Measurement Vantage Point Selection Using A Similarity Metric

Thomas Holterbach
ETH Zürich
thomahol@ethz.ch

Emile Aben
RIPE NCC
emile.aben@ripe.net

Cristel Pelsser
University of Strasbourg
pelsser@unistra.fr

Randy Bush
Internet Initiative Japan
randy@psg.com

Laurent Vanbever
ETH Zürich
lvanbever@ethz.ch

ABSTRACT

In a measurement platform with a wide selection of vantage points, it can be challenging to select the most appropriate points to source measurements from. One example of such platform is RIPE Atlas [2] which currently hosts over 9600 active vantage points, which can be selected based on categories, such as origin AS or country. When setting up a measurement, users are limited in how many vantage points they can use. This is not only due to limitations that the measurement platform imposes, but collecting data from a large number of vantage points would mean a large volume to analyse and store. It therefore makes sense to optimize for a minimal set of vantage points with a maximum chance of observing the phenomenon in which the user is interested.

Network operators may need to debug a network service with only limited information about the problem ("Our network is slow for users in France!"). A diversity metric would allow selection of the most *dissimilar* vantage points, in an attempt to explore the network phenomenon from as diverse angles as possible. If one finds an interesting network phenomenon, one could use the similarity metric to advantage by selecting the most *similar* vantage points to the one exhibiting the phenomenon, in an attempt to validate the phenomenon from multiple vantage points.

We propose a novel means of selecting vantage points, which is not based on categorical properties (such as origin AS, or geographic location), but rather on the topological (dis)similarity between vantage points. We describe a similarity metric across RIPE Atlas probes, and show how this performs better for the purpose of topology discovery than the default probe selection mechanism built into RIPE Atlas.

ACM Reference format:
Thomas Holterbach, Emile Aben, Cristel Pelsser, Randy Bush, and Laurent Vanbever. 2017. Measurement Vantage Point Selection Using A Similarity Metric. In *Proceedings of 2016 Applied Networking Research Workshop, Prague, Czech Republic, July 2017 (ANRW'17),* 3 pages.
DOI: http://dx.doi.org/-----

1 SIMILARITY METRIC

We aim to quantify the topological distance between RIPE Atlas probes. We argue that topological distance is more relevant than geographic distance from a measurement point of view, as it is directly based on network data which is network-agnostic.

While it is more useful, topological distance is also harder to capture in practice. For instance, probes in the same AS can actually see very different paths to external destinations (e.g., if they are connected to different egress routers), while probes in different AS can see similar paths to external destinations (e.g., if there are connected to the same IXP and the ASs have similar routing policies).

1.1 Measuring topological similarity

To measure how two probes are topologically similar, we define a metric based on the Jaccard similarity coefficient of the set of public IPs these probes are observing. More precisely, let x and y be two measurement vantage points, m a destination (IP or hostname) and $P_{x \to m}$ (resp. $P_{y \to m}$) the set of IPs in the path from x to m (resp. from y to m) using traceroute. The Jaccard similarity coefficient for the pair of probes (x, y) and the destination m is defined as: $d_{(x,y) \to m} = \frac{|P_{x \to m} \cap P_{y \to m}|}{|P_{x \to m} \cup P_{y \to m}|}$.

Intuitively, a result of 1 indicates both probes discover the same set of IP addresses and a result close to 0 indicates very few IP addresses in common. This metric is highly dependent on the destination m. To make it more robust, we consider M instead, which is the set of all the destinations that both x and y are targeting with traceroutes. We only calculate the Jaccard index for pairs of probes where $|M| \geq 17$. For each pair of probes, we only consider the 25th, 50th and 75th percentile of the list of coefficients computed (one for each common destination). Doing so makes the metric resistant to outliers even though the metric is sensitive to the set of common destinations between two probes. We calculate the metric for IPv4 and IPv6 separately, as these topologies are not congruent.

1.2 How topologically similar is RIPE Atlas?

We compute the metric over all probe pairs over one day (31 March 2016) of Atlas traceroute measurements.

Figure 1 plots the CDF of probes with respect to the maximum similarity to each other probe. In IPv4 about 10% (5% in IPv6) of probes have a median Jaccard index of 1.0 to at least one other probe, i.e.. on median, they discover the same set of IP addresses as some other probe. The large interquartile range results from the variability in the number of unique IPs encountered by each measurement.

ANRW'17, Prague, Czech Republic

———-…$15.00
DOI: http://dx.doi.org/-----

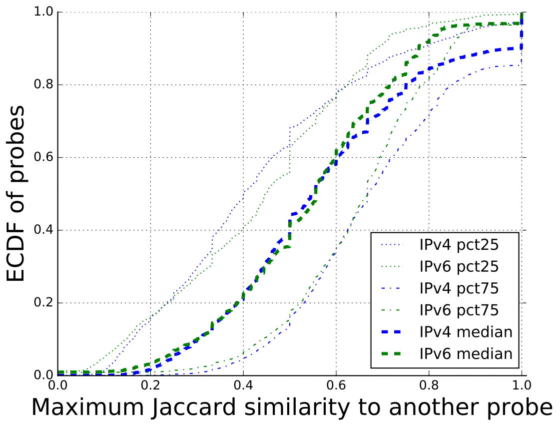

Figure 1: About 10% of the RIPE Atlas probes see a median Jaccard similarity of 1.0 in IPv4 to at least one other RIPE Atlas probe, In IPv6 that number is 5%.

Probe pairs in the same AS	IPv4	IPv6
# with Jaccard ind. ≥ 0.9	1805	56
25th percentile distance	7 km	0 km
50th percentile distance	40 km	2 km
75th percentile distance	104 km	17 km
Maximum distance	8,817 km	664 km
Probe pairs in the different ASes	IPv4	IPv6
# with Jaccard ind. ≥ 0.9	11	0
25th percentile distance	2 km	-
50th percentile distance	9 km	-
75th percentile distance	112 km	-
Maximum distance	532 km	-

Table 1: Pairs of probe with high Jaccard coefficient (≥ 0.90) tend to be geographically close to each other.

Validating the metric against physical distance and detecting geolocation discrepancies: Intuitively, we expect a probe pair with a result close to 1 to be physically close. Table 1 shows how physical distance or being part of the same AS affect the median Jaccard-index. We compute the distance between probes using geographical coordinates provided by the probe host.

As expected, the similarity metric is higher when the probe pairs belong to the same AS. In IPv4, 1,816 pairs of probes have a Jaccard metric higher than 90%, while 90.4% of these pairs are separated by less than 200 km.

We observe that few probes in different ASs show high similarity. Geographically distant probes that have a high Jaccard metric could indicate interesting topologies (e.g. long-distance tunneling) or incorrect geolocation data. Indeed, upon manual inspection of few geographically separated probes with a high Jaccard metric, and contacting the probe hosts, we found they forgot to update probe location after it moved, which was then corrected (including the maximum distance of 8,817 km in Table 1.

2 EXPLOITING SIMILARITY

Let's now look at how RIPE Atlas users could exploit probe similarity. Our metric could be used to:

Ensure a better measurement continuity RIPE Atlas probes typically have some downtime. For example, as of May 11, 2016, 32% were disconnected. To reduce the chances for gaps in a time series of measurement results, users could run their measurements on multiple similar probes instead of only one.

Improve measurements precision When launching a large set of measurements on only one probe, users may overload a probe and lose precision [4]. Dividing a set of measurements between similar probes could give users the ability to run a large number of measurements without losing precision. Similarily, if one probe is loaded by someone else, users could use a similar but not loaded probe so as to improve measurement precision.

Boost IP topology discovery by 25% An additional way to exploit our probe similarity metric is to (dis)cover as much of the IP topology address space as possible, given a limited probing budget [3]. To measure this, we conducted 1002 measurements, each one targeting a destination in a different and randomly selected routable IP prefix. We then compared the set of IP addresses discovered when the source probes are selected with the default RIPE Atlas selection mechanism or with our probe selection mechanism based on the Jaccard dissimilarity metric. For each experiment, the source probes are selected from a pool of probes. Each pool of probes is computed based on what the default RIPE Atlas probe selection mechanism can return (*e.g.* probes in the same AS, or the same country[1]). The number of source probes we select for each experiment is between 2 and 1/3 the number of probes in the pool (taking all the probes of the pool would lead to the same result). We designed our probe selection mechanism as follows: the first selected probe is one of the probes the default RIPE Atlas probe selection mechanism would select. Then, probes are selected one by one such that a new selected probe is maximally dissimilar with the already selected probes according to the Jaccard index.

We compared the total number of IP addresses discovered normalized by the number of probes that actually performed the measurements. In the median case, our probe selection mechanism based on the Jaccard index enables to discover 25% more IP addresses per probe than the RIPE Atlas probe selection mechanism.

3 CONCLUSION

We proposed a metric capturing the topological similarity between two active measurement vantage points. We showed that the RIPE Atlas platform is diverse in that only 10% of the probes are highly similar (in IPv4) to others according to our metric. In addition to proposing a set of practical uses for experimenters, we showed that selecting probes that are dissimilar increases the number of IPs discovered by 25% compared to the default RIPE Atlas probe selection. We plan to refine our similarity metric in future work, possibly by different selection of a set of destinations.

Example data of pairwise similarity metrics for 31 March 2016 is available at http://sg-pub.ripe.net/emile/probe-similarity/.

REFERENCES

[1] RIPE Atlas - User-Defined Measurements.

[2] RIPE NCC. RIPE Atlas.

[3] I. Cunha, P. Marchetta, M. Calder, Y.-C. Chiu, B. Schlinker, B. V. A. Machado, A. Pescapè, V. Giotsas, H. V. Madhyastha, and E. Katz-Bassett. Sibyl: A practical Internet route oracle. In *13th USENIX Symposiun on Networked Systems and Implementation (NSDI'16)*, pages 325–344, 2016.

[4] T. Holterbach, C. Pelsser, R. Bush, and L. Vanbever. Quantifying interference between measurements on the ripe atlas platform. In *Proceedings of the 2015 ACM Conference on Internet Measurement Conference*, IMC '15, pages 437–443, New York, NY, USA, 2015. ACM.