USING SIMULATION TO UNDERSTAND DYNAMIC CONNECTIVITY AT THE CORE OF THE INTERNET

DAVID NICOL, BRIAN PREMORE
MICHAEL LILJENSTAM

Dartmouth College
Hanover, NH 03755
{nicol,beej,mili}@cs.dartmouth.edu

ANDY OGIELSKI

Renesys Corporation
Hanover, NH 03755
ato@renesys.com

Keywords : BGP, routing, Internet, connectivity

Abstract Recent academic research on the connectivity of the Internet uses modeling and simulation to conclude that the Internet is vulnerable to directed attack. The so-called "Achilles heel of the Internet" is the fact that a few nodes are highly connected, and that removal of the most highly connected 3% of Internet nodes disconnects the network. We point out that these studies overlook two important characteristics of the Internet that have tremendous influence on its connectivity properties. Actual connectivity depends on routers that run the Border Gateway Protocol (BGP). We first observe that the "nodes" of the Internet with the supposed weakness are massive nationwide networks, not routers. The threat of having a backbone network (such as the ones operated by WorldComm, Sprint, and AT&T) failing entirely is significantly smaller than the threat of a single router failing. Secondly, most connectivity studies are content to assert that two network nodes are connected if there is a path through the network graph between them. We also point out that connectivity is based on an overlay network defined by BGP, not physical connectivity. This paper describes how connectivity is *actually* maintained at the core of the Internet, and revisits the issue of connectivity in the face of router failures. A focus on router failures rather than network failures shows that the actual threat of massive disconnection is smaller than prior work suggests, but that in the wake of router failures the disconnectivity measured by reference to BGP's overlay network is significantly larger than disconnectivity measured by reference to physical paths in the network. We conclude that the significant differences in system behavior between earlier studies and ours suggests that studies of the threats to the Internet need to be clear on the nature of the threats being studied, and need to account for connectivity as it is actually experienced through routing protocols.

1 INTRODUCTION

The Global Internet is a federation of cooperating networks. These Autonomous Systems (AS) are individually owned and managed. Global connectivity is achieved by agreements (called "peering") between ASes to transfer traffic between themselves. Large ASes owned by corporations like WorldComm, AT&T, and Sprint connect with many other ASes, and so form the "core" of the Internet (also known as the "backbone"). These 20 or so Tier 1 networks have national (and international) scale internal backbones and all peer with each other. Smaller Tier 2 networks have regional scale internal backbones, peer with other Tier 2 networks and Tier 1 networks (however not all Tier 2 networks peer with all other Tier 2 networks). Tier 3 networks are smaller still (e.g. metropolitan scale). The smallest networks peer only with larger networks that are their service providers. Traffic does not transit across these smaller networks as it does in the larger ones—their peering relations serve only to deliver traffic to destinations within the network, and to carry traffic originating in the network to other ASes. There are presently over 15,000 ASes in the Internet.

1.1 Background on BGP

The flow of traffic between ASes is governed by routers that use a protocol called BGP (Border Gateway

Protocol)[15]. On receiving an IP packet on one port, a router forwards the packet through another port that takes it one hop closer to its destination. The forwarding tables used in this decision specify blocks of consecutive IP addresses (called prefixes) as targets, and contain the entire anticipated path from that router to the AS containing the prefix. The route chosen carries the packet to "the next" AS on that path. Presumably (but not necessarily, owing to race conditions) that AS uses the same path to the prefix and makes its forwarding decision accordingly.

An AS may have many routers that "speak" BGP. All of an AS's BGP routers use the same AS paths for every announced prefix. All BGP speakers in an AS peer with each other. A packet entering an AS will be routed internally through the AS to a router that has a network connection to the next AS on the packet's path.

A BGP router's forwarding tables are built and dynamically maintained. When a router selects a route to a particular prefix, it announces that route to its peers. Receiving such an announcement for prefix p, a router decides whether the new path just announced is a better way of getting to p (through the sender), than its currently stored route to p. If it is, the router makes its own announcement about p to all its peers (except the one which provided the triggering announcement, and possibly excluding others, depending on the peering policy), prepending its AS identity to the route. Thus if a router r for AS A receives from AS B a path BCD to prefix p and r decides to use that path to get to p, it will announce to all its peers that policy permits, except B, that $ABCD$ is the path to p that it will henceforth use, until further notice. If r had previously announced a path to p, this new announcement serves as an *implicit withdrawal* of the path announced earlier. In any case, for every prefix p, r always saves the last path announced by each of its peers to p. Saved-but-unused routes can serve as backups, as we will see.

Explicit withdrawals of previously announced paths also happen. For example if it appears to router r in AS A that its peer in AS B is no longer operating, it will look for saved-but-unused routes to p announced by other peers. Finding one it will choose the best and make an announcement based on that route. This is an implicit withdrawal. Failing to find one A will send an *explicit withdrawal* message to all its peers concerning the route to p it last announced.

Every router runs timers, one per peer, that measure how long it has been since it last sent a message (a message which may contain multiple announcements) to that peer. The timer is scheduled when a message is sent, and is scheduled to fire after some MRAI (a BGP configuration parameter) seconds; MRAI is typically 30 seconds. When the router wishes to forward an announcement to a peer, it checks to see whether the timer is running and if so the announcement is simply buffered. If the timer is not running the announcement is sent. When the timer fires, the router checks whether there are any enqueued announcements for the peer, and if so it sends them in an aggregated message. The role of MRAI is to limit the flow of traffic between routers so as not to overburden them, and to reduce instability.

Every router runs timers, one per peer, that measure how long it has been since it last *received* a message from that peer. The BGP specification calls for a peer to send a "keep-alive" message to another peer if as many as 30 seconds go by since the last message sent. If peer A fails to hear anything from peer B after 90 seconds, A considers its session with B to have failed and sets about announcing new routes (or withdrawals) for all prefixes it announced whose routes took them from A to B.

Two things happen when a router r reboots. It sends messages to all the routers it peers with. Those messages announce every prefix managed by r's AS, which are essentially announcements of zero-hop routes. When a peer reestablishes a session with r, it then announces to r the routes to each prefix—this is called a table dump because the entire forwarding table is sent to r. Router r responds to each such announcement as it would any other—it determines whether the route advertised by that peer forms the basis of a better way to the prefix than it last advertised itself. If so, r prepends its AS identity to the route and announces the route to peers other than the one that sent it.

As we have seen, a router r processing a peer's announcement decides whether using that route is "better" than the last route r advertised to that prefix. The definition of "better" is configured into r using general policy rules. In practice those policies have more to do with business relationships between ASes than they do with algorithmic concepts such as shortest path.

Academic studies of BGP behavior have considered issues of "convergence", that is, how quickly route change information that propagates through the network stops propagation, into a converged state. Background reading in this area includes [13, 9, 12, 8, 6]. Other studies focused on instability dynamics of BGP include [2, 3, 16, 5, 14, 4, 10, 4]. Studies about BGP policies include [11, 7]. Ours is the only study we know of that examines how BGP affects dynamic connectivity.

1.2 Connectivity of the Internet

In a widely noticed paper published in *Nature* [1], Albert, Jeong, and Barabási consider so-called *scale-free* networks under random failure, and under directed attack. The key characterization of a scale-free network is that it has a few nodes that connect with many many other nodes, while most nodes have low connectivity. They note that the graph whose nodes are ASes and whose edges reflect peering relationships between ASes is scale-free. Using graph disconnectivity as a metric, they consider two models of simulating connectivity failures. In both models a fraction f of nodes are removed from the graph; two remaining nodes are considered to be connected if there is a path between them using only edges between remaining nodes. The first method models random "error" by choosing nodes uniformly at random, while the second models "attack" by rank ordering the nodes by degree of connectivity and removes the most highly connected nodes. They note that network connectivity is resilient to the error process, but that removal of only 3% of the most highly connected nodes disconnects the network.

While this is a seemingly alarming result, it is useful to revisit the assumptions of this model. "Nodes" in this model are entire networks. For example, the mostly highly connected network in the AS system is UUNET, a network that utilizes massive telecommunications resources, and geographically spans the globe. In this model UUNET is equivalent to the AS of a small college campus, or ISP serving a rural area. Not all AS nodes are created equal; at a minimum the vulnerability of an AS to being completely disconnected from the Internet must depend on its size. This is not to say that a large network like UUNET cannot be so removed. In a recent incident a flawed software upgrade to the routing software used by UUNET effectively disconnected it. Accidental mis-configurations can in principle also cripple an AS. In addition to the relative difficulty of disabling an entire network, there is an issue of scale. 3% of the Internet's approximately 15000 AS systems is 450 ASes. The threat of having the Internet's 450 mostly highly connected ASes simultaneously disabled by a cyber-attack is rather smaller than having 450 *devices* disabled through cyber-attack.

When we look at the problem from the point of view of routers at the core we get a different picture. As we have seen, the set of BGP speakers associated with an AS logically connect with each other. If one of these routers fails there are many ways to route around it within the AS, and because core ASes all peer with each other, many ways of reaching any other AS connected to that router, through other ASes.

2 CORE CONNECTIVITY REVISITED

Our interest is in how connectivity behaves at the core of the Internet as routers fail and recover. Routers are increasingly the target of direct cyber-attack; recent events have shown that network wide disturbances (such as the Code Red and nimda worms) can also induce temporary router failure. Our description of the Tier 1 backbone structure suggests an abstraction that is reasonably close to reality. We will assume that a Tier 1 AS has n BGP speaking routers, and that these routers all peer with each other. If routers are represented by nodes, and peering relationships by edges, then the graph of an AS's BGP speaking routers is an n-clique. Now all Tier 1 ASes peer with each other, so at the AS level we have an N-clique (where N is the number of ASes). Peering relationships between Tier 1 ASes are realized by peering relationships between their BGP speaking routers, creating what we call a "clique-of-cliques" (CoC). In most of our CoC simulation models we take $N \geq n$, assume that there is exactly one connection between any two ASes, and arrange so that every router peers with approximately $(N-1)/n$ routers in different Tier 1 ASes (and exactly $n-1$ routers in its own AS). In one experiment we do create models where there are more connections between every pair of ASes; if each AS peers k times with every other AS, then every router peers with approximately $k(N-1)/n$ routers in different ASes.

Of course in reality BGP speaking routers in Tier 1 ASes peer with many routers in non-Tier 1 ASes; for the purposes of this study we ignore these, focusing only on the Tier 1 core and the connectivity from router to router within that core. Parameters N, n, and k describe a network core of N ASes, each with n BGP speaking routers, with k BGP sessions between every pair of ASes.

Not only is our topology different from the Albert et al. model, our approach to failure is different as well. We model the core "under attack" but with repair processes at work as well. This corresponds with reality—network operators actively work to restore routers known to have failed. Correspondingly we look at dynamic connectivity rather than static connectivity such as considered by Albert et al. In the set of experiments reported now the attacks are not directed. The high degree of connectivity within and between ASes will make the directed attack of

Albert et al. less effective at partitioning the network, but we have yet to quantify this.

To understand how dynamic connectivity is different from static connectivity, it is essential to understand what happens after a router r fails. Each of its peers will take between 60 and 90 seconds to react, due to “keep-alive” timeouts. As long as a peer still points to r in its forwarding table, any packet it receives whose path carries it to the AS containing r will be forwarded to r, and hence lost. So immediately following r’s failure there is a period of time when there may well be physical paths able to circumvent r, but the mechanics of BGP have not yet engaged to find and use them. When a peer of r realizes it has been too long since r was last heard from, it looks through all of its announcements for ones that cause it to forward packets to r. For each such prefix the peer looks for backup routes, based on announcements by ASes other than r. It then either announces the best backup route, or, failing to find one, explicitly withdraws the route it formerly announced. The propagation of new routes through the Internet is limited by the MRAI timer. This allows the possibility for further disconnectivity as packets are forwarded in accordance with outdated forwarding tables.

After a router reboots there is a period of time during which its peers realize it is back, and rebuild and announce preferred paths that use it. In this case too propagation of announcements is limited by MRAI timers; prefixes that can *only* be reached through r remain unreachable from points in the network that are unaware that physical connectivity has been re-established, until the appropriate announcements reach them.

Thus we see that when connectivity is considered from the point of view of the routing infrastructure’s ability to deliver traffic, it is possible for there to be physical connectivity between two devices in the Internet, but not logical connectivity. These differences crop up when a router first withdraws from the network, and when it re-integrates itself into the Internet. The question remaining asks how significant those differences are. We next describe simulation experiments designed to address that issue.

3 EXPERIMENTS

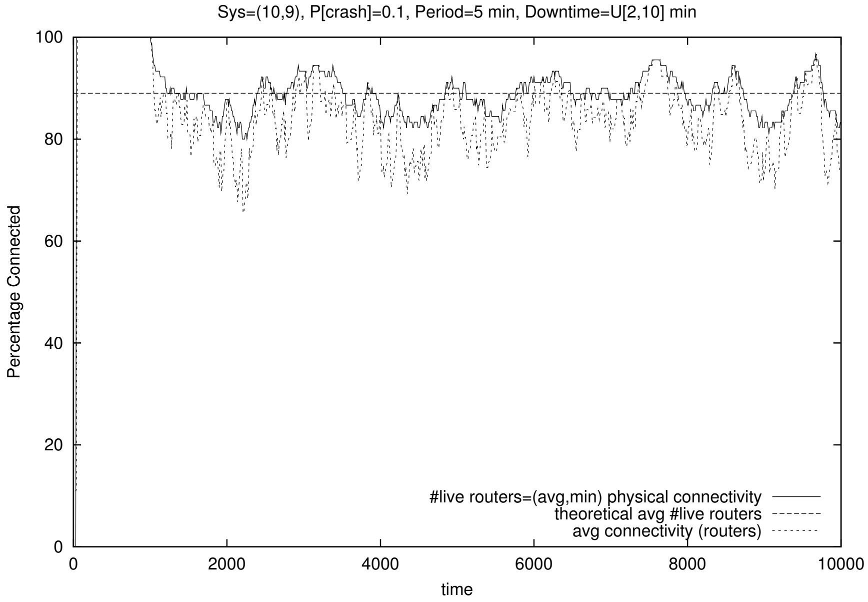
Our experiments concern core models with N ASes, each with n routers. The model sizes considered range from a few 10’s of routers, to nearly 400 routers. Unless otherwise noted, our experiments assume that two ASes peer through exactly one pair of routers, e.g., $k = 1$.

Each AS announces a prefix; each of the routers in an AS has an IP address from that block. To reduce complexity we put all of the model’s functionality into the routers; our model has no hosts. The router “devices” will also play roles needed to assess connectivity.

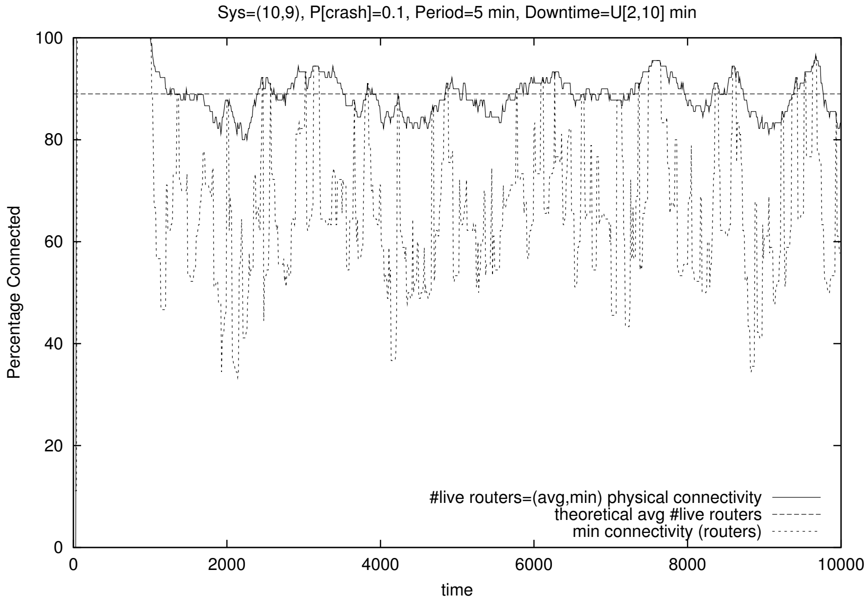
The routing of traffic within and between ASes involves a variety of complex of protocols : e.g. eBGP (external BGP), iBGP (internal BGP), TCP/IP, and OSPF. To assess the connectivity that results from subtle consequences within and between these protocols it is necessary to have detailed and faithful models of these protocols. In our study here we use SSFNet[17] as the simulation testbed, which satisfies these requirements. In SSFNet one can define very general protocol stacks for devices. In our model we construct devices that both route and serve as platforms for connectivity testing. The protocol stack for each device has internal and external BGP, TCP/IP, and ICMP. It also has a layer that implements the router’s failure model, as follows.

Every five minutes a running device decides whether to fail or not. The determination is random; with probability p_f it fails. On failure it samples a reboot time uniformly at random in the range 2 to 10 minutes; after this epoch has past, it re-inserts itself into the network by sending BGP “open” messages to all of its peers. This initiates the rebooting process we described earlier. Parameter p_f governs the intensity of the failure activity. In this model the fraction of devices that are running is the ratio of the mean time a device is up between reboots, divided by the sum of the mean time a device is up between reboots and the mean time it is down—$(5/p_f)/(5/p_f + 6) = (1 + 1.2p_f)^{-1}$. Because this network is so highly connected, under these assumptions the probability of the network being disconnected is vanishingly small. The fraction of physically connected pairs is then just the fraction of devices that are up.

Figure 1 plots observed physical and logical connectivity as a function of time. The network has 10 ASes, with 9 BGP routers per AS. p_f is 0.1, so that the long-term average fraction of physically connected pairs is 0.89%. We plot the measured fraction of live routers; in the first graph we also plot the measured fraction of logically connected pairs, and in the second graph we plot the measured fraction of devices that the least connected router can reach. Comparing the average connectivities we see a notable difference; comparing the minimum connectivities we see a marked difference, for under physical con-



Average Logical Connectivity



Minimum Logical Connectivity

Figure 1: Logical connectivity compared with physical connectivity

nectivity every live device is connected to all other live devices. Even if 90% of the devices are live, some routers at some times can reach fewer than half of all routers.

In a final set of experiments we considered the effect of adding redundant links between ASes, increasing model parameter k. To our initial surprise we found that with other parameters fixed, increasing k did not affect average connectivity. This made more sense to us after further study. Again because of the very high degree of redunancy in the clique-of-cliques architecture, for the ranges of models we studied, when $k = 1$ there is always a backup path for any announcement to a prefix whose preferred path included a recently failed router. Adding redundant links between ASes don't help.

4 CONCLUSIONS

Questions about the threat of the Internet being disconnected have many concerned, as they should. Some studies have examined the problem at abstract levels; we believe that different and more faithful (to reality) answers are found by conducting the study on more detailed models of the Internet, using more realistic measures of connectivity. Our results show that when attacks are considered at the level of a router, the high degree of connectivity within the core protects the network from massive disconnection. However we also see that measures of connectivity based on a physical link between devices can be quite different from connectivity that is experienced by devices within the network.

Important refinements to this work are needed. We need to develop a variety of attack models and reconsider questions about connectivity. We need to include geographical information that identifies which routers are physically close to others, to identify which peering sessions are supported by cables that are threaded through the same conduit. Consideration of geographic proximity will allow us to study the impact of physical attacks on the Internet core. We are currently working on issues such as these.

ACKNOWLEDGEMENTS

This research was supported in part by DARPA Contract N66001-96-C-8530, NSF Grant ANI-98 08964, NSF Grant EIA-98-02068, Dept. of Justice contract 2000-CX-K001, and Department of Energy contract DE-AC05-00OR22725. Accordingly, the U.S. Government retains a non-exclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes.

BIOGRAPHY

David M. Nicol is Professor of Computer Science at Dartmouth College, and Director of Research and Development at the Institute for Security Technology Studies. He earned a B.A. in mathematics at Carleton College in 1979, and M.S. and Ph.D. degrees in computer science from the University of Virginia in 1985. His research interests include high performance computing, performance analysis, simulation and modeling, and network security. He is co-author of the widely used text, *Discrete Event System Simulation, $3rd$ Edition* (by Banks, Carlson, Nelson, Nicol). He has been Editor-in-Chief of ACM Transactions on Modeling and Computer Simulation since 1997. He is a Fellow of the IEEE.

References

[1] R. Albert, H. Jeong, and A. Barabási. Error and attack tolerance of complex networks. *Nature*, 406:378–382, July 2000.

[2] Bilal Chinoy. Dynamics of Internet Routing Information. *Proceedings of SIGCOMM 1993*, pages 45–52, September 1993.

[3] Sally Floyd and Van Jacobson. The Synchronization of Periodic Routing Messages. *IEEE/ACM Transactions on Networking*, 2(2):122–136, April 1994.

[4] Lixin Gao and Jennifer Rexford. Stable Internet Routing Without Global Coordination. In *Proceedings of ACM SIGMETRICS 2000*, June 2000.

[5] Ramesh Govindan and Anoop Reddy. An Analysis of Internet Inter-Domain Topology and Route Sta-

bility. In *Proceedings of INFOCOM 1997*, pages 850–857, April 1997.

[6] Timothy G. Griffin and Brian J. Premore. An Experimental Analysis of BGP Convergence Time. In *Proceedings of ICNP 2001*, pages 53–61, November 2001.

[7] Timothy G. Griffin, F. Bruce Shepherd, and Gordon Wilfong. Policy Disputes in Path-Vector Protocols. In *Proceedings of ICNP 1999*, pages 21–30, October 1999.

[8] Timothy G. Griffin and Gordon Wilfong. An Analysis of BGP Convergence Properties. In *Proceedings of SIGCOMM 1999*, pages 277–288, August 1999.

[9] Craig Labovitz, Abha Ahuja, Abhijit Bose, and Farnam Jahanian. Delayed Internet Routing Convergence. In *Proceedings of SIGCOMM 2000*, pages 175–187, August 2000.

[10] Craig Labovitz, Abha Ahuja, and Farnam Jahanian. Experimental Study of Internet Stability and Wide-Area Backbone Failures. In *Proceedings of the International Symposium on Fault-Tolerant Computing*, June 1999.

[11] Craig Labovitz, Abha Ahuja, Roger Wattenhofer, and Srinivasan Venkatachary. The Impact of Internet Policy and Topology on Delayed Routing Convergence. In *Proceedings of INFOCOM 2001*, pages 537–546, April 2001.

[12] Craig Labovitz, G. Robert Malan, and Farnam Jahanian. Internet Routing Instability. In *Proceedings of SIGCOMM 1997*, pages 115–126, September 1997.

[13] Craig Labovitz, G. Robert Malan, and Farnam Jahanian. Origins of Internet Routing Instability. In *Proceedings of INFOCOM 1999*, pages 218–226, March 1999.

[14] Davor Obradovic. Real-time Model and Convergence Time of BGP. In *Proceedings of INFOCOM 2002*, June 2002.

[15] L. van Beijnum. *BGP*. O'Reilly, 2001.

[16] Kannan Varadhan, Ramesh Govindan, and Deborah Estrin. Persistent Route Oscillations in Inter-Domain Routing. Technical Report 96-631, USC/Information Sciences Institute, March 1996.

[17] `www.ssfnet.org`.