Internet Research 22.1 (2012): 29 – 56

DOI: http://dx.doi.org/10.1108/10662241211199951

Factor analysis of Internet traffic destinations from similar source networks

July 12, 2011

Abstract

Purpose – This study aims to assess whether similar user populations in the Internet produce similar geographical traffic destination patterns on a per-country basis.

Design/methodology/approach – We have collected a country-wide NetFlow trace, which encompasses the whole Spanish academic network, which comprises more than 350 institutions and one million users, during four months. Such trace comprises several similar campus networks in terms of population size and structure. To compare their behaviors, we propose a mixture model, which is primarily based on the Zipf-Mandelbrot power law to capture the heavy-tailed nature of the per-country traffic distribution. Then, factor analysis is performed to understand the relation between the response variable, number of bytes or packets per day, with dependent variables such as the source IP network, traffic direction, and country.

Findings – Surprisingly, the results show that the geographical distribution is strongly dependent on the source IP network. Furthermore, even though there are thousands of users in a typical campus network, it turns out that the aggregation level which is required to observe a stable geographical pattern is even larger. Consequently, our results show a slow convergence rate to the domain of attraction of the model, specifically, we have found that at least 35 days worth of data are necessary to reach stability of the model's estimated parameters.

Practical implications – Based on these findings, conclusions drawn for one network cannot be directly extrapolated to different ones. Therefore, ISPs' traffic measurement campaigns should include an extensive set of networks to cope with the space diversity, and also encompass a significant period of time due to the large transient time.

Originality/value – Current state of the art includes some analysis of geographical patterns, but not comparisons between networks with similar populations. Such comparison can be useful for the design of Content Distribution Networks and the cost-optimization of peering agreements.

Keywords Factor analysis, Geographical characterization, Heavy-hitters, Internet remote host location, Internet research, Zipf-Mandelbrot

Paper type Research paper

1 Introduction

The geographical characteristics of Internet traffic have a major impact on a wide range of applications, such as traffic engineering (Wasem et al., 1995), distributed network monitoring (Hofstede and Fioreze, 2009; Puzis et al., 2008; Bass, 2000), peering agreements with other ISPs (Lippert and Spagnolo, 2008; Laffont et al., 2001), or the design of Content Distribution Networks (CDN) (Erman et al., 2009; Qureshi et al., 2009). Consequently, knowledge of geographical traffic patterns has proven to be useful to perform routing and capacity planning in the Internet, even though routing

is mostly driven by economic and political decisions (Weis, 2010; Schwartz, 2010). However, once a routing and capacity planning decision is made, to which extent *can such decisions be extrapolated to other scenarios?* Moreover, regardless of the reasons that motivated a certain routing scheme in a network, is it possible to assure that adding new users, when the aggregation level is already very large, will not alter significantly the traffic destination distribution? In addition, an ISP may wish to know more about the geographical traffic pattern of its customers. To do so, a trace driven analysis must be performed. How long, in days worth of traffic, should it last?

It is also worth noting that two of the top questions in the CAIDA's Day in the Life of the Internet project (CAIDA, 2009) were precisely: "What are the traffic patterns and connectivity in different geographic regions?" And "for ISPs appearing in different geographic regions around the world, do peering relationships change depending on the location?"

In this paper, we address the above questions by studying whether similar user populations produce the same geographical traffic pattern on a per-country basis. To do so, we have performed a spatial and temporal diversity analysis of the geographical traffic pattern of destinations from different source IP networks, which share similar characteristics (population size, access link capacity, etc.). With respect to previous works, we provide a different approach by comparing the geographical traffic pattern of *similar user populations*. More specifically, we perform a country-wide measurement campaign that comprises the whole Spanish academic network. Then, we analyze the geographical traffic pattern per IP campus network and focus on whether similar campus networks provide the same per-country geographical traffic pattern or not.

Our findings can be used in any capacity planning or routing problem for which knowledge of the destination pattern is important. For example, let us assume that an ISP starts business with a population base which is similar to an existing population. Then, is the traffic destination pattern different, even though the customer population looks similar? Note that if the destination pattern differs so does the routing strategy (Subramanian et al., 2002) and possible peering agreements from the ISP (Norton, 2001a,b).

Furthermore, our study has direct application to the efficient design of content distribution mechanisms. The authors in (Erman et al., 2009) point out that a fundamental consideration for the performance of such mechanisms is the distance the data travels to reach the end user. From the user's point of view, increased travel distances affect the load time of any resource, such as web pages and file downloads, reducing throughput. In addition, it exerts a strong impact on the Quality of Experience (QoE) of multimedia applications, such as real-time video streaming and on-line gaming. From ISPs' viewpoint, the network miles data travel reflect the direct cost of transmitting data over their backbones. In this light, shorter distances entail lower costs, and conversely, larger distances imply an increase in the expenses. Consequently, knowledge of the geographical destinations is a key metric to dimension CDN or proxies. An ISP can decide what content to cache or where to deploy a content server taking into account, on the one hand, *the popularity of the destinations*, and on the other hand, the cost of delivering in "air miles", as defined in (Erman et al., 2009). We note that not only the geographical characterization is important when performing this task, but also the comparison between similar user populations. Such comparison is useful to assess to what extent the results can be extrapolated to other scenarios. Namely, our results serve to evaluate whether the same content distribution policy can be applied to similar populations, both in size and structure.

The rest of the paper is organized as follows: next sections presents a summary of our research objectives and a review of the related work. Then, Section 4 details the measurement set. Section 5 describes the methodology applied in Section 6, which is devoted to the results, and Section 7 comprises the discussion. Finally, Section 8 concludes the paper with a summary of the main findings and the future research directions.

2 Research Objectives and Research Design

The main research objective of this study is to perform an analysis of the space and time diversity (García-Dorado et al., 2008) of the geographical traffic patterns. To this end, *we have modeled the geographical distribution of Internet traffic on a per-country basis*, with connections both originated in or destined to a university in Spain. Specifically, we obtained that such distribution can be effectively modeled by means of a power law model, namely the Zipf-Mandelbrot (ZM) distribution. For each of the universities, we estimate the parameters of the model, and evaluate its accuracy through χ^2 goodness-of-fit tests. In addition to validating the model, *we have estimated the amount of measurement time which is required to reach stability in the model parameters.* This stability provides hints about the trace length which is required to obtain meaningful measurement results, i.e., avoiding the characterization of a particular behavior of the network in an arbitrary time frame. Our results show that at least 35 days worth of data are necessary to reach stability of the estimated parameters. Consequently, shorter measurements campaigns may result in misleading conclusions. Finally, *we have analyzed the space diversity issue*, i.e., whether similar source IP networks produce similar geographical traffic patterns. Namely, we would like to know to which extent there are invariants (Floyd and Paxson, 2001) in geographical traffic patterns, when the aggregation level is large. Hence, we could predict a routing behavior in a new population and update a network accordingly when new users are added. We performed a factor analysis to further explore this question. Specifically, we have adopted ANOVA (ANalysis Of VAriance) methodology to explain how the direction (incoming or outgoing), country and source IP network affect the geographical distribution of traffic. The results showed that the traffic distribution per country heavily depends on the source IP network, despite of the large number of users.

Our study is limited by the fact that we are considering academic users, who are different from residential users. However, we do not only pursue the characterization of the geographical traffic pattern, but to which extent it is homogeneous if the user populations are alike. The methodologies presented in this paper can also be applied to the case of residential networks, and provide valuable insight for a residential network operator.

3 Related Work

Despite of the importance of factor analysis of traffic destinations, as shown in the previous sections, it turns out that the state of the art does not feature any similar study. We believe that such lack of research effort is due to the difficulties in capturing contemporary traffic from many geographically disperse source IP networks. For instance, more than ten years ago, the authors in (Arlitt and Williamson, 1997) presented a detailed workload characterization study of Internet web servers, on attempts to find invariants (Floyd and Paxson, 2001) in the Internet behavior. One of the analyzed characteristics was the geographical distribution of document requests to several web servers. However, they only considered two possible options, whether the requests were local or remote to the web-server network, finding that most part of the requests were remote. In our case, we discriminate the per-country traffic distribution, and do not restrict the analysis to the web service only.

Similarly in (Feng et al., 2005), the authors analyzed the traffic received by a certain online-game server. Among other characteristics, the players' location is included in the study. Their results indicate a clear geographical dispersion with only 30% of the clients placed close to the online-game server. Again, the authors focus on the online-game service exclusively, i.e., the remaining of the traffic is not considered, and there is no comparison on a per-source IP network basis.

The authors in (Zink et al., 2009) compared the global popularity of YouTube videos obtained from the YouTube web portal and the video popularity of a campus network. Essentially, the

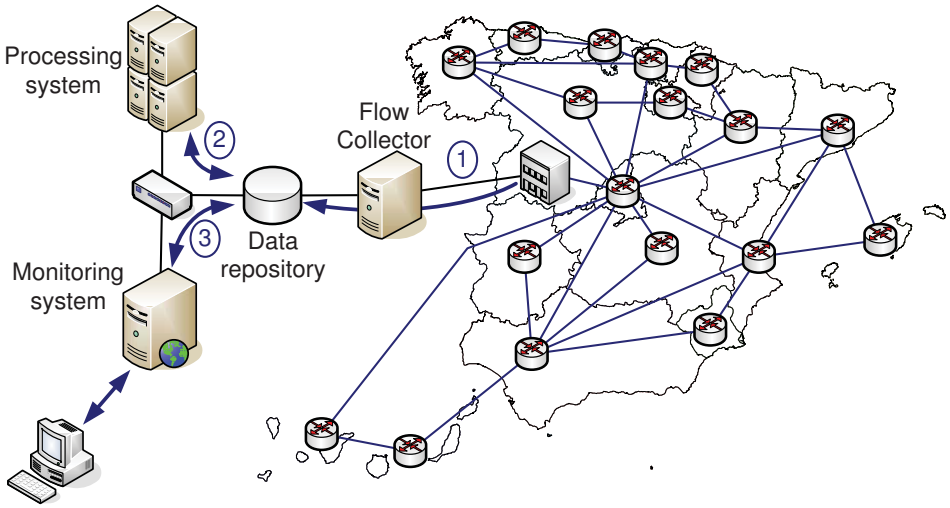


Figure 1: Measurement system architecture (left) and RedIRIS network topology (right)

authors analyzed if results from the entire Internet could be useful to make YouTube video caching decisions in a single campus. Given that they did not find significant correlation between both measurements, the answer was negative. Note that the users' profiles deeply varies according to life habits, different attitude towards technology, and other intangible cultural phenomena making very difficult to find any homogeneity by comparing Internet users. In this light, this paper goes one step further and aims to compare not a university and the Internet but an extensive set of similar IP subnetworks. In addition, we are not limited to any application and we have gathered traffic measurements for several months.

4 Measurement Set Description

The measurements available for this study have been kindly donated by RedIRIS (the Spanish National Research and Education Network, NREN) (RedIRIS, 2011) for research purposes[1]. The Spanish NREN comprises more than 350 institutions, mainly universities and research centers. It is connected to the rest of the Internet through commercial Internet exchange points (Telia, Level3, Cogent, etc.) and with the European Research and Education Network, GEANT. RedIRIS comprises 18 Points of Presence (PoP) across the country. Figure 1 shows the measurement system architecture and the RedIRIS network topology.

Since April 2007 to the present, RedIRIS is providing us with Netflow (Pras et al., 2009) records from each PoP. Essentially, Netflow provides summary records of IP flows traversing a PoP, which typically include the values of IP addresses and port numbers (source and destination), bytes transferred, flow start and finish times, and protocol. RedIRIS' Netflow records are being stored and processed in a central repository, located at Universidad Autónoma de Madrid premises. In the processing subsystem, these records are upgraded with geolocation information, i.e., the country to which each IP address belongs. The geolocation methodology will be described in the next section.

To make this information more manageable, we have computed daily aggregates of the num-

[1]The data is stored in isolated servers and never treated at the individual flow level, in full compliance with the Spanish regulation concerning privacy of electronic communications

Table 1: Measurement set summary

Field	Description
Source	University network/Country
Destination	Country/University network
Direction	Outgoing/Incoming
Bytes	Total number of bytes transferred from the source to the destination that day
Packets	Total number of packets transferred from the source to the destination that day
Percentage of bytes	Percentage of the bytes transferred from the source to the destination that day
Percentage of packets	Percentage of packets transferred from the source to the destination that day

ber of bytes and packets (and their corresponding percentages over the total of the day), in {university,country,direction} triples. Namely, for each day, we obtain the number of bytes and packets, and their corresponding percentages, per source IP network to each country, i.e., for the outgoing direction (from campus to the rest of the Internet). We also obtain the same information for traffic sourced in each country and destined to each campus network, i.e., for the incoming direction (from the rest of the Internet to campus). In what follows, we use the term "measured items" to refer generically to bytes or packets. The measurement set entries are presented in Table 1.

Following the methodology presented in (García-Dorado et al., 2008), we have carefully selected 12 universities out of the total set, for which the intrinsic network features, such as population size, bandwidth capacity, ratio students-staff, filtering policies (basically P2P applications), NAT capabilities and local cache/CDN mechanisms are very much alike. Table 2 provides some useful information about the selected universities, which are renamed to U_1, U_2, ..., U_{12} due to privacy concerns. This set of 12 universities is the largest set such that all universities share the mentioned features, specifically we remark that the use of local cache/CDN mechanisms is negligible and P2P applications are not banned.

In addition, it is worth noting that neither Network Address Translation (NAT) capabilities nor proxies have an influence on our measurements. NAT groups the traffic of several different hosts in a single public IP address. Nonetheless, this has no influence in the geographical location of hosts, neither remote nor local. Clearly, local proxies inside the campus network do not have any influence in the results. They are accounted for as local hosts, that concentrate traffic, in the incoming traffic measurements. However, remote proxies will be accounted for as end-hosts instead of the real end-hosts. Nevertheless, from an ISP standpoint, the traffic destination is the remote proxy not the real end-host.

We note that the sampling rate for Netflow records is the same throughout the measured routers, namely 1/100. We believe that the sampling error affects all measured campuses the same way and has no influence in our obtained percentages. Anyway, such sampling effect can be considered negligible for our analysis as shown in (Mai et al., 2006). On the other hand, the measured routers are configured differently with regard to the definition of flow. Namely, the maximum flow duration and inter-packet time are set to different values. Consequently, the per-flow analysis may

Table 2: User-base population size and networks' bandwidth capacity

University	Population	Ratio students/staff	Capacity access
U_1	38,000	8	1 Gb/s
U_1	50,000	11.2	1 Gb/s
U_3	38,000	11.1	1 Gb/s
U_4	46,000	8.8	1 Gb/s
U_5	31,500	10.3	1 Gb/s
U_6	40,000	9.6	1 Gb/s
U_7	33,500	11.0	1 Gb/s
U_8	38,500	8.6	1 Gb/s
U_9	31,000	10.8	1 Gb/s
U_{10}	31,000	8.9	1 Gb/s
U_{11}	36,000	12.2	1 Gb/s
U_{12}	30,000	11.7	1 Gb/s

be confusing and it is not included in this work.

Finally, we have eliminated weekends and holidays from the traffic sample. The behavior of academic networks during weekends is significantly different from weekdays. Weekends have a nearly flat underutilized daily traffic pattern, having minor impact for routing and capacity planning. However, it is worth remarking that we obtained equivalent conclusions when including weekends and holidays in the sample. This is consequence of the negligible volume of traffic during the weekends compared to the weekdays.

5 Methodologies

This section provides a brief description of the geolocation and statistical methodologies used in this study.

5.1 IP Geolocation Methodology

There are several ways to find the physical location of an IP address. The most straightforward approach is to use a name resolver and make a DNS reverse query, whereby the address location is obtained by parsing the retrieved name. A more accurate option is the database approach. In this study, we have used the free version of the *GeoIP Country* database of *MaxMind*, i.e., *GeoLite Country*, which has an accuracy of 99.5% as reported by the company (MaxMind, 2011) and outperforms other approaches (Poese et al., 2011). Such database has entries for the country code, country name and continent data. The shortcomings of this approach have been studied and reported (Gueye et al., 2007; Siwpersad et al., 2008), but anyway they seem adequate for our purposes, i.e., to perform grouping of destinations per country. Recently, there have been attempts to increase such accuracy (cf. (Padmanabhan and Subramanian, 2001; Gueye et al., 2006)). We have discarded these methods because the database approach is simpler and we do not need higher accuracy than country level. For a better understanding of geolocation procedures, the reader is referred to (Crovella and Krishnamurthy, 2006, Section 5.3.6) and references therein.

5.2 Statistical Methodologies

In this section, we introduce the statistical techniques applied in the paper. First, we present goodness-of-fit techniques that allow us to verify the traffic destinations geographical distribution. Second, we give a brief introduction to the ANOVA methodology, which allows us to measure the impact that factors such as the source IP network, country and direction have on the response variable, in this case, the measured number of bytes and packets.

5.2.1 Goodness-of-fit Techniques

To find a suitable model for the traffic destinations, we perform visual inspection first (Section 6.1). This visualization can only give us some insight on the shape of the distribution, and this is not sufficient for hypothesis testing. To this end, we adopt a goodness-of-fit technique over a hypothesized distribution. In our case, the hypothesized distribution is a mixture with Zipf-Mandelbrot distribution (Kvam and Vidakovic, 2007; Rayner and Best, 1989) and the goodness-of-fit test is the popular χ^2 test (D'Agostino and Stephens, 1986).

5.2.2 Factor Analysis

Factor analysis is a widely used statistical methodology whereby the observed variance of a given response or dependent variable is described in terms of explanatory factors. Such methodology has been typically applied in the social science area, but recently it has gained interest among the Internet community (Martínez-Torres et al., 2011). Specifically, in this work we use ANOVA. It provides a way to determine if such factors have any importance in explaining the variability of a response variable, and to which extent. ANOVA performs a contrast using the ratio between the adjusted sum of squares of samples that belong to each factor level, intra-level samples, and the total, inter-level samples. Such ratio is shown to follow a Snedecor's $\mathcal{F}$ distribution under the null hypothesis, provided that the samples are independent, fairly Gaussian, and exhibit homoscedasticity (i.e., share the same intra-level variance). However, the results of ANOVA are generally accepted provided that the number of elements in each group are similar (balanced ANOVA), and there is a non-excessive deviation from the homoscedasticity assumption (Glass et al., 1972).

The null hypothesis supports the homogeneity of means within factors. Basically, it contrasts, according to a given pre-defined significance level α (typically $\alpha = 0.05$), whether or not the intra-level variance values can be explained due to the randomness of measurements (generally, experimental errors) and not to differences in the population when grouped by categories (or levels). If the null hypothesis cannot be rejected, then the factor used to build the groups is statistically non-significant. Otherwise, the factor explains enough variance and it is considered as significant.

According to this, the simplest ANOVA univariate model for a response variable y with only one significant factor α is given by:

$$y_{iu} = \mu + \alpha_i + \epsilon_{iu}, \tag{1}$$

where y_{iu} represents the u^{th} observation on the i^{th} level ($i = 1, 2, \ldots, I$ levels), and μ represents the overall mean response (or intercept). On the other hand, α_i refers to the effect due to the i^{th} level of factor α and ϵ_{iu} is the deviation, random or experimental error, in the u^{th} sample on the i^{th} level. We also note that $\sum_{i=1}^{I} \alpha_i = 0$.

The resulting model in case of two significant factors is:

$$y_{iju} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{iju}, \tag{2}$$

and so forth in case of more than two factors. In this latter case, α_i and β_j represent the effect due to the i^{th} and j^{th} levels of factors α and β respectively. Similarly, $(\alpha\beta)_{ij}$ represents the interactions

between i^{th} level of factor α and j^{th} level of factor β. Finally, ϵ_{iju} represents the deviation in u^{th} sample to the overall mean of the samples within i^{th} level of factor α and j^{th} level of factor β. Again, note that $\sum_{i=1}^{I} \alpha_i = 0$ and $\sum_{j=1}^{J} \beta_j = 0$ being J the total number of levels of factor β. The reader is referred to (Dunn and Clark, 1974) for further details on the ANOVA methodology.

6 Results

Once the measurement set and methodologies have been shown, we study how to characterize the end-hosts locations. We will leverage on this characterization to compare the traffic destinations from similar campus networks.

6.1 Visual Inspection

Following the common practice in data analysis, we first provide a visual inspection of the main descriptive statistics. We order the destination countries by descending value of the measured item. Then, replacing the name of the country by its rank in the ordered list, and plotting the corresponding percentage of the measured item, we observed a power law model (for example, see Figure 2(a) for campus U_1). This observation is confirmed by the log-log plots of the same data (shown in Figure 2(b)) where the values approximate to a straight line. However, the first value in the rank seems to deviate from such a straight line. This first ranked country is Spain for almost all campus networks under study, as expected. If we remove Spain from the former figures, the data shows a better fit to a power law distribution (Figures 2(c) and 2(d)). Thus, we provide a mixture model for the whole dataset, we use a Dirac's δ function to represent the first ranked country and a power law model to fit the rest of the data (more details will be given in the next section).

Concerning population aggregates, Figure 3 shows the base 10 logarithm of the total number of bytes destined to/sourced from the top 15 contributing countries. This number of bytes is computed for the aggregate of $U_1 + \ldots + U_{12}$ in both directions, for a 3-month measurement period (between December of 2008 and March of 2009), from which weekends and holidays have been removed.

Regarding percentages, the majority of the bytes, around 40%, are sent and received within Spain. The United States (USA) comes in second place with 20% of the sent and received bytes, which is also expected because many of the most popular global brands are located in this country (Gill et al., 2008). In the third place, we find some of the most influential countries of the European Union such as United Kingdom, Germany, France, to name a few. They account for a range between 2.5 and 6% of the total number of bytes per country. In fourth place, we find Latin American countries such as Mexico, Argentina, Chile, etc., accounting for a range between 0.5 and 1.5% of the total share. These are Spanish-speaking countries and redirections to web pages in Latin America are usual. Also, there are many researchers from such countries visiting Spanish universities. Finally, we find that there is incoming and outgoing traffic from nearly all countries in the world, although their percentages of the total may be negligible.

In order to further inspect the data set and present such countries, we mapped each country with gray intensities according to the value of base 10 logarithm of the total number of bytes (Figure 4). To draw the maps, we used the Google's Visualization API that can be used directly as a gadget from Google docs.

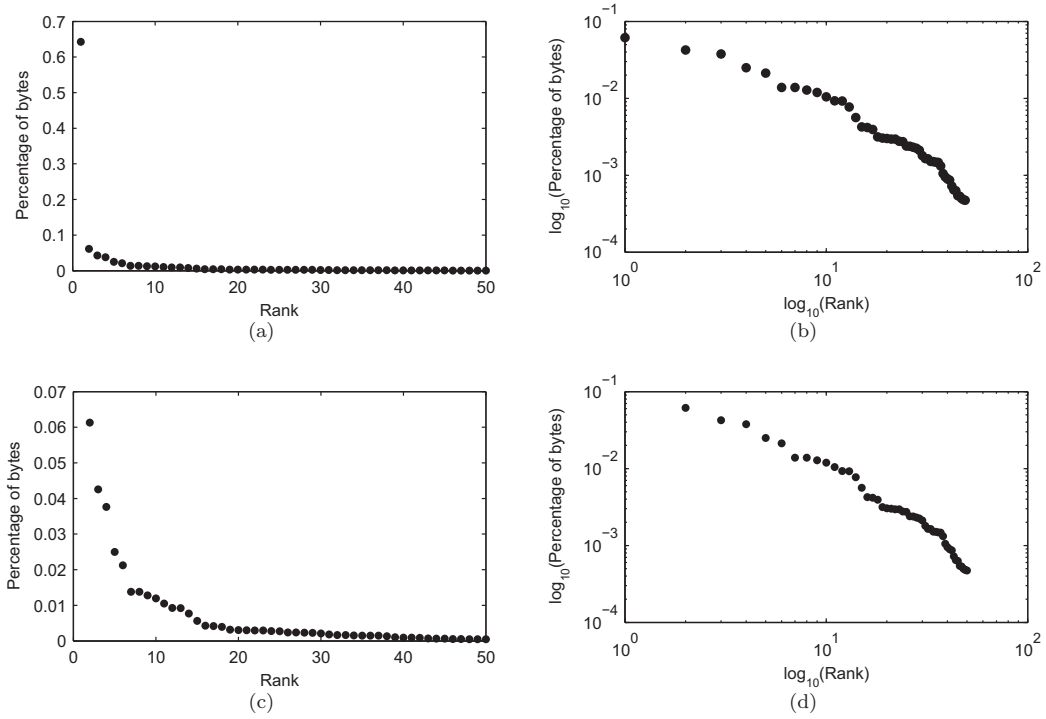


Figure 2: Visualization of the top 50 ranked countries for a day worth of measurements in the outgoing direction of U_1: (a) Percentage of bytes vs. rank. (b) $\log_{10}$ of the percentage of bytes vs. $\log_{10}$ of the rank. (c) Percentage of bytes vs. rank without the first ranked country. (d) $\log_{10}$ of the percentage of bytes vs. $\log_{10}$ of the rank without the first ranked country

6.2 Statistical Model

We propose a mixture model using a Dirac's δ function to model the top ranked country and a power law distribution to model the remaining countries. The most popular power law distribution with discrete support is Zipf's law, whereby the probability mass function of the element whose rank is k, z_k, is proportional to an inverse power a of k, i.e.,

$$z_k = \frac{q}{k^a}, \tag{3}$$

where $a > 1$ and q is a normalization positive constant (Johnson et al., 2005). Although this distribution has been widely used in Internet studies (Adamic and Huberman , 2002; Feldmann et al., 2001), we have chosen the Zipf-Maldelbrot (ZM) distribution, which is a generalization of the Zipf's law. The ZM distribution has three parameters instead of two, and shows better performance in terms of goodness-of-fit. The ZM probability mass function p_k is given by

$$p_k = \frac{c}{(k+b)^a}, \tag{4}$$

where $a > 0$, $b > -1$ and c is a normalization positive constant which is not necessarily equal to q. Consequently, our proposed mixture model has the probability mass function $P(Rank = k)$ given

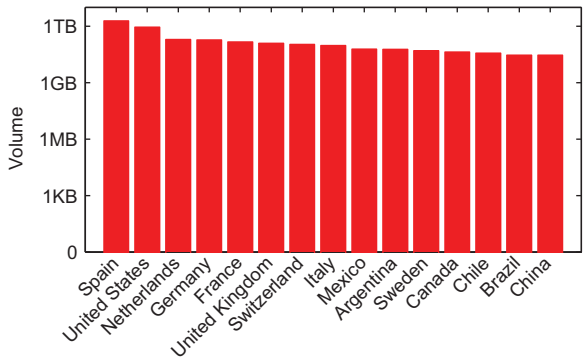


Figure 3: Base 10 logarithm of the total number of bytes sent and received by the top 15 contributing countries

Figure 4: Political map showing in gray scale the base 10 logarithm of the number of bytes sent and received by country

by:

$$P(Rank = k) = \begin{cases} p_0 & \text{if } Rank = 1, \\ c \cdot (k - 1 + b)^{-a} & \text{if } Rank \neq 1. \end{cases} \quad (5)$$

The estimate for p_0 is the percentage of traffic that is sent to the first ranked country, and c is set to make $\sum_{k=2}^{N} c \cdot (k - 1 + b)^{-a}$ to add up to $1 - p_0$, where N is the total number of countries in the model (we remove countries in the tail if their share is negligible). The Maximum Likelihood Estimation (MLE) procedure for the ZM distribution finds the parameters a and b that maximize the likelihood function for a random sample X of size n and it is given by

$$l(X; a, b) = \frac{n!}{n_1! n_2! \dots n_N} \prod_{k=1}^{N} \Big(\frac{c}{(k+b)^a} \Big)^{n_k}, \quad (6)$$

where n_k is the number of instances of the element in the k^{th} order. We note that in this procedure we have already removed the top ranked country, so the element $k = 1$ refers to the country ranked in second place. The numerical optimization of this function is a very challenging task, and several procedures have been studied to compute the multinomial coefficients involved in the likelihood function $l(X; a, b)$ in a precise and fast way. One option to circumvent the computation of the multinomial coefficients is presented in (Izsak, 2006), whereby coefficients are obtained through the probability mass function of a Binomial distribution:

$$l(X; a, b) = \prod_{i=1}^{N-1} B_{n^j}^{p_{a,b,j}/p_{a,b}^j}(n_j), \quad (7)$$

being

$$B_s^t(r) = \binom{s}{r} t^r (1 - t)^{s-r}. \quad (8)$$

In (Izsak, 2006) the calculation procedure for $p_{a,b,j}$ and $p_{a,b}^j$ is described, and this method can be easily implemented in a mathematical software package like Matlab, where the coefficients of (8) are optimally computed. However, this approach implies the rounding of the values of X, because the binomial coefficients have integer support. In addition, due to the limitations of the optimization functions in Matlab, n should be less than 3000 (Izsak, 2006). This is easy to achieve just by scaling the values of X (i.e., measuring X in tenths of megabytes instead of measuring them in bytes).

After obtaining the MLE parameters, we applied the χ^2 test to measure to which extent our model fits the data. As the ZM distribution is a discrete distribution, the buckets in the χ^2 test are defined by such discrete support. Even though it is recommended to have all buckets filled up with the same number of observations, this is unfeasible with our dataset, due to its power-law nature. However, we merged buckets with small number of samples in the tail of the distribution on attempts to have all the buckets with at least 5 samples on them. Finally, we would like to remark that the δ part of the model does not need to be accounted in the χ^2 goodness-of-fit test, because the expected value by the model and the observed one are the same.

We pursue a twofold objective in our analysis. On the one hand, we would like to assess the validity of the mixture distribution to model the data. Furthermore, we wish to find the smallest period of time such that the parameters of the model remain stable. It is worth noting that this stability check also provides hints about the trace length which is required to obtain meaningful measurement results.

Table 3: Results of the goodness of fit tests

University	Direction	Accuracy(%)	Mean p-value
U_1	Outgoing	100	0.9849
	Incoming	55	–
U_2	Outgoing	100	0.9794
	Incoming	45	–
U_3	Outgoing	0	–
	Incoming	100	0.8085
U_4	Outgoing	100	0.9888
	Incoming	78	–
U_5	Outgoing	77	–
	Incoming	99	–
U_6	Outgoing	100	0.6307
	Incoming	26	–
U_7	Outgoing	100	0.9120
	Incoming	100	0.8504
U_8	Outgoing	100	0.8569
	Incoming	12	–
U_9	Outgoing	100	0.8411
	Incoming	24	–
U_{10}	Outgoing	45	–
	Incoming	100	0.5213
U_{11}	Outgoing	100	0.9698
	Incoming	78	–
U_{12}	Outgoing	100	0.9976
	Incoming	100	0.2844

6.3 Goodness-of-fit Results

Table 3 shows the results of the χ^2 test for the number of bytes of the twelve campus networks for a period of 90 consecutive days between December of 2008 and March of 2009, from which weekends and holidays have been removed. Similar results were obtained with the other measured item and are not presented here for the sake of brevity.

The first column shows the (anonymized) university name as described in Section 4. The second column shows the direction of the traffic relative to the campus network. The accuracy in the third column is defined as the percentage of days in the sample for which the χ^2 test null hypothesis of goodness-of-fit cannot be rejected at the significance level $\alpha = 0.05$. Finally, the last column shows the average p-value from all the performed χ^2 tests. We show this average only for those pairs university-direction where the accuracy was 100%. It gives an estimate on how good the goodness-of-fit was in these cases, the larger the better. As can be seen in the table, except for a small number of university-direction pairs, the null hypothesis of goodness of fit cannot be rejected for 75% of the days of the measurement period. We think it is a reasonable value to support our assumptions and to validate a common model for the set of networks under study. Actually, it is not surprising to find model fitting discrepancies in some cases because goodness-of-fit tests are usually excessively demanding with real traffic measurements (van de Meent et al., 2006). Note that such discrepancies may arise from events such as network misuse, power cuts, temporal malfunctioning, etc., which differ from the typical network behavior, hence making the tests fail.

Remarkably, in most cases the ZM distribution fits the measurements better in the outgoing direction than in the incoming direction. In the incoming direction, the top ranked country is not as predominant as in the outgoing direction. This can be checked with the estimates of p_0 (see Table 4), and it implies that the distribution is more flat, and more days are needed to show a good fit. We hypothesize that this can be motivated by the asymmetry of the Internet applications and services, as well as by the higher activity of anomalous traffic in the incoming direction, as pointed out by (Jhon and Tafvelin, 2007). Note that traffic classification at the application layer is not possible relying on Netflow data (Moore and Papagiannaki, 2005). Further analysis of this issue is performed using factor analysis in Section 6.4.

To assess the stability of the estimated parameters, we form a time series of aggregated days and measure the relative error in the parameters for all the universities. The relative error $re_s(t)$ for a time series $s(t), t = 1, \ldots, N$ is defined as follows:

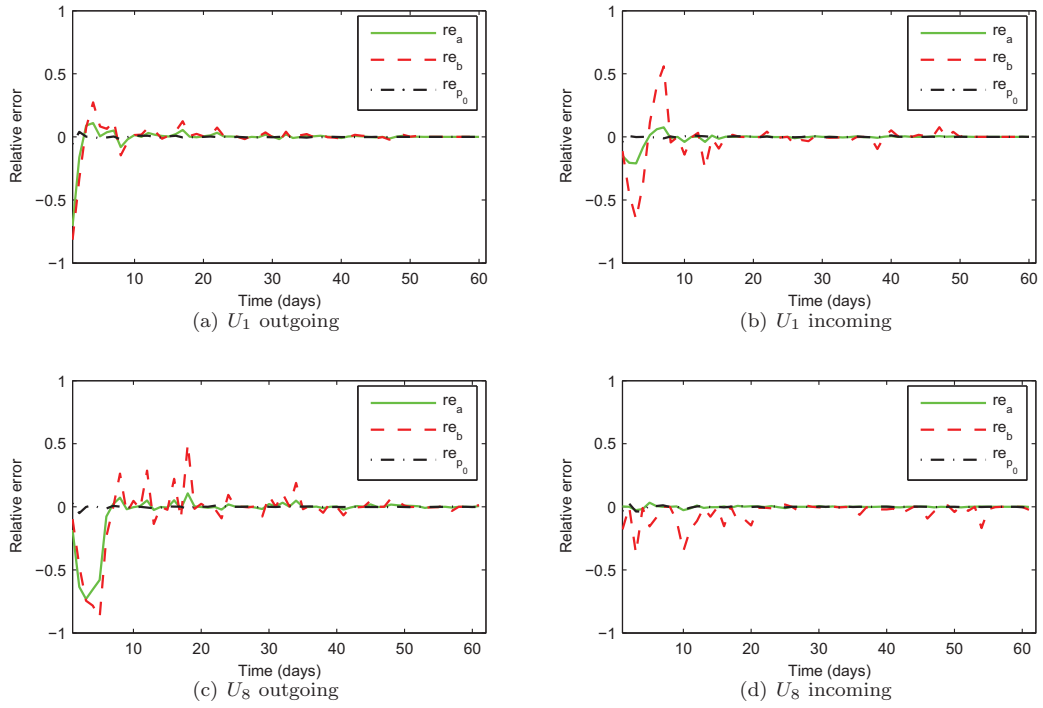
$$re_s(t) = \frac{s(t+1) - s(t)}{s(t)}, \qquad t = 1, \ldots, N-1. \tag{9}$$

In our case, t stands for the number of days used in the estimation and $s(t)$ is the estimated parameter a, b, c, p_0 in equation (5), for all the university-direction pairs showed in Table 3 using (9). Figure 5 shows the evolution of the relative error in time of a, b and p_0 parameters (note that c is function of a, b and p_0), for U_1 and U_8 networks.

As can be seen in Figure 5, parameter p_0 is the more stable, since its relative error is nearly 0 for all the estimation procedure. Regarding a and b, it can be seen an oscillation period in the beginning of the estimation procedure, when there are still few days aggregated. We measured the length of this transient period (i.e., number of measurement days aggregated until reaching stability) for each of the university-direction pairs where the χ^2 accuracy was above 75%. We removed those university-direction pairs with lack of fit because it makes no sense to consider the length of the transient period before stability when there is no goodness-of-fit. We used the convention that stability of the parameters is reached when there is a period of five consecutive days where the relative error is smaller than 5% (García-Dorado et al., 2008). Figure 6 shows a histogram of such aggregated number of days to reach stability. Note that there are 3 parameters for each university and direction (i.e., the total number of parameters is $72 = 3 \cdot 12 \cdot 2$), but the figure does not include those parameters related with the university-direction pairs ruled out according to the aforementioned criterion. However, we remark that this does not necessarily imply lack of stability for those university-direction pairs for which the χ^2 accuracy is below 75%. An example of university-direction pairs with χ^2 accuracy below 75% but remarkable parameter stability can be observed in Figure 5(b) and Figure 5(d).

The accumulation of values in the first bins of the histogram is due to the high stability of the estimation of parameter p_0, whereas the larger values are mainly due to the parameter b. In the worst case of the networks under study, at least 35 days worth of aggregated data are necessary to make the parameter estimation stable (e.g., Figure 5(c)). This result is valid for networks of similar size and user activity as the ones analyzed in this study, and other conclusions may apply at other aggregation levels. The stabilization of the parameters implies that the parameter estimation nearly does not change if we add one more day worth of data. According to it, Table 4 shows the stable parameter values for all the universities under study. It turns out that the parameter values differ from one university to another, even though, they are very much alike. Consequently, we find differences between campus networks, which, in principle, are similar in terms of population, access bandwidth, etc. This is the motivation for the factor analysis presented in the next section, which takes into account source IP network, direction and destination country.

Finally, Figure 7 shows the base 10 logarithm of the amount of traffic sent or received as a function of the base 10 logarithm of the rank, together with the fitted curve based on the estimations of the model for U_1 and U_8, showing remarkable goodness-of-fit.



(a) U_1 outgoing (b) U_1 incoming

(c) U_8 outgoing (d) U_8 incoming

Figure 5: Relative error for a, b and p_0 parameters

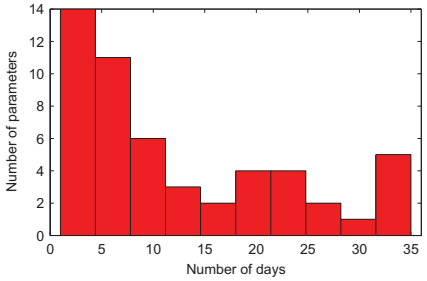
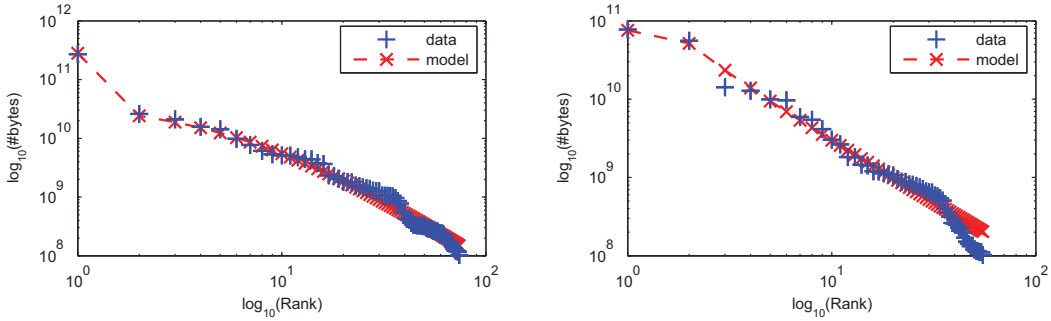


Figure 6: Histogram of the number of days aggregated to reach stability of all the parameter estimations

Table 4: Results of the maximum likelihood parameter estimation

University	Direction	Parameter estimate			
		a	b	c	p_0
U_1	Outgoing	2.37	8.36	0.62	0.60
	Incoming	1.56	0.49	0.42	0.34
U_2	Outgoing	2.09	5.28	3.53	0.55
	Incoming	1.44	0.22	0.33	0.30
U_3	Outgoing	53.48	0.03	2.31	0.54
	Incoming	1.86	0.72	0.61	0.45
U_4	Outgoing	2.46	6.74	13.82	0.49
	Incoming	2.15	1.88	2.27	0.30
U_5	Outgoing	1.43	0.74	0.26	0.55
	Incoming	2.13	1.14	1.28	0.38
U_6	Outgoing	1.25	0.00	0.17	0.47
	Incoming	3.05	2.67	15.80	0.29
U_7	Outgoing	1.43	2.39	0.50	0.49
	Incoming	1.18	-0.21	0.16	0.38
U_8	Outgoing	1.74	3.12	0.64	0.70
	Incoming	1.70	0.00	0.30	0.42
U_9	Outgoing	1.89	4.28	1.42	0.64
	Incoming	1.57	0.00	0.28	0.37
U_{10}	Outgoing	1.68	1.62	0.60	0.51
	Incoming	1.84	0.70	0.72	0.31
U_{11}	Outgoing	1.48	1.66	0.39	0.54
	Incoming	2.58	2.62	5.76	0.41
U_{12}	Outgoing	2.23	6.08	5.06	0.61
	Incoming	1.47	0.18	0.33	0.29



(a) U_1 outgoing (b) U_1 incoming

(c) U_8 outgoing (d) U_8 incoming

Figure 7: Examples of the goodness-of-fit of the model to the observations

6.4 Analysis of Variance

In this section, we apply ANOVA to our measurement set. The aim of such analysis is to assess the impact that the traffic direction (both incoming and outgoing), the source IP network under study (i.e., one of the selected university campuses) and the country in which the remote host is located, have in the response variable. In this study, the response variable refers to any of the measured items introduced in Section 4, specifically their percentages from the total. Note that we could also perform ANOVA analysis using the absolute value of the measured items instead of percentages. However, this analysis provides misleading results because small differences between the campus networks have a large influence in the response variable. For instance, the traffic load regardless the destinations, in absolute terms, within the set of universities is not identical. To avoid this overshadowing effect, we choose traffic percentages.

ANOVA allows to assess the impact of a specific country (i.e., knowing exactly which country it is) instead of using only its rank position as in the previous analysis. This enables to contrast (i) whether or not the set of countries under analysis accounts for similar popularity, (ii) whether or not the set of source IP networks under study connects to the same locations. Additionally, we also contrast (iii) whether or not the ratio incoming/outgoing traffic is similar between countries, and finally we check (iv) whether or not the networks connect to the same countries with similar ratio incoming/outgoing traffic.

Consequently, we define three fixed factors and their corresponding interactions (full factorial ANOVA): *Network*, that is the IP network under study, *Country* that represents the country in which the remote host is placed and *Direction*, either incoming or outgoing traffic. For instance, Figure 2 shows the percentages of traffic in bytes (response variable) that U_1 (factor *Network*) exchanges with top 50 contributing countries (factor *Country*) in outgoing direction (factor *Direction*) for a day worth of data. Thus, according to (2) we have the following complete model:

$$\begin{aligned} y_{ijku} = \mu &+ Network_i + Country_j + Direction_k \\ &+ (Network\&Country)_{ij} \\ &+ (Network\&Direction)_{ik} \\ &+ (Country\&Direction)_{jk} \\ &+ (Network\&Country\&Direction)_{ijk} \\ &+ \epsilon_{ijku}, \end{aligned} \tag{10}$$

where y represents any of the measured items and i, j, and k index the network, the country, and the traffic direction, respectively.

In the previous section, it was shown that at least 35 weekdays worth of data are necessary to obtain stability in the measurements under study. In this light, the ANOVA sample spans the month of January and some days of February 2009.

Regarding the number of countries, it makes sense to compare especially the head of the distribution, i.e., the most popular countries, given that the distribution follows a power law. Consequently, we have limited our study to the set of countries that account for 95% of the total traffic in terms of number of bytes. Such set is composed of 30 countries. As a result, we have a database for both directions, involving twelve networks, thirty countries and thirty-five days, that is, 25,200 samples for each measured item.

6.4.1 Assumptions

Regarding the ANOVA assumptions introduced in Section 5.2.2, Figure 8 shows the autocorrelation function (dots) along with its confidence intervals (dashed lines) applied to the averaged number

of bytes in both directions with U_1 and *Spain* as factors. It becomes apparent that the samples are not correlated. It is worth noticing that all levels showed similar results. Figure 9 shows the normal Q-Q plot for the same set of samples. If the data follows the Gaussian distribution, then it nearly fits to a straight line. In general, we have not found evidences of significant deviation from the gaussianity in the measurement set. Conversely, the homoscedasticity hypothesis was rejected by means of the Levene test. However, a non-significant deviation from the homoscedasticity assumption (Glass et al., 1972) can be accepted in case of balanced ANOVA with large number of samples, which is the case of our experimental design.

6.4.2 Effect of Network, Country and Direction in the Traffic

Table 5 shows the results of the ANOVA test with the percentage of bytes per day as the response variable. According to the results, the null hypothesis that supports the homogeneity of means cannot be rejected for the factors $Network$, $Direction$ and $Network\&Direction$, but it is rejected for $Country$, $Network\&Country$, $Country\&Direction$ and $Network\&Country\&Direction$ at the significance level $\alpha = 0.05$. Thus, we obtain the simplified model:

$$\begin{aligned} y_{ijku} = \mu &+ Country_j \\ &+ (Network\&Country)_{ij} \\ &+ (Country\&Direction)_{jk} \\ &+ (Network\&Country\&Direction)_{ijk} \\ &+ \epsilon_{ijku}, \end{aligned} \tag{11}$$

with $i = 1, 2, 3, \ldots, I = 12$ (number of networks), $j = 1, 2, 3, \ldots, J = 30$ (number of countries), $k = 1, 2$ (traffic directions) and $u = 1, 2, 3, \ldots, U = 35$ (number of analyzed days).

Several conclusions can be drawn from these results. The homogeneity of means when taking into account factor $Network$ implies that the traffic generated by the networks, ignoring destination and direction, has no influence in the measured items. This is a consequence of using traffic percentages. Similarly, the fact that factors $Direction$ and $Network\&Direction$ are not significant indicates that the traffic percentages are distributed similarly in both directions, regardless of the network. Intuitively, this means that the shape of the distributions of the percentages of traffic per country are similar between the source IP networks under analysis. However, the countries are not the same in each network. This is confirmed by the fact that the $Network\&Country$ is clearly significant. This means that the popularity of the countries depends on the source IP network that is being measured. More specifically, in the first section, we posed the following question regarding the deployment of content distribution networks: if CDN nodes are to be placed on two different campuses, with similar population size and structure, can we adopt the same content distribution strategy for both of them? ANOVA provides a negative answer. Essentially, ANOVA says that single-network measurements do not suffice for a meaningful characterization of the distribution of the remote hosts location, which supports the results presented in Table 4.

The results of the factor $Country$, which show a strong significance, state that the distribution of the popularity of the countries is clearly heterogeneous. That is, there are some countries that sent/received more traffic than others significantly. This ties in with what we expected taking into account the results of the previous sections and other works (for instance, (Giovannetti et al., 2005; Gill et al., 2008)).

In addition, the results of the factor $Country\&Direction$ shows strong significance, which implies that the relation incoming/outgoing traffic depends on the destination country under analysis. This is directly related to the peering agreement decision-making problem. Actually, one of the most typical peering agreements is the ratio-based paid peering (Norton, 2001b), in which peering is free of charge until traffic asymmetry reaches a certain ratio, commonly 4:1. With the ANOVA

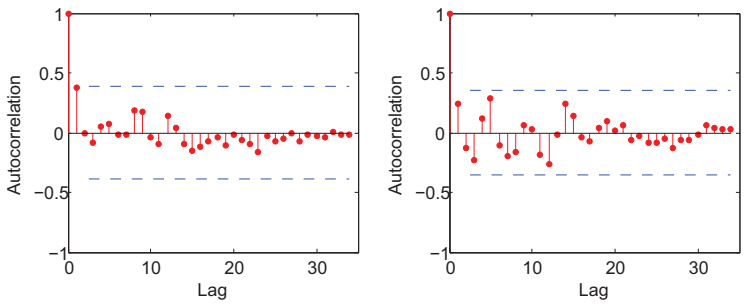


Figure 8: Autocorrelation function (dots) and 95%-confidence intervals (dashed lines) applied to the averaged number of bytes in both directions, first outgoing and then incoming, with U_1 and *Spain* as factors *Network* and *Country*

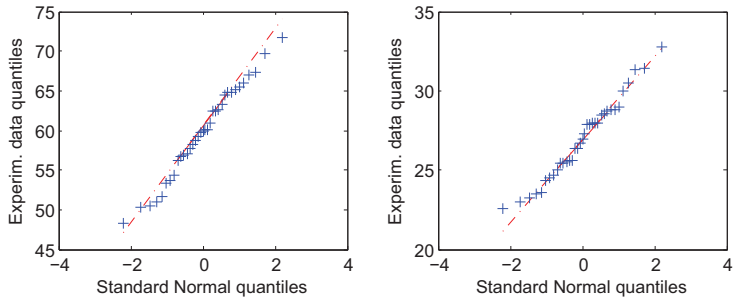


Figure 9: Q-Q plot diagram of the averaged number of bytes in both directions, outgoing and incoming, with U_1 and *Spain* as factors *Network* and *Country*

results we have shown that such ratio depends heavily on the destination country. In other words, this means that some countries in the trace behave as "consumers" for Spanish universities, whereas others show a balanced ratio, and finally others are "providers" of bytes. The explanation is likely to be found in the application layer. It is well known that the ratio incoming/outgoing traffic is a good discriminant to differentiate traffic applications (Liu et al., 2007). For instance, the ratio of HTTP protocol is usually high, i.e., more downloaded traffic that uploaded, whereas P2P applications has lower ratios. Bearing this in mind, ANOVA is detecting that typically the users of Spanish universities, for example, connect mostly to other Spanish-speaking countries using P2P applications, but they mostly access to Germany for Rapidshare (a popular one-click hosting service (Antoniades et al., 2009)). These applications present extremely different incoming/outgoing ratios.

However, the interaction factor of level 3 *Network&Country&Direction* reveals that the traffic ratio not only depends on the destination country, but also on the source IP network that generated such traffic. Surprisingly, this implies that it is not possible to label each country as a byte consumer/balanced/provider for the all set of Spanish universities under study, but each network behaves in a different fashion with respect to each country. Once more, this is closely related to ratio-based paid peering, because the ratio of incoming/outgoing traffic depends both on the country and on the network. Therefore, neither a single-network measurements nor the ratio incoming/outgoing traffic suffice for a meaningful characterization of the destination country popularity.

Table 5: ANOVA table with *Network*, *Country*, *Direction* and their interactions as fixed factors and average number of bytes as response variable

Dependent variable: Averaged number of bytes

Source	Sum of Squares	df	Mean Square	F	p-value
Network	25.399	11	2.309	0.532	0.883
Country	1741856.851	29	60064.029	13838.716	0.000
Direction	6.589	1	7.689	1.772	0.183
Network & Country	57404.851	319	179.953	41.461	0.000
Network & Direction	11.974	11	1.089	0.251	0.994
Country & Direction	299321.535	29	10321.432	2378.052	0.000
Network & Country & Direction	17330.482	319	54.328	12.517	0.000
Error	106250.279	24480	4.340		
Corrected Total	2226583.313	25199			

Ajusted R^2=0.961

Table 6: ANOVA table with *Network*, *Country*, *Direction* and their interactions as fixed factors and average number of packets as response variable

Dependent variable: Averaged number of packets

Source	Sum of Squares	df	Mean Square	F	p-value
Network	24.932	11	2.267	1.000	0.443
Country	1643499.135	29	56672.384	25002.452	0.000
Direction	0.021	1	0.021	0.009	0.922
Network & Country	104760.259	319	328.402	144.883	0.000
Network & Direction	1.355	11	0.123	0.054	1.000
Country & Direction	14029.105	29	483.762	213.424	0.000
Network & Country & Direction	1583.149	319	4.963	2.189	0.000
Error	55488.156	24480	2.267		
Corrected Total	1819386.111	25199			

Ajusted R^2=0.969

The same conclusions were obtained with packets as response variable, as shown in Table 6, and therefore the same simplified ANOVA model is reached.

Finally, Tables 5, and 6 also show the (adjusted) coefficient of determination $\bar{R}^2$. It represents a measure of the percentage of variation in the response variable that can be explained by the factors. As $\bar{R}^2$ is close to 1, we can conclude that the factors and their interactions model the measured items distribution accurately.

7 Discussion

We hypothesize that the heterogeneity in the geographical traffic pattern may be primarily due to the heavy-hitters, the P2P communities, and the malicious/unwanted traffic:

- Homogeneity in the geographical traffic pattern is expected whenever the networks population sizes are large enough. In this study, the smallest university under study has more than 30,000 students, which supports this hypothesis. Nonetheless, the Internet community has pointed out that most of the Internet traffic is generated by a small fraction of network users (Brownlee and Claffy, 2002; Papagiannaki et al., 2002), often referred to as *heavy-*

hitters (Feldmann et al., 2001). A heavy-hitter is typically defined as a user whose use of network resources exerts a significant impact on the aggregated traffic of the whole network. Thus, particular traffic patterns of a heavy-hitter user may have impact on the results of the whole campus network, explaining the different behavior that the analyzed universities have shown. We have found a number of examples of heavy-hitters in our datasets. For instance, we detected that a IP address of one of the RedIRIS' universities sent 122 GBytes to a certain external IP address located in Germany in only one day. Obviously, this heavy-hitter user's behavior had an important impact on the network measurements. Additionally, we note that this also explains the slow convergence rate to the domain of attraction of our model distribution with the number of days added to the sample. We had to add more and more days to diminish the impact of such heavy-hitter users.

- Traffic volumes related to P2P applications account for a sizable fraction of the total traffic (Sandvine, 2009). In addition, the closed P2P communities have emerged in popularity in recent years (Torres et al., 2009). In some of these communities, popular P2P applications have been modified to take advantage of peer locality as well as network topology. This implies that the location of the peers is not longer random, but it depends on the selection algorithm of the P2P application, which breaks with the assumption of homogeneity in the dataset. Note that the same applies to on-line games and other location-aware applications.

- There are a number of examples of how malicious traffic can exert considerable influence on the characterization of the remote host location. For instance, the authors in (Xie et al., 2008) identified 7721 botnet-based spam campaigns comprising 340,050 distinct IP addresses widespread across the Internet. Similarly, the authors in (Jhon and Tafvelin, 2007; Jin et al., 2007) showed the importance of the unwanted traffic, basically, worms, port scanning, and denial of service attacks. Specifically, in (Jhon and Tafvelin, 2007) it is shown that the 16 bit address ranges of the two universities they analyzed were scanned in their entirety ($2\times65{,}534$) in a 27-hour trace (20 minutes traces, four times a day, 20 days). This traffic would be computed as incoming traffic in our results. Additionally, they detected a large fraction of P2P non-malicious outbound connection attempts to non-existing hosts that resulted unsuccessful. This is often observed for P2P traffic, where unreliable file-sharing peers are common (Ruffo and Schifanella, 2007). It becomes apparent that the portion of malicious traffic over the total, often via collaborating (Puzis et al., 2009) or compromised hosts, can vary from one network to another according to each institution's network configuration and ability to protect from this traffic. This breaks again with the homogeneity of the measure set.

On the other hand, we believe that the careful process of selecting similar Spanish universities out of the total set of institutions and the sizable number of users of each institution is enough to assume that other factors like cultural issues or particular "brand loyalties" do not exert a significant impact on the results.

Finally, we note that abnormal results (non-typical or small countries) on the distribution of the most popular countries can be leveraged to detect malicious traffic. In this case, network managers should only take into account stationary results (as we have shown, more than one month worth of data).

8 Summary, Conclusions and Future Work

In this paper, we have modeled and performed a factor analysis of the Internet end-hosts location, from connections originated in an extensive set of campus networks for a long period of time. The analysis has been carried out using NetFlow records from the Spanish academic network.

First of all, we have visualized the results of the geolocation process. This visual inspection evidences that the country location of remote hosts follows a power law distribution. To confirm this hypothesis, we aggregated traffic from several universities for a month and repeated the visualization process. With the aggregated data, we confirmed the power law shape of the measurements, but the behavior was not the same per source campus network, even though the aggregation level was very high. Such observations motivated us to perform two different analysis of the data: a mixture model fitting of the measurements and a factor analysis to explain the impact of network, country, and traffic direction in our measured items.

In the modeling of the measurements, we have characterized the traffic volume according to its destination country, concluding that a small set of countries accounts for the most part of the traffic (that is properly captured by the Zipf-Mandelbrot distribution). This result was clearly expected as the location of the servers of the most popular global brands are also skewed distributed around the world (Gill et al., 2008; Sandvine, 2009). Such findings show how the content distribution networks and cache mechanisms inside a particular network can take advantage of the evident skewness of the traffic location. That is, only a small set of destinations must be taken into consideration in terms of traffic volume. As the distance should be considered of importance for the performance of such mechanisms (Erman et al., 2009), an ISP can deploy them in such a way that cuts the distance between the most remote and popular countries.

In addition, we have shown that model parameters a, b and p_0 need at least 35 days to be considered stationary, meaning that measurement campaigns should be long enough to be meaningful. Moreover, the characterization process evidenced that, while sharing the same model, all the networks have different values for the distribution parameters. In this light, it becomes necessary to collect the data all across the network and not just from selected measurement points. However, the Zipf-Mandelbrot characterization pays no attention to the countries themselves, but it only takes into account their position in the rank. This calls for the subsequent factor analysis.

As factor analysis, we have applied the ANOVA univariate methodology to assess the amount of variance of connection destinations that can be explained in terms of three factors, namely the traffic direction, the studied IP network and the destination country. The results show that the factor $Country$ is strongly significant, as well as its interactions with $Network$ and $Direction$ ones. The former interaction shows that the amount of traffic volume that each network exchanges with each country is different. That is, the ranking of the most popular countries in terms of traffic volume differs within the set of networks. Surprisingly, the latter result suggests that the sort of traffic, probably at application level, that the networks exchange depends both on the country in which the remote host is located and the network. That is, the ratio incoming/outgoing is not only different between countries, as the intuition may say, but also given a country the ratio is different within the set of networks. Thus, the relation network country, i.e., consumer/balanced/producer varies across the campus networks.

We note that we carefully selected 12 universities out of the total set of RedIRIS' institutions, which were objectively similar and with significant population size. However, it has been shown that there is no homogeneity in the parameters of the model for traffic destinations. Therefore, the routing policies, CDN designs, and peering agreements that may be good for an ISP may not apply to another ISP that serves a similar user population. As an example, let us compare university networks U_9, U_{10}, and U_{12} during the months of January and February 2009. From Table 2 we note that they are very much alike. However, the destination patterns are far from being similar. Just to mention some examples, more than 5% (in bytes) of the U_9 outgoing traffic is destined to Mexico, whereas this country represents less than 0.6% in the two other university networks. Similarly, 12% (in bytes) of the U_{10} incoming traffic comes from Germany, this amount is four times smaller in the other universities. Furthermore, about 20% of the bytes that U_9 and U_{10} networks sent were destined to USA. However USA accounts for nearly 40% of the U_{12} outgoing traffic. There are a number of similar examples. Thus, our findings show that serving new populations,

which in principle look similar, leads to dramatic changes in the connection destinations, and may call for totally different routing arrangements.

We have attributed this behavior to the heavy-hitter phenomena, i.e., a small set of users that accounts for the most of the traffic, closed community based P2P systems, and malicious traffic.

From a methodology point of view, we have also shown that the length of the measurement campaign needed to obtain a significant characterization of the end-host locations can involve a long period of time. Similarly, from the ANOVA results, we have learned that the granularity of such characterization should be very narrow, because each network connects to different countries in a different way, and the conclusions drawn for one network cannot be extrapolated to the other ones. Therefore, the ISPs' measurement campaigns should include an extensive set of networks to cope with the space diversity, and also encompass a significant period of time due to the large transient time.

Finally, as future work, we plan to perform a analogous study focusing on the remote ISP that the traffic is destined to. We believe that a study at the ISP-level will be attractive for operators and network managers, and could also pinpoint research directions to enhance traffic engineering and peering agreement establishments. We envisage that the remote ISP traffic distribution will also follow a power-law, but with heavier tail given that the number of ISPs in the world is several orders of magnitude larger than the number of countries. In addition, we plan to perform an in-deep analysis to the same data but paying special attention to the country ranking throughout the campus networks analyzed.

Acknowledgements

This work has been partially funded by the Spanish Ministry of Education and Science under project *ANFORA* (TEC2009-13385), European Union CELTIC initiative program under project *TRAMMS*, European Union project *OneLab*, and the F.P.U. and F.P.I. Research Fellowship programs of Spain. The authors would also like to thank the anonymous reviewers who helped us to improve the quality of the paper.

References

Adamic, L. A. and Huberman, B. A. (2002), "Zipf's law and the Internet". *Glottometrics*, Vol. 3, pp. 143–150.

Antoniades, D., Markatos, E. P. and Dovrolis, C. (2009), "One-click hosting services: A file-sharing hideout", In *Proceedings of the ACM SIGCOMM Internet Measurement Conference*, Chicago, IL, pp. 223–234.

Arlitt, M. F. and Williamson, C. L. (1997), "Internet web servers: workload characterization and performance implications", *IEEE/ACM Trans. Netw.*, Vol. 5 No. 5, pp. 631–645.

Bass, T. (2000), "Intrusion detection systems and multisensor data fusion", *Commun. of the ACM*, Vol. 43 No. 4, pp. 99–105.

Brownlee, N. and Claffy, K. (2002), "Understanding Internet traffic streams: dragonflies and tortoises", *IEEE Commun. Mag.*, Vol. 40 No. 10, pp. 110–117.

CAIDA (2009), What researchers would like to learn from the DITL project: The top questions and data types, available at: `http://www.caida.org/projects/ditl/questions/` (accessed 27 May 2011).

Crovella, M. and Krishnamurthy, B. (2006), *Internet measurement: infrastructure, traffic and applications*, John Wiley and Sons Inc., New York, NY.

D'Agostino, R. B. and Stephens, M. A. (1986), *Goodness-of-fit techniques*, Marcel Dekker, Inc., New York, NY.

Dunn, O. J. and Clark, V. A. (1974), *Applied Satistics: analysis of variance and regression*, John Wiley and Sons Inc., New York, NY.

Erman, J., Gerber, A., Hajiaghayi, M. T., Pei, D., and Spatscheck, O. (2009) "Network-Aware forward caching", In *Proceedings of World Wide Web Conference*, Madrid, Spain, pp. 291–291.

Feldmann, A., Greenberg, A. G., Lund, C., Reingold, N., Rexford, J. and True, F. (2001), "Deriving traffic demands for operational IP networks: methodology and experience", *IEEE/ACM Trans. Netw.*, Vol. 9 No. 3, pp. 265–279.

Feng, W.-C., Chang, F., Feng, W.-C. and Walpole, J. (2005), "A traffic characterization of popular on-line games", *IEEE/ACM Trans. Netw.*, Vol. 13 No. 3, pp. 488–500.

Floyd, S. and Paxson, V. (2001), "Difficulties in simulating the Internet", *IEEE/ACM Trans. Netw.*, Vol. 9 No. 4, pp. 392–403.

García-Dorado, J. L., Hernández, J. A., Aracil, J., López de Vergara, J. E., Montserrat, F. J., Robles, E. and de Miguel, T. P. (2008), "On the duration and spatial characteristics of Internet traffic measurement experiments", *IEEE Commun. Mag.*, Vol. 46 No. 11, pp. 148–155, 2008.

Gill, P., Arlitt, M., Li, Z. and Mahanti, A. (2008), "The flattening Internet topology: natural evolution, unsightly barnacles or contrived collapse?", In *Proceedings of Passive and Active Measurement Conference*, Cleveland, OH, pp. 1–10.

Giovannetti, E., Neuhoff, K. and Spagnolo, G. (2005), *Agglomeration in Internet co-operation peering agreements*, Cambridge Working Papers in Economics, Faculty of Economics, University of Cambridge.

Glass, G. V., Peckham, P. D. and Sanders, J. R. (1972), "Consequences of failure to meet assumptions underlying the fixed effects analysis of variance and covariance", *Review of Educational Research*, Vol. 42 No. 3, pp. 237–288.

Gueye, B., Ziviani, A., Crovella, M. and Fdida, S. (2006), "Constraint-based geolocation of Internet hosts", *IEEE/ACM Trans. Netw.*, Vol. 14 No. 6, pp. 1219–1232.

Gueye, B., Uhlig, S. and Fdida, S. (2007), "Investigating the imprecision of IP block-based geolocation", In *Proceedings of Passive and Active Measurement Conference*, Louvain-la-Neuve, Belgium, pp. 237–240.

Hofstede, R. J. and Fioreze, T. (2009), "SURFmap: A network monitoring tool based on the Google maps API", In *Proceedings of IFIP/IEEE International Symposium on Integrated Network Management*, Long Island, NY, pp. 676–690.

Izsak, F. (2006), "Maximum likelihood estimation for constrained parameters of multinomial distributions: application to Zipf-Mandelbrot models", *Computational Statistics & Data Analysis*, Vol. 51 No. 3, pp- 1575–1583.

Jhon, W. and Tafvelin, S. (2007), "Differences between in- and outbound Internet backbone traffic", In *electronic Proceedings of Terena Networking Conference*, Copenhagen, Denmark.

Jin, Y., Zhang, Z.-L., Xu, K., Cao, F. and Sahu, S. (2007), "Identifying and tracking suspicious activities through IP gray space analysis", In *Proceedings of ACM Workshop on Mining Network Data*, San Diego, CA, pp. 7–12.

Johnson, N. L., Kotz, S. and Kemp, A. W. (2005), *Univariate discrete distributions*, Wiley-Interscience, New York, NY.

Kvam, P. H. and Vidakovic, B. (2007), *Nonparametric statistics with applications to science and engineering*, Wiley-Interscience, New York, NY.

Laffont, J.-J., Marcus, S., Rey, P. and Tirole, J. (2001), "Internet peering", *The American Economic Review*, Vol. 91 No. 2, pp. 287–291.

Lippert, S. and Spagnolo, G. (2008), "Internet peering as a network of relations", *Telecommun. Policy*, Vol. 32 No. 1, pp. 33–49, 2008.

Liu, H., Feng, W., Huang, Y. and Li, X. (2007), "A peer-to-peer traffic identification method using machine learning". In *Proceedings of International Conference on Networking, Architecture, and Storage*, Guilin, China, pp. 155–160.

Mai, J., Chuah, C.-N., Sridharan, A., Ye, T. and Zang, H. (2006), "Is sampled data sufficient for anomaly detection?", In *Proceedings of the ACM SIGCOMM Internet Measurement Conference*, Rio de Janeiro, Brazil, pp. 165–176.

Martínez-Torres, M.R., Toral, S. R, Palacios, B. and Barrero, F. (2011), "Web site structure mining using social network analysis", *Internet Research*, Vol. 21 No. 2, pp. 104–123, 2011.

MaxMind (2011), GeoLite free country database, available at: `http://www.maxmind.com/app/geoip_country` (accessed 27 May 2011).

Moore, A. W. and Papagiannaki, K. (2005), "Toward the Accurate Identification of Network Applications", In *Proceedings of Passive and Active Measurement Conference*, Boston, MA, pp. 41–54.

Norton, W. B. (2001a), "A business case for ISP peering", Technical report, available at: `http://www.equinix.com` (accessed 27 May 2011).

Norton, W. B. (2001b), "Internet service providers and peering", Technical report, available at: `http://www.equinix.com` (accessed 27 May 2011).

Padmanabhan, V. N. and Subramanian, L. (2001), "An investigation of geographic mapping techniques for Internet hosts", In *Proceedings of the ACM SIGCOMM*, San Diego, CA, pp. 173–185.

Papagiannaki, K., Taft, N., Bhattacharyya, S., Thiran, P., Salamatian, K. and Diot, C. (2002), "A pragmatic definition of elephants in Internet backbone traffic", In *Proceedings of ACM SIGCOMM Internet Measurement Workshop*, Marseille, France, pp. 175–176.

Poese, I., Uhlig, S., Kaafar, M. A., Donnet, B. and Gueye, B. (2011), "IP geolocation databases: unreliable?", *SIGCOMM Comput. Commun. Rev.*, Vol. 41 No. 2, pp. 53–56.

Pras, A., Sadre, R., Sperotto, A., Fioreze, T., Hausheer, D. and Schonwalder, J. (2009), "Using NetFlow/IPFIX for network management", *J. Netw. Syst. Manage.*, Vol. 17 No. 4, pp. 482–487.

Puzis, R., Klippel, M. D., Elovici, Y. and Dolev, S. (2008), "Optimization of NIDS placement for protection of intercommunicating critical infrastructures", In *Proceedings of EuroISI*, Esbjerg, Denmark, pp. 191–203.

Puzis, R., Yagil, D., Elovici, Y. and Braha, D. (2009), "Collaborative attack on Internet users' anonymity", *Internet Research*, Vol. 19 No. 1, pp. 60–77.

Qureshi, A., Guttag, J., Weber, R., Maggs, B. and Balakrishnan, H. (2009), "Cutting the electric bill for Internet-scale systems", In *Proceedings of the ACM SIGCOMM*, Barcelona, Spain, pp. 123–134.

Rayner, J. C. W. and Best, D. J. (1989), *Smooth tests of goodness of fit*, Oxford University Press, New York, NY and Oxford, United Kingdom.

RedIRIS (2011), What is RedIRIS, available at: `http://www.rediris.es/index.php.en` (accessed 27 May 2011).

Ruffo, G. and Schifanella, R. (2007), "Fairpeers: Efficient profit sharing in fair peer-to-peer market places", *J. Netw. Syst. Manage.*, Vol. 15 No. 3, pp. 355–382.

Sandvine: Global Broadband Phenomena (2009), available at: `http://www.sandvine.com` (accessed 27 May 2011).

Schwartz, D. G. (2010), "The Internet in six words or less", *Internet Research*, Vol. 20 No. 4, pp. 389–394.

Siwpersad, S. S., Gueye, B. and Uhlig, S. (2008), "Assessing the geographic resolution of exhaustive tabulation for geolocating internet hosts", In *Proceedings of Passive and Active Measurement Conference*, Cleveland, OH, pp. 11–20.

Subramanian, L., Padmanabhan, V. N. and Katz, R. H. (2002), "Geographic properties of Internet routing", In *Proceedings of USENIX Annual Technical Conference*, Monterey, CA, pp. 243–259.

Torres, R., Hajjat, M., Rao, S., Mellia, M. and Munafo, M. (2009), "Inferring undesirable behavior from P2P traffic analysis", In *Proceedings of ACM SIGMETRICS*, Seattle, WA, pp. 25–36.

van de Meent, R., Mandjes, M. R. H. and Pras, A. (2006), "Gaussian traffic everywhere?", In *Proceedings of IEEE International Conference on Communications*, Istanbul, Turkey, pp. 573–578.

Wasem, O. J., Gross, A. M. and Tlapa, G. A. (1995), "Forecasting broadband demand between geographic areas", *IEEE Commun. Mag.*, Vol. 33 No. 2, pp. 50–57.

Weis, A. H. (2010), "Commercialization of the Internet", *Internet Research*, Vol. 20 No. 4, pp. 420–435.

Xie, Y., Yu, F., Achan, K., Panigrahy, R., Hulten, G. and Osipkov, I. (2008), "Spamming botnets: signatures and characteristics", *SIGCOMM Comput. Commun. Rev.*, Vol. 38 No. 4, pp. 171–182.

Zink, M., Suh, K., Gu, Y. and Kurose, J. (2009), "Characteristics of YouTube network traffic at a campus network - measurements, models, and implications", *Comput. Netw.*, Vol. 53 No. 4, pp. 501–514.