Consolidated Review of
Growth Analysis of a Large ISP

1. Strengths

The Cogent naming scheme is quite revealing--it gives information about ports on line cards, and the speed of the port. The methodology and the gathered dataset are interesting contributions.

The use of examples throughout the text makes it easy to read and follow.

This is a novel contribution that makes a large and interesting dataset available to the community that could be used for different kinds of topological studies. The dataset obtained looks solid thanks to the consistent naming conventions adopted by Cogent.

The idea is neat. It extends previous attempts to glean information from naming conventions. Maps like this are useful. I've seen numerous papers within the last year that still use Rocketfuel maps, even though those are quite old. It will be great to have your maps available. Great that authors gathered data over time before publishing.

This allows them to draw some interesting conclusions, specific to Cogent but interesting. Cogent has about 4500 routers, and adds about 11 per week. The diameter of their network and the density of edges seem constant. Most of the ports now being added are 10G. The bit about inferring links and using Cogent's looking glass to corroborate seems the most useful.

2. Weaknesses

The authors are quite clear that there is no guarantee that this method will work on other networks. It looks like they didn't even try to do this for other networks. If they had tried the tier-1 networks (even if they didn't succeed), it would have been useful. As it is, this looks too much like "a first look at Cogent".

The success of the method is interesting but not a big surprise. The data reported is also interesting but not a big surprise. There is not much in the way of concrete results or generalizable knowledge, or discussion of implications of the results.

Mentioning the implications of the observations (rather than stating the facts), and a discussion on the scope of the results, would have added a lot of value to this work.

The paper could be a lot better written. The abstract should give the densification result on page 6, which seems the most important outcome, even if it is still an early result that needs more data points over time. Use of dataset [6] is controversial, both from an ethical perspective and since it is not validated while published anonymously.

Considering the richness of the data, the analysis is limited and perhaps not as interesting as other potential uses of the data. (What you did is plenty for a good short paper, but you could have had a great short [or perhaps even long] paper) Validation is limited.

3. Comments

This is cool - I'm happy you did it, it should clearly be accepted as a short paper, I'm very glad you gathered data for a long time before submitting, I'm happy you're making the data available, and I hope you'll continue gathering and making data available. The feedback below is to make a good paper even better, and I hope you'll take the time to do it. I hope you'll consider taking this data and furthering the analysis before finalizing any publication.

The paper presents interesting observations on the evolution of a top level ISP network. It might be interesting to see more details on the nature of this growth (which should be readily available from the datasets collected). For instance, is there any significant difference in 10G interface additions to the core and edge? How does min/median/max cross-section BW is changing over time? Etc.

I found myself wondering what else one might learn from 17M DNS names derived from reverse lookup. But the lack of structure and accuracy may make this question not worth asking.

Page 1 is dense with IP addresses and names. Please put them in a table, each row containing IP, name, and inferences made. For example, it took me way too long to figure out how you derived that Tetratech is connected at 100Mbps; the interpretation has to be gleaned through the /30 mate and the fa->100MBps inference supplied two paragraphs earlier. How many interfaces inferred to be on cogent routers in CAIDA's ITDK releases are missed by your DNS-based assignment? I believe this is an important validation step. Running iffinder is only going to help you with what you see. While ITDK will probably not give such a deep view of Cogent as you have obtained, (and the ITDK releases are only made periodically) it is nonetheless a dataset obtained from a different angle that should overlap with your dataset at least once. Why was your measurement infrastructure unavailable for weeks 21-27? Please use dates on X-axis, rather than week numbers. Can you say anything about other ASes you tried the naming method for, if any? I have used DNS-names in a similar way for cross validation and have been burned by naming conventions that don't identify if an interface is on a third-party router or not. Cogent is nice because it puts those in .demarc., but I'd be interested to know how applicable the method is to any other ASes you considered, and why/why not.

I didn't find the growth rates over time to be particularly compelling. Can you instead look into the nature of the expansion? (To what degree...) is Cogent increasing capacity in places they already have coverage? Upgrading technologies (replacement of old infrastructure with new)? Adding customers? Expanding into new geographic regions? Can you tie any of it to publicly known Cogent plans/announcements? Is any of it a surprise? Are different technologies used in different places (either regions or, say, core vs. customer links etc.)? Why iffinder, which tends to have very limited coverage, and not also use other available alias sets or techniques?

The authors use iffinder for validation, but with the assumption that no interface is removed in the network. It might be possible to have reverse DNS entries for removed interfaces, and iffinder can be used to detect such removals.

4.1 needs to talk about how many aliases iffinder identifies and how many (if we assume your maps are accurate) it misses. The intro mentions inferring information about customers / peers. It would be interesting to expand on this. Are most of these connections already known, or can you infer some that aren't in known datasets? How much capacity do the peering links tend to have?

How do we know your codes in Table 1 are right?

Does the public data in 4.2 include any links you miss? Why do you think you miss a few cities that appear in it? Consider using pings from PlanetLab with speed of light constraints to identify bad locations (ISI had a geolocation paper at IMC last year that may have the data you would need).

4.3: The traceroute seeding methodology is not clear. Neat use of the census dataset.

5.3: Can you take a large set of traceroutes (Ark or iPlane, say) and confirm that you infer most of the Cogent links seen in these traceroutes?

Fig 6: This is neat, but it would also be neat to see it on a world map (using the airport code locations). Consider making this available with your dataset online, even if there is no room in the paper. What were the shortest paths you used to create this path: all links with unit weight and all pairs paths? The approach to infer and verify (or, rather, double-check in routing tables) links is neat. I was wondering how you would get the links in.

4. Summary from PC Discussion

The paper was accepted w/o discussion.