfortunately, most of the publicly available traces of the Internet do not include payloads mainly due to privacy concerns. This makes the study of the factor Application a challenge. However, Wide network did share some of their traces with payloads for research purposes, and we have analyzed them. In particular, we have applied deep packet inspection (DPI) tools to classify flows per application, and then, recalculated the ANOVA table adding such Application factor. We have entrusted the task of classifying to l7filter tool [43]. Such tool uses ports and DPI techniques to detect more than 100 protocols, including HTTP, HTTPS, BitTorrent, email clients, among others.

The ANOVA results are shown in Table 4 where the factors Timeout and Protocols are first considered, and then Applications. As previously, the fourth column represents the percentage of explained variance, and, in particular, for Application factor it is below 5%. That means that Application explained 10 times less variance than, simply, the layer-4 protocol, while its estimation is clearly more complicated. We believe that this lack of a significant contribution of factor Application comes from the dominant share that HTTP/HTTPS exhibits, the few examples found for some of the applications, and the significant traffic volume that classifier was unable to label. In more detail, only applications/protocols HTTP/HTTPS, DNS, VoIP, and P2P achieved a significant number of flows. The rest of the flows were categorized as others or unknown when no application matched. Roughly, this latter category was slightly below 30%, so limiting the contribution of the factor. Finally, the HTTP and DNS represent the dominant protocols for TCP or UDP respectively, so their behaviors were largely captured by factor (layer-4) Protocol, and only the traffic from VoIP and P2P are resulting useful to explain some previously unexplained variance. Nevertheless, such variance, as noted previously, remains below 5%.

All these downsides, in addition to the fact that a priori estimation for application shares is more difficult than estimations for layer-4 popularity, suggest that the extra effort does not worth the price.

6 Related work

The study of the characteristics of the Internet flows started shortly after Netflow born given its applicability to a number of network tasks [44]. Since then, the research community has described a number of flows metric such as the flow size, flow duration, flow inter-arrival and bandwidth variation during flows life, among others in some Internet networks.

In particular, in [45] and [46], the Internet flows of one day in the Universities of Auckland and California at San Diego are studied, and further divided into categories according to size and lifetime. They are, in turn, related such categories to animals. This way, heavy flows are related to elephants and short flows are related to dragonflies among other analogies. They found a heavy-hitter distribution of flows in traffic aggregates, in such a way that a small fraction of flows tends to carry the most of the traffic, while a large fraction of flows are short and carry only some bytes. In general, they pointed out that this is likely related to the underneath application layer.

This latter point was further studied in [47] with traces from Los Nettos, a regional-area network in Los Angeles. In particular, the flows belonging to each of the above animal-related classes are linked to applications. In such a way, elephants are typical in P2P applications, whereas tortoises (long flows) are more common in DNS zone transfers, and web connections tend to be more related to cheetah and porcupine classes (i.e., high-rate and high-burstiness flows, respectively). These conclusions motivated the study included in Section 5 of this paper, where we showed how some information at application layer can help to model concurrency and flow ratios.

Specifically, focusing on the flow concurrence metric, the authors in [48] depicted 1-minute timescale series of active flows, bandwidth and packet rates in a university link in Korea according to layer-4 protocols. By visual inspection, it becomes apparent that the ratio between bandwidth and active flows is larger in UDP than in TCP, which is coherent with we have found on the Internet. They found that the rationale behind this is the existence of flash flows, i.e., short and light flows (or dragonflies and mice), and that such a phenomenon is more common in UDP.

Then, the authors in [42] analyzed the flow concurrence in the Swiss national research and education network for five years. They found a significant impact of idle timeouts in the concurrency in flows, which is coherent with our results and motivates to handle them independently. In addition to this, they showed that the number of concurrent flow increased non-linearly during their measurement campaign (2003-2011), which is not the same that the flow ratio. Similarly, the authors in [49] searched for correlation between time series of traffic rates in Mb/s, flows/s and packets/s in the Tsinghua University. Among them, the most relevant for our study is the correlation between the first two. They found a significant correlation between them, which intuitively means that the more traffic, the more concurrent flows coexist.

However, we note that all these increases in the number of concurrent flows were not normalized by the bandwidth. That is, we remark that in this paper we do not focus on the concurrence in absolute terms as these previous works did but on its bandwidth ratio. Coherently, if the use/capacity of the network or the number of users increase, it is expected that both the bandwidth and the number of concurrent flows increase in absolute terms, which explains why both

measurements are correlated over time. However, our point is to study the ratio between them and, particularly, in the most diverse set of scenarios possible, to provide planning rules for other networks on the Internet or, even, networks in deployment phase. Hence, these previous works do not provide a tool, link or model to help networks managers in capacity-planning tasks.

Actually, the authors in [2] have already pointed out that to determine when a router with Netflow capabilities must be upgraded, the number of active flows is a key metric. Moreover, they allude to excursions over the typical number of active flow as a relevant indication that something is not working fine. However, and although they performed some stress tests with real traffic, they did not shed light on which "typical" refers. We note that providing an intuition of how many concurrent flows one can expect in a network is essentially our target.

As another distinguishing characteristic, we propose to explain the diversity of the phenomenon by factors, and so modeling more precisely. Finally, the factor analysis will provide general results if applied in a diverse enough set of scenarios and for a long period of time, this way we have studied hundreds of traces during eight years which has not been so common in the literature.

7   Conclusion

This paper has studied the number of flows per Mb at one-second granularity that both network managers and practitioners may expect in a diverse set of scenarios as a useful tool to foresee load for flow-based systems. Such diverse set has allowed us to provide results as general as possible in an attempt to result applicable to other scenarios on the Internet.

We have found that the variance of flow ratio follows a lognormal distribution, which allows us to describe the process with parameters mean and variance as well as applying further analysis assuming normality. This way, after the initial results, where the estimates included all the measurements, suggested a large domain for the ratio, we proposed to improve the study by applying a factorial approach in a progressive fashion (from simpler to more particular factors). Between them, the factors Timeout and Protocols stood out as they explain a significant fraction of variance and keep the modeling simple. Finally, we have also studied the inclusion of some knowledge at application layer, however, this has not resulted in a significant contribution.

As a conclusion, we have modeled the metric from a top-down approach providing the research community with a model and their parameter estimates according to the chosen timeout and fractions of layer-4 protocols to be used in their own scenarios, given the generality of the measurements studied. The evaluation of the approach has demonstrated its effectiveness with errors ranging between 10% and 30%, reasonable figures compared to the current lack of any intuition about the expectable number of flows given a network or link to monitor.

As future work, we plan to pay attention to factors such as the size of the network, and, especially to the type of the network. To do so, several examples per type of network (e.g., backbone links) are needed but, unfortunately, this depends on the availability of more traces. Finally, another point to focus on is the potential impact of Google's QUIC protocol. Such a protocol is gaining popularity due to the omnipresence of Google's services and it constructs flows

at application layer.

Acknowledgment

References

References

[1] M. Molina, A. Chiosi, S. D'Antonio, G. Ventre, Design principles and algorithms for effective high-speed IP flow monitoring, Computer Communications 29 (10) (2006) 1653–1664.

[2] H. Zang, A. Nucci, Traffic monitor deployment in IP networks, Computer Networks 53 (14) (2009) 2491–2501.

[3] R. Hofstede, P. Ĉeleda, B. Trammell, I. Drago, R. Sadre, A. Sperotto, A. Pras, Flow monitoring explained: From packet capture to data analysis with NetFlow and IPFIX, IEEE Communications Surveys Tutorials 16 (4) (2014) 2037–2064.

[4] E. Miravalls-Sierra, D. Muelas, J. Ramos, J. E. López de Vergara, D. Morató, J. Aracil, Online detection of pathological TCP flows with retransmissions in high-speed networks, Computer Communications 127 (2018) 95–104.

[5] A. Feldmann, A. Greenberg, C. Lund, N. Reingold, J. Rexford, F. True, Deriving traffic demands for operational IP networks: methodology and experience, IEEE/ACM Transactions on Networking 9 (2001) 265 – 280.

[6] A. Sperotto, G. Schaffrath, R. Sadre, C. Morariu, A. Pras, B. Stiller, An overview of IP flow-based intrusion detection, IEEE Communications Surveys Tutorials 12 (3) (2010) 343–356.

[7] A. Aguiar Amaral, L. de Souza Mendes, B. Bogaz Zarpelã, M. Lemes Proença Junior, Deep IP flow inspection to detect beyond network anomalies, Computer Communications 98 (2017) 80–96.

[8] A. Callado, C. Kamienski, G. Szabo, B. P. Gero, J. Kelner, S. Fernandes, D. Sadok, A survey on Internet traffic identification, IEEE Communications Surveys Tutorials 11 (3) (2009) 37–52.

[9] V. Carela-Español, P. Barlet-Ros, A. Cabellos-Aparicio, J. Sol-Pareta, Analysis of the impact of sampling on Netflow traffic classification, Computer Networks 55 (5) (2011) 1083–1099.

[10] A. Monemi, R. Zarei, M. N. Marsono, Online NetFPGA decision tree statistical traffic classifier, Computer Communications 36 (12) (2013) 1329–1340.

[11] C. Fraleigh, S. Moon, B. Lyles, C. Cotton, M. Khan, D. Moll, R. Rockell, T. Seely, C. Diot, Packet-level traffic measurements from the Sprint IP backbone, IEEE Network 17 (6) (2003) 6–16.

[12] J. L. García-Dorado, J. A. Hernández, J. Aracil, J. E. López de Vergara, F. J. Montserrat, E. Robles, T. P. de Miguel, On the duration and spatial characteristics of Internet traffic measurement experiments, IEEE Communications Magazine 46 (11) (2008) 148–155.

[13] K.-C. Lan, J. Heidemann, A measurement study of correlations of Internet flow characteristics, Computer Networks 50 (1) (2006) 46–62.

[14] S. Floyd, E. Kohler, Internet research needs better models, ACM Computer Communication Review 33 (1) (2003) 29–34.

[15] W. Wanrooij, A. Pras, Data on retention, in: IFIP/IEEE International Workshop on Distributed Systems: Operations and Management, 2005, pp. 60–71.

[16] R. Sommer, A. Feldmann, Netflow: information loss or win?, in: Proceedings of ACM SIGCOMM Workshop on Internet Measurment, 2002, pp. 173–174.

[17] L. Avramov, M. Portolani, The policy driven data center with ACI: architecture, concepts, and methodology, Cisco Press, 2014.

[18] CISCO, Netflow services and appications, http://www.cisco.com/en/US/products/ps6601/prod_white_papers_list.html.

[19] B. Trammell, E. Boschi, An introduction to IP flow information export (IPFIX), IEEE Communications Magazine 49 (4) (2011) 89–95.

[20] V. Moreno, P. M. Santiago del Río, J. Ramos, D. Muelas, J. L. García-Dorado, F. J. Gomez-Arribas, J. Aracil, Multi-granular, multi-purpose and multi-Gb/s monitoring on off-the-shelf systems, International Journal of Network Management 24 (4) (2014) 221–234.

[21] C. Estan, G. Varghese, New directions in traffic measurement and accounting: Focusing on the elephants, ignoring the mice, ACM Transactions on Computer Systems 21 (3) (2003) 270–313.

[22] K. Claffy, Ten things lawyers should know about the Internet, http://www.caida.org/publications/papers/2008/lawyers_top_ten.

[23] AMPATH, Anonymized Internet traces, https://ampath.net/.

[24] HPCN, High performance computing and networking, http://www.hpcn-uam.es.

[25] CAIDA: Center for Applied Internet Data Analysis, Anonymized Internet traces, https://www.caida.org/data/.

[26] R. Barbosa, R. Sadre, A. Pras, R. van de Meent, Simpleweb/University of Twente Traffic Traces Data Repository, no. TR-CTIT-10-19 in CTIT Technical Report Series, Centre for Telematics and Information Technology (CTIT), 2010.

[27] USC/LANDER project, Traces from day in the life of the internet, `http://www.isi.edu/ant/lander`.

[28] MAWI Working Group traffic archive, Wide project, `http://mawi.wide.ad.jp/mawi/`.

[29] S. Floyd, V. Paxson, Difficulties in simulating the Internet, IEEE/ACM Transactions on Networking 9 (4) (2001) 392–403.

[30] R. Van De Meent, M. Mandjes, A. Pras, Gaussian traffic everywhere?, in: IEEE International Conference on Communications, 2006, pp. 573–578.

[31] O. J. Dunn, V. A. Clark, Applied Satistics: Analysis of Variance and Regression, New York: John Wiley and Sons Inc., 1974.

[32] J. L. García-Dorado, Bandwidth measurements within the cloud: Characterizing regular behaviors and correlating downtimes, ACM Transactions on Internet Technology 17 (4).

[33] M. R. Martínez-Torres, S. L. Toral, B. Palacios, F. Barrero, Web site structure mining using social network analysis, Internet Research 21 (2) (2011) 104–123.

[34] D. Parra, C. Trattner, D. Gómez, M. Hurtado, X. Wen, Y.-R. Lin, Twitter in academic events: A study of temporal usage, communication, sentimental and topical patterns in 16 computer science conferences, Computer Communications 73 (2016) 301–314.

[35] A. Bianco, G. Mardente, M. Mellia, M. Munafò, L. Muscariello, Web user-session inference by means of clustering techniques, IEEE/ACM Transactions on Networking 17 (2) (2009) 405 –416.

[36] Y.-s. Lim, H.-c. Kim, J. Jeong, C.-k. Kim, T. T. Kwon, Y. Choi, Internet traffic classification demystified: on the sources of the discriminative power, in: Proceedings of CoNEXT, 2010, pp. 9:1–9:12.

[37] Í. Cunha, F. Silveira, R. Oliveira, R. Teixeira, C. Diot, Uncovering artifacts of flow measurement tools, in: Passive and Active Measurement Conference, 2009, pp. 187–196.

[38] The ANT lab: Analysis of network traffic, `https://ant.isi.edu`.

[39] T. W. Anderson, An Introduction to multivariate statistical analysis, John Wiley, 1958.

[40] C. Estan, K. Keys, D. Moore, G. Varghese, Building a better Netflow, in: Proceedings of SIGCOMM, 2004, pp. 245–256.

[41] MAWI Working Group traffic archive, traffic trace info for wed apr 8 2015, `http://mawi.wide.ad.jp/mawi/samplepoint-F/2015/201504081400.html`.

[42] B. Trammell, D. Schatzmann, On flow concurrency in the Internet and its implications for capacity sharing, in: ACM Workshop on Capacity Sharing, 2012, pp. 15–20.

[43] Linux layer 7 packet classifier, `https://sourceforge.net/projects/l7-filter/`.

[44] B. Li, J. Springer, G. Bebis, M. Hadi Gunes, Review: A survey of network flow applications, Journal of Network and Computer Applications 36 (2) (2013) 567–581.

[45] N. Brownlee, K. Claffy, Understanding Internet traffic streams: dragonflies and tortoises, IEEE Communications Magazine 40 (10) (2002) 110–117.

[46] A. Soule, K. Salamatia, N. Taft, R. Emilion, K. Papagiannaki, Flow classification by histograms: Or how to go on safari in the Internet, ACM Performance Evaluation Review 32 (1) (2004) 49–60.

[47] K.-C. Lan, J. Heidemann, A measurement study of correlations of Internet flow characteristics, Computer Networks 50 (1) (2006) 46–62.

[48] M.-S. Kim, Y. J. Won, J. W. Hong, Characteristic analysis of Internet traffic from the perspective of flows, Computer Communications 29 (10) (2006) 1639–1652.

[49] J. H. Wang, C. An, J. Yang, A study of traffic, user behavior and pricing policies in a large campus network, Computer Communications 34 (16) (2011) 1922–1931.