Consolidated Review of

AS-level Topology Collection Through Looking Glass Servers

1. Strengths:

The paper explores how a new dataset can reveal additional links in the AS-level topology that were not previously visible.

Characterizing datasets like this can be helpful to others. The use of looking glass servers gives very up-to-date sets of links. This paper demonstrates that it also provides a large data set, whereas prior work appears to have only used LG servers to verify data collected from other sources. The study highlights the utility of the looking glass servers for collecting this additional information and assesses the utility of other existing datasets for AS-level topologies, and explores how this missing data affects various conclusions (e.g., concerning AS degree).

2. Weaknesses

The main weakness of this paper is the lack of insights and conclusive results. What causes the unique AS links seen by LG servers? Why would LG servers also miss some AS links that were seen by other methods? Is the use of LG servers a better way to generate AS topology? Is there any lesson learned here lead to a better way to get AS topology?

Should have used [10] as a comparison point.

The paper is of limited scope and ambition. The IRL work and the IXPs:Mapped? work also used LG servers, so it is partly just the accounting here that might be new.

3. Comments

This paper explores the effects of augmenting existing BGP datasets to uncover additional ASes and inter-AS links in the Internet topology. Interestingly, the paper finds that significant portions of the topology are not visible from the existing data sources, such as RouteViews, RIP, PCH, and the IRRs.

This is a worthwhile study, and it will be useful for the community to learn from this study. Maintainers of some of the other topology datasets may even benefit from reading this and choose to augment their topologies with the looking glass data as a result of this study.

While I find nothing wrong with this study, I must say that I found the paper to be a rather dry read. Essentially, the conclusion is that adding another dataset yields more ASes and more inter-AS links. That's good to know, but there's no insight into why that's the case. Where are the missing links and ASes coming from? Why do the other datasets not see these links? Thus, while the result is still useful, the paper could be a lot more interesting if it offered some insights into why the LG data makes the AS-level topologies more complete.

Neat that you found "new" ASes!

Do you have any intuition for why ASes that are willing to provide these services have yet to offer feeds to route collectors.

It would have been nice to include the Ono / Sidewalk Ends data from NWU ([10]) as a comparison point.

In Fig 1, it is hard to see the differences between the traces. It would help if not every point had a symbol on it, so that the symbols do not overlap. Moreover, the font is too small. The font on a figure should be comparable to the size of the font in the caption.

It would be nice to see Table 3 at other granularities in addition to matching IPs (which is what I think it shows). Two routers could have different IPs but be basically equivalent, right? So it's hard from that section to tell how much of the view is unique.

Table 3 suggests that the BGP feeders of the different data sets have very different IP addresses. Is that because the feeders themselves have little overlap, or simply because they are accessed via different interfaces?

I'm struck by how low the numbers are in Table 5, even with your new links. Google and many of the others listed have essentially open peering policies at major IXPs. So, I'm not sure how meaningful it is to see 31 more links, when hundreds/(more likely) thousands are still missing.

I'm suspicious of the comparison to IRL. As you note, they are having problems with their data (going back into 2012, from what I can tell). This isn't your fault, but it is unfortunate. Similarly, it makes me worry about the results in Table 5 "which suggests that these IP prefixes are possibly aggregated by their provider ASes." Can't you easily check whether the IP prefixes are aggregated by seeing how the addresses appear in, say, RouteViews route collectors?

Given all the peering links that all the views are still missing, are the degree distributions meaningful?

"We believe that it is due to the maintenance of no-probing list by CAIDA Ark, by which ASes can request CAIDA not to probe their networks" Did you confirm this with CAIDA? It seems surprising that large ASes would bother to opt out. I've rarely / never seen ASes opt it, and, when they do, it is rarely the big ones.

I found the discussion on the inadequacy of other AS-level topologies somewhat underwhelming. Effectively, the conclusion is that a lot of other datasets are not well maintained. While that may be true, it is not fundamental, and it does not lend any insight into why the LG data would yield a non-overlapping set of nodes and links.

I am impressed on the large-scale measurement the authors did to generate AS topology. However, the paper did not provide any insights on whether LG server is better way for learning AS topology and the rationale behind it. The paper simply reports

their comparison numbers. Based on the results, it is not clear to me why someone should use LG servers alone for AS topology generation. It requires a lot of commands/queries. But the results are not any better than existing methods (e.g., they capture some unknown AS links but also missed some known ones). I don't think this is surprise at all.

It would be most helpful if you could redo this on an ongoing basis and make your data and scripts available to the community, so that it isn't t a onetime snapshot.

How do you get locations for the LG routers? Do the providers always give them?

I'm a bit skeptical of the analysis on Section 4. Even if you do not observe the same IP, couldn't two adjacent routers exist, one offering LG services and one offering essentially an equivalent view to Routeviews?

You characterize where the new links are. It would be nice to add (a) where the LGs are, (b) (the distribution of) how many new links each LG provides (is it a small number that give most of the benefit)? Similarly, for the unique ASes, are most customers of just a small number of providers?

I like this paper. However, it should explain more clearly the drawbacks of this approach. This would be suitable in the first paragraph of Section 5.1, where the discrepancies with other data sets are being discussed.

In the expression P(k) = n(k)/n, no definition for "n" has been given. Is it the sum of n(k)?

Since the number of routers the LG servers access (2.6K) seems much larger than RV or RIPE, the fact that they only observe a small number of new ASes should probably tell us something, shouldn't it? It would possibly be interesting to compare the number of observed ASes to the total number of assigned AS numbers to see what fraction of the previously unobserved are now accounted for.

In summary, the results in this paper could be useful to a relatively small set of people who spend a lot of time on Internet topology. For that reason, I think the paper should be published. Yet, the paper could have been a lot stronger if it explained the intuition behind some of the results.

The only way that I can think of to use this data is to combine the topology that derived from other methods (i.e., fill the gap of existing method). From this perspective, I think this paper is useful to be included in the IMC program.

4. Summary from PC Discussion

PC meeting discussion summary: The PC felt it will be useful to the community to know what LG servers can add to our view of the topology and appreciate the hard work that likely went into gathering the data. Given that, in addition to the detailed reviews, two high level suggestions:

- It would be great to see more discussion in the paper of why the views differ (from LGs vs. from other datasets)
- It would be a nice service to the community if you can automate the analysis and make fresh datasets available on an ongoing basis

5. Authors' Response

We are very thankful to the reviewers for their constructive feedback, which helped in enhancing the paper contents.

We acknowledge the need to add more discussion to the paper regarding the reasons of differences between LG servers and other AS topology datasets. In Section 5.1, we have added a paragraph to answer "Why do the LG servers miss AS links observed in other datasets". This can be due to the following reasons. First, BGP feeders may provide a full feed to the router collector projects such as RouteViews while they may share only partial feed to LG servers due to economic relationships (such as Peer-to-Peer (P2P)) with ASes operating the LG servers. Second, LG servers also suffer from vantage point bias. More specifically, depending on the view of a BGP feeder and its location in the Internet, a specific portion of AS topology can be discovered by an LG server. Third, it is not clear whether all the AS links published by the traceroute and IRR datasets are correct.

To elaborate more on why do the other datasets not see AS links discovered using LG servers?, we note in Section 5.1 that the reasons are different depending on different AS topology datasets. First, incompleteness of current widely used BGP-based datasets such as RouteViews. Second, traceroute-based datasets suffer from limited vantage points, selectively probing prefixes, various inaccuracies due to IP-to-AS mapping issues. Thus, topology collection from LG servers result in discovering otherwise unobserved part of the Internet as LG servers provide new BGP feeders from geographical diverse locations in the Internet.

We have added the ASN (and router IP) wise comparison of route collector projects to Table 3 as suggested by the reviewers.

To answer why ASes that are willing to provide these LG services have yet to offer feeds to route collectors", we have added a paragraph at the end of Section 4.

To give the reason for not using the Ono [10] dataset, we have added the following to Section 3.2. We have decided to use the recently (and regularly) published AS topology datasets only, not including ones such as Ono [11], which had been collected using BitTorrent P2P clients in 2007-2008. Since we cannot quantify how much of this AS topology dataset is outdated, we decided not to use it. Similarly, we exclude DIMES [15] as it has not been updated since Apr. 2012.

We do not suggest that other AS topology collection methods such as Traceroute and IRR is not needed. The stand point of this paper is that LG servers help in augmenting the current AS topology collection efforts reliably as BGP based methods are less error prone as compared to traceroute-based ones. Moreover, collecting BGP traces from the LG servers can help widen the narrow view of BGP observed from the current BGP collector projects such as RouteViews, RIPE-RIS, and PCH [22].

While we have incorporated most of the suggestions by reviewers to revise our paper for camera ready submission, a few suggestions are left to be investigated as part of our future work such as more in-depth analysis behind the contribution of different LG servers to the overall AS topology collected from LG servers.