A Study on the Routing Convergence of Latin American Networks

Xiaoliang Zhao, Daniel Massey
Information Science Institute
University of Southern California, USA
{*xzhao, masseyd*}*@isi.edu*
Dan Pei, Lixia Zhang
Computer Science Department
University of California, Los Angeles, USA
{*peidan, lixia*}*@cs.ucla.edu*

Abstract

BGP is known to suffer from slow routing convergence after network connectivity changes. In this paper we identify the impact of network connectivity on the routing convergence delay and discuss its implication for networks in the Latin American region. BGP routing table snapshots showed that some networks in Latin American region are directly attached to large Internet service providers, while others attached to regional services providers. Our study shows that, when an edge network loses some of its connectivities, its attachment point to the Internet has great impacts on the BGP convergence delay. Our analysis shows that proximity to large Internet service providers improves the convergence time. We confirm our analysis through both simulation experiments and BGP routing log analysis for specific Latin American destinations.

I. Introduction and Previous Work

The Internet is composed of thousands of Autonomous Systems (ASes), loosely defined as networks and routers under the same administrative control. BGP[13] is the *de facto* inter-AS routing protocol. It is well-known that, when a topological change occurs, BGP routers often take a long time to explore a large number of transient routes before converging on a new stable route. Measurements in [6] found that the delay in Internet inter-AS path failover averages to 3 minutes, and some non-trivial percentage of failovers trigger routing table oscillations that may last up to 15 minutes.

Figure 1 shows an example of BGP slow convergence as observed from a single router's view. This example occurred in the Internet on Jan. 25, 2003, and the prefix involved in this example was originated by an AS registered in Latin American and Caribbean Internet Addresses Registry (LACNIC) [8]. As Figure 1 shows, at 5:32, AS 4777 changed its path, which had been used for hours, to a longer path. This message is an indication that the old path was

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No DABT63-00-C-1027 and by National Science Fundation(NSF) under Contract No ANI-0221453. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the DARPA or NSF.

LANC'03 October 3-5, 2003, La Paz, Bolivia
Copyright 2003 ACM 1-58113-789-3/03/0010.....$5.00

TIME	AS Path announced by AS 4777
03:07:45	4777 2516 3561 1916 10715
05:32:32	4777 2516 1239 3561 1916 10715
05:34:56	4777 2497 701 4230 8167 10715
05:36:20	4777 2497 2914 701 4230 8167 10715
05:36:47	4777 2497 1 701 4230 8167 10715
05:37:15	4777 2516 209 701 4230 8167 10715
05:37:42	Withdrawal
06:03:45	4777 2516 3561 1916 10715

Fig. 1. Slow convergence example in the Internet

no longer available, and AS 4777 attempted a new available path. However after trying several paths, the router ended with sending a withdrawal message, which suggests that the destination may have been unreachable from the beginning of the search. This example illustrates that, after a route failure, it can take time for the network to converge to the final view. This delay is commonly referred as “delayed convergence problem” [6], [7]. In this paper, we will show that even when an alternative path to a destination network exists, switching from a failed primary path to the alternative path, common referred to as *route failover*, may also suffer from similar slow convergence delay.

During the convergence of the routing system, the data packets may be delayed, lost or reordered, which may adversely impact end-to-end communication performance. The slower the routing system converges, greater would be the impact. [6] measured the packet loss and latency during routing convergence and found that the packet loss grows by a factor of 30 and latency grows by a factor of four. According to [4], packets can also be trapped in a loop during routing convergence, which wastes the network resources such as routers’ CPU time and bandwidth.

What are the major factors that lead to BGP slow convergence? In [6], [7] Labovitz et al concluded that BGP convergence time is bounded by the network diameter. In this paper, we show that the exact attachment point of an edge network plays an important role in the routing convergence delay, when the edge network connection status changes. As the topological connectivity among Internet service providers continues to increase, the impact of the attachment point on the convergence delay becomes even more pronounced. Our simulation results clearly demonstrate that, if an edge network is attached to a regional service provider with limited connectivity to the backbone providers, then the loss or change of reachability to the edge network may cause long convergence delay. Analysis of real BGP routing table revealed that most of Latin American networks have limited connectivity to the Internet, and our analysis of BGP routing logs show that some Latin American networks suffered the long convergence delay.

A few studies [12], [2], [3] have addressed the delayed convergence problem before. [12] proposed to let a node fully utilize the local information, and remove inconsistent information. By removing the outdated information faster, the assertion approach also helps prevent outdated information from propagating further. The *Ghost Flushing* approach proposed in [2] requires that a node send a withdrawal when this node has changed to a longer path than the previously sent path but this new path has to be delayed due to MRAI timer. Since a withdrawal message is not limited by the MRAI timer, such a “flushing withdrawal” message can quickly flush outdated path information.

The paper is organized as the following: First, we analyze topological characteristics of Latin American networks in Section II. Then we study two real world cases of convergence in Section III. Section IV analyzes the impact of the attachment point of an edge node on convergence. Section V presents the simulation results. Section VI concludes this paper.

II. Latin American network topology study

According to Internet Assigned Numbers Authority [5], the IP prefix 200/8 was assigned to LACNIC on Nov. 2002. We collected daily BGP routing table snapshots from Route Views server on January 2003. For each routing table, we searched for those prefixes which were in form of 200.x.x.x and which were originated by registered LACNIC ASes [1]. Then we studied the AS paths reaching those prefixes to examine the topological property of Latin American networks. In the following, we will present the numbers by averaging the daily result over 31 days period.

Averagely 4110 LACNIC prefixes were presented in the routing table per day. And there were 13629 unique AS paths to reach those IP prefixes, which were composed by 489 unique ASes. Out of 489 ASes, 424 ASes were registered LACNIC ASes, *LAC ASes* for short. From all AS paths, by searching the "neighboring and non-LAC ASes" of LAC ASes, we found that there were 36 non-LAC ASes serving as the first transit stop for any LAC AS to reach the rest of the Internet. Those 36 non-LAC AS are richly connected to the Internet with the median node degree of 24, and the list includes Sprint, UUNet, Global Crossing, Telefonica, AT&T, etc, which provide data transit services for LAC networks.

We are more interested to know where the LAC ASes are connected to the Internet. The AS path information reveal that 157 (37%) LAC ASes are connected to the Internet via the direct connections to the non-LAC ASes, 240 (56.6%) are one hop away from non-LAC ASes. Only 27 (6.4%) LAC ASes are more than one hop away from non-LAC ASes. In addition, the AS paths information also reveal that there are 182 (42.9%) LAC ASes are single-homed and 141 (33.3%) LAC ASes are dual-homed. There are only 101 (23.8%) ASes that are multi-homed to more than 2 other ASes. Thus there are two topological characteristics of LAC ASes:

1) Flat Topology: most of LAC ASes (93.6%) are very close to the richly connected ASes in the Internet (directly connected or one hop away).
2) Limited Connectivity: most of LAC ASes (76.2%) are connected to the Internet via no more than two connections.

III. Convergence Case Studies

Here we will study two real-world BGP convergence cases obtained from real BGP data.

A. BGP Data

We primarily used data which are collected from the Oregon Route Views server [14]. Currently, the Oregon Route Views server collects route updates from 37 BGP routers at geographically dispersed locations. The Route Views server provides a diverse view of routing states so that we are able to study Latin American networks from a number of different vantage points. However, because of policies, not all routers export their view, if any, of Latin American networks to the observation points. In this study, we monitored 27 routers which reports routing activities of studies networks.

To study the convergence time, we have to know the starting time and ending time of the convergence process. However, to precisely determine the starting time and ending time requires the external knowledge of the routing event which triggered the convergence process. Such external information usually are not available. Furthermore, if two or more events happened quite closely in time, like temporary link failure [2], it will further complicated the measurement. This work is interested in studying the topological impacts on the convergence time, so those multievents' effects are rather annoying, because multiple events could lengthen the convergence time to be very long regardless of variations of topologies.

Therefore, the best way we can do is to estimate the convergence time. In this paper, such estimation is based on the measurement of the duration of BGP update burst. As introduced in section I, when a routing event occurs, BGP routers will start to exchange the new routing information and recompute the best paths. Due to its distributed and asynchronized nature of BGP, a router may go through couple of cycles

[1] Registered AS numbers are documented at LACNIC website [9]

[2] The temporary link means a link failed for a very brief time, saying for seconds, but recovered quickly. In this case, BGP will react to the temporary link failure as two events occurred, one is *link down* event, the other is *link up* event.

of receiving new information, computing the best path, then sending the new path, before it finds the final path. Therefore, a router may continuously advertises different paths during convergence, which is referred as the *update burst*, or termed as *instability burst* in [11]. Here we are particularly interested in the burst triggered by a single event. As an attempt to separate the events, the following heuristics was used to filter the bursts:

1) No route changes occurred before the burst for a while, which is trying to eliminate the possibility that the current burst was triggered by a previous event.
2) No route changes occurred after the burst for a while [3], which is trying to estimate the ending time.
3) Most monitored routers changed paths during the burst, which is trying to focus on the routing events occurred close to the studied Latin American networks.

After the filtering and further manually examined the burst to filter out the bursts possibly triggered by multiple events, we found two interesting cases in January 2003.

B. Case One

In this case, we study the prefix 200.135.0.0/16, which is originated by AS 10715, a registered LACNIC AS. BGP logs show consistently that after 7:53am January 23, all 27 monitored BGP routers started to send BGP updates for the said prefix. Before that, the prefix had remained stable for about 7.8 minutes. Around 7:58am, all monitored routers had withdrawn the prefix and remained stable for 24 minutes thereafter. Such consistent views from diversely located BGP routers strongly indicate some routing changes, such as link failure, had occurred very close to the origin AS.

Depending on the timing, the topologies between monitored routers and the studied prefix, the local policies of different ASes along the AS path and many other factors, different monitored routers may take different time to learn the final path individually. Figure 2 shows some sample results obtained from 27 routers. In this case, there are 11 monitored routers took longer than 100 seconds to stabilize, and the longest delay is 283 seconds. For example, AS8121 experienced the longest delay and it tried six different paths during the course of path exploration.

We could estimate the convergence time by measuring the convergence time of monitored network topology, as partially drawn in Figure 4. Basically, we use the first message which a monitored router sent as an earliest indication of on-going routing changes, so the time of the first message will be used as the estimated starting time for the convergence process. Similarly, we use the last message a monitored router sent as the estimated ending time. The first message was sent at 7:53:42 and the last message was sent at 7:58:25. So it took at least 4.7 minutes for the whole Internet to converge [4].

C. Case Two

This case studied the prefix 200.33.143.0/24, which was originated by LACNIC AS 6332. From around 9:02am on January 25, all routers started to send updates for this prefix. At 9:05am, all routers had withdrawn the prefix, and 15 minutes later, the prefix was announced reachable again.

In this case, there are 16 routers took less than 60 seconds to stabilize, and the longest delay is 150 seconds. For AS8121, it immediately withdrew the prefix without exploring any paths in this case. Figure 3 shows some sample results.

We could estimate the convergence time in the similar way as case one. The first message was sent at 9:02:13 and the last message was sent at 9:05:15. Thus, it took at least 3 minutes for the network to converge in this case.

[3] In this paper, we empirically set the "quiet time" before and after the burst as five minutes.

[4] Obviously, we underestimate the convergence time because the starting time we used is later than the actual starting time. The real convergence has already taken place before the first update were seen by the observation points. In addition, the ending time we used may be earlier than the real one because we could not know if all Internet routers had converged when the only monitored routers converged.

Peer	Time of first update	Time of last update	Paths exploited	Longest backup path length
8121	07:53:42	07:58:25	6	9
1	07:54:51	07:56:43	1	5
6539	07:54:59	07:56:52	1	6
7911	07:55:06	07:57:24	2	6
5511	07:55:11	07:57:24	2	6
3561	07:57:02	07:57:02	0	5
7018	07:57:21	07:57:21	0	6

Fig. 2. Convergence of Case One

Peer	Time of first update	Time of last update	Paths exploited	Longest backup path length
7018	09:02:20	09:02:20	0	3
3561	09:02:43	09:03:11	1	3
5511	09:02:52	09:04:37	1	4
1	09:02:55	09:03:24	1	3
7911	09:02:55	09:05:15	2	5
6539	09:02:56	09:03:23	1	4
8121	09:04:06	09:04:06	0	7

Fig. 3. Convergence of Case Two

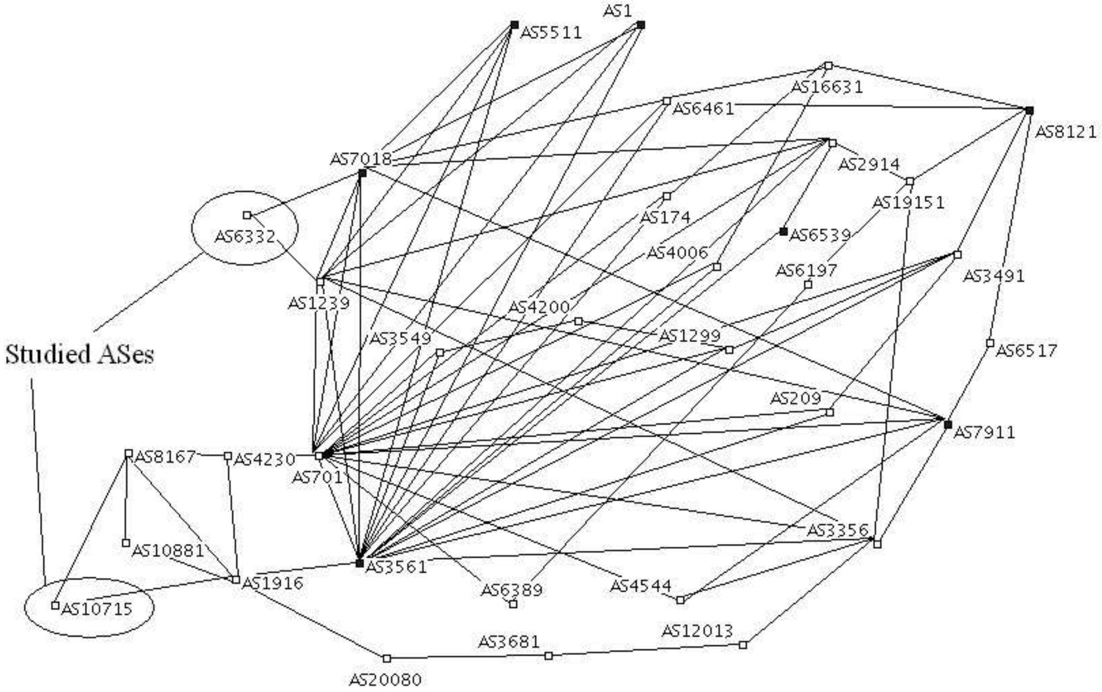


Fig. 4. Topology example. The nodes with dark color are the ASes where the monitored routers are located.

D. Analysis

Labovitz et al observed that the convergence time averaged three minutes [6]. The first case shows that the studied Latin American prefix experienced relatively long convergence delay ($\geq$ 4.7 minutes). Figure 4 draws a graph of the topology between monitored routers and the studied prefix. We could see that the prefix's origin AS (AS10715) was first connected to two intermediate providers, which are believed to be the Latin American regional providers according to the LACNIC WHOIS database [10]. Then the two intermediate providers connected to the richly connected ASes, such as AS3561 and AS701, which are non-LAC ASes and they are large international Internet service providers according to the WHOIS database from American Registry for Internet Numbers (ARIN) [1]. We believe the topological characteristics of how AS10715 connects to the Internet is a major factor which causes such long convergence delay, as analyzed in Section IV.

For case two, data show that it took shorter time to converge from both individual routers' point of view and the estimation of network convergence time. Figure 4 shows that the second prefix's origin AS was directly connected to two richly connected ASes, which are also large international Internet service providers according to the ARIN WHOIS database. We argue that different attachment points result in the differences in convergence delays between two cases, as we will analyze in the next section.

IV. Edge Node's Attachment Point and Convergence Time

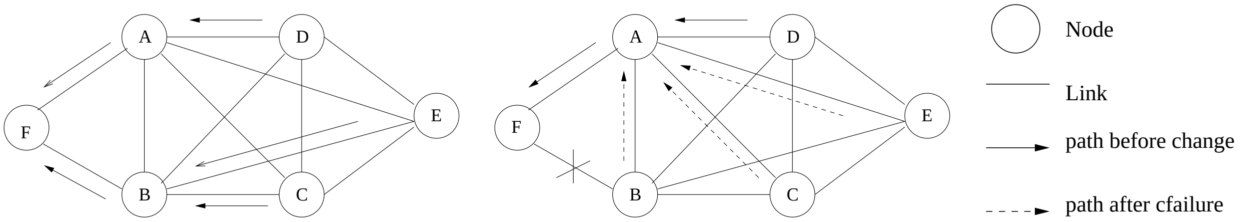
In this section, we mainly use examples to explain why the location where the edge nodes are attached to the network is one of the major factors of convergence time.

As shown in Figure 5 and 6, five nodes A, B, C, D and E connect with each other and form a clique, which attempts to mimic the network "core" in which the large providers are richly connected with each other. In Figure 5, node F represents an edge node, which directly connects to the clique via two links. In Figure 6, node I represents an edge node which has a direct connection with the clique, but the another connection to the clique is via several nodes. When the connectivity of edge nodes fails or changes, other nodes in the network may experience the slow convergence as described in Section I. We next will show that the convergence delay may be different for node F and I because of the different attachment points they are connected to.

For the first example, as shown in Figure 5, when node F's link $<F,B>$ fails, node B will detect the link failure and search its local routing table to find a new best path. For clarity, we only consider the simplest case: selecting the shortest path as the best path [5]. In such case, node B will select the shortest path (B,A,F) as the new best path. And node B will send its new path to all its neighbors, including node C and E, which use node B as their next hop. Upon receipt of the new routing information from node B, node C and E will realize node B now reaches the destination via a longer path. Consequently, node C and E will re-select the shortest path from their local routing table respectively and they will find going through node A is the best path now to reach node F. In this case, note that the new best paths have equal length as the old one, and all other paths containing the failed link have longer length. Therefore, according to the shortest path policy, the new shortest path, as well as a "valid" path containing no failed link, will be selected *first*, which results in the fast routing convergence. This example illustrates that when an edge node directly connects to multiple richly connected nodes, one link failure may not trigger slow convergence, as described in section I.

However, for the second example, as shown in Figure 6, when node I's link $<I,B>$ fails, eventually, all nodes will failover to the path using the path segment (A,F,G,H,I). Before node A selects this path, node A will try shorter paths first, such as (A,D,B,I), which is "invalid" because it contains the failed link. Similarly, other nodes like C and E will also try some shorter path first, and such path exploration

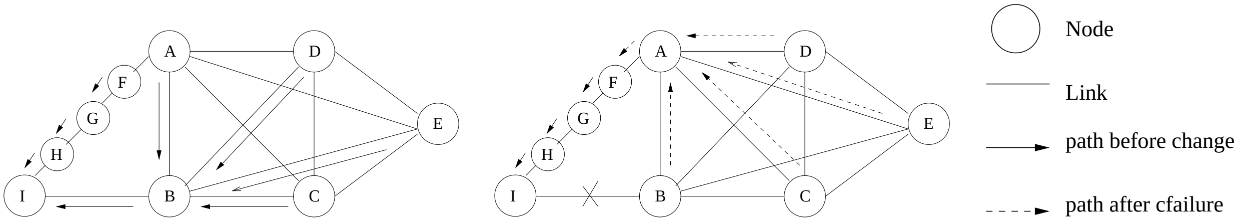
[5] In reality, one may have to take the various local policies and route selection criteria into account.



(a) Network before route change

(b) Network after route change

Fig. 5. Fast convergence example



(a) Network before route change

(b) Network after route change

Fig. 6. Slow convergence example

will take some time. After explored all possible shorter path, node *A* starts to announce the "valid" path, which results in other nodes to converge eventually. In this example, the finally chosen path has a relatively significant path length difference from the old best path, and there are some "invalid" paths with the path length in between. Again, according to shortest path policy, those "invalid" paths will be selected first, which results in slow convergence. This example illustrates that when an edge node is attached to the network via poorly connected nodes, there are good chances to introduce the path length differences between different connections, which, in turn, may cause the edge node to suffer the slowness of routing convergence.

Similar analysis has been conducted in [7], which mainly analyzed the case that the origin AS was completely disconnected from the network, defined as a T_{down} event. This study shows that in the case when the origin AS only partially lost its connectivity, the slow convergence may also occurr and the attachment point has a great impact on the convergence delay.

Also note that our case studies in Section III probably are resulted from a failure close to or at origin AS, which may be different from the examples we analyze in this section. However, Labovitz et al [6] found that *failure* and *failover* form equivalent class in terms of convergence behavior, which means if the convergence time caused by failure is long, the convergence time due to failover probably will be long as well.

V. Simulation Results

This section will present the simulation results to confirm our analysis described in the previous section.

A. Simulation Settings

In this work, we used the SSFNET [15] and a built-in BGP simulator to simulate BGP behavior with different topologies. SSFNET was designed to model and simulate the behavior of various network protocols in large networks.

The topologies used in the simulations include simple topologies, such as Clique-like and B-Clique-like. A Clique-like topology of size *n* used in this simulation is a clique (full mesh) connecting with an extra node via two links. An example of Clique-like topology with size 6 is shown in Figure 5. A B-Clique-like topology consists of $n+4$ nodes. Nodes $1,\cdots,4$ constitute a chain topology of size 4, and nodes $5,\cdots,n+4$ constitute a clique topology of size n. Furthermore, node 1 is also connected to node $n+4$, and node 5 is connected to node 4. An example of B-Clique-like topology is shown in Figure 6. In our simulation, we simulated various size of Clique-like and B-Clique-like topologies with the size from 6 to 63.

Empirically, in all our simulations, the Minimum Route Advertisement Interval (*MRAI*) timer value is configured to be 30 seconds with a random jitter, which is normally seen in real network operations. The link delay is set to be 2 milliseconds. The processing delay of each routing message was randomly generated during simulation to be between 0.1 and 0.5 second. Note the *MRAI* timer value can play a more dominant role during convergence, since it is typically 30 seconds, significantly larger than the link delay and processing delay.

The *Convergence Time* is measured in our simulation as the difference between the starting time and ending time of convergence process. The starting time is measured as the time when the failure happens, and the ending time is measured as the time when the last BGP messages is sent.

When a routing event occurs, it will trigger the process of BGP convergence. Labovitz et al [6] defined four routing events: T_{up}, T_{down}, T_{short}, and T_{long}. We are only interested in T_{long} to match our analysis in Section IV. T_{long} events will be injected into the simulations after the network has been stabilized. For simple topologies, we compare the convergence time between Clique-like and B-Clique-like topologies, similar to the examples in Section IV. For both two kind of topologies, we fail the link directly connecting the clique and the node outside of the clique. We run 10 simulation runs for a particular size n then average the results.

B. Results

Figure 7 shows the results for simple topologies, where the solid line represents the convergence time for a B-Clique-like topology with size n and the dotted link shows the convergence time for a Clique-like topology with the same size. Clearly, for B-Clique-like topologies, when the link failure occurs between the clique and the node outside the clique, the network converges much slower than Clique-like topologies. It confirms our analysis in Section IV. Furthermore, when the network becomes larger, the convergence time will be even longer for B-Clique-like topologies. Thus it may imply that for the Internet such a large scale network, the convergence time for those nodes connecting to small providers might be quite longer than those with large providers.

The simulation results clearly demonstrate that the convergence time is closely correlated to the location where an edge node attached to the network. When an edge node is connected to the Internet, multi-homing to the richly connected providers may help to improve the routing convergence.

VI. Summary

In this work, we studied the BGP routing behavior of Latin America networks, especially their topological characteristics and routing convergence time. From real BGP data, we found that some Latin American networks are connected to large Internet service providers, while others are connected via regional service providers. Our studies show that when link failure occurs, the networks attached to different providers experienced different delay of routing convergence. We conclude that rich connectivity among the ISPs is an important factor in the convergence behavior and proximity to large Internet service providers improves the convergence time.

References

[1] Whois database of the american registry for internet numbers (arin). whois.arin.net.

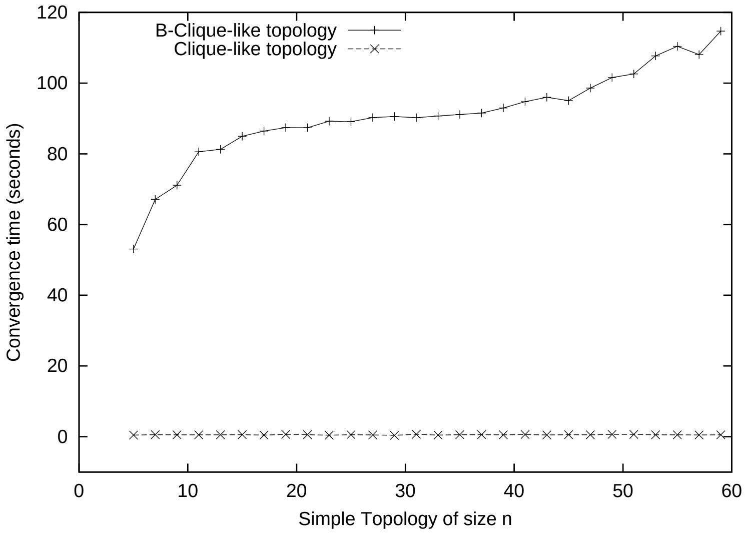


Fig. 7. Simulation results of B-Clique topology

[2] A. Bremler-Barr, Y. Afek, and S. Schwarz. Improved BGP Convergence via Ghost Fluching. In *Proceedings of the IEEE INFOCOM*, April 2003.

[3] D. Pei and M. Azuma and N. Nguyen and J. Chen and D. Massey and L. Zhang. BGP-RCN: Improving BGP Convergence Through Root Cause Notification. submitted for publication.

[4] D. Pei and X. Zhao and D. Massey and L. Zhang. A Study of Transient Loops in BGP. submitted for publication.

[5] Internet Assigned Numbers Authority. http://www.iana.org.

[6] C. Labovitz, A. Ahuja, A. Bose, and F. Jahanian. Delayed Internet Routing Convergence. In *Proceedings of ACM Sigcomm*, August 2000.

[7] C. Labovitz, R. Wattenhofer, S. Venkatachary, and A. Ahuja. The Impact of Internet Policy and Topology on Delayed Routing Convergence. In *Proceedings of the IEEE INFOCOM*, April 2001.

[8] The latin american and caribbean internet addresses registry (lacnic). http://www.lacnic.net.

[9] Lacnic ftp site. ftp://ftp.lacnic.net/pub/stats/lacnic/.

[10] Lacnic whois database. whois.lacnic.net.

[11] O. Maennel and A. Feldmann. Realistic BGP Traffic for Test Labs. In *Proceedings of the ACM SIGCOMM '02*, August 2002.

[12] D. Pei, X. Zhao, L. Wang, D. Massey, A. Mankin, F. S. Wu, and L. Zhang. Improving BGP Convergence Through Assertions Approach. In *Proceedings of the IEEE INFOCOM*, June 2002.

[13] Y. Rekhter and T. Li. Border Gateway Protocol 4. RFC 1771, SRI Network Information Center, July 1995.

[14] The Route Views Project. http://www.antc.uoregon.edu/route-views/.

[15] The SSFNET Project. http://www.ssfnet.org.