Border Gateway Protocol (BGP) and Traceroute Data Workshop Report

kc claffy
CAIDA
kc@caida.org

This article is an editorial note submitted to CCR. It has NOT been peer reviewed. The author takes full responsibility for this article's technical content. Comments can be posted through CCR Online.

ABSTRACT

On Monday, 22 August 2011, CAIDA hosted a one-day workshop to discuss scalable measurement and analysis of BGP and traceroute topology data, and practical applications of such data analysis including tracking of macroscopic censorship and filtering activities on the Internet. Discussion topics included: the surprisingly stability in the number of BGP updates over time; techniques for improving measurement and analysis of inter-domain routing policies; an update on Colorado State's BGPMon instrumentation; using BGP data to improve the interpretation of traceroute data, both for real-time diagnostics (e.g., AS traceroute) and for large-scale topology mapping; using both BGP and traceroute data to support detection and mapping infrastructure integrity, including different types of of filtering and censorship; and use of BGP data to analyze existing and proposed approaches to securing the interdomain routing system. This report briefly summarizes the presentations and discussions that followed.

Categories and Subject Descriptors

C.2.3 [**Network operations**]: Network monitoring; C.2.5 [**Local and Wide-Area Networks**]: Internet; C.2.6 [**Internetworking**]: Standards; C.4.2 [**Performance of Systems**]: Measurement techniques—Active

General Terms

Measurement, Management, Performance

Keywords

Internet measurement techniques, routing, topology, data analysis, validation, censorship, filtering

The world (of BGP dynamics) is flat!

Geoff Huston offered some eye-opening data analysis [2] regarding the long-term evolution of BGP dynamics, in particular the rate of growth of observable BGP updates over time compared to rate of growth of BGP-speaking networks themselves. For years many network operators and engineers have predicted that BGP would not scale with the continued growth in Autonomous System (AS) numbers and would eventually melt down, causing significant reachability failures. Following an IAB-sponsored workshop and recommendations on this issue [16], IETF and IRTF working groups have made efforts to build replacement protocols [23, 8]. However, based on examination of actual BGP updates over the last 15 years, Geoff has discovered that despite a doubling in the number of BGP prefixes in publicly observable core routing tables over the last 15 years (now at 300K), the median number of updates per day has not changed! Although, the pool of noisy prefixes constantly changes – on any given day, between 20,000 - 40,000 prefixes send updates – the average number of updates is relatively constant. Why are updates not growing as quickly as the number of networks attached to the Internet? Is it a natural constraint or possibly the way we use BGP that does not push its limits? There may exist some relationship between topology (e.g., number of unprepended path lengths attached to an AS) and the level of dynamic updates as seen by any given AS.

Inferring interdomain routing policies

Recognizing that one of industry's strongest criticism of Inter- net AS topology research is the lack of accurate AS relationship data, Vasileios and his advisor Shi Zhou have investigated the use of BGP policy attributes, in particular communities and local preference, to provide another data input that may improve the accuracy of AS relationship inference [24]. They used Routeviews and RIPE-NCC's Routing Information Service (RIS) to gather AS connectivity information and BGP communities, and used `peeringdb.com` and `traceroute.org` to find 32 route servers and 50 looking glass servers that provided supplementary community and local preference information. They inferred AS routing relationships using the 32-bit community attribute and publicly available documentation of the interpretation of specific community values found on the Internet Routing Registries (`www.irr.net`) and PeeringDB `peeringdb.com` web sites. The local preference field does not express policy but often provides insights into AS relationships. They believe they have used these attributes to credibly infer about 40% of the relationships (of 110,000 observed inter-AS links), which they believe are mostly Tier 1 and Tier 2 links, i.e., the core of the Internet. They hope to extend their application to infer more of the inter-AS links, integrate traceroute data into their inferences, and study IPv6 relationships [3].

There was group consensus that there might be additional value in the use of community information, especially as researchers attempt to refine traditional heuristics such as the "valley-free" rule as complex AS business relationships become increasingly popular. There was also agreement that IPv6 relationships should be studied separately, since the economics driving IPv6 are vastly different (and in some cases non-existent). Geoff pointed out that IPv6 BGP analysis starkly reveals the presence of "ghosts" of bad information, i.e., prefixes that are announced and soon withdrawn but not before others ASes have propagated the prefix's reachability. Although the same phenomenon occurs in IPv4, it is seldom highlighted because it does not impact forwarding. In other words, consumers assume all paths seen are equally believable, but an arbitrary snapshot of a routing table will have relationships that do not exist because people do not look for withdrawals.

Realtime BGP data access

Dan Massey presented the status of his real-time BGP routing instrumentation project BGPMon [15], including a live demonstration of the latest version [14]. BGPMon is designed to scalably monitor BGP updates and routing tables from many BGP routers simultaneously, while providing a consolidated user-friendly interface. BGPmon uses XML to represent BGP messages, handling all attribute and element types, and various classes of data (2-byte or 4-byte ASN, v4 or v6 peer, etc.). Compared with other BGP monitoring software such as Zebra and Quagga, BGPMon has the following advantages: real-time access to BGP data, streamed to clients in flexible and extensible XML format; periodic route refresh to keep monitor in sync; no BGP complexity to worry about (no risk of emitting routes!); scalable single user interface even when chaining to 100's of peers; and easy configuration via a Cisco-like command line interface. BGPMon is not a full-fledged BGP implementation, and does not have to peer directly with a BGP-speaking router; it gathers data from several existing Multithreaded Routing Toolkit (MRT) routers, and there is a patch to the Quagga software collector that allows it to ship data to BGPMon.

The main job of the BGPMon route collector is to gather updates and feed them into the peer queue and into the RIB table for classification and labeling. Additional features include multithreading, negotiating keep-alives, and route refresh capability. The current release (7.22) introduced the concept of chains as a dynamic structure, allowing the BGPMon front-end system to act as a gateway and a layer of protection from upstream data providers.

A lively discussion ensued regarding what happens when clients cannot keep up with the BGPMon XML feed. Dan confirmed that clients that can read at line speed should not lose any data, but even the most capable clients may fall behind at times, so the server supports a notification to clients of how many messages are being skipped to bring the client up to date if they fall behind. This feature does introduce a risk if a client monitors for origin hijacking, but at least the client knows when they missed something, and it keeps slower clients from affecting others. (Geoff was not impressed, and gently reminded us that if we want to know how BGP views the network, we need to speak real BGP.)

For clients that want to focus on specific subsets of routing data, e.g., prefixes, the BGPMon team offers Hermes, a monitoring/filtering mechanism to scale down the data to only relevant records based on regular expressions applied to any BGP attributes or elements. Their next steps include linking to data plane (traceroute) information, starting with support for real-time queries to the BGPMon system for all reachability information related to a specific prefix.

Geoff Huston raised the point that the routes seen by Routeviews/RIS collectors depend on whether the peering AS treats a Routeviews/RIS collector as a customer or peer. As follow-up to the workshop, Dan confirmed that Routeviews always requests full tables, not a peer view. Ultimately the ISP decides its own policy and can theoretically change the policy at any time without notifying the collector. However, since the ISP is volunteering the data, it seems unlikely that the ISP would intentionally falsely report their policy, or apply unique filters to a feed contributed to a community monitoring platform. Earlier this year Dan's group conducted a separate study comparing reachability seen by a set of ISPs providing data to Routeviews [6], finding little difference in overall reachability among peers providing in excess of 300,000 routes. As another data point, the BGPmon site at Colorado State peers with six ISPs for testing and evaluation: five peers reported that they provide full tables, which varied in size from 356,408 to 368,301 routes. The peer that is not providing a full table has a table size of 144,224. The distribution of table sizes has a sufficiently low set of ranges, researchers can be reasonably confident that tables over 350,000 routes are full tables. This and other studies suggest that Routeviews data is predominantly full views and furthermore that one can reasonably infer the peering policy at a coarse granularity based on the size of the BGP table contributed by a given peer.

Dan's group plans to enhance the BGP data collection software so it reports a peer's table size as a parameter in each BGP update, which will facilitate inferences regarding whether the peer is providing a full table or partial view.

Measuring interdomain routing policies

Amogh Dhamdhere presented ongoing work on measuring interdomain routing policies. The goal of this work is to measure how often the rule of thumb for ISP routing policies (valley-free, prefer customer, then prefer peer) is used in practice. This assumption has implications for AS-relationships and various modeling/simulation work. The approach is to use Routeviews/RIS datasets and multiple AS relationship inference algorithms to identify the set of prefixes that each AS reaches via its customers and peers. Then identify routing anomalies where an AS reaches a prefix in its customer tree via peers/providers, or reaches a prefix in the customer tree of its peer via a provider.

In the discussion following the talk, Vasileios Giotsas suggested using community values to determine if customers request their providers to prefer an alternate route. Such cases should not be considered anomalies. Several participants also mentioned the SIGCOMM 2007 paper "In search for an appropriate granularity to model routing policy", as relevant to this topic [17]. The authors investigated how and where to configure per-prefix policies in an AS-level model of the Internet, such that the selected paths in the model are consistent with those observed in BGP data from multiple vantage points. They discovered that popular locations for filtering corresponded to valleys where no path should be propagated according to inferred business relationships, supporting the validity of the valley-free property used for business relationships inference. They also introduce a new abstraction to help model how ASes choose their best paths: next-hop atoms. Next-hop atoms capture the different sets of neighboring ASes an AS uses for its best routes. Many next-hop atoms correspond to per-neighbor path choices, but a non-negligible fraction of path choices correspond to hot-potato routing and tie-breaking within the BGP decision process.

A few days after the workshop an interesting email thread about ISP routing policies appeared on the NANOG mailing list, initiated by a negative review of a recently published academic model of secure routing deployment [18]. One of the complaints [20] was that all ISPs do not use the "valley-free, prefer customers, then prefer peers" model as their default routing policy. Disagreement ensued on the list regarding the underlying reality: how many ISPs really prefer customer routes over peers as a matter of policy? Randy Bush of IIJ claimed that multiple large global providers preferred peers over customers as their default policy. Patrick Gilmore of Akamai claimed that preferring peers over customers was the exception rather than the default policy for transit-free networks. Furthermore, the exceptions were cases where the customer explicit requested such behavior using BGP communities. Gilmore offered to conduct a survey of ISP routing policies and share anonymized results with the community. Sharon Goldberg, an author on the disputed paper, also posted a similar survey on the web [19].

Improving an AS traceroute tool

Despite several attempts over the last decade at AS-level traceroute utilities [1, 25, 13, 12], we still do not have a convenient standalone tool for performing accurate AS-level traceroute measurements, mostly due to the difficulty of accurately mapping the IP addresses in the traceroute output to their respective owning AS. Matthew Luckie presented his latest ideas for developing a fairly robust algorithm for IP to AS mappings, using public BGP data for initial mappings and then filling in gaps with alternative data sources such as RIR allocation data, PCH IXP mappings, and router aliases from CAIDA's ITDK. Initial results improve paths with a low error count, but there is significant room for improvement, particularly in identifying and removing ASes from the AS path that are in the control but not the data plane.

Measuring macroscopic censorship

Minaxi Gupta presented some early thoughts on how to measure certain types of Internet censorship, e.g., blocking websites, IPs, ports, protocols, keywords, removing content from flows, and blocking access to the web entirely. She conducted preliminary measurement experiments this summer, systematically accessing different types of Internet resources via proxies within countries known to censor, and comparing the responses with accesses from the U.S. to infer censorship behavior. Her first experiments focused on China and Iran, using 20 free proxies in each country. She found that censorship behavior was much more prevalent than she imagined. Also, while Iranian censorship focused more on blocking entertainment-related websites, Chinese censorship revolved more around websites related to Tibet, Taiwan and certain religions.

Censorship was implemented in multitude of ways, with the types of errors shown to users ranging from connection reset by peer, timeout, 403 forbidden and even timeouts. Interestingly, all ISPs in Iran appeared to either filter a website or not, but results varied significantly across Chinese ISPs, consistent with the observations in [27] that while most Chinese filtering occurs at the border, some choke points exist in many provincial networks. Another interesting observation was that censorship seemed to change over time, particularly in China. Though no clear trends could be inferred due to the short range of these measurements, her observations convinced her that we need a broader scientific study of censorship, which will require a cooperative community research platform that supports standard continual measurements of who censors what, when, and how. KC mentioned an example of such an effort: Rob Beverly's spoofer project, which is trying to empirically assess the prevalence of one type of (beneficial) Internet filtering [21]. Minaxi expressed interest in testing censorship at finer granularities, including based on keywords, ports, protocols, as well as the role of DNS poisoning. She also emphasized the importance of understanding the mechanics, location and statefulness of censoring devices, with the goal of finding avenues for successful circumvention.

Alberto Dainotti then summarized relevant details of a study that will be presented this year at IMC, related to censorship of Internet communications in Libya and Egypt in response to civilian protests and threat of civil war. His team analyzed several Internet measurement data sources available to the Internet research community – BGP updates and traceroute topology data as well as traffic to unassigned address space – to reconstruct the dynamics of the outages in Egypt and Libya. The goal was to characterize the nature and extent of the filtering, and to ascertain how different measurements can help detect and document future censorship activity. They used RIR delegation files and MaxMind's geolocation database to determine which IP address ranges in each country to monitor, and then mapped these prefixes of interest to BGP-announced prefixes and origin ASes using publicly available BGP data repositories in the U.S. and Europe (Routeviews and RIS). Using both control plane and data plane data sets in combination allowed them to narrow down which form of Internet access disruption was implemented in a given region over time. Among other insights, they detected what they believe were Libya's attempts to test firewall-based blocking before they executed more aggressive BGP-based disconnection. The methodology could be used, and automated, to detect outages or similar macroscopically disruptive events in other geographic or topological regions.

Discussion during and after these talks included mention of related work including the recent USENIX Workshop on Free and Open Communications on the Internet (FOCI'11) [22], Google's Transparency Report [4], the OpenNet Initiative [7], and Freedom House's "Freedom on the Net 2011" report [9].

Effects of RPKI deployment scenarios

Benno Overeinder of NLnet Labs used modeling and simulation [10, 26] techniques to study the effect of deploying Resource Public Key Infrastructure (RPKI) on security deployment scenarios to find out whether order of deployment (by AS) matters to maximize benefit to local infrastructure. Using abstract models of BGP behavior on top of CAIDA's AS topology data, he examined possible impacts of securing Tier 1 and Tier 2 ISPs first. He also considered the importance of securing CDN networks. Their simulation results showed that deploying RPKI within a small fraction of this top tier (5 to 10% of the top-tier ASes) could protect more than 98% of networks from prefix-hijacking, in the sense that those networks would receive only legitimate route announcements because the rogue announcements would be filtered/dropped by the top-tier transit ASes that deploy S*BGP. In contrast, their model predicts that without top-tier ASes involved in RPKI deployment, a RPKI-wise disconnected graph with "islands" that have valid origin ASes for a given prefix would lie in an "ocean" of vulnerability to hijacks for this prefix. Their conclusion was that top tiers must be involved in successful deployment of secure routing: no mid-tier deployment can outweigh a small top-tier deployment percentage.

Their modeling and simulation approach has been tested and validated against (convergence time distribution) results obtained from studies based on real-world experiments and analysis [11]. Further simulation experiments using different AS-level topologies as input have revealed that the number of BGP messages and BGP convergence times in different topologies are almost directly proportional to each other [5].

Putting secure BGP data in reverse DNS

Dan Massey presented recent work investigating an alternative path to BGP security, making use of the existing reverse DNS hierarchy, enhanced with DNS security. He and collaborators designed a scheme that can use the existing reverse DNS system, with two new record types, to embedded routing security-related information. A scheme that does not require fundamental changes to either the DNS or BGP protocols, but still allows address owners to manage their own certifications, is essential today since the IPv4 address grey market is emerging already – we do not have another decade to wait to deploy secure routing technology. To evaluate the feasibility and practicality of the scheme, they did a preliminary comparison of reverse DNS entries and BGP routing tables, to see how congruent the mapping might be. They found that of the 3.78M zones currently in the reverse DNS, BGP only needed 137K of them (3.6%) to map all of its prefixes, and up to 98% of the reverse DNS zones could deploy this new resource certification framework today. It would be helpful if other sites provided independent measurements of the reverse DNS tree for confirmation.

An even more open problem remains regarding how to map BGP prefixes (e.g 129.82.0.0/16) into reverse DNS zone names (e.g. m16.82.129.in-addr.arpa). Ideally, a single reverse zone would only contain prefixes from a single organization. A reverse zone that contains prefixes for both Colorado State and CAIDA raises questions about who owns the zone keys used to sign data, whether Colorado State can impact CAIDA, and vice-versa. The integrity of the naming convention for converting BGP prefixes into DNS names will likely determine the viability of this approach.

Dan also noted the need to study the likely extent of the path hijacking threat under this approach, specifically, at which hop of the path hijacks tend to occur. There are two types of route hijacking events; a competing route attack and non-competing route attack. In a competing route attack, both the real origin and the attacker announce the same route. For example, Colorado State announces 129.82/16 and an attacker also announces the same prefix in attempt to hijack Colorado State traffic. Some ASNs will prefer the valid path announced by Colorado State, but others will prefer the invalid path announced by the attacker.

In a non-competing attack, the attacker announces a more specific prefix that is not being announced by the legitimate owner. For example, Colorado State announces 129.82/16 and the attacker announces 129.82.128/17 and 129.82.0/17. By announcing these routes, the attacker hijacks the entire 129.82/16 address space. But the more specific attacker routes do not compete with the less specific valid route from Colorado State. A router does not need to choose between the valid and invalid routes, it stores both and then uses only the attacker's route when forwarding packets since the attacker's route is more specific.

Rather than direct efforts at path hijacking, imagine that one could block all non-competing routes. In other words, the attacker could only announce routes that the valid origin also originates, which is feasible with the DNS approach and may be feasible with RPKI as well. If one uses RPKI or DNS to secure the origin in last hop, a malicious actor could still lie about the third hop in the path, such as happened in the YouTube Pakistan scenario. But if most AS paths on the Internet are four AS hops or less – which most measurements indicate is the case – how significant of a threat is path hijacking at the third or later hop?

Open questions and work items

Interesting open questions and work items included:

1. Why are BGP updates not growing as quickly as the number of networks attached to the Internet?
2. Can we integrate traceroute data into inferences of AS routing relationships to further improve their accuracy?
3. What is the best way to study IPv6 routing relationships, given the very different economics in play?
4. The BGPMon team is enhancing the system to support real-time queries for reachability data related to a specific prefix, and add meta-data indicating a contributing peer's table size in order to facilitate inferences regarding whether the peer's view is full or partial.
5. The next step to improving AS-level traceroute methodology is to identify and remove ASes from the path that are in the control but not the data plane.
6. Progress on scientific study of censorship and filtering would greatly benefit from a community research platform that supports standard continual measurements.
7. Recent study of RPKI deployment using realistic AS topology scenarios as input suggests that is critical to get top-tier ASes in the hierarchy to be early deployers.
8. A recent alternative proposal for secure routing that relies on the DNS to embed route certification information would benefit from confirmation of empirical assumptions, including how congruent the BGP prefix tree is with the reverse DNS tree, and how long most Internet AS paths are.

Workshop participant list and material

The following participants contributed material to the workshop and to this final report: Emile Aben (RIPE-NCC), kc claffy (CAIDA, UCSD), Alberto Dainotti (University of Napoli Federico II), Amogh Dhamdhere (CAIDA, UCSD), Marina Fomenkov (CAIDA, UCSD), Vasileios Giotsas (University College London), Minaxi Gupta (Indiana University), Bradley Huffaker (CAIDA, UCSD), Geoff Huston (APNIC), Alistair King (CAIDA, UCSD), Matthew Luckie (University of Waikato, NZ), Dan Massey (Colorado State University), Benno Overeinder (NLnet Labs), Josh Polterock (CAIDA, UCSD), Shi Zhou (University College London), and Young Hyun (CAIDA, UCSD).
Links to presentation slides and this report are available at http://www.caida.org/workshops/bgp-traceroute/.

ACKNOWLEDGMENTS. The workshop was supported by the National Science Foundation via grant CNS-0958547. We thank all participants for their insights.

1. REFERENCES

[1] Internet Systems Consortium. prtraceroute tool, 2008. http://irrtoolset.isc.org/wiki/prtraceroute.

[2] Geoff Huston. BGP 2010-2011, 2011. http://www.caida.org/workshops/bgp-traceroute/slides/bgp-traceroute1108_flatbgp.pdf.

[3] V. Giotsas and S. Zhou. Detecting and Assessing the Hybrid IPv4/IPv6 AS Relationships. In *ACM SIGCOMM*, 2011. http://doi.acm.org/10.1145/2018436.2018501.

[4] Google. Google transparency report: Government reports, 2010. http://www.google.com/transparencyreport/governmentrequests.

[5] Shaza Hanif. Impact of Topology on BGP Convergence, 2010. http://www.nlnetlabs.nl/downloads/publications/bgp-topology-thesis.pdf.

[6] He Yan and Benjamin Say and Brendan Sheridan and David Oko and Christos Papadopoulos and Dan Pei and Daniel Massey. IP Reachability Differences: Myths and Realities. In *IEEE Global Internet Symposium*, 2011.

[7] OpenNet Initiative. Opennet initiative, 2011. http://opennet.net/.

[8] Internet Research Task Force. IRTF Routing Research Group. http://irtf.org/rrg.

[9] Sanja Kelly and Sarah Cook. Freedom on the Net 2011: A Global Assessment of Internet and Digital Media. http://www.freedomhouse.org/images/File/FotN/FOTN2011.pdf, Apr 2011.

[10] NLnet Labs. BGP Dynamics Modeling and Simulation, 2011. http://www.nlnetlabs.nl/projects/bgpsim/.

[11] Z. Morley Mao, Randy Bush, Timothy G. Griffin, and Matthew Roughan. Bgp beacons. In *ACM SIGCOMM*, IMC '03, pages 1–14, New York, NY, USA, 2003. ACM.

[12] Z. Morley Mao, Lili Qiu, Jia Wang, and Yin Zhang. On AS-level path inference. In *Proceedings of the 2005 ACM SIGMETRICS international conference on Measurement and modeling of computer systems*, SIGMETRICS '05, pages 339–349, New York, NY, USA, 2005. ACM.

[13] Z. Morley Mao, Jennifer Rexford, Jia Wang, and Randy Katz. Towards an accurate AS-level traceroute tool. In *SIGCOMM*, pages 365–378, Karlsruhe, Germany, September 2003.

[14] D. Massey, D. Matthews, H. Yan, and Y. Chen. BGPMon Demonstration. http://bgpmon.netsec.colostate.edu/stat_ver7-2.html.

[15] D. Massey, D. Matthews, H. Yan, and Y. Chen. BGPMon Monitoring system. http://bgpmon.netsec.colostate.edu/people.html.

[16] D. Meyer, L. Zhang, and K. Fall. Report from the IAB workshop on routing and addressing. IETF, RFC 4984, 2007.

[17] W. Muehlbauer, S. Uhlig, B. Fu, M. Meulle, and O. Maennel. In Search for an Appropriate Granularity to Model Routing Policy. In *SIGCOMM*, Kyoto, Japan, Aug 2007.

[18] Phillipa Gill and Michael Schapira and Sharon Goldberg. Let the Market Drive Deployment: A Strategy for Transitioning to BGP Security. In *ACM SIGCOMM*, 2011.

[19] Phillipa Gill and Michael Schapira and Sharon Goldberg. Routing Policy Survey, 2011. http://www.cs.toronto.edu/~phillipa/measurement/opsurvey/survey.php.

[20] Randy Bush. Do Not Complicate Routing Security with Voodoo Economics, 2011. http://mailman.nanog.org/pipermail/nanog/2011-September/039660.html.

[21] Robert Beverly and Arthur Berger and Young Hyun and kc claffy. Understanding the Efficacy of Deployed Internet Source Address Validation Filtering. In *ACM SIGCOMM)*, November 2009.

[22] Wendy Seltzer. Infrastructures of Censorship and Lessons from Copyright Resistance. In *USENIX FOCI Workshop*, August 2011.

[23] Cisco Systems. Locator/id separation protocol (lisp), 2010. http://lisp.cisco.com.

[24] Vasileios Giotsas and Shi Zhou. Inferring Internet AS Relationships Based on BGP Routing Policies, 2011. http://arxiv.org/abs/1106.2417.

[25] VOSTROM Holdings, Inc. Pwhois, 2007. http://pwhois.org/lft/.

[26] Maciej Wojciechowski. Border Gateway Protocol Modeling and Simulation, 2008. http://www.nlnetlabs.nl/downloads/publications/thesis_bgpsim.pdf.

[27] Xueyang Xu, Z. Morley Mao, and J. Alex Halderman. Internet Censorship in China: Where Does the Filtering Occur? In *PAM*, 2011.