Biases in top selected ASN in RIPE Atlas user measurements

by Theodoros Diamantidis and Pavlos Sermpezis,
Data & Web Science Lab (https://datalab.csd.auth.gr/),
Aristotle University of Thessaloniki

In this report we analyze the findings of comparing the top 10, 100 and 800 most used Autonomous System Numbers (ASNs) from a sample of measurements we gathered from the RIPE Atlas API. We have analyzed these ASNs using the AI4NetMonweb app.

In doing so we also provide an interesting example of how a multi-dimensional bias analysis for Internet measurements can provide useful insights, especially to researchers that are not very well versed in network science.

Methodology

For our analysis, we sampled ping and traceroute measurement data from the RIPE Atlas API.
For each measurement, we kept the following fields from the response of the API:

- **af:** IPv4 of IPv6 Address family of the measurement.
- **description:** User-defined description of the measurement.
- **id:** The unique identifier that RIPE Atlas assigned to this measurement.
- **is_oneoff:** Indicates this is a one-off or a recurring measurement.
- **participant_count:** Number of participating probes.
- **probes:** Probes involved in this measurement.
- **probes_scheduled:** Number of probes actually scheduled for this measurement.
- **tags:** Array of tags to apply to the measurement.
- **target:** The FQDN (if it was requested) or the ip address of the target of this measurement.
- **target_asn:** The number of the Autonomous System the IP address of the target belongs to.
- **target_ip:** The IP Address of the target of the measurement.
- **target_prefix:** Enclosing prefix of the IP address of the target.
- **type:** The type of the measurement (we only kept "ping" and "traceroute").

https://atlas.ripe.net/api/v2/measurements/ping/?page_size=50&fields=af,description,id,is_oneoff,participant_count,probes,probes_scheduled,tags,target,target_asn,target_ip,target_prefix,type&page=15

URI used for the get request to the RIPE Atlas API for a collection of 50 ping measurements lying on page 15 of the response.

```
{
  "af": 4,
  "description": "poison-art.ru",
  "id": 1008188,
  "is_oneoff": false,
  "participant_count": 200,
  "probes": [
    {
      "id": 2,
      "url": "https://atlas.ripe.net/api/v2/probes/2/"
    },
    {
      "id": 4,
      "url": "https://atlas.ripe.net/api/v2/probes/4/"
    },
    {
      "id": 5,
      "url": "https://atlas.ripe.net/api/v2/probes/5/"
    },
    {
      "id": 7,
      "url": "https://atlas.ripe.net/api/v2/probes/7/"
    },
    {
      "id": 9,
      "url": "https://atlas.ripe.net/api/v2/probes/9/"
    },
    {
      "id": 12,
      "url": "https://atlas.ripe.net/api/v2/probes/12/"
    },
    {
      "id": 15,
      "url": "https://atlas.ripe.net/api/v2/probes/15/"
    },
    {
      "id": 17,
      "url": "https://atlas.ripe.net/api/v2/probes/17/"
    },
    ...
    {
      "id": 3074,
      "url": "https://atlas.ripe.net/api/v2/probes/3074/"
    },
    {
      "id": 3110,
      "url": "https://atlas.ripe.net/api/v2/probes/3110/"
    },
    {
      "id": 3120,
      "url": "https://atlas.ripe.net/api/v2/probes/3120/"
    }
  ],
  "probes_scheduled": 200,
  "tags": [],
  "target": "poison-art.ru",
  "target_asn": 8402,
  "target_ip": "93.81.240.99",
  "target_prefix": "93.80.0.0/15",
  "type": "ping"
}
```


Example of a measurement returned from the RIPE Atlas API. This is the measurement with ID 1008188. Not all probes included in the image.

For a given measurement type, the RIPE Atlas API returns all measurements in different pages of the response, where each page contains 50 measurements by default. The pages of the response appear in chronological order with the first page having the least recent measurements.

To make sure that we have a random sample, we sampled 50 measurements from 30 randomly selected pages of the response. Upon completion of the sampling process, we collected a total of 1500 measurements of each type. Our goal is to gather 1000 measurements of each type, but given that some measurements in the API contain no or too little information, we sampled 1500 of each to have some redundancy.

As we can see in the fields above, the API returns the list of probes that took part in each measurement. Our analysis, though, is made at the level of the Autonomous System (AS) each probe belongs to. That's because the bias analysis in the AI4NetMon project is made at that granularity. Therefore, for each measurement, we find the ASN each probe in that measurement belongs to.

As there are two different address families, IPv4 and IPv6, each AS can have an IPv4 ASN and/or an IPv6 ASN. For our analysis, we kept the ASN that corresponds to the address family of the measurement itself, i.e. if the measurement was an IPv4 measurement we kept the corresponding IPv4 ASNs, and similarly for IPv6 measurements. It is rare that the IPv4 and IPv6 ASNs are different; such is the case for 3% (1395) of all the RIPE Atlas probes, none of which appear in our sample.

Another case that appears in the data is the case when a probe does not have a corresponding ASN (IPv4, IPv6 or both). This is the case for 27.71% of all RIPE Atlas probes, while for our measurement sample, this percentage was 14.4%. This is normal, and can happen for various reasons.

For example, for the ping measurement with measurement ID 1008188, which is an IPv4 measurement, we have a list of 200 probes from which we find their corresponding (IPv4) ASNs, as shown in the table below for the first few probe IDs of this measurement:

Probe ID	ASN
2	1136
4	3265
5	3265
7	33915
9	3333
12	-
15	3333
17	3333

Probe IDs and their corresponding IPv4 ASNs for the first 8 probes of the measurement with ID 1008188.

We can see that the same ASN can appear more than once in the same measurement depending on the probe selection for that measurement. Furthermore, we see that the probe with ID 12 has no IPv4 ASN. In fact, it doesn't have an IPv6 ASN either.

It should be noted here that some measurements did not contain the list of probes that were used for it, in which case we simply discarded these measurements (hence the aforementioned redundancy in the measurement data we collected from the RIPE Atlas API).

Having the list of ASNs that were used for each measurement, we calculated the bias for each measurement. This was done by taking the Kullback–Leibler divergence between each measurement's ASN list, and the list of all the ASNs that exist, across various dimensions which we will be referring to as "bias dimensions". These bias dimensions are a group of 23 properties that each AS has. They are organized into Location, Topology, Network Size, Interconnection/IXP and Network Type -related attributes for all ASes. For more details on the bias dimensions as well as the bias calculation you can refer to the ai4netmon project documentation.

	RIR region	Location (country)	Location (continent)	Customer cone (#ASNs)	Customer cone (#prefixes)	...	Network type (PeeringDB)	Traffic ratio (PeeringDB)	Traffic volume (PeeringDB)	Scope (PeeringDB)	Personal ASN
1008188	0.11891	0.402729	0.174522	0.200032	0.265912	...	0.083693	0.061924	0.125684	0.147269	0.000704

Bias values across some bias dimensions for measurement with ID 1008188.

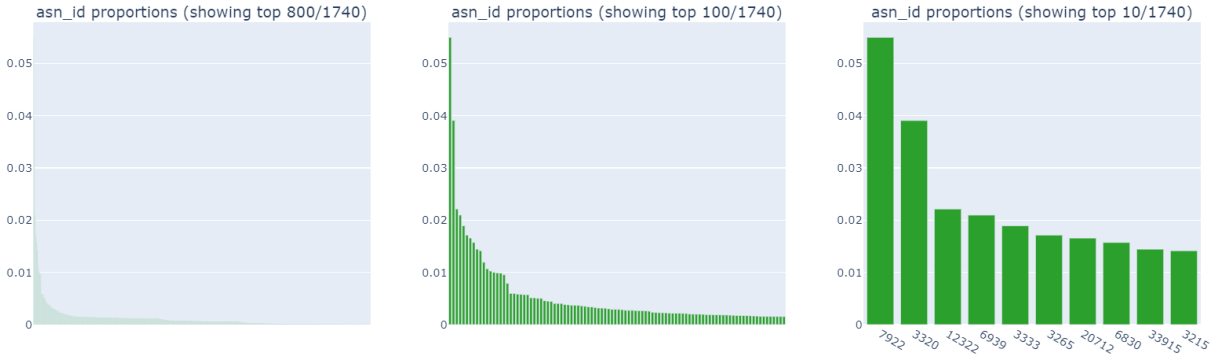
At the end of this process, we had a dataset of 2000 measurements (1000 ping and 1000 traceroute), where for each measurement we had the initial fields we mentioned plus the bias values for all bias dimensions.

Finally, before proceeding with our analysis, we cleaned this dataset removing duplicate columns, univariate columns as well as any rows that contained null values. Our final dataset consisted of 1551 ping (633) and traceroute (918) measurements.

Motivation

As we have mentioned, one characteristic of the dataset we created was the ASNs hosting the probes that were used in each measurement. In each measurement, we may have 0, 1 or any number of ASNs appearing and in addition, an ASN can appear in more than one measurements and also more than once in a single measurement. This means that, a priori, it is quite likely that a lot of duplication exists for the ASNs, and some ASNs are used more frequently than others.

To check that assumption, we plot the counts of ASNs, and what we see is that the form of the resulting distribution resembles a power law distribution, i.e. there seem to be a few ASNs that appear much more than the rest, which tend to appear a small and constant number of times (long tail).



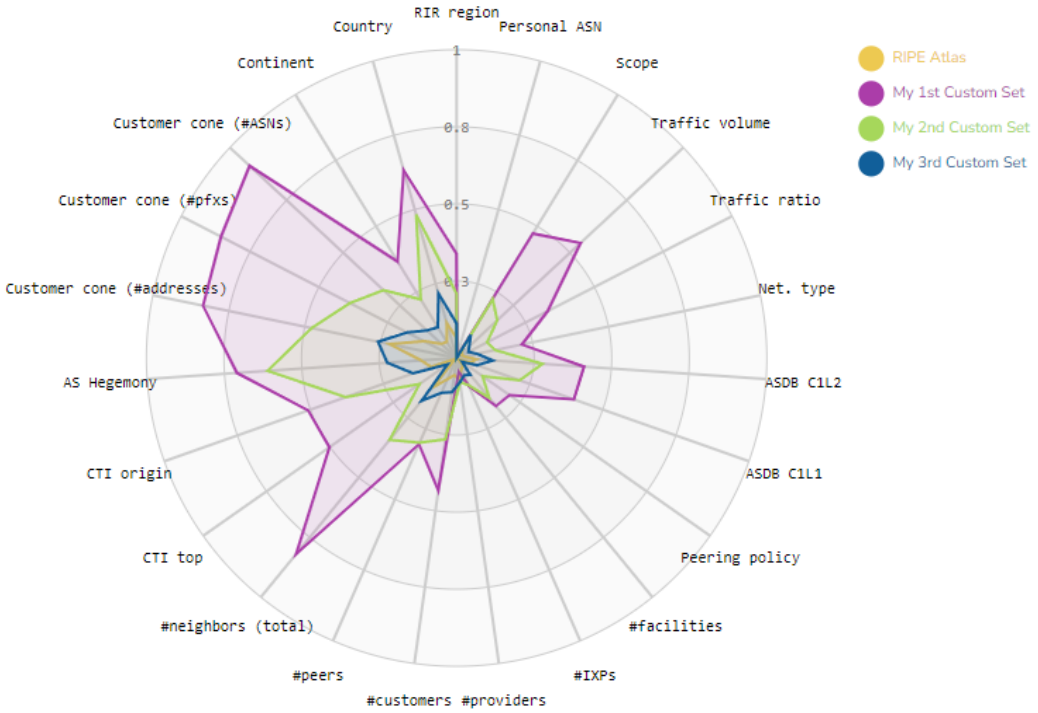
Frequency distributions of the top 800, 100 and 10 most frequent ASNs in our measurement sample.

What are the characteristics of the most frequent ASNs?

To answer that question, we explore the bias and bias causes for the top 10, 100 and 800 ASNs and compare them to all ASes as well as the ASes that belong to RIPE Atlas. For the rest of this report, we follow the following naming conventions:

- 1st Custom Set = Top 10 ASNs
- 2nd Custom Set = Top 100 ASNs
- 3rd Custom Set = Top 800 ASNs

Bias overview



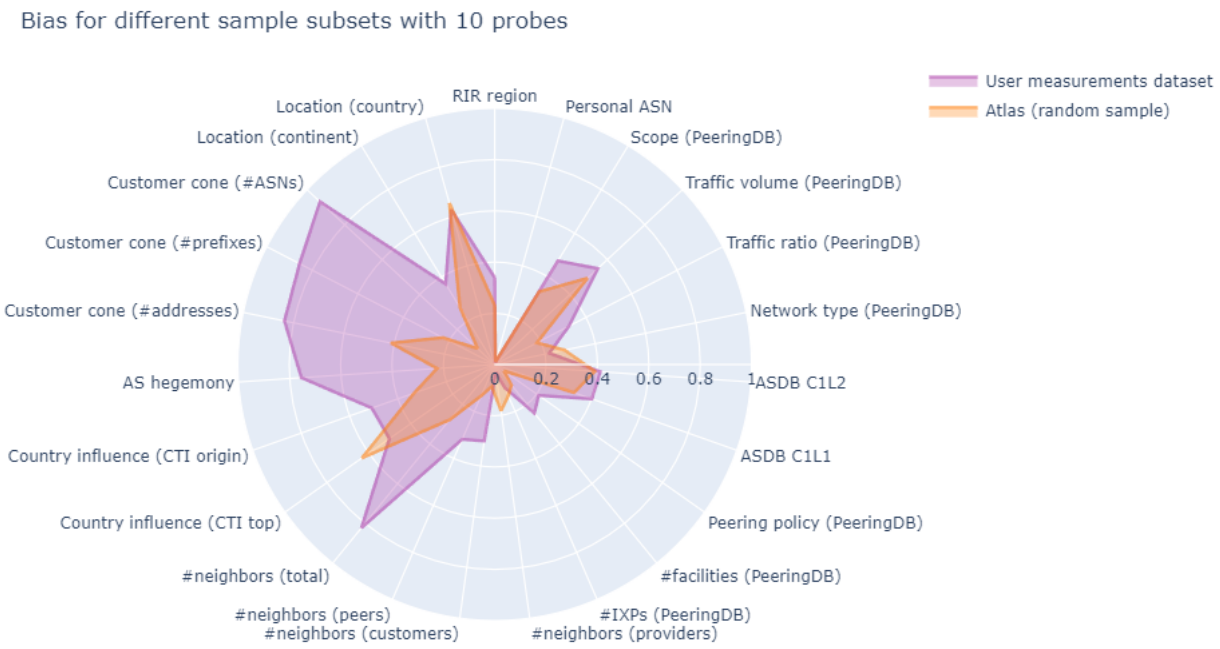
Bias distribution of ASes belonging to RIPE Atlas (yellow), and the top 10 (purple), 100 (green) and 800 (blue) most frequent ASNs in our measurement sample.

From this plot we can see that the distribution of bias across the different bias dimensions tends to be relatively similar for the different number of ASes we consider, and as that number increases from 10 to 100 to 800, this distribution resembles the distribution of the RIPE Atlas ASNs more and more.

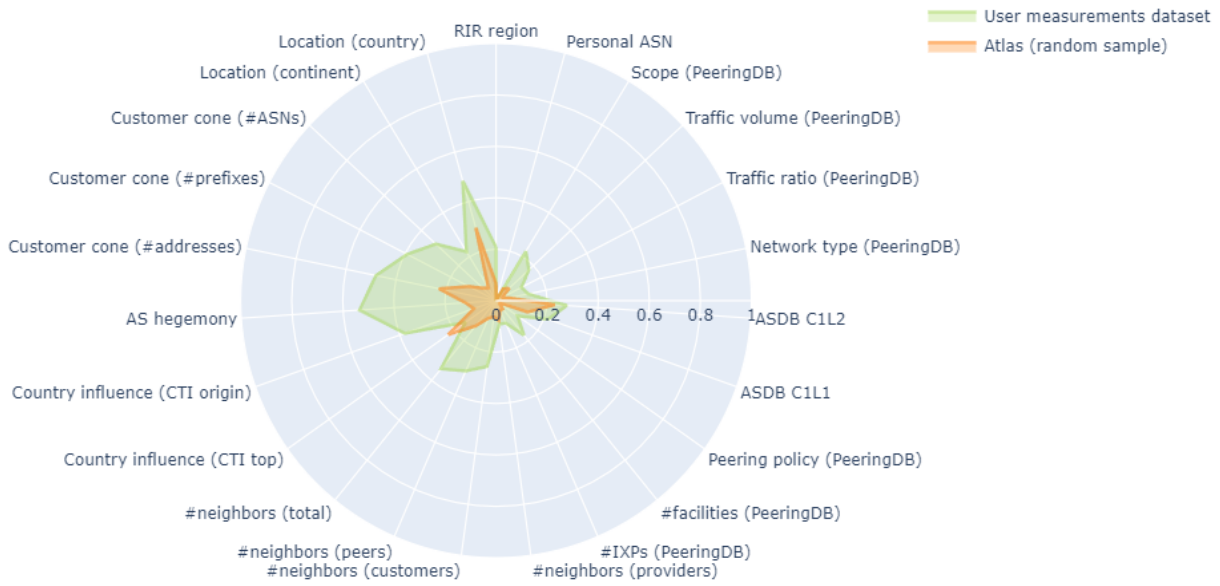
More specifically, we see that the top 10 ASes seem to be very biased in terms of the Network Size (customer cone) and Topological (#neighbors) dimensions. Significant bias is also present in the “Country” and “Traffic Volume” dimensions.

What’s interesting is that most of these high bias values shrink quite significantly as we consider a larger number of frequent ASes (Top 100 and 800). Namely, the bias value in the “Customer cone (#ASNs)” dimension drops from 0.85 to 0.33 as we go from the Top 10 to the Top 100 ASes, which corresponds to a 61.2% relative decrease. The same thing (with slightly different values) happens in the case of the “#neighbors (total)” and the “Traffic ratio” dimension.

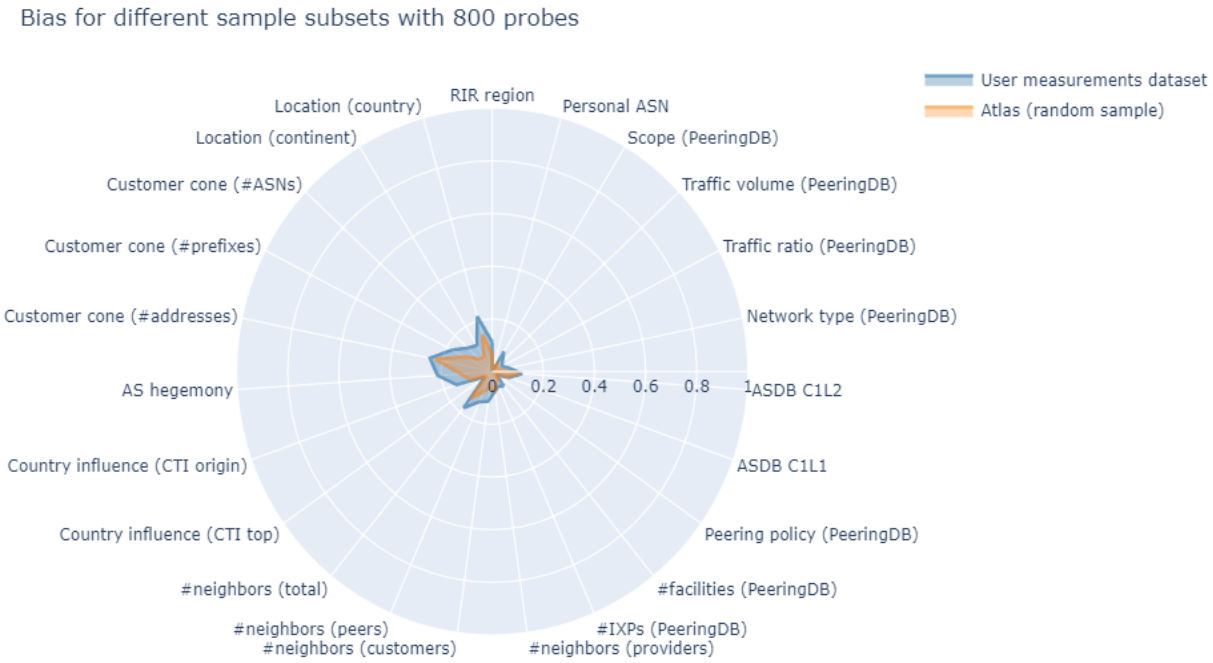
We can compare the bias distribution of each of our Top ASN samples with that of randomly sampled sets of 10, 100 and 800 ASNs of RIPE Atlas. For these random sets, we sampled 10, 100 and 800 probes 10 different times, and calculated the mean bias across all dimensions to get the final result which we plot below:



Bias distributions of Top 10 most frequent ASNs in our sample (purple) and 10 randomly selected RIPE Atlas ASNs (orange).



Bias distributions of Top 100 most frequent ASNs in our sample (green) and 100 randomly selected RIPE Atlas ASNs (orange).



Bias distributions of Top 800 most frequent ASNs in our sample (blue) and 800 randomly selected RIPE Atlas ASNs (orange).

From the above plots, we can see that as the number of probes increases, the bias values decrease for all samples, and the distribution of the Top ASNs seems to look more and more like the distribution of the randomly sampled ASNs.

The largest difference is between the distribution of the Top 10 most frequent ASNs and the corresponding randomly sampled ASNs. Namely, we see that there is a very large difference in the bias values of Network Size (Customer Cone), as well as Topology-related (#neighbors) dimensions. There are similarities too though, with the most prominent one being the bias on the Location (country) dimension followed by Traffic Volume (PeeringDB) and ASDB C1L2.

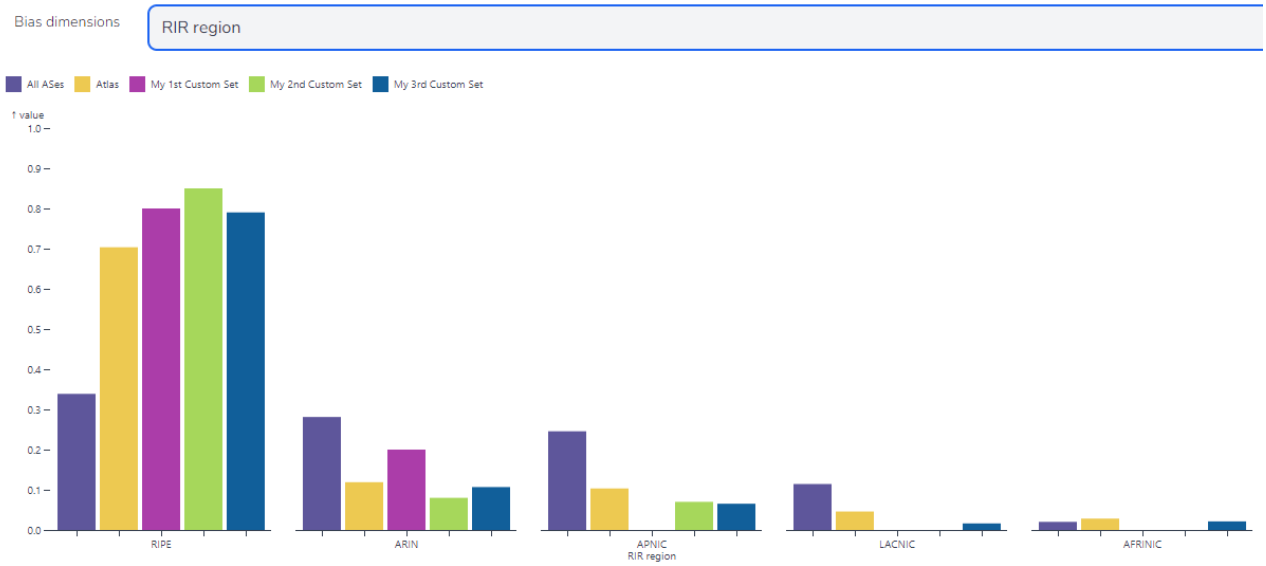
What causes these high bias values though, and why is there such a large drop as we consider a larger number of frequent ASes? Let us examine the bias causes in more detail by taking a deeper look into the distribution of the ASes for our different samples for each bias dimension.

Bias Dimensions Breakdown

Location Dimensions

By examining the Location-related dimensions, we see that most of the ASes (~80%) in our sample come from the RIPE region. This does not agree with the distribution of all ASes, where European countries contain approximately 30% of the ASes, and North America (ARIN) and East Asian countries and Australia (APNIC) containing roughly 30% each.

RIR Region

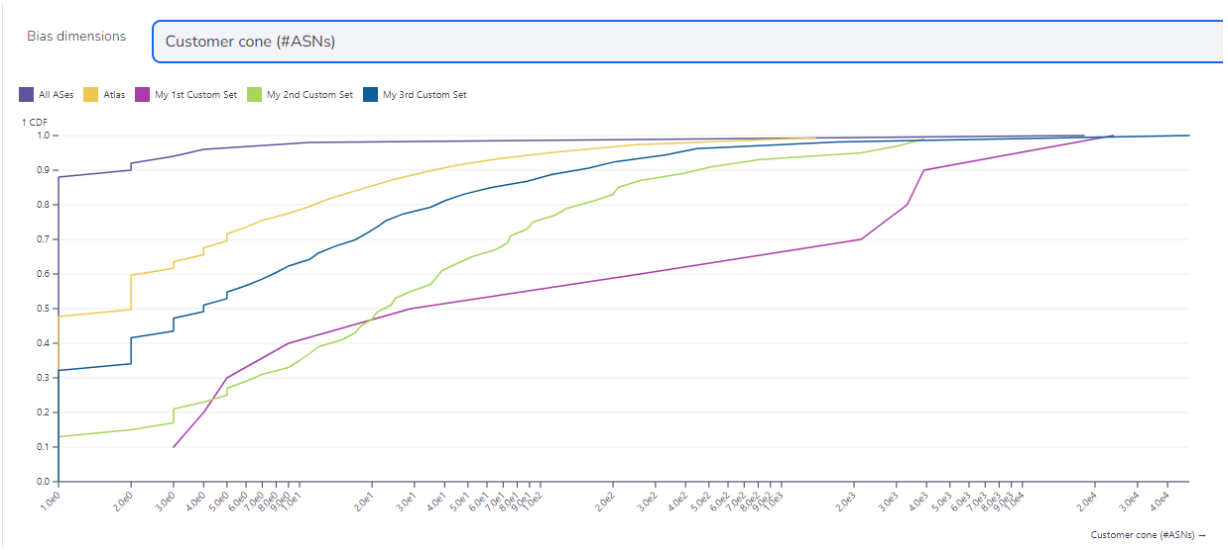


Distribution of ASNs across the RIR Region dimension for different samples.

Network Size Dimensions

By analyzing the dimensions related to Network Size, we can see that the most frequent ASes in our sample have a larger customer cone (here expressed as #ASNs, but similar findings hold for the #prefixes or #addresses in the customer cone). Similar findings hold when we use the AS Hegemony as a metric indicating the network size / importance (instead of the customer cone metric) .

Customer Cone (#ASNs)



CDF of ASNs across the Customer Cone (#ASNs) dimension for different samples.

In the figure, we can see that from the top 10 ASes in our sample, 50% of them have a reach of up to 30, while 90% of them have a vast reach of 400. Compared to the distribution of all ASes, where 90% of them can reach up to 2 other ASes in their customer cone, we can see how the more popular ASes have a larger reach.

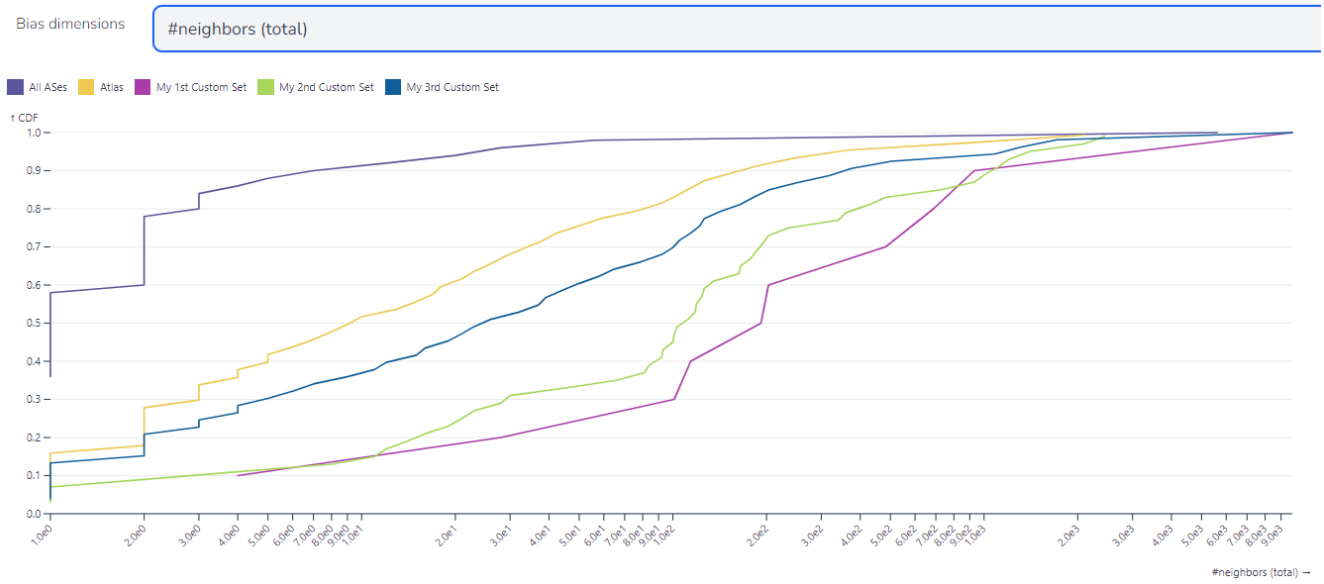
Topology Related Dimensions

Upon examination of topology-related dimensions, we can see that the most frequent ASes in our sample have a very large number of neighbors. For example, 50% of the top 10 most frequent ASes have around 200 neighbors, while 60% of all ASes have up to 1 neighbor.

By further analyzing the types of neighbors (peers, customers, providers) for the most frequent ASes in our sample, we see that they have a lot more peers and customers and a lot less providers compared to all ASes.

The large number of peers indicates that the most frequently used ASNs are well connected, and probably have significant presence at IXPs, which we analyze (and confirm below).

#neighbors (total)

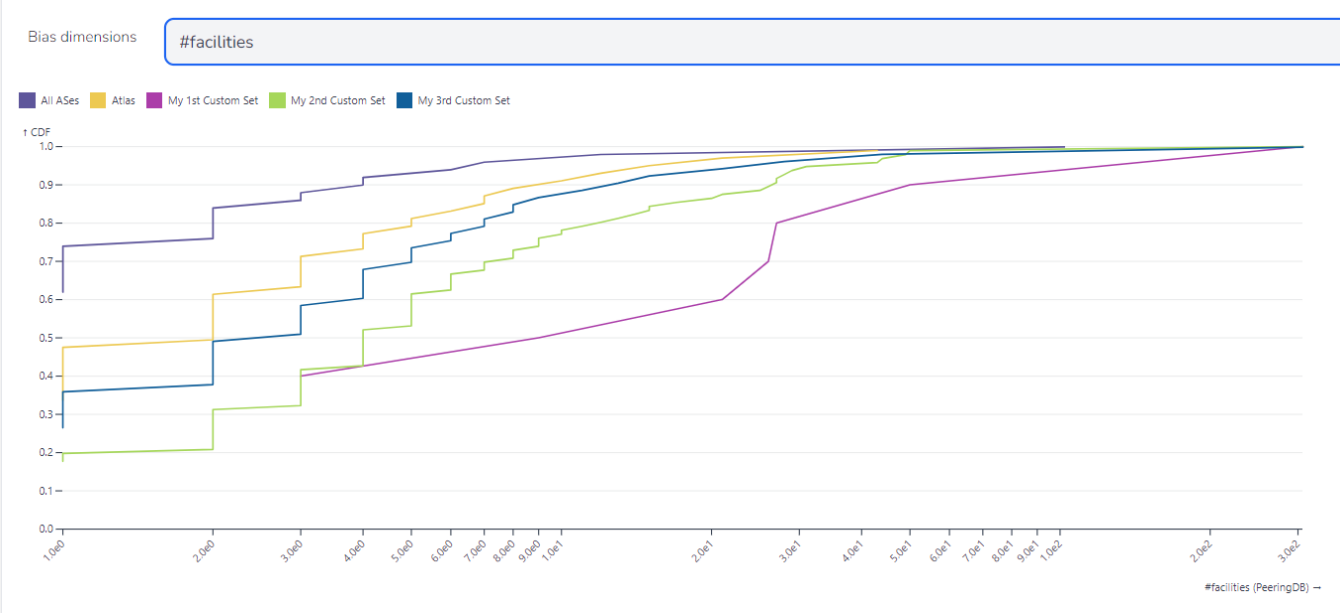


CDF of ASNs across the #neighbors (total) dimension for different samples.

Interconnection/IXP-related

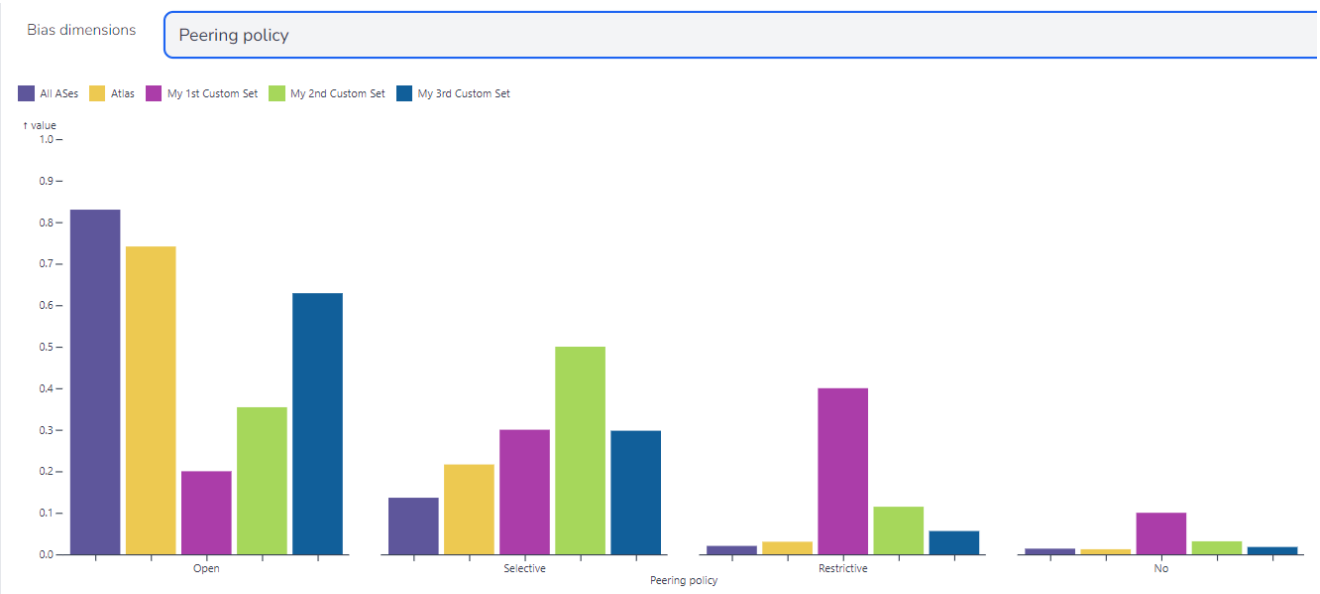
A close look at dimensions related to Interconnection and IXPs, we see that most ASes have few connections to IXP locations, while the top 10 most common in our sample, have a large number of connections.

#facilities



CDF of ASNs across the #facilities dimension for different samples.

Peering Policy



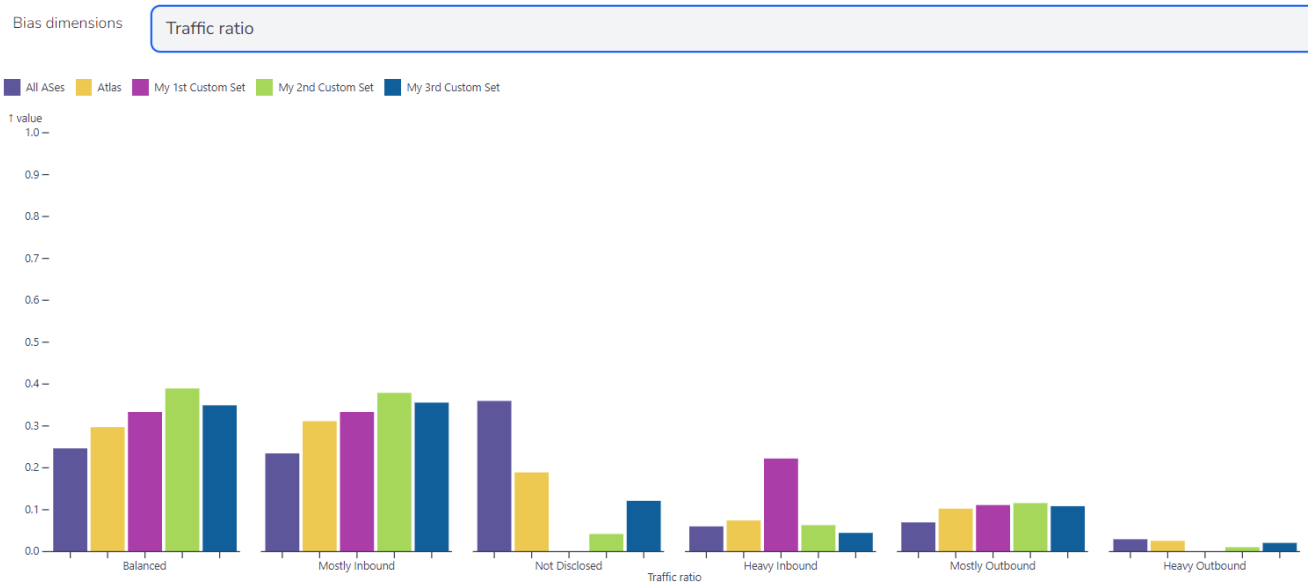
Distribution of ASNs across the Peering Policy dimension for different samples.

In addition, while most ASes have an “Open” peering policy (over 80% of all ASes), the most frequent ASes in our sample have “Restrictive” (more than 40%) and "Selective" (30%) peering policies. These types of policies are common among large networks (e.g., ISPs), which is in line with our findings with respect to network size.

Network Type related

A closer look at Network Type related dimensions reveals that the most frequently appearing ASes in our sample consistently have the most traffic compared to all our other samples. Furthermore, we also see that the top 10 most frequent ASes have global scope, while the top 100 and 800 have mostly European scope as opposed to the vast majority of all ASes which have a regional scope.

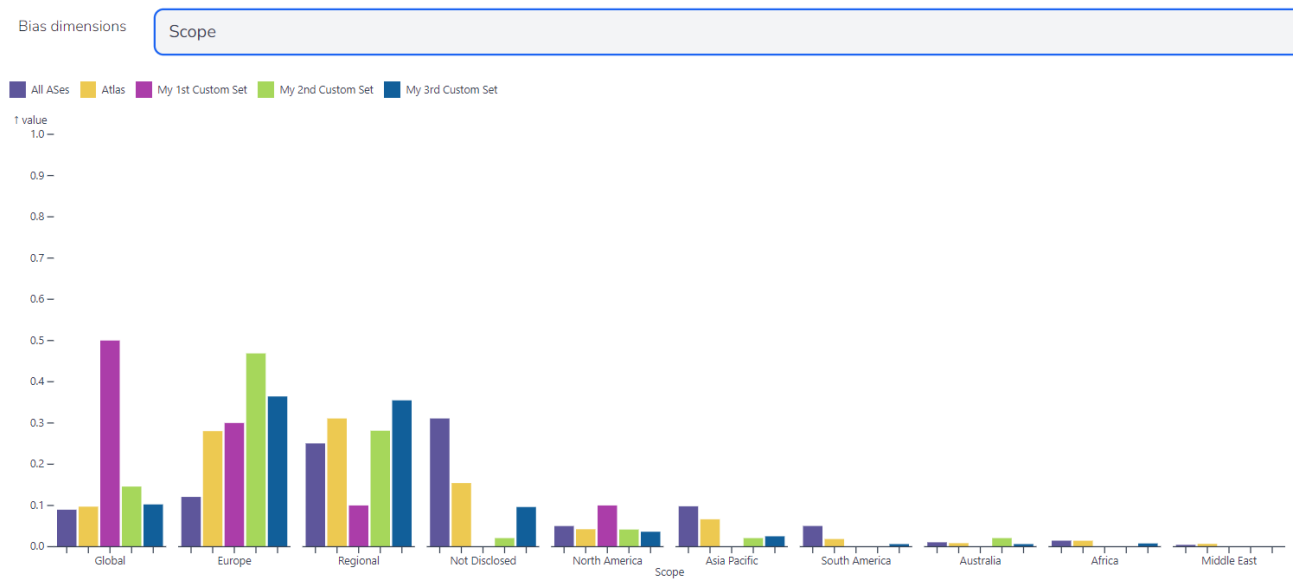
Traffic Ratio



Distribution of ASNs across the Traffic Ratio dimension for different samples.

Regarding traffic ratio, we can see that there is an almost uniform distribution between ‘Balanced’ and ‘Mostly Inbound’ traffic for all our samples. An interesting observation is that our top 10 most common ASes have vastly more ‘Heavy Inbound’ traffic compared to the rest of the samples.

Scope



Distribution of ASNs across the Scope dimension for different samples.

It is interesting to see that our samples for the top 10, 100 and 800 ASes, as well as the sample representing the ASes in RIPE Atlas, have mostly European scope. Excluding the top 10, we see that the rest of the samples also have mostly Regional scope, but for 50% of our top 10 ASes we see that they have ‘Global’ scope.

Conclusion

To summarize the above results, we have seen that there is a power-law distribution in the ASes that appear in the measurements of our sample. Taking a look at the bias distribution for the top 10, 100 and 800 most frequent ASes in our sample and comparing it to the bias of all the ASes as well as the ASes belonging to RIPE Atlas, we find some interesting results.

There is significant bias in Location, Topology, Network Size and Network type related dimensions for the top 10 most frequent ASes in our sample which then shrinks significantly as we go to the top 100 and 800 ASes, the bias distribution of which are more in par with the bias in the RIPE Atlas ASes.

We have also examined the distributions of ASes for each bias dimension and compared them to those of RIPE Atlas as well as all the ASes, and we have come up with the following explanations for the aforementioned bias values:

1) **Location:** Our sample is biased towards the RIPE region as this is a bias that exists within RIPE Atlas itself and, therefore, presents itself in our sample as well.
2) **Network size:** We have seen that the top 10 most frequent ASes have a significantly larger reach (in terms of the customer cone) than even the top 100 most frequent ASes, and a vastly larger reach compared to all ASes, but also compared to the entire set of probes in the RIPE Atlas platform.
3) **Topology:** The top 10 most frequent ASes have a significantly larger number of neighbors and these neighbors are mostly peers and customers as opposed to the other ASes which have mostly providers as neighbors.
4) **Interconnection:** The top 10 most frequent ASes have a larger number of connections to IXPs as opposed to most other ASes.They also have mostly “restrictive” and “selective” peering policies while most other ASes have “open”.
5) **Network Type:** The top 10 most frequent ASes have the highest levels of traffic.

Based on the above, we can see that the most frequently used ASNs tend to be large ISPs: they have large customer cones, a lot of neighbors which are mostly peers and customers, presence at many IXPs, mostly “selective” and “restrictive” peering policies and heavy inbound traffic. This is further confirmed by taking a look at the table below.

As we take into account more ASNs (top 100 and 800 most frequent ASes in our sample), they tend to have characteristics that are closer to an average AS belonging to RIPE Atlas platform, which is of course expected since all of our measurements are taken from RIPE Atlas.

ASN	Sample %	AS Name	Organization	Country	Type
7922	5.5	COMCAST-7922	Comcast Cable Communications, LLC (aka Xfinity)	USA	Tier 1 ISP
3320	3.9	DTAG - Deutsche Telekom AG	Deutsche Telekom AG	Germany	Tier 1 ISP
12322	2.2	PROXAD - Free SAS	Free SAS	France	Top ISP in France
6939	2.1	HURRICANE	Hurricane Electric LLC	USA	Technically Tier 2 ISP, but considered Tier 1
3333	1.9	RIPE-NCC-AS - Reseaux IP Europeens Network Coordination Centre (RIPE NCC)	Reseaux IP Europeens Network Coordination Centre (RIPE NCC)	Netherlands	European RIR
3265	1.7	XS4ALL-NL - KPN B.V.	KPN B.V.	Netherlands	Top ISP in Netherlands
20712	1.7	AS20712 - Andrews & Arnold Ltd	Andrews & Arnold Ltd	UK	Top ISP in UK
6830	1.6	LIBERTYGLOBAL - Liberty Global B.V.	Liberty Global B.V.	Netherlands	Top ISP in Netherlands
33915	1.4	TNF-AS - Vodafone Libertel B.V.	Vodafone Libertel B.V.	Netherlands	Top ISP in Netherlands
3215	1.4	AS3215 - Orange S.A.	Orange S.A.	France	Top ISP in France

Top 10 most frequent ASNs in our measurement sample.

Personal Opinion

We believe this to be a useful example of how a multi-dimensional bias analysis can give some insights, especially to researchers or engineers with little network science experience. It tells us that a random sample of measurements will be biased against ASes that belong to large ISPs, and how exactly this is reflected in the different bias dimensions. This can be useful to an IMP user to make a more careful selection of probes/ASes when conducting their measurements if the aforementioned bias can affect their research goals.

We are excited to further explore this topic and bring forth more and deeper insights of how bias affects internet measurements in the future.