Towards a Statistical Characterization of the Interdomain Traffic Matrix

Jakub Mikians[1] Amogh Dhamdhere[2] Constantine Dovrolis[3]
Pere Barlet-Ros[1] Josep Solé-Pareta[1]

UPC BarcelonaTech[1] CAIDA[2] Georgia Tech[3]
{jmikians,pbarlet,pareta}@ac.upc.edu
amogh@caida.org dovrolis@cc.gatech.edu

Abstract. Identifying the statistical properties of the Interdomain Traffic Matrix (ITM) is fundamental for Internet techno-economic studies but challenging due to the lack of adequate traffic data. In this work, we utilize a Europe-wide measurement infrastructure deployed at the GÉANT backbone network to examine some important spatial properties of the ITM. In particular, we analyze its sparsity and characterize the distribution of traffic generated by different ASes. Our study reveals that the ITM is sparse and that the traffic sent by an AS can be modeled as the LogNormal or Pareto distribution, depending on whether the corresponding traffic experiences congestion or not. Finally, we show that there exist significant correlations between different ASes mostly due to relatively few highly popular prefixes.

Keywords: Internet, interdomain traffic, measurements

1 Introduction

The knowledge of interdomain traffic characteristics is important for a number of reasons, particularly related to economics and policy, as the flow of money on the Internet typically follows the flow of traffic. Even though interdomain traffic patterns significantly impact the evolution of interdomain topology and economics, Internet pricing, and policy considerations (e.g., network neutrality), we have little knowledge of the global Internet Interdomain Traffic Matrix (ITM) and of its dynamics. The major obstacle to infer interdomain traffic characteristics has been lack of data, at least in the research community. As such, accurately measuring the complete ITM is likely to remain an elusive goal. Even if direct measurements of the ITM are unlikely to be available, there is value in measuring qualitative properties of the ITM that can then be used to better inform Internet economics and policy research.

In this paper, we take first steps towards inferring some statistical properties of the interdomain traffic matrix. We rely on passive flow data from the GÉANT network, the largest academic/research backbone in Europe that connects hundreds of universities and research organizations to the global Internet. Using this data, we directly measure the ITM elements that are routed via the GÉANT network. *We emphasize that our goal is not to accurately measure each entry of the ITM.* Instead, we aim to *infer statistical*

properties of the ITM from the elements that we can observe at GÉANT. We believe that such properties of the ITM can yield a better understanding of its nature and can be used to generate synthetic, but realistic ITMs for simulation and modeling purposes. We are aware of the limitations of the analysed dataset: GÉANT, as a European academic network, it is not representative of the whole Internet. Nevertheless, it is one of the most complete datasets of interdomain traffic available to the research community, and we hope that the findings presented here will serve as first steps towards a better understanding of interdomain traffic. Verifying our findings with data from other sources, including commercial ones, is a part of our current work.

In this preliminary work, we focus on *spatial* properties of the ITM, leaving the study of temporal aspects and longitudinal evolution for future work. In particular, we characterize the visible portion of the AS-to-prefix traffic matrix. We confirm previous results about the sparsity and low effective rank of the ITM. We find that the distribution of traffic sourced by ASes is heavy-tailed, but the exact nature of the distribution can be between Pareto and LogNormal, depending on the source AS. We conjecture that the exact shape of the distribution could be related to congestion within the source AS. We also find significant correlations across different rows of the ITM, mostly due to relatively few highly popular prefixes.

2 Datasets

2.1 Traffic data

Our approach relies on using traffic data collected from a "network in the middle", i.e., a network that provides transit services to edge networks. To this end, we use traffic data from the GÉANT network [13], a Europe-wide backbone provider spanning 34 countries and connecting over 30 million researchers and students, with an overall throughput of about 50 Gb/s. GÉANT customers are mainly universities and national research networks; consequently, the traffic at GÉANT does carry an academic bias. Nevertheless, approximately half of the traffic is directed to commercial networks. For most of the connected entities, GÉANT is not the only network provider, so only a part of their traffic can be observed. Also, ASes connected to GÉANT are usually not stub networks, but can contain many subnetworks, e.g., National Research and Education Networks (NRENs) connecting many national universities. In the rest of this paper, all ASes for which we analyse traffic are research and academic networks that GÉANT is serving.

We collect NetFlow traffic summaries from 18 routers at GÉANT points of presence (POPs) for all traffic entering the GÉANT network. As GÉANT is a transit network and the traffic is neither locally produced nor consumed, we measure all traffic entering and leaving the network by combining the information from the 18 POPs. Because the GÉANT NetFlow data is sampled at the rate of 1/100, we estimate bytes and packets by dividing them by the sampling rate[1]. We determine the source and destination ASes

[1] We do not estimate the number of flows, because packet sampling does not sample flows uniformly.

by mapping the source and destination IP addresses from NetFlow records to the corresponding ASes. Previous work defined an ITM at the AS-to-AS granularity [4, 2], i.e., ITM element $T_{i,j}$ measures the traffic sent by a source AS i to destination AS j. However, as ASes do not necessarily route all their traffic through GÉANT, we do not observe traffic to all prefixes originated by the same destination AS. An AS-to-AS ITM would underestimate the traffic to such destination ASes. Consequently, we work with an AS-to-prefix ITM, i.e., we characterize the visible traffic sent from a source AS to each destination prefix over a certain aggregation interval, where a row of the matrix indicates the traffic *produced* by an AS, and a column indicates the traffic *consumed* by the prefix. In the rest of the paper we will concentrate mostly on the *rows*, as characterizing ASes (rather than prefixes) is more relevant in the context of Internet economics. Table 1 describes our traffic data. For `trace W` we observe traffic for about 8×10^6 ITM elements, that is only about 0.06% of the total number of elements in the AS-to-prefix matrix. During that week, the matrix consisted of 36k rows (ASes) and 349k columns (prefixes).

Working with an AS-to-prefix definition of the ITM, we can classify ITM elements into three groups. *Unknown* elements are those that we do not observe in the NetFlow data, as the routing path $i \to j$ does not cross GÉANT. *Visible non-zero* elements are the ITM elements for which we observe some traffic, so $TM_{i,j} > 0$. Finally, we have *visible zeros*, the elements $TM_{i,j} = 0$ for which the routing path $i \to j$ crosses GÉANT, but they see no traffic in the aggregation interval over which the ITM is constructed. In Section 3.1, we describe how we identify visible elements.

	`trace W`	`trace M`	`trace Y`
period	1 week Nov 22–28, 2010	1 month Nov 1–30, 2010	52 weeks from Jan 4, 2010
flows	3.91×10^9	1.99×10^{10}	2.17×10^{11}
packets	3.61×10^{12}	1.74×10^{13}	1.70×10^{14}
bytes	3.26×10^{15}	1.55×10^{16}	1.45×10^{17}
NetFlow data volume	111 GB	476 GB	5.75 TB

Table 1: Parameters of the GÉANT NetFlow traces.

We also collect NetFlow data from the UPC[2] access link. We see all traffic from UPC in that data because this is the only access link at UPC. We use UPC data to validate the sparsity results in Section 3.1.

2.2 Routing stability and snapshot length

As described in Section 2.1, the ITM is estimated over a certain time interval. If the interdomain routing is stable during that interval, we can be certain that if we observed some traffic for an element $T_{i,j}$, then this reflects *all* traffic sent from i to j in that time interval. If, however, routing is not stable, then $TM_{i,j}$ may reflect only a portion of the

[2] Universitat Politècnica de Catalunya, BarcelonaTech

traffic sent from i to j during this interval. We need to find an appropriate aggregation period that, on one hand, catches a significant volume of the traffic, and, on the other hand, is affected by routing instability as little as possible.

To examine routing stability, we use BGP data from RouteViews [14] collectors that peer with several hundred ASes to collect BGP tables and updates. We analyzed BGP table dumps from 4 collectors over one month. We are interested only in the routes that cross GÉANT, and so we extracted 9000 AS-to-prefix paths, each of which crossed GÉANT[3] at least once in that month. For each path we examined if it is *stable*, i.e., if it is routed via GÉANT in all BGP snapshots. Note that a path may be seen by one BGP collector as crossing GÉANT, but not crossing GÉANT by another collector.

We define routing stability ρ as the probability that a path through GÉANT does not change during a specified time interval. We find that for a day $\rho = 0.999$, for a week $\rho = 0.952$, and for a month $\rho = 0.750$. We conclude that an aggregation interval of one week provides a good trade-off between the volume of traffic captured by the ITM snapshot and route stability.

3 Properties of the ITM

In this section we examine the statistical properties of the measured ITM, particularly sparsity (Section 3.1), statistical distribution of ITM rows (Section 3.2), and possible causes for the differences across distributions for different source ASes (Section 3.3 and 3.4).

3.1 Sparsity

For a given ITM snapshot, we estimate the sparsity S as the ratio of the number of visible zeros (defined in Section 2.1) to the number of all visible elements. In the case of our data this is problematic, since we cannot directly distinguish visible zeros from unknown elements. We next describe an approach to estimate a *lower bound* on the sparsity of the AS-to-prefix ITM.

Assume, initially, that the routing path between source i and destination prefix j is stable. Let T refer to the AS-to-prefix ITM measured over a certain aggregation interval, for which we estimate the sparsity. Let R be another instance of the AS-to-prefix ITM, aggregated over a larger time interval. We refer to R as a *reference ITM*. If $T_{i,j} = 0$ and the same element $R_{i,j} > 0$ then $T_{i,j}$ is a visible zero - we are sure that $i \rightarrow j$ is routed via GÉANT (because we saw some traffic in the reference ITM). If the aggregation interval for snapshot R is larger than (and overlaps with) that of T, we can identify *some* of the visible zeros in T. Let n_R be the number of visible non-zeros in R, and n_T the number of visible non-zeros in T. Then $n_0 = n_R - n_T$ is the number identified visible zeros in T. The lower bound of the sparsity of T is then $S = n_0/n_R$. This is a lower bound, because not all visible non-zeros in T can be identified (we cannot identify the elements that are visible zeros both in R and T).

The longer the aggregation interval for R, the more visible zeros in T we can identify. However, the longer the aggregation interval, the lower the routing stability ρ (see

[3] GÉANT's AS number appears in the AS path.

Sec. 2.2). If path $i \rightarrow j$ is not stable, then we could see that $R_{i,j} > 0$ and $T_{i,j} = 0$, but the cause is that this path was routed via GÉANT for R and not routed via GÉANT for T. The real number of visible zero elements in T is lower bounded by $\rho(n_R - n_T)$. Therefore, the lower bound of the sparsity is $S = \rho(n_R - n_T)/n_R$.

We estimate the sparsity for ITM snapshots aggregated over each week in `trace Y` and over each day in `trace W`. In the former, we constructed the reference snapshot by aggregating over one month, while in the latter the reference snapshot was over one week. The average estimated lower bound of the sparsity for the weekly snapshots in `trace Y` is 0.26, which means that *at least* 26% of the ITM elements are always zero. For the daily snapshots in `trace M`, the lower bound of the sparsity is 0.47. We also observed weekly trends in the sparsity – the estimated sparsity of the daily ITM is higher during weekends (we omit the graphs due to space constraints).

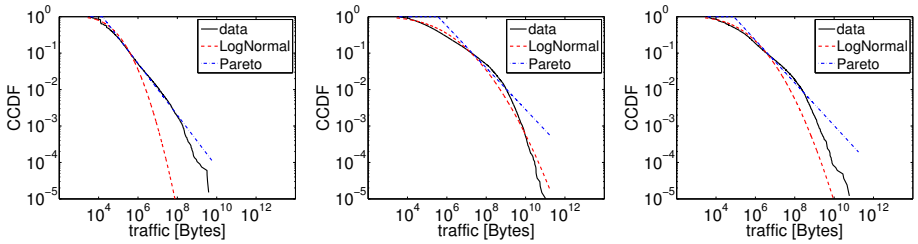
We also examined the traffic measured at the UPC access link, which is equivalent to observing one fully visible ITM row. For a single week, we observed no traffic to 45% of the destination prefixes, i.e., 45% of elements in this row were visible zeros. The results we report here corroborate the observations by Gadkari et al. [9]. Those authors observed that for the traffic sent from a regional ISP, during a single day, 49% of the destination prefixes were not used.

3.2 Distribution of traffic generated from each AS

Heavy-tailed distributions are commonly observed in the Internet [3, 5, 1]. It is not surprising that we also see heavy-tailed distributions for the generated traffic from each AS in the AS-to-prefix ITM. We analysed the distribution of generated traffic in ITM snapshots for each week in `trace Y`, selecting only those ASes (rows) for which traffic to a significant number of prefixes is routed via GÉANT (we set this threshold to 10k prefixes). In total, we analyze 3189 rows (119 distinct ASes in all 52 weeks). We find evidence for heavy-tailed distributions in the majority of the rows (94%) – the top 15% of the destination prefixes account for over 95% of the traffic. For the remaining 6% of the rows, the top 15% of prefixes account for over 71% of the traffic. In the remainder of the paper, we refer to the "tail of the traffic distribution" as the traffic sent to the top 15% of destination prefixes by the corresponding AS.

Figure 1 shows the distribution of the traffic generated by three ASes, as an example. The tail of the distribution in Figure 1a can be modeled as Pareto (the CCDF in log-log scale resembles a straight line), while the distribution in Figure 1b can be modeled as LogNormal. This confirms previous observations of the heavy-tailed nature of sourced traffic distributions [1, 6, 11] with a more recent dataset. On the other hand, the distribution in Figure 1c decays faster than Pareto but slower than LogNormal. The values in the tail refer to "heavy" prefixes, i.e., destinations that receive the largest fractions of traffic. The tail of LogNormal decays faster than the tail of Pareto, and so there is a higher probability of observing heavy destination prefixes at source ASes that follow the Pareto distribution than the LogNormal. We analyze a potential cause for this difference in the distribution shape in Sec. 3.4.

We next describe a method to determine whether the distribution of ITM elements for a row follows the LogNormal or Pareto distributions. We could use the Kolmogorov-Smirnov (K-S) or other goodness-of-fit tests. However, we are mainly interested in



(a) Pareto-like ($D = 0.88$) (b) LogNormal-like ($D = 0.27$) (c) In the middle ($D = 0.43$)

Fig. 1: Instances of the generated traffic distribution. The tail of the distribution varies between the "straight" Pareto-like to the "bent" LogNormal-like.

characterizing the *tail of these distributions*, ignoring the values in the main body of the distribution. This is because, due to NetFlow sampling, the body of the distribution consists of small values that are noisy.

Let X be the examined sample and F be the empirical distribution of X. The tail of X consists of all values in the top 15-percentile[4] of the distribution, i.e., the values above some "tail threshold" τ. Let F' be a candidate distribution (LogNormal or Pareto) that we try to fit to the tail of F. From the candidate distribution F' we generate a sample X'. We then generate a sample $\hat{X}$ by combining the tail of X and the body of X'.

$$\hat{X} = \{X' : X' < \tau, X : X \geq \tau\}$$

We now apply the K-S test under the null hypothesis H_0 that $\hat{X}$ is drawn from the same distribution F'. By construction, both $\hat{X}$ and X' have the same bodies and they differ only in their tails. Therefore, the differences between $\hat{X}$ and X' reported by the K-S test should be caused by the differences in the tails. If H_0 is rejected for a LogNormal candidate distribution and not rejected for Pareto, we assume that the tail of the data fits Pareto. In the opposite case the tail is modeled as LogNormal.

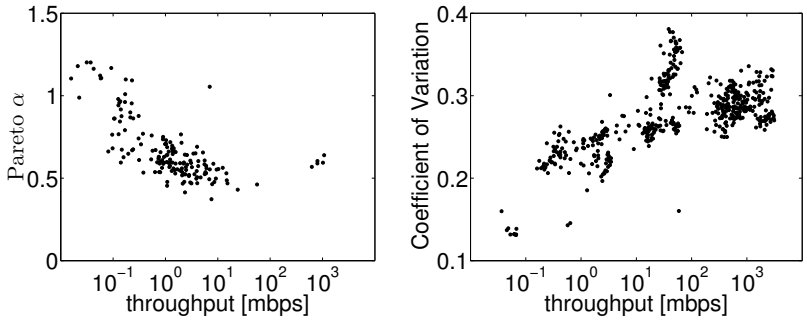
We applied this method on the traffic distributions of 3189 ASes, of which 504 were classified as LogNormal and 162 as Pareto. Our method does not classify the majority of ASes as either Pareto or LogNormal. In those cases, the empirical distribution seems to be between the previous two models.

3.3 Distribution parameters

To generate synthetic distributions of sourced traffic, we need to know the nature of the distribution (Pareto or LogNormal) and the associated parameters. In particular, we are interested in the "shape" parameter of these two distributions.

We investigated whether the shape of the measured distributions depends on the AS traffic throughput, i.e., on the total traffic generated by that AS. The shape of the Pareto distribution is represented by the α parameter; lower values of α indicate a heavier tail.

[4] We examined different values of the threshold τ and obtained qualitatively similar results.



(a) Pareto α (b) LogNormal, coefficient of variation

Fig. 2: Distribution parameters as a function of throughput.

For LogNormal, we characterize the shape of the distribution using the coefficient of variation (CoV); a higher CoV indicates a heavier tail.

Figure 2a shows α and Figure 2b shows CoV as a function of the average AS throughput. Clearly, in both cases, an increasing throughput causes a change in the shape parameter – the tail becomes heavier, and the more popular destinations receive even more traffic. This is not obvious, because the increasing traffic could cause only changes in the *scale*, but not necessarily in the *shape* of the distribution. The values of the Pareto α parameter are between 0.37 – 1.20, while the LogNormal CoV varies between 0.13 – 0.38.

3.4 What determines the shape of the tail?

In this section we investigate why the generated traffic distribution follows a LogNormal tail for some ASes, and a Pareto for others. We also show that this difference could be related to congestion within the corresponding AS.

Shape and throughput To compare the shape of the previous distribution, we define a metric D that indicates if the tail is LogNormal-like or Pareto-like. Let F be an empirical CDF of the sample, and let F_P and F_L be the CDFs of the Pareto and LogNormal distributions that fit the tail of the sample. We measure the difference in the tail using the Kolmogorov-Smirnov metric: $KS(F_1, F_2) = max|F_1(x) - F_2(x)|$ *only for values of* x *that are in the tail*. We define D as

$$D = \frac{KS(F, F_L)}{KS(F, F_L) + KS(F, F_P)} \tag{1}$$

where $D = 0$ indicates that the tail follows a LogNormal distribution, $D = 1$ indicates a Pareto distribution, and values in between represent how close the sample is to each of those two distributions.

In Figure 3 we plot the metric D and the overall throughput for each examined AS in a single week (`trace W`). The dot size indicates the number of visible-non zero prefixes for that AS. Visually, we see that ASes with lower throughput are more Pareto-like and ASes with larger throughput have a more LogNormal-like tail.

The reader may be concerned that the relation seen in Fig. 3 is an artifact of visibilty – the fact that we do not observe traffic from each source AS to the same set of destination prefixes. We investigated this possibility by performing the following experiment. Let AS_P be an AS with Pareto-like distribution and let AS_L be an AS with LogNormal distribution. Let Q be the set of prefixes that are visible non-zeros for *both* AS_P and AS_L. We determine whether the traffic sent from AS_L to prefixes Q follows the distribution of AS_P or AS_L. If it follows the distribution of AS_L, then it means that the distribution does not depend on the number of observed prefixes. We selected 4 Pareto-like ASes (with between 19k and 57k visible non-zero prefixes) and 10 LogNormal-like ASes (with between 120k and 260k visible non-zeros) and examined all 40 pairwise combinations. Interestingly, in all cases the distribution of the traffic sent by AS_L to prefixes in Q retained the properties of AS_L. Note that ASes denoted as AS_L have a larger number of visible prefixes than those in AS_P so, in general, Q overlaps more with AS_P. If the distribution of the traffic was dependent on the set of prefixes, we would expect the distribution of the traffic sent to Q to be more similar to AS_P. We thus reject the possibility that the shape of the generated traffic distribution is a function of the number of observed prefixes.

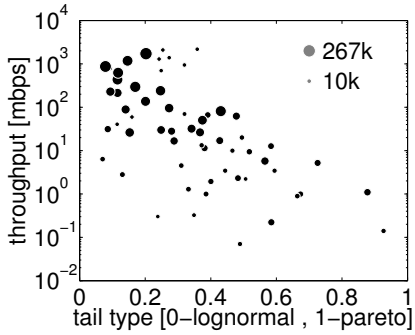
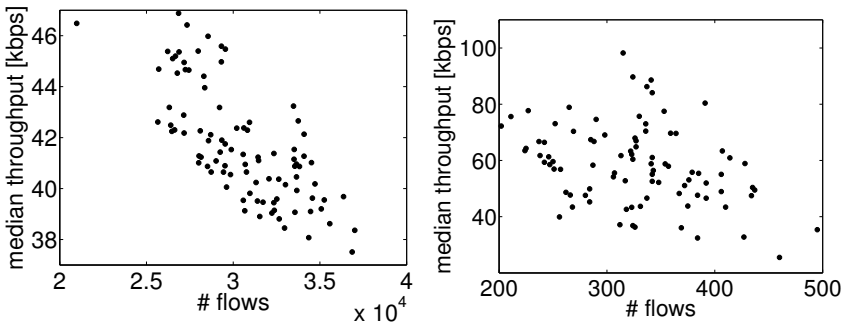


Fig. 3: Type of the distribution tail and average throughput. Each dot is a separate AS. The dot size indicates the number of visible non-zero prefixes.

Congestion In this section, we investigate a possible reason why some ASes follow the LogNormal distribution and others the Pareto distribution. Cha et al. [3] show that Pareto "tail truncation" effect can be caused by bottlenecks. In the case of interdomain traffic, we suppose that tail truncation is caused by bandwidth bottlenecks. Specifically, we conjecture that congestion can "push" the generated traffic distribution from the Pareto distribution towards the LogNormal distribution. It would mean that congestion affects large ASes more than the small ones. Finding evidence, and explanation,

of congestion inside networks is a challenging task, as we do not have any direct information about the ASes connected to GÉANT. We only have NetFlow data *collected at GÉANT* for a subset of destination prefixes; we plan to confirm these observations with more exact traffic samples as part of future work.

To detect congestion, we follow the intuition that during periods of congestion, every additional connection at the link will compete for throughput with existing connections. Consequently, we should see a negative correlation between the number of active connections at a link and the median throughout of each connection. We analyzed NetFlow data for two ASes (one LogNormal-like and the other Pareto-like) over three days at the time period that congestion is most likely (10:00–20:00), with bins of 20 minutes. To reliably estimate flow throughput and to discard TCP control flows, we only consider flows with at least 5 sampled packets, at least 100B each. Figure 4 shows the number of flows and the median throughput per flow for both ASes. For both ASes, we measured the Spearman correlation coefficient for each day. For the LogNormal-like AS, the daily correlations are -0.85, -0.77 and -0.82. For the Pareto-like AS we do not see any significant correlation. In summary, there is some evidence that ASes with LogNormal traffic distribution are subject to congestion, at least for certain time periods, while ASes that follow the Pareto distribution are not subject to congestion.



(a) Congested (b) Not congested

Fig. 4: Number of flows and the median throughput for a LogNormal-like (a) and Pareto-like (b) AS. 22–24 Nov 2010, 10:00–20:00. A few extreme outliers in (b) are not drawn.

4 AS correlations and popular prefixes

In this section we show that the ITM rows are not independent. For example, this can be the case when a set of destinations is popular for several source ASes. Correlations across rows are important in matrix completion techniques that attempt to estimate unknown elements in one row using known values in other rows. The correlations are also useful for generating synthetic ITMs.

4.1 Correlations

The nature of our dataset makes it challenging to directly measure correlations between rows, as two ITM rows can observe different sets of destination prefixes. Even if we could observe two complete ITM rows, we should not expect to see very high correlation between them, as each row consists of only few large values, with the bulk of the distribution consisting of small and highly noisy values. Hence, we restrict ourselves to studying correlations only for the set of heaviest prefixes in each row. To measure correlations between two rows of the ITM, we retain the top 15% of prefixes in each row, and calculate the Spearman correlation across prefixes that are present in both rows. We calculate pairwise correlations in this manner for each pair of rows in `trace W`. To obtain more accurate results, we only consider rows with at least 3000 visible non-zero elements. To calculate the correlation between two rows, we require that the overlap between them is at least 100 prefixes.

Using this method, we measure the correlations between 15146 pairs of rows. 10931 pairs give statistically significant correlations ($p<0.01$). 99% of the correlations are positive; the average correlation is 0.28. The highest correlation is 0.85 and 408 pairs of rows have a correlation larger than 0.5. Interestingly, for 135 pairs of rows with an overlap of more than 10000 prefixes, we observe an average correlation of 0.44.

4.2 Popular prefixes

The previous section raises the question of whether there are some globally "significant" (or popular) prefixes, i.e., prefixes that account for a large fraction of the total traffic generated by each AS. We define a prefix p as *significant* for source AS i if p is in the top-q quantile of the visible non-zero elements in row i. If $n(p)$ is the number of ASes that send traffic to p via GÉANT and $n_S(p)$ is the number of ASes for which p is significant, then the *significance* of p is $I(p) = n_S(p)/n(p)$. For the sake of accuracy, we consider only rows with at least K visible non-zero elements, and prefixes with $n(p) > 20$. We experiment with different values of K and q. Figure 5 shows, for each prefix p the significance value $I(p)$, for different values of K and q. The curves for different values of K and q are similar, at least in shape. Interestingly, there are some prefixes that are significant for most ASes (I values close to 1). For instance, for K=3000 and q=0.85, 460 out of 61000 prefixes have significance value of 0.8 or higher, and those very popular prefixes receive on average 32% of the total traffic produced by the corresponding ASes. 8800 prefixes with $I(p) > 0.5$, account for about 78% of the traffic.

This implies that there is a small group of prefixes which are significant for almost all source ASes. We found by manual inspection that more than 25% of these very popular prefixes ($I(p) > 0.8$) belong to well known large Internet entities (such as Google, NTL Virgin, OVH, Level 3, to name a few).

4.3 Low effective rank

A matrix that has *low effective rank* can be approximated by a linear combination of a small number of independent rows or columns. Some techniques to estimate invisible

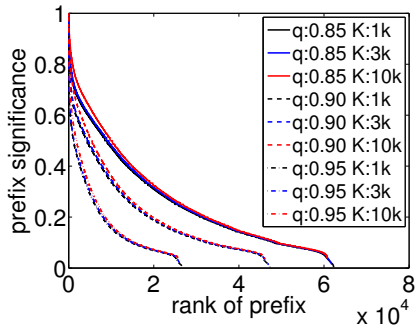


Fig. 5: Significance of prefixes, ordered. Only prefixes observed in at least 20 rows are considered.

elements of the ITM (e.g., matrix completion [2, 15]) rely on the fact that the ITM has low effective rank.

To study whether the AS-to-prefix ITM has a low effective rank, we used an ITM snapshot from `trace W`, identifying visible zeroes using a monthly reference snapshot (see Sec. 3.1). To examine the rank of the matrix, we adapted the methodology used in [2]. Using a simple heuristics based on sorting, from the observed ITM we extracted square visible submatrices of various sizes, and calculated the eigenvalues for these submatrices. Figure 6 shows the normalized (sum to 1) and averaged eigenvalues across the extracted submatrices. Clearly, only about 10 eigenvalues are significant (even for the submatrix that is 236-by-236 elements) , meaning that the ITM can be approximated with a relatively small number of independent vectors. This observation remains independent of the size of the submatrix, indicating that the global ITM is also likely to be of low rank.

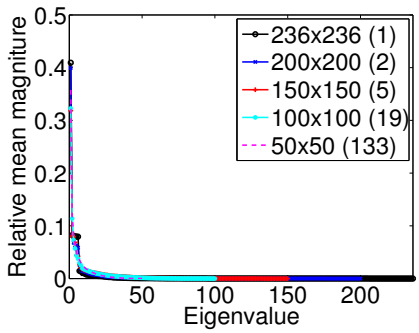


Fig. 6: Eigenvalues of the submatrices (relative magnitudes). Only a small number of the values is significant, what indicates a low effective rank.

5 Related Work

Given the importance of characterizing interdomain traffic demands, there has been surprisingly little prior work on estimating the characteristics of the interdomain traffic matrix. The major reason for this, unfortunately, has been the lack of publicly available data [13] to enable such a study by the research community. An early study by Fang et al. [7] showed that interdomain traffic distributions are highly non-uniform, an observation that has since been confirmed by others [1, 11]. Feldmann et al. [8] described a method to estimate web traffic demands using data from server logs at a large content delivery network. Chang et al. [4] propose a method to estimate interdomain traffic demand by estimating the importance of an AS in various roles – residential access, business access and web hosting. In contrast, our work aims to extract relevant statistical properties of the ITM from direct measurements, which resembles the approach in [11] for intradomain traffic. Sen et al. [12] analysed P2P traffic in large networks. A recent study from Arbor networks [10] revealed some important characteristics of interdomain traffic, such as the increasing dominance of large content providers. That study does not, however, measure a traffic matrix. Gadkari et al. [9] study prefix activity from a source AS, discovering that only a small fraction of destination prefixes receive traffic during a day, indicating that the ITM is sparse. Bharti et al. [2] also report on the sparseness of the ITM, and propose methods to infer the invisible elements of the ITM. Our work confirms the sparsity and low effective rank of the ITM seen in previous work [15].

6 Conclusions and future work

In this paper we took some first steps towards characterizing the interdomain traffic matrix. We are exploring several directions in our ongoing work in this area. First, we plan to expand our study by utilizing data from other sources. In this paper, we studied traffic characteristics from the perspective of GÉANT ASes, which has an academic and geographic (European) bias. In order to make general conclusions about the global traffic matrix, we plan to corroborate our findings using data from other academic/research networks such as Internet2, SWITCH, and commercial transit providers. Second, we plan to study temporal aspects of the ITM using historical data from GÉANT and Internet2, with the goal of studying historical trends in the evolution of the ITM. Third, an eventual goal of this work is to extract statistical properties of the ITM that will enable us to generate synthetic ITMs that retain statistical properties of the measured ITM. Synthetic traffic matrices, though important for models of interdomain topology and economics, have been hard to obtain. To this end, we are working on techniques to *infer missing elements* from the global ITM using the measured dependencies between different rows. We are also working on techniques to scale the traffic matrix to arbitrary sizes for use in simulations, while retaining the statistical properties of the measured ITM.

7 Acknowledgements

We gratefully thank DANTE, the GÉANT network operator, for kindly providing us with the data.

J. Mikians was funded by FI Grant 2010FI_B 00512 from Generalitat de Catalunya. The research was funded by the Spanish Ministry of Science and Innovation under contract TEC2011-27474 (NOMADS project) and by the *Comissionat per a Universitats i Recerca del DIUE de la Generalitat de Catalunya* (ref. 2009SGR-1140).

A. Dhamdhere and C. Dovrolis were financially supported by the National Science Foundation (grant CNS-1017139).

References

1. Alderson, D., Chang, H., Roughan, M., Uhlig, S., Willinger, W.: The many facets of internet topology and traffic. Networks and Heterogeneous Media (2006)
2. Bharti, V., Kankar, P., Setia, L., Gürsun, G., Lakhina, A., Crovella, M.: Inferring invisible traffic. In: Proceedings of the 6th International Conference. ACM (2010)
3. Cha, M., Kwak, H., Rodriguez, P., Ahn, Y., Moon, S.: I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system. In: Proc. of the 7th ACM SIGCOMM conference (2007)
4. Chang, H., Jamin, S., Mao, Z.M., Willinger, W.: An Empirical Approach to Modeling Inter-AS Traffic Matrices. In: Proceedings of the Internet Measurement Conference (IMC) (2005)
5. Downey, A.: Evidence for long-tailed distributions in the internet. In: Proceedings of the 1st ACM SIGCOMM Workshop on Internet Measurement. ACM (2001)
6. Erramill, V., Crovella, M., Taft, N.: An independent-connection model for traffic matrices. In: Proceedings of the 6th ACM SIGCOMM conference on Internet measurement (2006)
7. Fang, W., Peterson, L.: Inter-AS Traffic Patterns and Their Implications. In: IEEE Global Telecommunications Conference (GLOBECOM) (Dec 1999)
8. Feldmann, A., Kammenhuber, N., Maennel, O., Maggs, B., De Prisco, R., Sundaram, R.: A Methodology for Estimating Interdomain Web Traffic Demand. In: Proceedings of ACM SIGCOMM Internet Measurement Conference (IMC). ACM (2004)
9. Gadkari, K., Massey, D., Papadopoulos, C.: Dynamics of prefix usage at an edge router. In: Passive and Active Measurement. Springer (2011)
10. Labovitz, C., Iekel-Johnson, S., McPherson, D., Oberheide, J., Jahanian, F.: Internet inter-domain traffic. In: Proceedings of the ACM SIGCOMM 2010 conference (2010)
11. Nucci, A., Sridharan, A., Taft, N.: The problem of synthetically generating ip traffic matrices: initial recommendations. ACM SIGCOMM Computer Communication Review (2005)
12. Sen, S., Wang, J.: Analyzing peer-to-peer traffic across large networks. IEEE/ACM Transactions on Networking (ToN) (2004)
13. Uhlig, S., Quoitin, B., Lepropre, J., Balon, S.: Providing public intradomain traffic matrices to the research community. ACM SIGCOMM Computer Communication Review (2006)
14. University of Oregon Route Views Project: Http://www.routeviews.org
15. Zhang, Y., Roughan, M., Willinger, W., Qiu, L.: Spatio-temporal compressive sensing and internet traffic matrices. ACM SIGCOMM Computer Communication Review (2009)