Public Review for

Looking for Hypergiants in PeeringDB

T. Bottger, F. Cuadrado, S. Uhlig

In 2009 Arbor Networks used the term "Hyper Giants" to refer to those large companies (about 30) that at that time generated and consumed a disproportionate 30% of all Internet traffic. In this paper, the authors highlight that – despite their influence in the evolution of Internet topology, infrastructure, traffic, etc. – the research community still lacks a clear definition of "hypergiants" and, more importantly, a precise understanding of what are the characteristics that separate them from other organizations operating on the Internet.

The authors analyze PeeringDB data to extract characteristics of organizations taking part in public traffic exchange at IXPs and identify a set of properties that are distinctive of hypergiants (i.e., potentially not exhaustive to provide a definition but useful to improve our understanding of the role hypergiants play in the Internet). They then explore how these organizations reach the global IPv4 space through the IXP ecosystem.

Reviewers found the topic very relevant and appreciated the effort to come up with a characterization approach that would expose interesting distinctive properties. They found that the use the authors made of unsupervised clustering was effective given that there is no precise definition of a hypergiant and some of them argued that they are not convinced that it would be possible to obtain a crisp definition at all. Reviewers also felt that more can probably be discovered by manually analyzing the characteristics of the organizations identified and the discriminating power of the features used in the clustering. The authors share the data snapshots and the code used for this paper in the hope to stimulate further research, which is to be applauded.

Public review written by
Alberto Dainotti
CAIDA

Looking for Hypergiants in PeeringDB

Timm Böttger
Queen Mary University of London
timm.boettger@qmul.ac.uk

Felix Cuadrado
Queen Mary University of London
felix.cuadrado@qmul.ac.uk

Steve Uhlig
Queen Mary University of London
steve.uhlig@qmul.ac.uk

ABSTRACT

Hypergiants, such as Google or Netflix, are important organisations in the Internet ecosystem, due to their sheer impact in terms of traffic volume exchanged. However, beyond naming specific instances, the research community still lacks a sufficiently crisp understanding of them. In this paper we analyse PeeringDB data and identify features that differentiate hypergiants from the other organisations. To this end, we first characterise the organisations present in PeeringDB, allowing us to identify discriminating properties of these organisations. We then use these properties to separate the data in two clusters, differentiating hypergiants from other organisations. We conclude this paper by investigating how hypergiants and other organisations exploit the IXP ecosystem to reach the global IPv4 space.

CCS CONCEPTS

• **Networks** → **Public Internet**; *Very long-range networks*; Network performance analysis;

KEYWORDS

Hypergiants, PeeringDB, Internet eXchange Points

1 INTRODUCTION

The Internet research community has commonly accepted that a significant fraction of today's Internet traffic relates to so-called hypergiants like YouTube or Netflix [5]. While their importance is known for some time now, the research community still falls short of a definition of hypergiants. Most evidence on their behaviour and existence is anecdotal, or self-reported but lacking sufficient detail [3, 11, 12]. The current way the community understands and defines hypergiants is mainly by naming examples, which we believe is unsatisfactory. This is surprising as hypergiants not only are a massive source of traffic, but they also are believed to be one of the driving forces behind the observed flattening of the Internet hierarchy. The reason for the observed flattening indeed is their approach to peering, reaching customers via direct peering links instead of using and paying transit providers. The amount of traffic they carry is so significant that it has shifted traffic away from the traditional hierarchy of the Internet, and thus asked the research community to revisit their mental model of the Internet [5].

To obtain a better understanding of the role hypergiants play in the Internet, we first analyse PeeringDB data to get a better understanding of the *organisations* taking part in public traffic exchange at IXPs. We then use the results of this analysis to identify features from the data available in PeeringDB to differentiate hypergiants from other organisations.

In this paper we make the following contributions:

(1) We characterise the *organisations* in PeeringDB, looking at several features: their geographical scope, provisioned port capacity and potential reach.
(2) We exploit a natural split in the data across those features to differentiate hypergiants from other *organisations*.
(3) We then explore how these hypergiants and other organisations reach the global IPv4 space through the IXP ecosystem.

Code and Data sharing. We make the PeeringDB data snapshot and code used for this paper available to the research community, in the hope that this stimulates and facilitates further research.

2 PEERINGDB DATA SET OVERVIEW

PeeringDB[1] curates data to facilitate the exchange of information related to peering, by letting *organisations* and IXPs advertise themselves. In this paper we use the term *organisation* to refer to an entity participating in traffic exchange through the public Internet, with a record on PeeringDB. Google, Netflix and Yahoo are examples for *organisations*.

As of writing this paper, more than 600 IXPs and more than 10,000 *organisations* are present in PeeringDB. However, only 6,910 of them have at least one presence at a public IXP recorded. Refer to Table 1 for more details.

Data in PeeringDB is voluntarily reported by *organisations* and IXPs. Despite the self-reported nature of this data, making it potentially unreliable, its public nature and popularity for the Internet peering ecosystem guarantees that significant scrutiny is applied to it. Therefore, as already established by previous work [7], we argue that the biases have to be comparably small and the data set is thus reliable enough to allow us to derive insights into the peering ecosystem for two reasons. Firstly, it has a very good standing in the network operators community, which naturally has very big interest in having reliable peering information available. Some of the biggest, and arguably most important *organisations* (e.g., Google, Netflix or Cloudflare), rely on PeeringDB. The first two refer to PeeringDB as authoritative and sole information source regarding peering capabilities[2,3]. Cloudflare even automatically provisions their configuration from PeeringDB, expressing a high level of trust in PeeringDB's data[4]. Furthermore, PeeringDB is sponsored by a multitude of large *organisations* (e.g., Facebook, Microsoft, Akamai), stressing the importance and usefulness of it for their network operations. Secondly, recent studies have found that PeeringDB data is consistent with BGP derived information [7] as well as with other publicly available data sources on IXPs [4]. In this paper, we are thus going to treat data from PeeringDB as a ground-truth for our analysis.

[1] https://www.peeringdb.com
[2] https://peering.google.com/#/infrastructure
[3] https://openconnect.netflix.com/en/peering-locations/
[4] https://www.peeringdb.com/asn/13335

Entity	Count
IXPs	643
Organisations (total)	11,918
Organisations (at IXPs)	6,910
AS numbers (total)	11,596
AS numbers (at IXPs)	7,171

Table 1: Number of entities listed in PeeringDB.

The data snapshot used in this paper was retrieved on Jan 10, 2018 through the PeeringDB REST API.

3 HYPERGIANTS OF THE INTERNET

In this section we dig into the PeeringDB data to identify characteristics that differentiate today's hypergiants from other *organisations*.

The challenge we face is that there is neither a definition of a hypergiant nor an authoritative list of hypergiants that could be used as ground truth. Despite this limitation of the current state-of-the-art, we will nevertheless attempt to reach a set of sufficiently convincing characteristics by relying on the fact that hypergiants are the largest among today's Internet *organisations*. This should make them significantly different from the majority of networks, at least across some metrics. If they were identical on all metrics, then it would either mean the data we rely on is not appropriate, or that they surprisingly are not actually the largest among today's Internet *organisations*, which would then highly question the very term hypergiant originally coined in [5].

To identify such metrics, we first look at the port capacity, geographic footprint and traffic profiles of all *organisations* participating in the public peering landscape. We then combine these three dimensions into a single, more coherent picture and employ a clustering algorithm to separate possible hypergiants and other *organisations*. The cluster made of these *organisations* matching what is expected from hypergiants will therefore be considered as the current set of hypergiants, as visible through the PeeringDB data snapshot we used.

3.1 The Peering Landscape

In this subsection we use 3 dimensions to obtain a first characterisation of Internet *organisations*: port capacity, geographic footprint and traffic profile. Port capacity is a proxy metric for the actual amount of traffic, being a likely higher bound on the actual amount of traffic exchanged. Geographic footprint reflects the geographic deployment of *organisations*. Traffic profile finally is the self-declared directionality of the traffic by *organisations*, which differs strongly between eyeballs, transit networks, and content-heavy players of the Internet ecosystem.

Port capacity. Based on PeeringDB data, we extract for each *organisation* the IXPs it is present at, along with the corresponding router port sizes. We then sum up those port sizes to obtain the aggregated provisioned port capacity. The total aggregate port capacity across the data amounts to 290 Tbps, with an average port capacity of 40.45 Gbps per *organisation*.

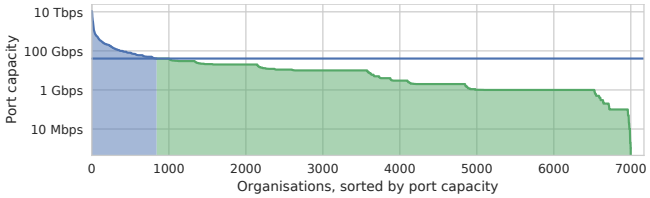

Figure 1: Total provisioned IXP port capacity for each *organisation*. The horizontal line depicts the average port capacity (40.45 Gbps). *Organisations* having more than the average capacity provisioned are depicted in blue, *organisations* having less than average in green. Note the log-scale of the y-axis.

One would naturally expect that hypergiants should be amongst the *organisations* with the highest provisioned port capacity. Figure 1 shows the port capacity provisioned by each *organisation* present in PeeringDB. The distribution of provisioned port capacity is strongly non-uniform, with an average of 40.45 Gbps but a standard deviation of 311.6 Gbps. Figure 1 exposes that a few *organisations* are responsible for a significant, way above-average port capacity (blue bars), while the overwhelming majority of them declares below-average capacity (green bars). The top five largest *organisations* represent 16.3% of the total port capacity, the top 80 covering half. In contrast with these massive *organisations*, the majority provisions significantly less port capacity. This result is in line with the traffic figures from [5].

Geographic footprint. We now turn to the geographic footprint, by looking at the number of continents[5] where an *organisation* is present at IXPs. We expect that hypergiants will aim to have wide, if not global, geographic presence, publicly exchanging data in IXPs across multiple continents. Figure 2 shows in its columns the distribution of continent presence in the PeeringDB data set. The largest share of *organisations* (6,623) are present on only one continent. On the other hand, there are only 89 with presence across four or more continents. Among those 89 *organisations* with most port capacity, the top ones, Apple (4 continents), Twitch (5 continents), Amazon (6 continents) and Google (7 continents), are strong hypergiant candidates.

Traffic profile. Organisations do not only differ in their geographic footprint and total port capacity, but also in their purpose and thus traffic profile. Some, such as content providers, are expected to have a predominantly outbound traffic profile, whereas ISPs connecting eyeballs to the Internet are expected to have an inbound traffic profile. PeeringDB defines five different profiles ranging from (Heavy) Inbound to (Heavy) Outbound, with Balanced in the middle. *Organisations* not wishing to expose their traffic profile have the option 'Not Disclosed' as well.[6] Figure 2 shows in its rows the traffic profiles of the *organisations* in the data set. Besides a small fraction who hide their traffic profile, we see that the majority are inbound oriented or balanced, likely referring to eyeballs and

[5] PeeringDB recognises the following continents: Africa, Asia Pacific, Australia, Europe, Middle East, North America and South America. We adopt this non-textbook definition of a continent to maintain comparability to other works using PeeringDB data.
[6] A few *organisations* chose to leave the corresponding database field empty. We treat these the same as 'Not Disclosed'.

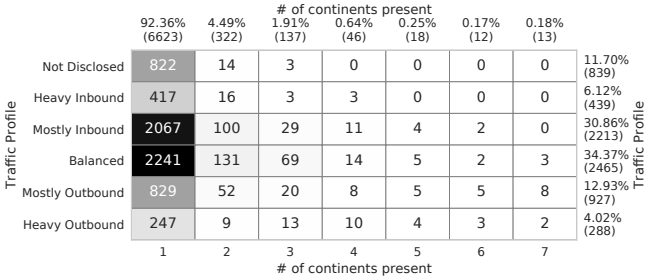

Figure 2: Traffic profile and continent coverage for each *organisation*. Continent presence means an *organisation* is present at at least one IXP of this continent. Secondary axes show the distributions of *organisations* across traffic profiles and continent presence.

transit networks respectively. There are more than twice as many *organisations* with an inbound traffic profile than with an outbound profile. Among the 288 *organisations* with a Heavy Outbound profile, the majority (247) have presence in a single continent, making them very unlikely to be a hypergiant. Only 41 *organisations* declare a Heavy Outbound profile and are present at multiple continents. 139 *organisations* declare an Outbound (Heavy or not) profile and are present at multiple continents. These numbers suggest that finding hypergiants among sufficiently large *organisations* that have extensive footprint should be possible.

3.2 The Whole Picture

After having discussed the three dimensions in isolation, we now put them together to obtain a more comprehensive picture of the *organisations* participating in the peering ecosystem, as seen from PeeringDB. Figure 3 shows a tree-map combining the three dimensions: continent presence, traffic profile, and aggregate port capacity. In this tree-map, the area of each rectangle is proportional to the aggregated port capacity it represents. *Organisations* are first grouped by number of continents (one to seven) at which they maintain IXP presence, enclosed by a white border. The on-print shows the number of continents of each group, and the aggregate port capacity of all its members. Each group is then subdivided by the traffic profiles of the group's *organisations*.

First, we observe that *organisations* present at a single continent account for 45% of the overall port capacity. The remaining capacity is spread almost evenly across the other groups in terms of continent presence, with between 8-13% for each group, except for the group of five continents that has only 3%. While 92% of all *organisations* are present at a single continent, they are only responsible for 45% of the total provisioned port capacity. In contrast, the 1% of them with presence on four continents or more are responsible for 38% of provisioned port capacity. This implies that the many *organisations* with a local geographic scope tend to have little port capacity (hence little expected traffic) at IXPs. In contrast, there are a few with large geographic scope, combined with large port capacity (hence large expected traffic) at IXPs.

Second, within each group of *organisations* in terms of continent presence, their composition differs in terms traffic profile. Within the single continent group, more than 75% of the port capacity

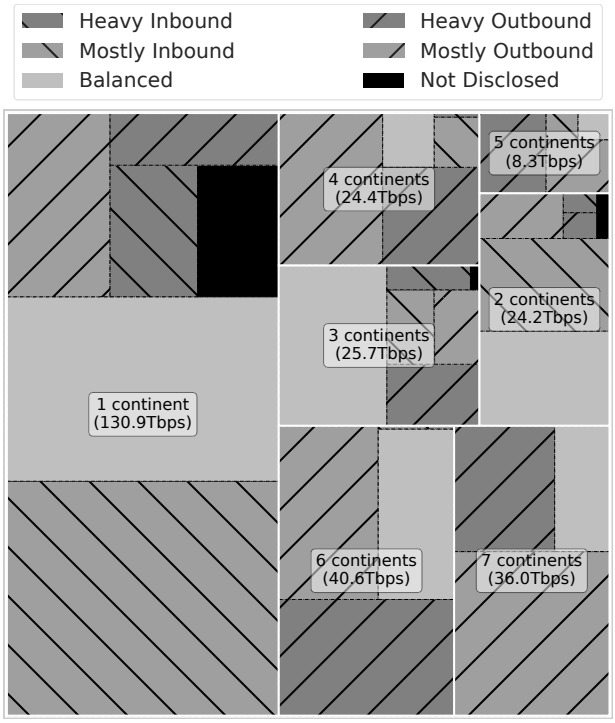

Figure 3: Distribution of aggregated port sizes over traffic profiles and continent presence. An *organisation* is present on a continent if it is present at an IXP at this continent. The area of each rectangle is proportional to the aggregated port size it represents. *Organisations* are grouped by number of continents and then by traffic profile. The on-print depicts the number of continents *organisations* are present at and aggregated port size of the *organisations* in each group.

belongs to balanced (30.6%) or inbound dominant (45.9%) *organisations*. Among the *organisations* in this group with an outbound traffic profile, we find content and hosting providers with a local audience, like BBC, Hetzner, Strato, VKontakte and Baidu.

Looking at the groups with presence in multiple continents, we see a smaller contribution from inbound traffic profiles to the total port capacity. While inbound dominant *organisations* still have a notable share in the groups of two, three and four continents, they play no role in the groups of five, six or seven continents. In those groups, *organisations* with an outbound traffic profile are dominant. Balanced *organisations* with presence at four or more continents are those with a data-centric business model, that do not only deliver but also consume content, such as Dropbox, Amazon (AWS), Hurricane Electric and Microsoft.

In this subsection, we have seen how a relatively small group of global *organisations* gather a substantial amount of port capacity. Moreover, they mostly declare an outbound or balanced traffic profile. This is expected, as large content providers strive to deliver their content to a global audience of end customers. Based on what we observe in this section, content providers rely on a wide IXP presence to serve traffic to the eyeball *organisations* that operate smaller networks with a local footprint and have an inbound traffic profile. Further, this strong concentration of port capacity strongly

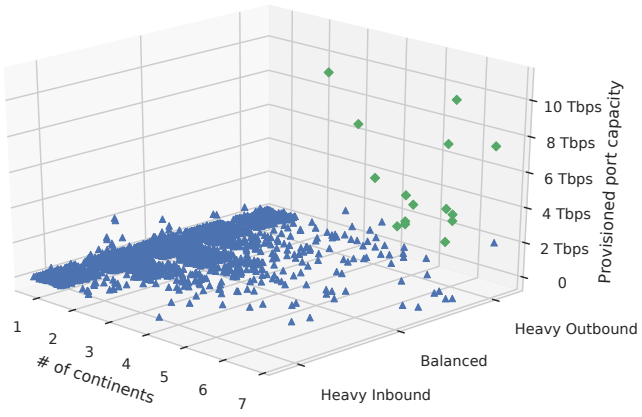

Figure 4: Results of the k-means clustering. *Organisations* in the smaller group are depicted by the green diamonds, those in the other group by the blue triangles. We added a small jitter on the x- and y-axis to make markers easier to discern.

hints at hypergiants, which are quite likely to be in this small group of global *organisations*.

3.3 Hypergiants of the Internet

Depending on their business model, hypergiants should exhibit different characteristics. Intuitively, content hypergiants are expected to be heavy on (outbound) traffic, with a large geographic reach to cater for a world-wide customer base. Cloud hypergiants will have similar characteristics, however their traffic profile might be more balanced. In general, we expect hypergiants to fall within the group of *organisations* with an outbound or balanced traffic profile and presence on many continents. In the following, we will try to identify a small subset of *organisations* fulfilling these characteristics, while being significantly different from the remaining ones.

We use the k-means algorithm [6] to split the *organisations* from the data set in two clusters, expecting that hypergiants are different enough to form a cluster on their own. We first use the k-means algorithm as provided by the Python module scikit-learn [10] with default values for all parameters except for the number of clusters, which we set to two. Data is preprocessed and normalised using scikit-learn's RobustScaler. The clustering yields one cluster with 15 *organisations* and a second cluster that contains the remaining 7,156. Figure 4 visualises the resulting clustering, with the blue triangular markers depicting the larger cluster, and the green diamond shaped markers depicting the smaller cluster. We added a small jitter on the x- and y-axis to make markers easier to discern.

To ascertain that this split in the data set is not an artifact of the clustering method we employed, we apply further clustering and outlier detection algorithms to the data set. In contrast to k-means, these methods do not directly cluster the data, merely providing a score for each data point. A threshold is required to translate this scoring into a clustering, but choosing the threshold value is difficult, since we do not know a priori how many outliers should be expected and there is no ground truth to be compared against.

We show results for principal component analysis (PCA) and the k-nearest neighbour algorithm (k-NN). For PCA, we chose to

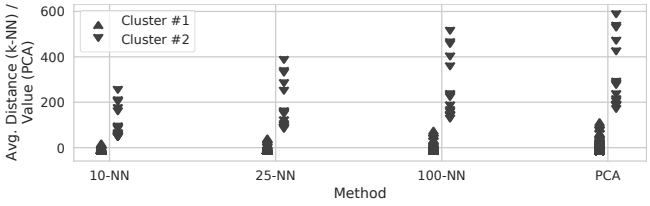

Figure 5: Data point scores assigned by k-NN and PCA. We plot data point with small horizontal offsets representing the k-means clustering we derived previously. For each method the scores assigned to cluster #2 are always higher than the scores for datapoint in cluster #1.

reduce to one principal component only and use the resulting values directly as scores, since this dimension alone captures more than 99.5% of the variance. For k-NN, we use the average distance to the 10, 25 and 100 nearest neighbours. Data was preprocessed in the same way as for the clustering. The resulting scores are shown in Figure 5. We marked data points by the original clustering obtained through k-means, and offset the two groups in the figure to ease distinguishing the clusters. We observe that the resulting scores are consistent with the original k-means clustering: the top 15 scores in every ranking belong to the data points from the small cluster. This makes it possible to choose a threshold such that the resulting clustering is identical to the one obtained through k-means.

Before looking at the resulting clustering in more detail, we first assess the robustness of the clustering against misrepresentations of port capacity values in PeeringDB. We first fix the above mentioned clustering as reference. We then add normal distributed multiplicative random noise to the port capacity and rerun the k-means algorithm. We choose a mean of 1 and standard deviations of 0.01, 0.05 and 0.1 for the noise. We run 1,000 iterations of drawing noise and running k-means for all three standard deviation values. We obtain the exact same clustering in more than 95%, 60% and resp. 30% of all runs. However, in all cases, even with 10% of noise, we end up clustering together at least 10 of the above identified *organisations* in almost 90% of all runs. Given that a 10% deviation in port capacity is multiple 100Gbps for the biggest *organisations*, we conclude that the clustering approach is relatively robust against misrepresentations of port capacity in PeeringDB.

We now take a closer look at the resulting clustering. The 15 *organisations* in the smaller cluster represent only 0.2% of all *organisations*, yet they account for more than 30% of the provisioned port capacity. This smaller group does not only accumulate a disproportionately large share of port capacity, but all its members are also present on at least four continents and have either a heavy outbound or balanced traffic profile. This is in contrast with the other cluster, whose members are on average present on one continent only, and most of them have an inbound (2,652) or balanced (2,463) traffic profile. This suggests that the smaller cluster captures *organisations* fulfilling the expectation we have for hypergiants.

Table 2 lists the 15 *organisations* that were singled out by the clustering algorithm. These fifteen networks indeed typically are considered to be hypergiants; Google, Akamai, Microsoft and Limelight are also explicitly mentioned as hypergiants in [5]. We explain the absence of Tier-1 ISPs in this list by their typical reluctance

to participate in public peering, while our data obtained through PeeringDB focusses on such peerings.

Since these 15 *organisations* are naturally separable from the remaining data set, we conclude that, given the dataset we used, these are the hypergiants in the Internet, at the time the dataset was taken.

4 THE REACH OF HYPERGIANTS

So far, our focus has been on the specific information present in PeeringDB, in a way that would help us identify hypergiants. We found out that the geographic presence was a strong aspect differentiating aspect. Combined with the traffic profile and port capacity, this led to a ranking of *organisations* on PeeringDB that exposes hypergiants.

Now, we slightly shift the focus onto IXPs: we ask how hypergiants rely on IXPs to build their interconnection footprint. More specifically, we would like to answer: what it is that hypergiants are looking for with their IXP presence?

Quite naturally, a hypergiant should have a strong interest to reach eyeball IP address space, as they have built their business model around providing services to end users. While this might be less critical to cloud hypergiants that are more focused on hosting networked applications and services, this is definitely very important to content hypergiants like Netflix, who generate their revenue through end-users paying for their services.

We define the *potential reach* of an *organisation* as the number of potentially reachable IP addresses through its IXP presence, by peering with the other *organisations* also present at the same IXPs. To compute this metric, we combine the IXP membership information from PeeringDB with Routeviews routing information and customer cones from CAIDA [8]. For every *organisation*, we extract all the IXPs it is present at, and then for each IXP extract all the ASes present. We then use the routing information and customer cones to map ASes to customers and IPv4 prefixes, and then calculate the number of unique IPs covered by those prefixes.

The boxplots in Figure 6 show the distribution of potential reach among the *organisations*. Whenever IPs are reachable through members with different peering policies, we assume they are reached through the peer with the most open peering policy, i.e., open > selective > restrictive. In all boxplots, the whiskers indicate the full range of the data. The left figure shows the number of directly reachable IPs of *organisations*, by peering at the IXPs they are present with all the other members. We only consider IP space of the IXP members, not the customer cones. The center figure shows the potential reach, by peering with all other IXP members, assuming these members give full access to their customer cone. The right figure show the potential reach, by peering with all other IXP members, also considering peering policies of the other members. We assume that members with an open policy give access to the full customer cone; members with a selective policy to 66.6% of it; and members with a restrictive policy to only 33.3% of it. When no peering policy is stated, we assume access to 50% of the customer cone.

When we focus on reachable IP addresses for the 15 hypergiants previously identified by k-means clustering, we observe that they can indeed reach a significant amount of the address space. In all

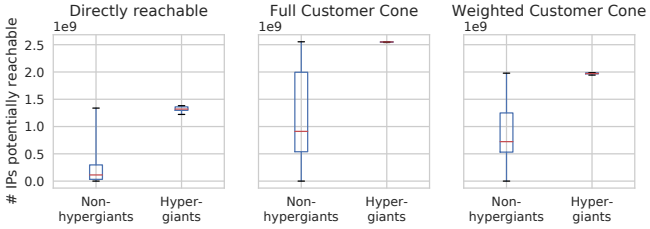

Figure 6: Potentially reachable IP space through peerings at IXPs. Whiskers show full range of the data.

three cases we consider, all of them are among those *organisations* with the highest reach. While we observe that the majority of non-hypergiants have a smaller reach than the identified hypergiants, some *organisations* that are not clustered as hypergiant also have a similar reach. From these boxplots, we can conclude that hypergiants are in the set of *organisations* with the biggest reach, while at the same time there also are other *organisations* with a comparable reach. Additional aspects, such as port capacity (or traffic if available), and footprint are necessary to differentiate hypergiants from other organisations.

5 DISCUSSION

Hypergiants. In this paper we focused on coming up with a set of characteristics that allows to identify the hypergiants coined by Labovitz et al. [5]. Because we relied on global reachability as seen through PeeringDB as a way to find these hypergiants, we limited our study to the largest of them. However, there is a variety of organisations that operate on a less global scale than these specific hypergiants, which still exchange a significant amount of traffic, without relying on a global footprint due to the nature of their business, e.g., BBC. Also, some organisations that are not considered as such yet will become hypergiants in the future. Our results only apply to the time at which the dataset we use was collected, the Internet is a fast-changing ecosystem. Further work into the diversity of hypergiant-like organisations and their evolution is needed if we are to truly understand the Internet ecosystem and its diversity.

Public vs. private. Despite the unique and rather trustworthy information provided by PeeringDB, it misses an important part of the Internet network interconnection ecosystem, namely private peerings. Some large hypergiants, such as Facebook, rely heavily on private interconnection to deliver their traffic [11, 12]. Fortunately, despite not showing the private part of the network interconnection ecosystem, PeeringDB appears to provide sufficient information to still see the largest hypergiants. However, PeeringDB provides a view that (largely) underestimates the network interconnection ecosystem of the Internet. This bias is similar to the one of the AS-level topology, for which publicly available BGP routing data misses a large fraction of the AS-level connectivity [9], especially due to the rich worldwide IXP ecosystem [1, 2].

6 RELATED WORK

In their seminal work, Labovitz et al. [5] were the first to coin the term hypergiant. They observed a shift over time of traffic being diverted away from large Tier-1 and Tier-2 backbone networks and

	Organisation name	ASN	Continents	Port. Cap.	Traffic Profile
1	Apple Inc	714	4	10.960 Tbps	Mostly Outbound
2	Amazon.com	16509	6	9.991 Tbps	Balanced
3	Facebook	32934	6	9.840 Tbps	Heavy Outbound
4	Google Inc.	15169	7	8.741 Tbps	Mostly Outbound
5	Akamai Technologies	20940	7	7.854 Tbps	Heavy Outbound
6	Yahoo!	10310	6	5.310 Tbps	Mostly Outbound
7	Netflix	2906	7	5.170 Tbps	Mostly Outbound
8	Hurricane Electric	6939	7	5.037 Tbps	Balanced
9	OVH	16276	4	4.270 Tbps	Heavy Outbound
10	Limelight Networks Global	22822	6	3.840 Tbps	Mostly Outbound
11	Microsoft	8075	6	3.680 Tbps	Mostly Outbound
12	Twitter, Inc.	13414	6	3.401 Tbps	Heavy Outbound
13	Twitch	46489	5	3.340 Tbps	Heavy Outbound
14	Cloudflare	13335	7	3.320 Tbps	Mostly Outbound
15	Verizon Digital Media Services	15133	6	3.030 Tbps	Heavy Outbound

Table 2: The fifteen hypergiants sorted by port capacity.

instead being directly exchanged between networks without any intermediary. This observation forced the research community to significantly revisit their mental model of the Internet. Our work is motivated by their use of the word hypergiant, which is currently lacking a precise definition. In contrast to their work, we do not use traffic measurements but information within PeeringDB augmented by routing information to characterise hypergiants.

Previous works have used PeeringDB as an information source, extracting insights about the peering ecosystem and assessing its usability to better understand the Internet ecosystem. Lodhi et al. [7] made a first step in assessing the reliability and thus usability of PeeringDB for Internet research. They assessed the plausibility of PeeringDB data by comparing the information in PeeringDB against Local Internet Registries (LIRs) and BGP data. They found that while the data exhibits some biases, overall it appears to be reliable. They also made a first attempt at characterising the participating *organisations*. In contrast to our work, their focus is more on an overall assessment of PeeringDB than on analysing *organisations* specifically. Klöti et al. [4] compared the data in PeeringDB against data from other publicly available IXP data sets. They linked together the data sets available from PeeringDB, Euro-IX and PCH to assess their degree of complementarity and completeness. While they found biases in every data set, caused by its sourcing and intended usage, they nevertheless concluded that the data sets present similar views of the Internet.

7 SUMMARY

In this paper we combined PeeringDB and Routeviews BGP data to obtain a better understanding of today's hypergiants. Starting with a characterisation of the *organisations* taking part in public traffic exchange, we identified features differentiating hypergiants from the other *organisations*. Based on these features, we identified fifteen hypergiants. We then explored whether the approach those hypergiants take to make use of IXPs to reach their global customer base is unique. While it is different to many of the other *organisations*, on its own it not is sufficient to differntiate hypergiants from all other *organisations*. All these steps identified and discussed important characteristics of hypergiants, a set of *organisations* which has a significant impact on the Internet, due to the massive amount of traffic they are responsible for.

ACKNOWLEDGMENTS

This research is supported by the UK's Engineering and Physical Sciences Research Council (EPSRC) under the EARL: sdn EnAbled MeasuRement for alL project (Project Reference EP/P025374/1).

REFERENCES

[1] Bernhard Ager, Nikolaos Chatzis, Anja Feldmann, Nadi Sarrar, Steve Uhlig, and Walter Willinger. 2012. Anatomy of a large european IXP. In *Proc. of SIGCOMM*. ACM.
[2] Nikolaos Chatzis, Georgios Smaragdakis, Anja Feldmann, and Walter Willinger. 2015. Quo vadis Open-IX? *Computer Communication Review* 45, 1 (2015).
[3] Sushant Jain, Alok Kumar, Subhasree Mandal, Joon Ong, Leon Poutievski, Arjun Singh, Subbaiah Venkata, Jim Wanderer, Junlan Zhou, Min Zhu, Jon Zolla, Urs Hölzle, Stephen Stuart, and Amin Vahdat. 2013. B4: experience with a globally-deployed software defined wan. In *Proc. of SIGCOMM*. ACM.
[4] Rowan Klöti, Bernhard Ager, Vasileios Kotronis, George Nomikos, and Xenofontas A. Dimitropoulos. 2016. A Comparative Look into Public IXP Datasets. *Computer Communication Review* 46, 1 (2016).
[5] Craig Labovitz, Scott Iekel-Johnson, Danny McPherson, Jon Oberheide, and Farnam Jahanian. 2010. Internet inter-domain traffic. In *Proc. of SIGCOMM*. ACM.
[6] Stuart Lloyd. 1982. Least squares quantization in PCM. *IEEE transactions on information theory* 28, 2 (1982).
[7] Aemen Lodhi, Natalie Larson, Amogh Dhamdhere, Constantine Dovrolis, and kc claffy. 2014. Using peeringDB to understand the peering ecosystem. *Computer Communication Review* 44, 2 (2014).
[8] Matthew J. Luckie, Bradley Huffaker, Amogh Dhamdhere, Vasileios Giotsas, and kc claffy. 2013. AS relationships, customer cones, and validation. In *Proc. of IMC*. ACM.
[9] Ricardo V. Oliveira, Dan Pei, Walter Willinger, Beichuan Zhang, and Lixia Zhang. 2010. The (in)completeness of the observed internet AS-level structure. *IEEE/ACM Transactions on Networking* 18, 1 (2010).
[10] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.
[11] Arjun Roy, Hongyi Zeng, Jasmeet Bagga, George Porter, and Alex C. Snoeren. 2015. Inside the Social Network's (Datacenter) Network. In *Proc. of SIGCOMM*. ACM.
[12] Brandon Schlinker, Hyojeong Kim, Timothy Cui, Ethan Katz-Bassett, Harsha V. Madhyastha, Ítalo Cunha, James Quinn, Saif Hasan, Petr Lapukhov, and Hongyi Zeng. 2017. Engineering Egress with Edge Fabric: Steering Oceans of Content to the World. In *Proc. of SIGCOMM*. ACM.