Inter-Domain Traffic Estimation for the Outsider

Mario A. Sanchez† Fabian E. Bustamante† Balachander Krishnamurthy‡
Walter Willinger⋆ Georgios Smaragdakis◦ Jeffrey Erman‡
†Northwestern University ‡AT&T Labs Research ⋆Niksun, Inc. ◦MIT / TU Berlin

ABSTRACT

Characterizing the flow of Internet traffic is important in a wide range of contexts, from network engineering and application design to understanding the network impact of consumer demand and business relationships. Despite the growing interest, the nearly impossible task of collecting large-scale, Internet-wide traffic data has severely constrained the focus of traffic-related studies.

In this paper, we introduce a novel approach to characterize inter-domain traffic by reusing large, publicly available traceroute datasets. Our approach builds on a simple insight – the popularity of a route on the Internet can serve as an informative proxy for the volume of traffic it carries. It applies structural analysis to a dual-representation of the AS-level connectivity graph derived from available traceroute datasets. Drawing analogies with city grids and traffic, it adapts data transformations and metrics of route popularity from urban planning to serve as proxies for traffic volume. We call this approach *Network Syntax*, highlighting the connection to urban planning Space Syntax. We apply Network Syntax in the context of a global ISP and a large Internet eXchange Point and use ground-truth data to demonstrate the strong correlation (r^2 values of up to 0.9) between inter-domain traffic volume and the different proxy metrics. Working with these two network entities, we show the potential of Network Syntax for identifying critical links and inferring missing traffic matrix measurements.

Categories and Subject Descriptors

C.2.5 [**Communication Networks**]: Local and Wide-Area Networks—*Internet*; C.4 [**Performance of Systems**]: Measurement techniques

General Terms

Measurement, Traffic

IMC'14, November 5–7, 2014, Vancouver, BC, Canada.

http://dx.doi.org/10.1145/2663716.2663740.

Keywords

Traceroute; Inter-domain traffic; AS-level path

1. INTRODUCTION

Studies on the Internet inter-domain system have focused on network connectivity and dynamics and have ranged from exploring techniques to measure and generate AS-level graphs [10, 17, 48] to examining the properties of topology snapshots [23, 40]. There is, however, a growing consensus on the need to shift focus beyond connectivity towards understanding Internet traffic.

Knowledge of inter-domain traffic characteristics is important in a number of different contexts, from capacity planning to anomaly detection, and performance analysis. The major impediment to Internet traffic research has been the scarcity of publicly available traffic data. Researchers have typically had to choose between fine-grained data on a small slice of the network [7, 11, 14, 33], or publicly available, but coarse-grained and sparse datasets [13]. While detailed studies of important network entities such as Internet eXchange Points (IXPs) [7] and Content Providers [6] can improve our understanding of the inter-domain traffic, enlisting the cooperation of Internet Service Providers (ISPs), Content Providers or IXPs requires personal connections and are thus hard to replicate or scale. On the other hand, analysis of individual networks for which traffic data is available, seriously limits researchers to a handful entities.

In this paper, we introduce a novel approach to characterize inter-domain traffic by reusing the many publicly available traceroute datasets. Our key observation is that the popularity of a route on the Internet can serve as an informative proxy for the volume of traffic it carries. While traceroute measurements allow us to draw the paths taken by packets when traversing the Internet, the routes identified by a large number of traceroutes can be used to infer the popularity of a path.

Building on this observation, we introduce a new abstraction of AS-level path and apply structural analysis to a dual-representation of the AS-level connectivity graph, derived from traceroute datasets. Drawing analogies with city grids and traffic, we adapt metrics of route popularity from urban planning to serve as *proxies* for network traffic. We call this approach *Network Syntax*, highlighting the connection to Space Syntax [26, 41], an urban-planning graph-based approach to study human and vehicular flows by leveraging the strong correlation between traffic and the morphological property of streets. Network Syntax (as the related Space Syntax) builds on known abstractions and techniques from graph theory, and adapts them to our problem domain to derive new insights on inter-domain traffic from traceroute data.

We leverage publicly available traceroute datasets and apply Network Syntax in the context of a global ISP and a large Internet eXchange Point (IXP).[1] To the best of our knowledge, our work is the first to point out and capitalize on the strong correlation between Internet route popularity and the volume of traffic it carries, showing how this popularity can be derived from easy-to-perform traceroute campaigns and available datasets. In this context, we present:

- An approach, Network Syntax, that leverages traceroute datasets to tackle a problem that can currently only be studied by a few researchers with access to proprietary data.
- A demonstration of the strong correlations between inter-domain traffic volumes and the different Network Syntax metrics applied to traceroute-based AS-level connectivity graphs (and, in contrast, the weak correlations that result from applying them to BGP-derived connectivity graphs).
- An analysis of the robustness of Network Syntax to inherent idiosyncrasies of the underlying traceroute data (e.g., IP alias resolution problem and inability to trace through layer 2 clouds) and the particulars of the measurement platform used (e.g., number and network location of the vantage points).
- An illustration of the potential of Network Syntax with two use cases – the prediction of missing traffic link volumes in a connectivity graph, and the ranking of AS-links based on traffic volume.

For validation we rely on traffic ground-truth data from an ISP and an IXP; the fact that, as in most Internet studies, we cannot reveal the sources or share this data, further motivates our approach.

2. BACKGROUND

There is a large body of work focused on generating, modeling and analyzing the inter-domain topology. These include efforts that examine graph properties of the AS topology as a logical construct [19, 23, 40], techniques to measure and infer AS-level connectivity [10, 17, 48], approaches to model and characterize the Internet topology [40], or concentrate on the IXP substrate and its topological importance [8, 48]. Other efforts have used traceroute measurements to augment intra-domain router-level ISP maps by deriving OSPF link weights that are consistent with routing [35].

Beyond topology, inter-domain traffic has been an active research topic given its importance in a wide range of contexts, from network engineering to application design. However, while some research projects have made selected network traffic traces available to vetted researchers [45], the nearly impossible task of collecting large-scale, Internet-wide traffic data has seriously restricted the focus of traffic-related studies. Previous efforts have thus investigated traffic estimation and characterization (e.g., [6,22,38,42,46]), but have to rely on close collaboration with ISPs, content providers or IXPs [16, 20–22, 33] to gain access to the necessary traffic data or be limited by the coarse-grained nature of publicly available datasets [13].

Some related efforts have explored techniques and methodologies for inferring traffic matrix elements that are either not directly measurable [11,24] or missing [38,50–52]. What distinguishes our approach from these methods is their reliance, in one way or another, on link measurements obtained from either proprietary data or publicly available traffic measurements.[2]

3. NETWORK SYNTAX

In this section we expand on our descriptions of Network Syntax, its methodology and metrics. Network Syntax applies structural analysis to a dual-representation of the AS-level connectivity graph, derived from publicly available traceroute datasets, and uses different metrics to capture the popularity of a network path as a proxy for the volume of traffic it carries.

Drawing analogies with city grids and traffic, our approach adapts metrics of route popularity from urban planning's Space Syntax [26,41] for the analysis of inter-domain traffic. The following paragraphs presents a short overview of Space Syntax. We refer the reader to Hillier et al. [26] for a more in-depth description.

3.1 Space Syntax Overview

Space Syntax is a configurational analysis methodology first introduced in 1984 [25] for predicting pedestrian movement in urban settings based on an analysis of the urban grid. The key observation behind Space Syntax is that the configuration of space is the driving force behind how cities operate. Over the years, this observation has been leveraged to draw correlations between topological accessibility of spaces and urban features: from pedestrian and vehicular flows to land use, and the geographic distribution of various types of crime [27].

In Space Syntax, cities are represented as "axial maps" and then transformed into graphs. Axial maps of cities are obtained by drawing the smallest number of straight lines (called axial lines) that pass through all open spaces. These maps are then transformed into graphs by representing the axial lines as nodes and interconnecting the nodes that intersect in the map. This dual representation of the graph, where nodes are streets and edges are intersections, focuses on the connectivity of the streets irrespective of their width, length and location, and enables the identification of concrete metrics for each street. The centrality of a street or space in this graph is an indication of its importance in the city operation. Figure 1 shows an example of such a transformation.

Space Syntax introduces four core syntactic metrics, three of which can be mapped to an equivalent graph-theory metric but described using its own terminology. *(1) Connectivity* – also known as degree centrality in graph theory – is the simplest metric for assessing ranking of nodes within a connectivity graph, it equals the number of directly linked or neighboring nodes. A closely related metric *(2) Control value* measures the degree to which space controls access to its immediate neighbors by taking into account the number of alternative connections that each of

[1] We are making our own traceroute dataset and Network Syntax scripts publicly available.

[2] From networks such as Internet2 or GEANT.

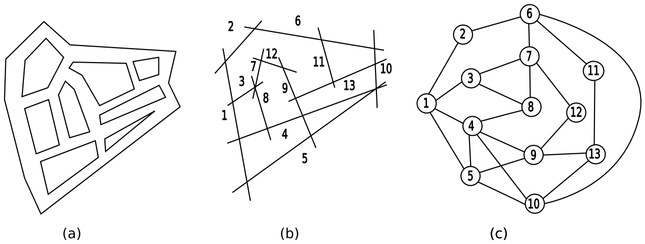


Figure 1: A sample urban system (a), its axial map (b) and connectivity graph (c).

these neighbors has [31]. It is calculated by summing the reciprocals of connectivities between neighbors. This metric can be linked to clustering coefficient in graph analysis. *(3) Global choice* – also known as betweenness centrality in graph theory – captures how often each line is used on topologically shortest paths from all lines to all other lines in the system. Finally, *(4) Integration* – a type of normalized closeness centrality metric in graph analysis – measures the mean distance between every segment and all other segments in the system [32]. The more integrated segments are those that are closest on average to all other segments, while the more segregated segments are those that are furthest on average from all other segments. Many empirical studies have shown that the integration metric is accurate in determining which segments are favored by the configuration [26, 32].

3.2 Network Syntax

We argue that the construction of AS-level connectivity graphs as carved out by probes of large traceroute campaigns contain valuable information that can be leveraged through the use of similar metrics to those of Space Syntax. Carefully vetted traceroute measurements allow us to derive partial AS-level connectivity graphs that highlight the actual routes traversed by data packets. Given that the popularity of a route on the Internet can serve as an informative proxy for the volume of traffic it carries, we argue that by concentrating on all the AS-level paths that traverse a specific AS, the application of graph structural metrics can help us identify the popular links connected to those networks.

We would not expect, however, that the direct application of these metrics to just *any* undirected AS-level graph of the Internet (such as one derived from BGP data) would yield similar results. Indeed, as we show in Section 7.1, using a non-uniform line representation of AS-level connectivity and analyzing it by measures that are essentially topological ignores too much contextual information to be useful. The mere direct connectivity between two ASes says little about the utilization of those links for carrying traffic between different parts of the Internet, especially when many important deciding factors such as routing policy are ignored. After all, as has been pointed out, there are dangers in taking available data at face value while ignoring domain-specific context [47].

3.3 Connectivity Graphs and Metrics

At a high-level, Network Syntax metrics are applied to connectivity graphs generated for specific ASes by extracting the set of AS-level paths present in a traceroute dataset that traverse each individual AS. Each of the different AS-level

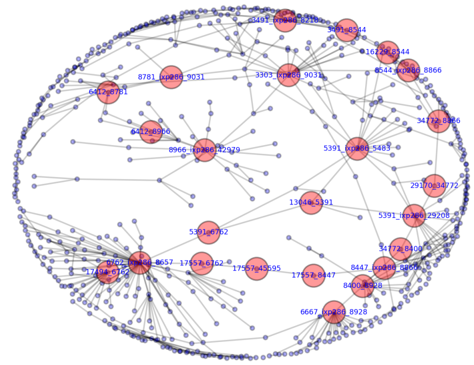


Figure 2: Dual representation of a partial AS-level connectivity graph. Each AS-link is represented by node, while the intersection between AS-links (the Autonomous Systems they interconnect) are represented as edges.

paths is broken down into pairs of hops which represent a link between those ASes. The connectivity graph is generated by adopting a dual representation where each AS-link is transformed into a node, while the intersection between AS-links (the Autonomous Systems they interconnect) are represented as edges. Figure 2 shows a connectivity graph for a subset of important AS-links.

Metrics. The different metrics take a specific meaning in the context of the dual representation of the AS-level connectivity-graph.

Connectivity. In this context, captures the number of different AS-links that precede or succeed each AS-link in the graph. A high connectivity indirectly captures the diversity of the different end-to-end AS-level paths that traverse the link.

Control value of an AS-link is defined as the sum of the reciprocal of its neighbors' connectivity. Similar to connectivity, it captures the diversity of the AS-level paths that traverse the AS-link by considering the different AS-links that precede or follow its directly connected links.

Global choice measures the popularity of a link by looking at how likely it is to be passed through on all shortest paths from all other AS-links in the network. Important links will lie on a high proportion of paths between other AS-links in the network.

Integration attempts to capture link popularity by looking at the average distance from each AS-link to every other AS-link in the connectivity graph. Using this metric, important links are identified as those that are typically "close" (on average) to other AS-links in the network.

ALTP-frequency. In our context, the popularity of an AS-link can also be captured by the number of different end-to-end AS-level paths (discovered by the probes of a traceroute campaign) that traverse the particular link. That is, the frequency of AS-link Traversing Paths, or ALTP-frequency. This metric is specific to Network Syntax and seems to have no obvious or commonly known parallel in either graph theory or Space Syntax. We describe how to compute ALTP-frequency in Section 4.3.

Dataset	Unique VPs	Src ASes	Dst ASes	Probes
Ono 2011	116,978	2,095	12,010	12.9M
Ono 2013	51,884	1,351	12,592	13.8M

Table 1: Number of unique vantage points, unique source ASes and probes for each resulting dataset after applying the different heuristics described in Section 4.2.

4. FROM NETWORK SYNTAX TO INTER-DOMAIN TRAFFIC

In this section, we start with a description of our datasets: a collection of traceroutes launched from topologically diverse vantage points and the traffic datasets we rely on for ground truth. We describe then how we leverage the AS-level paths gathered by the probes of traceroute datasets to derive a connectivity graph upon which core Network Syntax metrics can be applied.

4.1 Datasets

We evaluate Network Syntax using a traceroute dataset and traffic data, our ground-truth, from two large network entities: a large European IXP and a global ISP.

Traceroutes. Our traceroute datasets consist of data collected in two difference campaigns by topologically diverse vantage points. It contains the probes launched towards randomly selected IP addresses from the Ono BitTorrent extension [18]. Ono peers perform measurements to randomly selected destinations from the set of connections established through BitTorrent. The datasets consist of all the measurements gathered between two 30-day periods in two different years – April 1 to 30, 2011 and April 1 to 30, 2013. Table 1 shows a summary of both datasets. The first one consists of ≈12.9 million probes launched by 116,978 distinct vantage points located in 2,143 unique ASes. The second dataset includes ≈13.8 million probes launched from 51,884 different vantage points located in 1,351 different networks.

Traffic. To validate our approach, we perform our analysis in the context of a large European IXP (IXP) and a global Tier-1 Internet Service Provider (ISP). Ground-truth traffic data for IXP consists of sFlow [30] records, collected over 1-week periods in April 2011 and April 2013, capturing the traffic exchanged over the public peering fabric of the IXP. Using a random sampling of $1/16K$ packets, the resulting *traffic matrix* contains the estimated number of bytes exchanged between pairs of ASes peering at the IXP. This detailed information allows us to rank the peerings based on the volume of (bi-directional) traffic they exchange.

The ground-truth data for ISP consists of the traffic exchanged between the ISP and all its customer ASes. The data contains per-customer link utilization from SNMP records for April 2011 and April 2013, and includes the 95th percentile utilization during the course of the month of the hourly port utilizations. Traffic data is summarized on a per-customer basis using link aggregation across different physical interfaces for customers with multiple links.

4.2 Methodology

We generate the partial connectivity graph for a particular ASX from the total set of AS-level paths that include ASX, i.e., paths that include AS-level links that connect to ASX. This traceroute-derived AS-level graph is "partial" in that it does not capture the complete connectivity graph. Since an AS-level path is a sequence of AS-level links, we call such AS-level paths ALTPs, for AS-link Traversing Paths.

The following paragraphs formally define ALTPs and describe how they are extracted from traceroute datasets and how they are used to generate a partial connectivity graph.

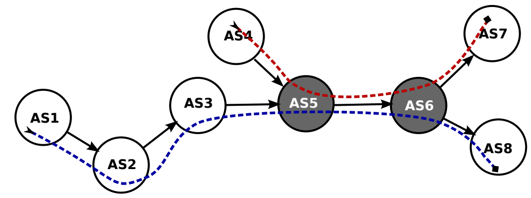


Figure 3: Two instances of AS-link Traversing Path $ALTP(5,6)$: $\langle as_1, as_2, as_3, AS_5, AS_6, as_8\rangle$, $\langle as_4, AS_5, AS_6, as_7\rangle$

AS-link Traversing Paths (ALTPs). Formally, an AS-level path can be defined as a sequence of unique AS-level links, each connecting a pair of ASes. We denote an AS-level path as $\langle as_1, as_2, ..., as_k\rangle$. We define an $ALTP(x, y)$ as an AS-level path that traverses the AS-link $\langle x, y\rangle$, i.e., $\langle as_1, ..., x, y, ..., as_k\rangle$. Note that $\langle as_1, ..., x, y, ..., as_k\rangle$ and $\langle as_k, ..., y, x, ..., as_1\rangle$ are considered two different ALTPs. Figure 3 shows an example of such a path. The ALTP abstraction is directly applicable to paths that traverse an IXP whose presence is identified by the prefix assigned by the responsible Internet Registry. The set of all unique ALTPs found in a dataset that traverse a specific AS-link is called the ALTP-set of that link.

From probes to AS-level paths. We extract AS-level paths from different traceroute datasets using public IP-to-AS mapping and correcting for inconsistencies with BGP information [17]. For paths that traverse an IXP, we follow [10] to assign confidence levels to the discovered IXP-peerings.

We first prune our dataset by eliminating loops and cycles and filtering private and reserved IP ranges and remove the associated hops if they appear at the ends of the probe. We convert IP-level to AS-level paths using the AS mapping derived from publicly available BGP information. From the derived set we then conservatively remove any probes with unknown AS-hops in the path, probes for which the source or destination AS cannot be mapped[3], and probes for which the final AS-level path is too short (probes within the same AS).

We then apply known heuristics [17] to correct the set of AS-level paths, but preserving the IP addresses belonging to known IXP hops. IXP mapping requires a complete list of prefixes assigned to the different IXPs and the list of their AS members which we obtain from centralized databases like PCH [2], PeeringDb [3] and EuroIX [1], and websites maintained by the IXPs themselves. We remove obsolete

[3]Since the removed links appear only in the discarded paths, they are not popular and their removal has no impact on our findings.

records (inactive IXPs) and match IXPs with different names that represent the same entity. We assign confidence levels to the discovered IXP peerings following the approach in [10] labeling as high-confidence those peerings for which we have probes traversing the peering in both directions or those for which the ASes at both sides of the peering have been identified as members of the IXP from BGP data collected at the IXP. For our analysis we consider only these high-confidence peerings.

AS-Link connectivity graphs. To generate the connectivity graph for the various Autonomous Systems we extract the set of ALTPs present in the dataset that traverse links connected to each specific AS. As described in Section 3.2, each different ALTP is broken down into pairs of hops which represent a link between those ASes, and a dual-connectivity graph is generated where each AS-link is transformed into a node, while the intersection between links (the Autonomous Systems they interconnect) are represented as edges.

4.3 Computing ALTP-frequency

Having introduced partial connectivity graphs and formally defined ALTPs, we can now describe how to compute the Network Syntax ALTP-frequency metric. Using the derived set of AS-level paths, we compute the ALTP-sets for all the AS-links of interest found in those paths.

The relative cardinality of the ALTP-set of an AS-link, in a set S, is the cardinality of its ALTP-set divided by the sum of the cardinalities of the ALTP-sets of every AS-link in S. This relative cardinality is what we refer to as *ALTP-frequency*.

5. EVALUATION

In this section, we apply the approach outlined in Sections 3 and 4 to the links of IXP and ISP present in our 2011 and 2013 traceroute datasets, we then analyze the relation between the different Network Syntax metrics and the links' traffic volumes. Note that the idea of route popularity applies equally well to physical and logical AS-links between ASes. As a result, the Network Syntax approach can be applied in both our IXP and ISP scenarios for which the available measurements provide ground truth about traffic traversing physical and logical AS-links, respectively.

In our analysis, we rely on a number of statistical techniques that assume normally distributed variables. However, as we show next, this assumption does not necessarily hold for all the variables we are interested in. A well-known remedy for failure of normality expected by many parametric tests [9, 28, 44] is to apply appropriately-chosen transformations. In our analysis, whenever necessary, we use the most appropriate transformation for the different variables.

5.1 Traffic distribution

We first look at the traffic distribution for the subset of AS-links discovered by the different traceroute campaigns for both ISP and IXP. We extract from the corresponding traffic matrix (derived from the ground-truth traffic data) the amount of traffic each link carries, and generate a histogram of their traffic distribution. Figure 4 shows this histogram for the links of IXP discovered by the April 2011 traceroute campaign. As it is clear from the figure, the links' traffic follows a power-law distribution with several heaviest

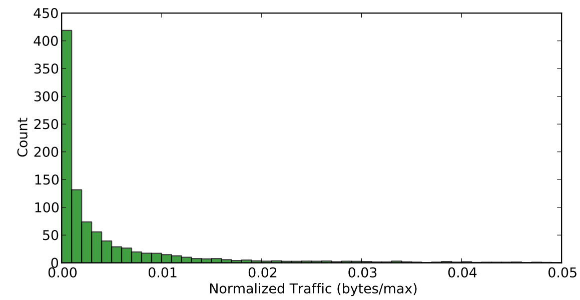


Figure 4: Distribution of traffic for all peerings in traffic matrix for IXP, April 2011.

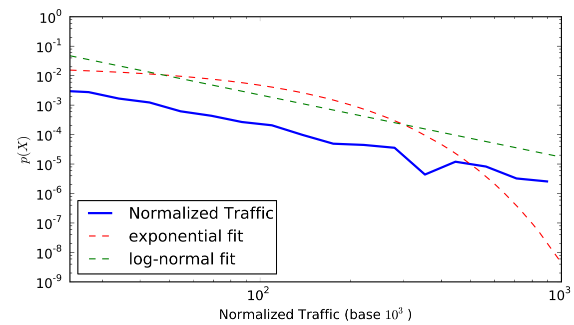


Figure 5: Comparing the goodness of fit of normalized traffic (blue line) with log-normal distribution. Dashed green line: log-normal fit starting from the optimal xmin=23.41, alpha=2.1. Dashed red line: exponential fit starting from the same xmin.

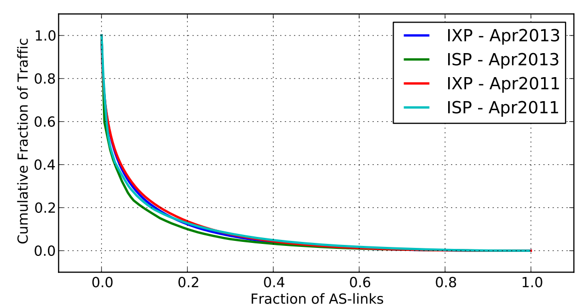
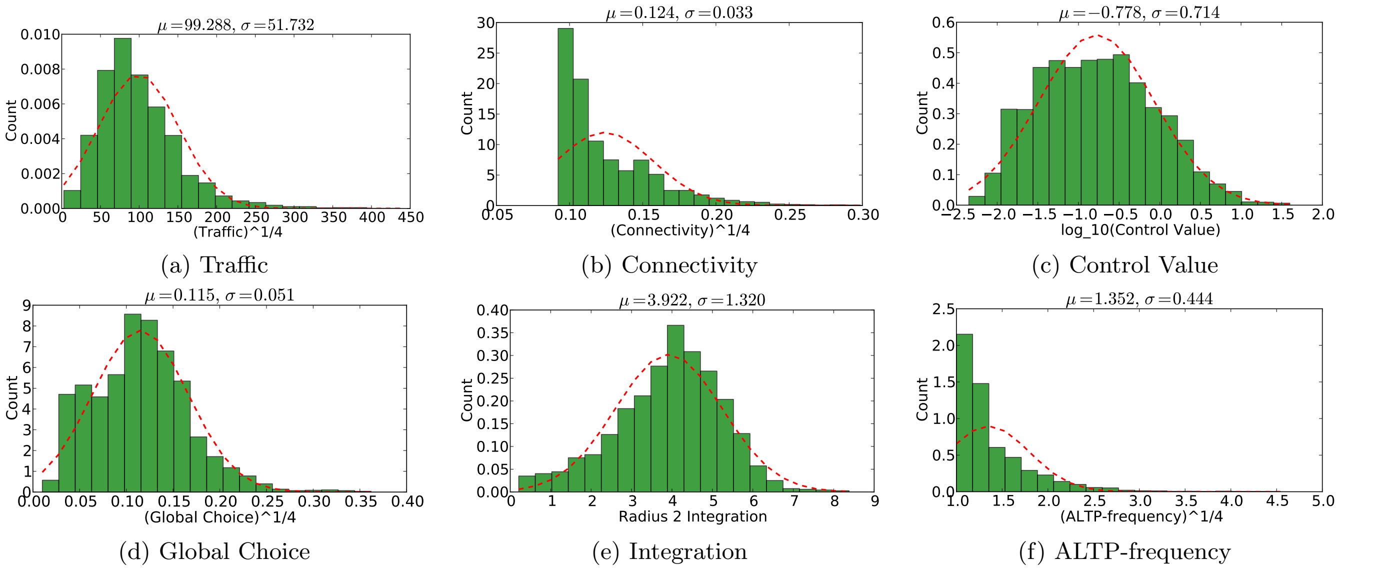


Figure 6: Complimentary CDF of the cumulative fraction of traffic for all AS-links of ISP and IXP discovered by our traceroute datasets.

links carrying most of the traffic and the majority of the links carrying very little traffic.

Figure 5 shows the goodness of fit statistical confirmation that indeed the distribution of per-peering traffic (showed in blue) follows a log-normal distribution (represented by the dashed-green line), which is consistent with the previously observed property of intra-domain traffic matrix snapshots [13, 39].

We observe similar distributions in the traffic carried by the links discovered by our April 2011 and April 2013 traceroute datasets for both ISP and IXP. This can be seen in Figure 6, the complimentary CDF of the cumulative fraction of traffic for all the discovered AS-links. In all



(a) Traffic (b) Connectivity (c) Control Value

(d) Global Choice (e) Integration (f) ALTP-frequency

Figure 7: Distribution of the transformed different metrics and traffic volume for all AS-links discovered in the April 2011 traceroute datasets for IXP. Similar distributions are observed for the other datasets and the two studied network entities.

cases, a small fraction of the links carries the majority of the traffic. To approximate normality in the cases where the dependent variable denotes traffic volume, we use the fourth root transformation for its highly skewed distribution (see Figure 7a).

5.2 Network Syntax analysis

We start by selecting, from our traceroute datasets, the ALTP-sets for all relevant AS-links that traverse IXP and ISP. We generate separate AS-link connectivity graphs for each dataset and compute the different Network Syntax metrics for each link present in the graph. Finally we plot each AS-link against the volume of traffic it carries as indicated by our traffic matrix.

To reduce the potential noise on the correlations introduced by particular AS-links (e.g., due to sampling issues), we use the different metrics to order AS-links, breaking ties based on connectivity, and cluster them in equal sized groups of ten links [4]. For each group, we compute its value for both the relative corresponding metric and carried traffic, as the average of the individual values of the AS-links within the group.

The different Network Syntax metrics present varied distributions. As shown in Figure 7e, the integration metric is already close to normal and is used without transformation. Furthermore, while the control metric (Figure 7c) is approximately normalized with the help of a logarithmic transformation, for the rest of the metrics, we achieve the desired approximate normality using the fourth root transformation (see Figures 7b, 7d and 7f). Having transformed the distributions of the dependent (i.e., traffic volume) as well as independent variables, we can study the relationships between traffic volume and the different Network Syntax metrics.

Figures 8 to 11 show the correlations between the five Network Syntax metrics and traffic volume for the links of *ISP* and *IXP* found in our traceroute datasets. The figures are presented side by side to facilitate horizontal and vertical comparisons between metrics and across datasets.

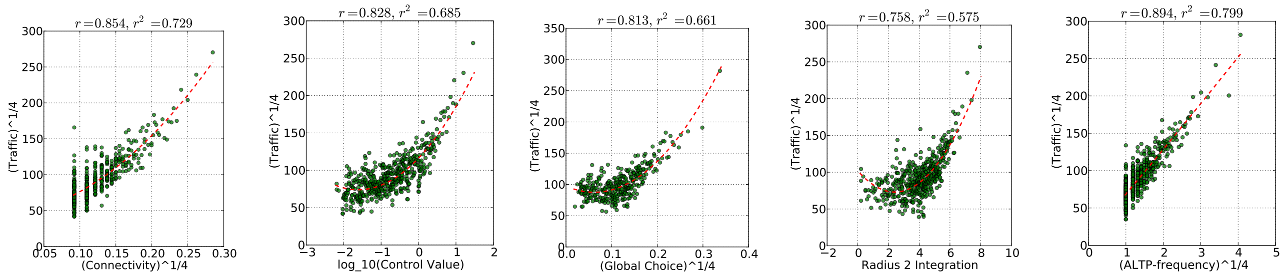
ALTP-frequency and connectivity have the strongest correlation coefficients, with the integration metric having the weakest one of all. The correlation with ALTP-frequency has r^2 values as high as 0.95 (*ISP* in April 2013) with the lowest value at 0.71 (for IXP in April 2013). We argue that this strong correlation comes from the fact that ALTP-frequency more directly captures the popularity of the high-traffic links.

The connectivity metric shows consistently strong correlations as well, with r^2 values ranging between 0.61 (for IXP April 2011) to 0.95 (for ISP April 2013). Recall that this metric captures the degree of each node in the connectivity-graph which correspond to a different AS-link in our dual representation of the AS-level connectivity-graph. As such, a large connectivity value captures the number of different AS-links that precede or succeeds it on the ALTPs identified. The connectivity metric, then, captures the diversity of the ALTPs that traverse through the link; it indirectly captures the ALTP-frequency of the link.

The correlation between traffic volume and the control value metric, while still strong, is comparably lower with a minimum r^2 value of 0.52 (for IXP April 2013) and a maximum value of 0.76 (for ISP April 2011). Although this metric is based on the connectivity values of a link's neighbors, it can overestimate the popularity of an AS-link given that the high connectivity of a neighbor can be partially due to AS-links traversed by paths that never cross the link in question.

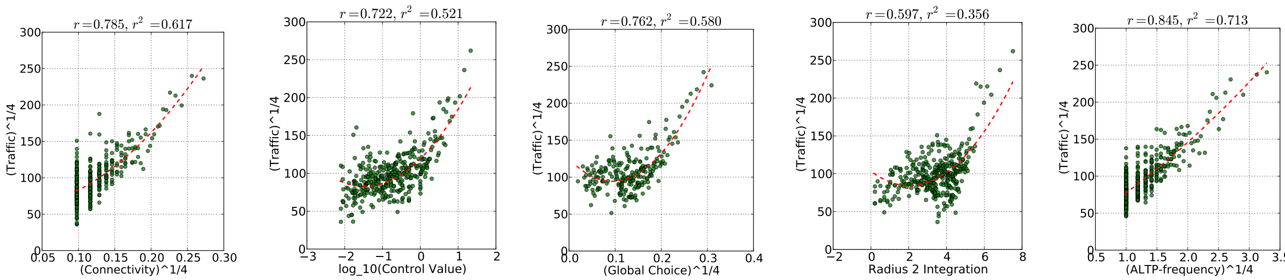
The integration metric highlights the AS-links that have the shortest average path to every other AS-link in the network. The results for this metric present the largest variations in terms of correlation for the different datasets, ranging from 0.356 (for IXP April 2013) to 0.826 (for ISP 2011).

[4] Different clusters sizes yield similar trends.



(a) Connectivity (b) Control Value (c) Global Choice (d) Integration (e) ALTP-frequency

Figure 8: Correlation between *connectivity*, *control value*, *global choice*, *local integration* metric (integration radius 2) and *ALTP-frequency* with traffic volume for IXP for April 2011.



(a) Connectivity (b) Control Value (c) Global Choice (d) Integration (e) ALTP-frequency

Figure 9: Correlation between *connectivity*, *control value*, *global choice*, *local integration* metric (integration radius 2) and *ALTP-frequency* with traffic volume for IXP for April 2013.

Finally, recall that global choice takes into account the relations between each node and the whole system. It indexes how often each line is used on topologically shortest paths from all lines to all other lines in the system. It thus, finds the AS-links that are necessary conduits for information that must traverse disparate parts of the network. The figures show that the correlation of this metric with traffic volume is also significant, with r^2 values between 0.58 (for IXP April 2013) and 0.90 (for ISP April 2013).

	IXP		ISP	
	Apr 2011	Apr 2013	Apr 2011	Apr 2013
Connectivity	0.729	0.617	0.789	0.954
Control Value	0.685	0.521	0.759	0.750
Global Choice	0.661	0.580	0.653	0.903
Integration	0.575	0.356	0.826	0.629
ALTP-freq	0.799	0.713	0.965	0.958

Table 2: r^2 values of the different metrics for ISP and IXP.

The values in Table 2 show that, although the different datasets vary in their degree of correlation, the regression lines are more or less coincident. While varying with context, a correlation coefficient greater than 0.5 is generally considered strong, and values greater than 0.8 as very strong correlation. The table shows the values of the coefficient of determination resulting from our regression analysis – nearly all (19/20) the r^2 values are above 0.5 and the ALTP-frequency values range between 0.7 and 0.96.

While the ALTP-frequency metric outperforms the rest of the metrics, there is a strong correlation between the different metrics and traffic. Exploring the relationship between different variables is part of future work.

6. USE CASES

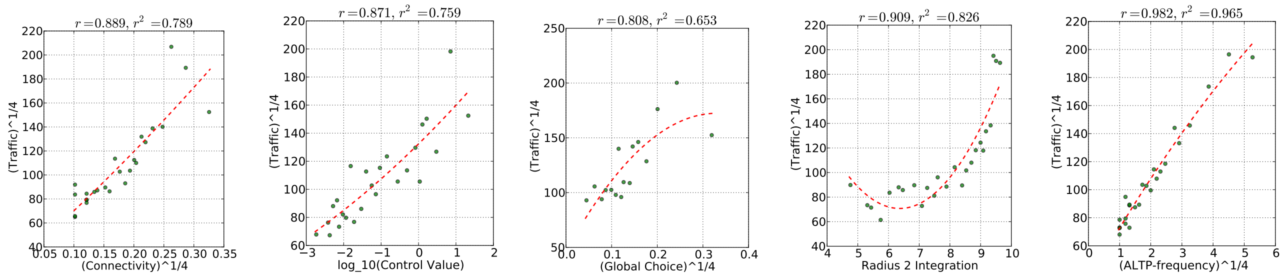
In this section, we illustrate the potential uses of Network Syntax using two use-cases: (i) predicting missing traffic link volumes in a connectivity graph and, (ii) ranking AS-Links based on their traffic volume.

6.1 Predicting link traffic

We have shown in Section 5.2 that the fraction of traffic carried by the AS-links identified in massive traceroute datasets strongly correlates with the different Network Syntax metrics when those links are clustered in small groups. We now show that it is possible to leverage this strong correlation to estimate the traffic volume of arbitrary links, in the *absence* of ground-truth traffic data as long as we have information about the traffic for a subset of the remaining links in the connectivity graph. We do this using the April 2013 datasets for ISP; similar results were obtained using the remaining datasets.

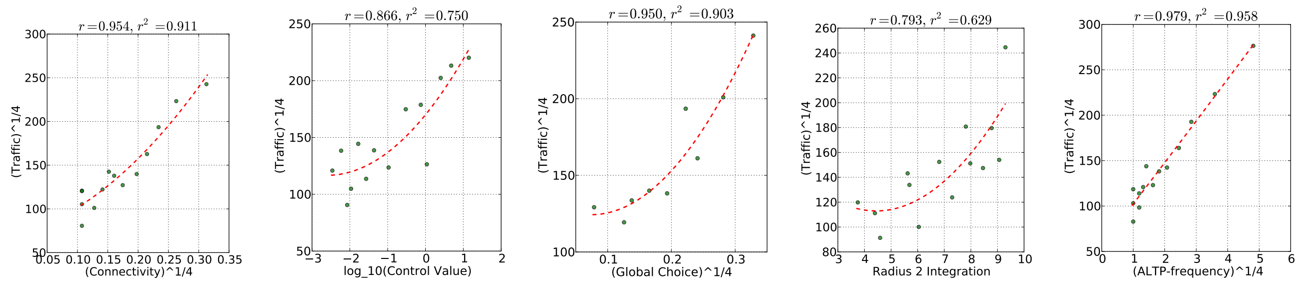
In this analysis we employ a subset of the clusters of AS-links to compute the correlation and corresponding regression line between traffic volume and the Network Syntax metric with the strongest correlation: ALTP-frequency. We then use the computed parameters to estimate the traffic volume of the remaining clusters of links by using their ALTP-frequency as proxy.

To reduce the number of links per cluster as much as possible, we start by generating clusters of size ten as described in Section 5.2. We then remove, from each individual cluster, the AS-links that diverge from the median



(a) Connectivity (b) Control Value (c) Global Choice (d) Integration (e) ALTP-frequency

Figure 10: Correlation between *connectivity*, *control value*, *global choice* , *local integration* metric (integration radius 2) and *ALTP-frequency* with traffic volume for *ISP* for April 2011.



(a) Connectivity (b) Control Value (c) Global Choice (d) Integration (e) ALTP-frequency

Figure 11: Correlation between *connectivity*, *control value*, *global choice*, *local integration* metric (integration radius 2) and *ALTP-frequency* with traffic volume for *ISP* for April 2013.

cluster ALTP-frequency by at least one standard deviation (a total of 25 links out of 147). This allows us to use the remaining links to generate clusters of half their original size while maintaining a similarly strong correlation.[5] Figure 12 shows this correlation and corresponding regression line for *ISP* for clusters of size five for the remaining 122 links.

We vary the fraction of clusters used to compute the regression line from 65% to 85% of the available clusters (in increments of 5%) and compute the difference, in orders of magnitude, between the median estimated traffic values and the actual traffic values (from the ground-truth) for the remaining clusters.[6] If the estimated and real value fall within the same order of magnitude (say, between 0 and 10MB or between 10 and 100MB), then the difference is zero. A difference between the estimated and real value of 1, on the other hand, means we may have under/over-estimated the traffic volume by one order of magnitude (e.g., declaring it to be in 50MB when it is closer to 5MB).

Figure 13 shows the result of our analysis after repeating our random selection for each percentage of clusters, 500 times. The figure plots the median difference between estimated and actual traffic volumes for each of the different fractions. For the median case, ≈80% of the estimated values fall within the same order of magnitude as the ground-truth values.

To characterize the size estimation errors for the link clusters with predicted and real values within the same order of magnitude, we compute the normalized mean absolute error between the median estimated and ground-truth traffic volumes. Figure 14 plots the mean estimation error for the different fractions of clusters used to compute the regression line. The figure shows the standard deviation logically varies depending on the fraction of clusters used. That is, increasing the number of points used to compute the fit decreases the number of clusters left for value estimation, making estimation errors more noticeable for smaller clusters. However, although for some clusters of links the traffic prediction estimates can be off by a large margin (e.g., predicting traffic to be 5MB when it was close to 1MB yields an error of 500%), the figure shows that the mean value remains relatively stable around 0.5, indicating that, on average, the median estimated traffic (in megabytes) differs from the cluster's real value by ≈50%. While we acknowledge that order of magnitude is a coarse approximation, we argue this is a valuable first step at inferring traffic volumes that are not directly measurable at scale or without access to a collection of proprietary data (e.g., Arbor Network's collection of inter-domain traffic data from some 110 commercial networks [33]). For instance, the approximate nature of alternative methods that formulate inter-domain traffic estimation as a matrix completion problem [24] is largely unknown.

[5]We observed similar results using clusters of size ten, without removing any links from the original sets.

[6]We considered the use of OC-based bucketing for this case study, but decided against it as our analysis compares groups of links rather than individual ones.

6.2 Ranking AS-Links based on Traffic Volume

As our second use case, we show that ranking AS-links based on different Network Syntax metrics can be used as a proxy for the traffic-volume based ranking of those links.

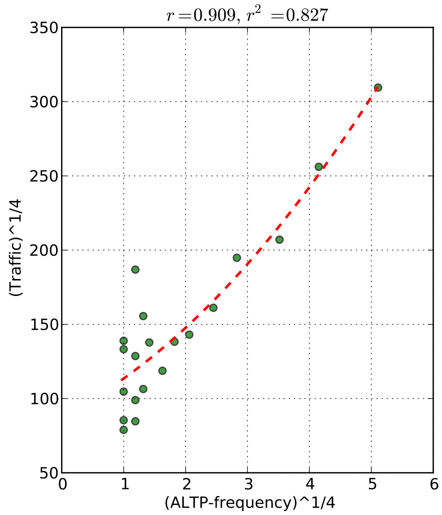


Figure 12: Correlation between ALTP-frequency and traffic volume for ISP for April 2013 with clusters of size 5.

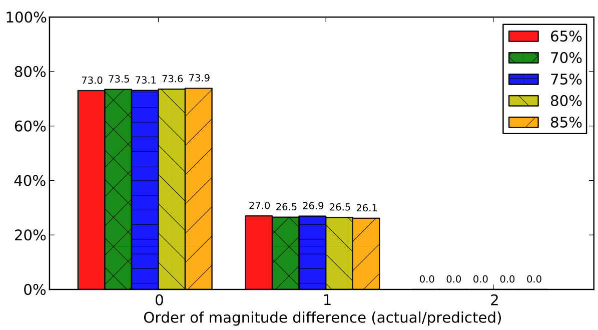


Figure 13: Traffic prediction using ALTP frequency.

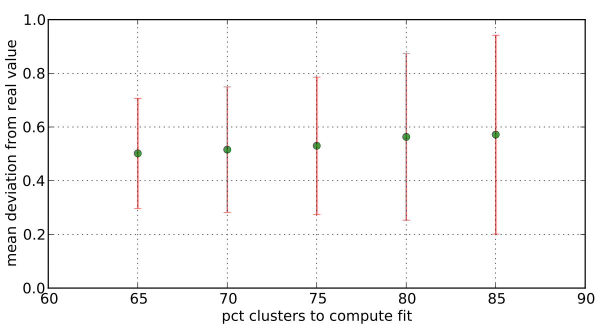


Figure 14: Traffic prediction using ALTP frequency.

For this analysis, we start by ranking the AS-links based on the amount of traffic they carry using our ground-truth traffic. We then select the subset of links identified on our traceroute dataset and rank those links a second time, based on the selected Network Syntax metric. As before, to reduce the potential noise introduced by the ranking of individual AS-links, we cluster them in equal-sized groups (ten in this case). For each group, we compute its ranking for both the relative Network Syntax metric and carried traffic, as the average of the individual rankings of the AS-links within the group. Given the strong correlation between the different Network Syntax metrics and traffic volume, potentially any of the metrics could be used to rank the links. For this analysis we select the two metrics with the highest degree of correlation: connectivity and ALTP-frequency, and compare their results.

Figures 15 and 16 show the correlation between traffic-based ranking and connectivity or ALTP-frequency respectively, for both IXP and IXP. The figures show strong r^2 for all four dataset using both Network Syntax metrics. However ALTP-frequency r^2 values are slightly higher than their connectivity counterpart, with values as high as 0.95 in the case of ISP and 0.75 in the case of IXP. Regardless, the results from this analysis show that using Network Syntax metrics to rank AS-links can be effectively used to rank links based on the amount of traffic they carry.

7. DISCUSSION

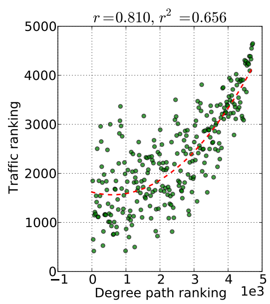
In this section, we elaborate on three critical issues: (i) wether the application of Network Syntax analysis to AS-level connectivity graphs derived from BGP data works as intended, (iii) the robustness of the approach to the known pitfalls of IP-to-AS level mapping for AS topology inference when using traceroute datasets, and (iii) the impact of different traceroute dataset characteristics on the results from Network Syntax metrics.

7.1 BGP-derived connectivity-graphs

As we discussed in Section 3.2, Network Syntax can not be applied to just any AS-level connectivity graph but depends on the information embedded in the graph inferred from traceroute datasets. To illustrate this we apply Network Syntax to the connectivity graph for ISP derived from the subset of AS-level paths contained in the public BGP view [5] for April 2011. Specifically, we extract all the BGP announcements that contain the AS number for ISP, derive their corresponding AS-level paths, and generate the connectivity-graph. We then compute the different Network Syntax metrics and evaluate our findings in the context of the ground-truth traffic data for ISP for same time period. As in Section 5, we first rank the links based on the corresponding metric and create clusters of ten links before examining the correlation.

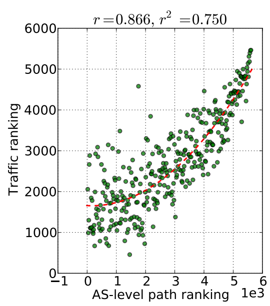
Figure 17 shows the results of our analysis for the subset of 2,016 links identified in the dataset for the connectivity, control value, global choice and integration metrics. Since this analysis is based on AS-level paths extracted from BGP announcements, no traceroute probes were available to compute the ALTP-frequency metric.

The figure shows that, as anticipated, none of the metrics are strongly correlated with traffic volume. In most cases, links are clustered together either on the lower left side of the plot (corresponding to low traffic volume and nearly identical Network Syntax metric) as is the case in Figures 17c, 17b and 17c, or mostly grouped on the right lower side of the plot (which corresponds to high Network Syntax metric and low traffic volume) in the case of 17d. The seemingly moderately strong r^2 values are mostly driven by a few outliers, but it is apparently clear from the figures that the cluster of links are not cleanly distributed around the regression line (in contrast to Figures 10 and 11). This



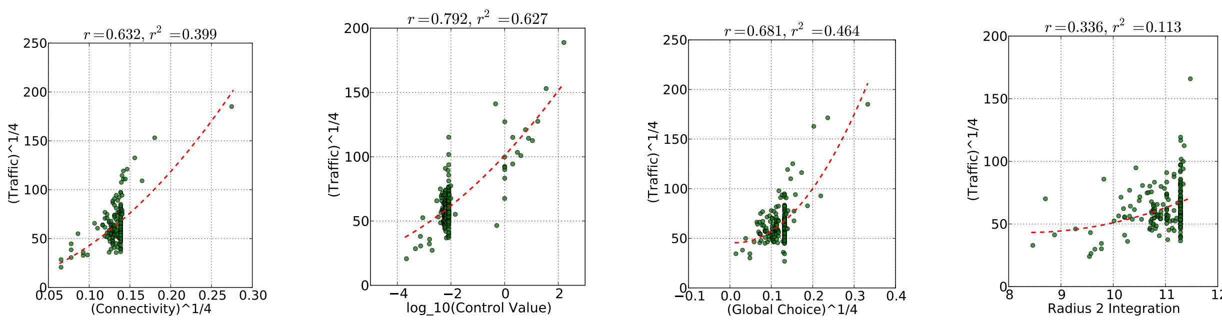
(a) IXP April 2011 (b) IXP April 2013 (c) ISP April 2011 (d) ISP April 2013

Figure 15: Ranking based on connectivity and traffic volume.



(a) IXP April 2011 (b) IXP April 2013 (c) ISP April 2011 (d) ISP April 2013

Figure 16: Ranking based on ALTP-frequency and traffic volume.



(a) Connectivity (b) Control Value (c) Global Choice (d) Integration

Figure 17: Correlation between Network Syntax metrics and traffic volume (BGP dataset) for ISP April 2011.

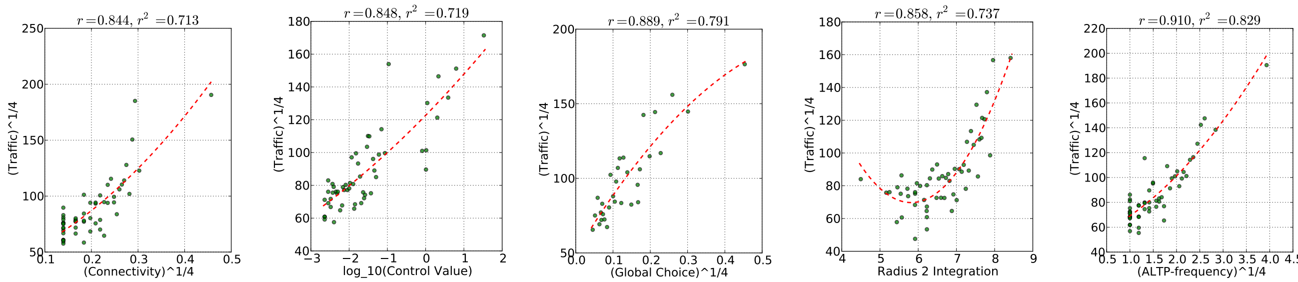
analysis shows that it is the data-plane and not the control-plane perspective that is relevant for Network Syntax.

7.2 Errors in traceroute-to-AS mappings

The pitfalls of IP-to-AS level mapping for AS topology inference are well-known. The common approach of using longest prefix matching to map the routers' IP addresses of a traceroute to AS numbers is known to generate potentially false AS links [49]. Several previous research efforts have studied these pitfalls [15,17,29,36,37] and identified common causes for the mismatch which range from the incompleteness of IP-to-AS mappings gathered from publicly available BGP feeds, to the constraints inherent to the traceroute measurement itself (e.g., routers silently dropping probes or not altering packets' TTL). We correct our datasets to avoid these pitfalls, as described in Section 4.2.

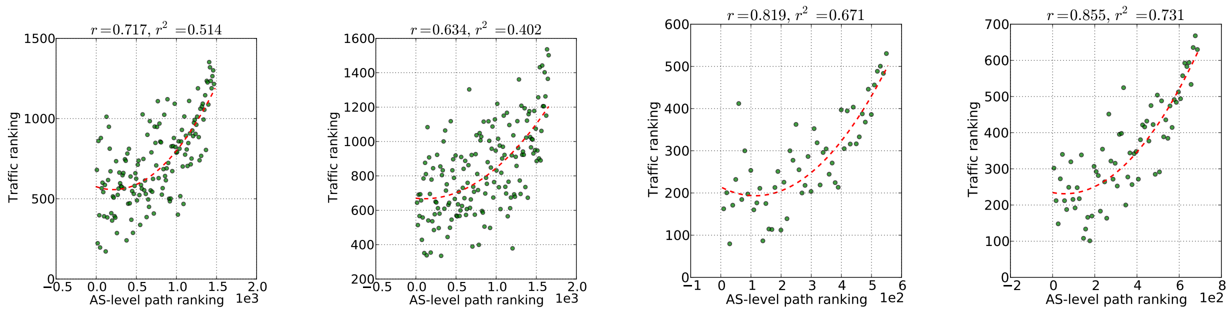
In this section we explore the robustness of Network Syntax when applied to traceroute datasets with some of these well known problems. We do this by computing the different Network Syntax metrics on the un-corrected April 2011 traceroute dataset for ISP and comparing the resulting correlation with the corrected version.

To generate this alternative dataset we apply the same basic traceroute sanitation process described in Section 4.2 but stop before the application of the correction heuristics described in [17], wich we summarize in Table 3. The



(a) Connectivity (b) Control Value (c) Global Choice (d) Integration (e) ALTP-frequency

Figure 18: Correlation between *connectivity*, *control value*, *global choice* , *local integration* metric (integration radius 2) and *ALTP-frequency* with traffic volume for *ISP* for April 2011 using the CAIDA traceroute dataset.



(a) IXP April 2011 (b) IXP April 2013 (c) ISP April 2011 (d) ISP April 2013

Figure 19: Correlation between ALTP-frequency and traffic volume for *IXP* and *IXP* using CAIDA datasets.

False AS links problems
Internet eXchange Points (IXPs)
Sibling ASes
Unannounced IP addresses
Using outgoing interface IPs
Private peering interface IPs

Table 3: Summary of problems within traceroute-inferred AS-level paths addressed by the filtering heuristics proposed in [17].

	Ono April 2011
1. Repeated Hop	19%
2. Unk Hop in Path	47%
3. Unk Src or Dst ASN	5%
4. Path too Short	30%

Table 4: Percentage of probes dropped by each of the extraction rules.

basic sanitation process conservatively discards ≈61% of our initial dataset and reduces the number of probes to 12.9M. Table 4 presents a summary of the fraction of traceroutes dropped by each of our basic sanitation rules.

Table 5 presents a comparison of the r^2 values for the different metrics when Network Syntax is applied to the connectivity graphs that result from both the un-corrected and corrected versions of the dataset. The results show that the application of the different correction heuristics does improve correlation between traffic volume an the different metrics, but most metrics (with the exception of global choice) still show significant r^2 values when computed using the un-corrected dataset.

When comparing both the corrected and un-corrected datasets, we note that the amount of traceroutes modified by the different correction heuristics accounts for ≈16% of the probes; which ultimately map to approximately 6% of the end-to-end AS-level in the dataset. The results from Table 5 highlight the robustness of Network Syntax when identifying popular AS-links in the presence of mis-mapped paths in the underlaying connectivity-graph.

7.3 Other datasets

We have shown that applying Network Syntax metrics to AS-level connectivity graphs using paths extracted from massive traceroute datasets reveals a strong correlation between traffic volume and the different metrics. We now apply the same technique to a different dataset, one

	Un-corrected	Corrected
Connectivity	0.759	0.789
Control Value	0.708	0.759
Global Choice	0.314	0.653
Integration	0.749	0.826
ALTP-frequency	0.919	0.965

Table 5: r^2 values of the different metrics for ISP using the April 2011 dataset.

collected from CAIDA's Ark monitors for the same time periods. This tracerotue dataset consist of probes launched towards randomly selected IP addresses from CAIDA's Ark monitors [12] which probe IP addresses from every routable IPv4 /24 prefix in cycles of approximately 48 hours. For this analysis, we combined data from three different probing cycles completed by different sets of Ark monitors between April 1-7, 2011 and April 1-7, 2013.

Figure 18 shows a comparison of the correlation between traffic volume and the different Network Syntax metrics for the CAIDA 2011 dataset for *ISP*[7]. The observed trends are similar to those seen in Section 5.2, where the amount of traffic carried by the clustered links strongly correlates with the different Network Syntax metrics. Additionally, Figure 19 shows a comparison of the correlation of AS-link ranking based on traffic volume versus ranking based on ALTP-frequency, using the CAIDA dataset for both *IXP* and *ISP*. The same trends as those seen in Section 6.2 can be observed for both network entities, with clusters of links carrying larger amounts of traffic corresponding to higher ALTP-frequency links.

Even though in both cases the correlations can still be observed, the r^2 values are smaller than their Ono-dataset counterpart. We argue that this is due to a fundamental difference on how the underlaying traceroutes were collected. As Table 6 shows, although the tracerotues of both the Ono and CAIDA datasets contain millions of traceroutes launched against a large number of different destination ASes; there is a *2-order of magnitude difference* in the number of source ASes from where the probes were launched.

Even though the correlation is still present, the r^2 values are smaller than their Ono-dataset counterpart. We argue that this is due to a fundamental difference on how the underlaying traceroutes were collected. As Table 6 shows, although the tracerotues of both the Ono and CAIDA datasets contain millions of traceroutes launched against a large number of different destination ASes; there is a *2-order of magnitude difference* in the number of source ASes from where the probes were launched.

As discussed in [34], one consequence of taking measurements using a small number of sources and relying on an end-to-end strategy, is that edges are selected disproportionately, so bias arises when edges incident to a node in the underlying graph are sampled disproportionately. Thus, an edge is much more likely to be visible if it is close to the vantage point that discovered them. To explore this potential issue, we focus our analysis on the CAIDA 2011 dataset and look at the discovered AS-links as the intersection of the number of probes that crossed through them vs the corresponding number of ALTPs that contain it. We concentrate on the top AS-link with the largest number of ALTPs which is seen by almost 10K ALTPs and was discovered by almost half a million probes. Closer inspection shows that the link connects AS195 (San Diego Supercomputer Center) and AS2152 (the California State University Network) at SD-NAP (an IXP located in San Diego, CA), and that all the probes responsible for discovering this link correspond to a CAIDA Ark monitor placed in the San Diego Supercomputer Center. A similar scenario was observed for the other top three links. This highlights that the location of the vantage points can lead to erroneous inferences about a link's popularity.

[7]Similar trends were observed for both *IXP* and *ISP* using CAIDA's 2011 and 2013 datasets.

Dataset	Unique VPs	Src ASes	Dst ASes	Probes
CAIDA 2011	53	52	36,034	26.9M
CAIDA 2013	65	60	42,440	48.9M

Table 6: Number of unique vantage points, unique source ASes and probes for each dataset.

8. CONCLUSIONS AND FUTURE WORK

We advance the state-of-the-art in traffic characterization by presenting a novel technique to infer traffic volumes from AS-level routing graphs carved out by massive traceroute campaigns. Our Network Syntax approach builds on the observation that the popularity of a route on the Internet can serve as an informative proxy for the volume of traffic it carries. Drawing analogies with city grids and traffic, Network Syntax applies structural analysis and metrics to predict with high accuracy the inter-domain traffic volume carried by different links. We demonstrated the effectiveness of our approach using two months of data (collected two years apart) from a Tier-1 Internet Service Provider and a large Internet eXchange Point, by identifying traffic-critical links and inferring missing traffic matrix measurements.

We evaluated four different publicly available traceroute datasets, but selected two for inclusion (due to space constraints) representing different standpoints with respect to type and location of vantage points. Multiple other datasets collected over the years could be leveraged by our technique, such as RIPE's Atlas project [4], or those collected from the DIMES project [43].

Going forward, Network Syntax opens a rich research agenda from specific methodological aspects (e.g., different metrics and datasets attributes) to applications of a better-understood flow of Internet traffic. Could we identify high-traffic links in the context of arbitrary AS-links on the Internet, i.e. can we establish the relative importance of a pair AS-links not tied to a specific network? This could be used, for instance, to augment existing AS-topology maps with the relative importance of links based on traffic carried. Could Network Syntax, perhaps in combination with existing techniques, be used to complete partial-traffic matrixes? Given our new ability to leverage long-available traceroute datasets, what could this approach tell us about the variability and evolution of the Internet over time?

Acknowledgements

We would like to thank our shepherd, Paul Barford, and the anonymous reviewers for their valuable feedback and assistance. Georgios Smaragdakis was supported by the EU Marie Curie International Outgoing Fellowship "CDN-H" (PEOPLE-628441). This work was supported in part by the National Science Foundation through Awards CNS 0644062, CNS 0917233 and CNS 0855253 and by a generous Google Faculty Research Award.

9. REFERENCES

[1] European Internet Exchange Association. `https://www.euro-ix.net/`.

[2] Packet Clearing House. https://www.pch.net/.

[3] PeeringDB. https://www.peeringdb.com/.

[4] Ripe. http://www.ripe.net/project/ris.

[5] University of Oregon Route Views Project. http://www.routeviews.org/.

[6] Adhikari, V. K., Jain, S., and Zhang, Z.-L. Youtube traffic dynamics and its interplay with a Tier-1 ISP: an ISP perspective. In *Proc. of IMC* (2010).

[7] Ager, B., Chatzis, N., Feldmann, A., Sarrar, N., Uhlig, S., and Willinger, W. Anatomy of a large european IXP. In *Proc. of ACM SIGCOMM* (2012).

[8] Ahmad, M. Z., and Guha, R. Understanding the impact of Internet eXchange Points on Internet topology and routing performance. In *Proc. of the ACM CoNEXT Student Workshop* (2010).

[9] Atkinson, A. *Plots, transformations, and regression.* Oxford science publications. Clarendon Press, Oxford, 1985.

[10] Augustin, B., Krishnamurthy, B., and Willinger, W. IXPs: Mapped? In *Proc. of IMC* (2009).

[11] Bharti, V., Kankar, P., Setia, L., Gürsun, G., Lakhina, A., and Crovella, M. Inferring invisible traffic. In *Proc. of ACM CoNEXT* (2010).

[12] CAIDA. The CAIDA UCSD IPv4 Routed /24 Topology Dataset - 1-7 April 2011 and 1-7 April 2013. http://www.caida.org/data/active/ipv4_routed_24_topology_dataset.xml.

[13] Cardona Restrepo, J. C., and Stanojevic, R. A History of an Internet eXchange Point. *SIGCOMM Comput. Commun. Rev. 42*, 2 (2012), 58–64.

[14] Chang, H., Jamin, S., Morley, Z., and Willinger, M. W. An empirical approach to modeling Inter-AS traffic matrices. In *Proc. of IMC* (2005).

[15] Chang, H., Jamin, S., and Willinger, W. Inferring AS-level internet topology from router-level path traces. In *Proc. of SPIE ITCom* (2001).

[16] Chatzis, N., Smaragdakis, G., Boettger, J., Krenc, T., and Feldmann, A. On the benefits of using a large IXP as an Internet vantage point. In *Proc. of IMC*.

[17] Chen, K., Choffnes, D. R., Potharaju, R., Chen, Y., Bustamante, F. E., Pei, D., and Zhao, Y. Where the sidewalk ends: Extending the Internet AS graph using traceroutes from P2P users. In *Proc. of ACM CoNEXT* (2009).

[18] Choffnes, D. R., and Bustamante, F. E. Taming the torrent: A practical approach to reducing cross-ISP traffic in peer-to-peer systems. In *Proc. of ACM SIGCOMM* (2008).

[19] Dhamdhere, A., and Dovrolis, C. Ten years in the evolution of the Internet ecosystem. In *Proc. of IMC* (2008).

[20] Fang, W., and Peterson, L. Inter-AS traffic patterns and their implications. In *Global Telecommunications Conference, 1999. GLOBECOM'99* (1999).

[21] Feldmann, A., Greenberg, A., Lund, C., Reingold, N., Rexford, J., and True, F. Deriving traffic demands for operational IP networks: Methodology and experience. *IEEE/ACM Transactions on Networking (ToN)* (2001).

[22] Feldmann, A., Kammenhuber, N., Maennel, O., Maggs, B., De Prisco, R., and Sundaram, R. A methodology for estimating interdomain web traffic demand. In *Proc. of IMC* (2004).

[23] Gao, L. On inferring autonomous system relationships in the internet. *IEEE/ACM TON 9*, 6 (2001).

[24] Gursun, G., and Crovella, M. On traffic matrix completion in the internet. In *Proc. of IMC* (2012).

[25] Hillier, B., and Hanson, J. *The Social Logic of Space*. Cambridge University Press, 1984.

[26] Hillier, B., Penn, A., Hanson, J., Grajewski, T., and Xu, J. Natural movement: or, configuration and attraction in urban pedestrian movement Environment and Planning B: Planning and Design, 1994.

[27] Hillier, B., and Sahbaz, O. High resolution analysis of crime patterns in urban street networks: an initial statistical sketch from an ongoing study of a London borough. In *Space Syntax Symposium* (2005), p. 451âĂŞ478.

[28] Howell, D. *Statistical Methods for Psychology.* Thomson Wadsworth, 2007.

[29] Hyun, Y., Broido, A., and claffy, k. On Third-party Addresses in Traceroute Paths. In *Proc. of PAM* (2003).

[30] InMon. Inmon sFlow. http://sflow.org.

[31] Jiang, B. A space syntax approach to spatial cognition in urban environments. In *Workshop on Cognitive Models of Dynamic Phenomena and Their Representations* (1998).

[32] Kostakos, V. Space Syntax and Pervasive Systems. *Geospatial Analysis and Modeling of Urban Structure and Dynamics* (2010), 21–52.

[33] Labovitz, C., lekel Johnson, S., Oberheide, J., and Jahanian, F. Internet inter-domain traffic. In *Proc. of ACM SIGCOMM* (2010).

[34] Lakhina, A., Byers, J. W., Crovella, M., and Xie, P. Sampling biases in IP topology measurements. In *Proc. Joint Conference of the IEEE Computer and Communications Societies* (2003).

[35] Mahajan, R., Spring, N., Wetherall, D., and Anderson, T. Inferring link weights using end-to-end measurements. In *Proc. ACM IMW* (2002).

[36] Mao, Z. M., Johnson, D., Rexford, J., Wang, J., and Katz, R. H. Scalable and accurate identification of AS-level forwarding paths. In *INFOCOM* (2004).

[37] Mao, Z. M., Rexford, J., Wang, J., and Katz, R. H. Towards an accurate as-level traceroute tool. In *Proc. of ACM SIGCOMM* (2003).

[38] Medina, A., Taft, N., Salamatian, K., Bhattacharyya, S., and Diot, C. Traffic matrix estimation: existing techniques and new directions. In *Proc. of ACM SIGCOMM* (2002).

[39] Nucci, A., Sridharan, A., and Taft, N. The problem of synthetically generating IP traffic matrices: Initial recommendations.

[40] Oliveira, R., Z. B., and Zhang, L. Observing the evolution of Internet AS topology. In *Proc. of ACM SIGCOMM* (2007).

[41] Penn, A., Hillier, B., Banister, D., and Xu, J. Configurational modelling of urban movement networks. *Environment and Planning B 25* (1998).

[42] Peterson, L. Inter-AS traffic patterns and their implications. In *in Proc. IEEE GLOBECOM* (1999).

[43] Shavitt, Y., and Shir, E. DIMES: Let the Internet measure itself. *ACM SIGCOMM Computer Communication Review 35*, 5 (October 2005).

[44] Tabachnick, B. G., and Fidell, L. S. *Using Multivariate Statistics (5th Edition)*. Allyn & Bacon, Inc., Needham Heights, MA, USA, 2006.

[45] The PREDICT Repository. Protected repository for the defense of infrastructure against cyber threats. `https://www.predict.org/`, Aug 2014.

[46] Uhlig, S., and Bonaventure, O. Implications of interdomain traffic characteristics on traffic engineering. *European Transactions on Telecommunications* (2002).

[47] Willinger, W., Alderson, D., and Doyle, J. C. Mathematics and the Internet: A Source of Enormous Confusion and Great Potential. *Notices of the AMS 56*, 5 (May 2009).

[48] Xu, K., Duan, Z., Zhang, Z.-L., and Chandrashekar, J. On properties of Internet eXchange Points and their impact on as topology and relationship. *Networking 3042* (2004).

[49] Zhang, Y., Oliveira, R. V., Zhang, H., and Zhang, L. Quantifying the pitfalls of traceroute in as connectivity inference. In *PAM* (2010), A. Krishnamurthy and B. Plattner, Eds., vol. 6032, Springer, pp. 91–100.

[50] Zhang, Y., Roughan, M., Duffield, N., and Greenberg, A. Fast accurate computation of large-scale ip traffic matrices from link loads. In *Proc. of ACM SIGMETRICS* (2003).

[51] Zhang, Y., Roughan, M., Lund, C., and Donoho, D. L. Estimating point-to-point and point-to-multipoint traffic matrices: An information-theoretic approach. *IEEE/ACM Trans. Netw.* (2005).

[52] Zhang, Y., Roughan, M., Willinger, W., and Qiu, L. Spatio-temporal compressive sensing and internet traffic matrices. In *Proc. of ACM SIGCOMM* (2009).