Impact of Topology on BGP Convergence

by
Shaza Hanif

Dissertation for
Masters in Parallel and Distributed Computer Systems
Vrije University Amsterdam

Supervised by:
Benno Overeinder, NLnet Labs
Guillaume Pierre, Vrije Universiteit Amsterdam

Date: 23rd August 2010

Abstract

For a few decades, the Boarder Gateway Protocol (BGP) realized Internet to be robust and acceptably stable despite its overwhelming growth. Although it is a fairly simple peer to peer protocol, yet because of the scale of Internet at which BGP is deployed, it is very difficult to understand behavior of BGP. It is considered a complex task to state what *trends* BGP may show if different internal factors (i.e. routing table size, parameter setting at BGP routers etc.) or external factors (such as size and shape of network, flow of BGP traffic, underlying protocols, performance and capacity of BGP routers, etc.) are varied. In addition certain emergent behaviors may only arise when protocol is simulated at Internet wide scale.

We attempt to understand how the underlying topology; one of the factors effecting BGP; at which protocol is deployed influences BGP performance. We use a highly scalable simulator, capable of simulating current full scale AS-level Internet, and give diverse topologies as input to it. The topologies that we use have operational semantics like the real Internet. We found that BGP is sensitive to certain topological characteristics of Internet, while remains completely unaffected on variation in some other characteristics.

Key words

BGP, BGP Simulation, Convergence Delay, Internet Topology, Topology Generation, Autonomous Systems

Dedicated
To

My Parents,
Husband,
Brother (late)
And all those Unknown Souls
Who Strive to Make this World
a Better and Peaceful
Place to Live.

Acknowledgements

All praise is due for Allah who bestowed us power and wisdom to nourish our lives in the best way possible. Its my pleasure to pay my gratitude to my day to day supervisor Benno Overeinder for his thorough guidance and my VU supervisor Guillaume Pierre for his upper hand and supervision throughout the project span. Without their cooperation and timely direction, it might have been a worst project of my life. I want to admit that I tried my best to be a good researcher under their umbrella of guidance, knowledge and experience.

I specially acknowledge the technical support given by Ahmed from Simula Technologies by providing the topology generator tool. I am thankful for his direction in using the tool and his concern to complete the project.

I am immensely thankful to my husband Shahab, without his cooperation and willingness it was almost impossible to successfully complete not only this project but my whole Masters. I have no doubt in admitting that I was able to perform well only because I was not only fully relaxed and comfortable but encouraged from home. Special thanks to him for bearing me and for his encouragement and patting at the moments when I was discouraged and dishearten.

I pay my gratitude to my father, choti Ami jan and bari Ami jan who suffered being without our company and always gave us warmth of their lap through unending prayers. Many a times, I specially requested my both Ami's for prayers and always felt strength afterwords. I am also very thankful to my sisters and brother who continually knocked (via telephone) and inquired "whats the progress Sister?" It was due to their encouragements and pushes that enabled me to proceed during ups and downs in different phases of the project.

It would be a complete injustice if I didn't thank my old chaps; Rabail, Samia, Shahida, for bearing my weekly routine talkative sessions, and my new friends Munazza baji for accepting my apologies at 80% of tea parties she invited and Amal Baji for her patience and relaxation from home tasks during my last month. Special thanks to brother Saleem who was always there to quench my thirst of technical discussion throughout my Masters. Thanks to my students Niemel and Ghasia who not only willingly gave their help in making some diagrams but it always relaxed me when I taught them even at my rough times.

Shaza Hanif

Table of Contents

Chapter 1

Introduction

Despite the inexorable growth of Internet in last two decades, overall Internet has proven to be remarkably robust and acceptably stable. A considerable credit goes to the core inter-domain routing protocol, the Border Gateway Protocol (BGP) [1, 2]. Since its introduction in 1989 as a protocol for the exchange of network reachability information between Autonomous Systems (AS) [18], it has gone through several updates until its current version's recent specification [1]. Coping with the upcoming challenges that resulted due to Internet growth, it served Internet for a few decades and today BGP is de facto standard for inter-domain routing.

However, along with the success stories of BGP, there exists a darker side as well, exposing unexpected emergent behavior resulting from large scale protocol deployment across Internet. Several studies have revealed certain pathological behaviors of BGP. For example, it is possible that temporary or permanent losses of transit data can result from paths generated by BGP [6, 12]. Furthermore, with the saturation of IPv4 address space, IPv6 is to be deployed in coming 3 to 4 years, increasing the BGP table sizes dramatically [13]. A recent study showed that the rate of BGP update messages (churn) increases at a much faster rate than the routing table size [21]. The current trends in expansion of Internet worries the network operators and vendors [11].Consequently, many ideas were proposed to increase the robustness of the inter-domain routing, enabling it to meet the upcoming as well as current challenges. These include adjustment of BGP's parameters [14, 36], BGP protocol modifications [15, 35], or entirely new inter-domain routing architectures [16, 33, 37, 38].

However, very few proposals have been accepted the internet community and experienced widespread adoption. For instance, none of the technical modifications in latest BGP specification [1] are a result of new proposed research. In addition, none of the leading router manufacturer [17] has neither adapted the new proposals [15, 16] for BGP nor any parameter settings for the installed equipment for BGP speakers [14]. We suggest that this widespread unacceptability of new proposals is due to the inability to demonstrate the benefits of a research at Internet wide scale. Since the new technologies cannot be evaluated in the deployment environment (Internet), networking test beds and simulations need to play a significant role. These facts increased the desire to understand global BGP dynamics advantageously and motivated several studies using active [6, 7] and passive [3, 4, 5] measurement techniques upon both deployed Internet and on simulations. This project is a contribution towards better understanding of dynamic behavior of BGP using a highly scalable BGP simulator.

Before any new proposal for BGP parameter settings, modification in its specification, or new Internet architectural paradigm are accepted, there is an inherent requirement of understanding the behavior of BGP in its different specifications and its sensitivity on underlying layers/technologies. Furthermore the emergent behavior resulted from individual and component level settings of routers could be unexpected and unpredictable when tried on the magnitude of Internet [22]. It is also equally essential to figure out whether BGP behave with similar robustness in near future with expected growth of Internet.

Research to identify, analyze and answer the following questions is timely:

- Why does BGP behave the way it behaves? Why certain pathological, undesired and unexpected messages are observed [4, 5, 6, 7, 34] .
- How sensitive is BGP performance to underlying protocols and technologies?
- Which aspects should be put into consideration in case a BGP update or modification proposal is to be globally deployed?
- What is the influence of local AS level decisions about different parameter settings by router vendors and AS administrators on global behavior of BGP [14, 36]?

Although it is necessary to answer all these questions, it is equally important to realize that such an abstruse protocol with a complex behavior cannot be studied from all aspects at once. The protocol can be analyzed in steps and one aspect can be studied at a time.

1.1.Research Question

This research effort converge its focus on analyzing the dependence of BGP performance upon the underlying topology. It attempts to comprehend the response of BGP if Internet faces unavoidable substantial growth and if the current topology changes in characteristics other than its magnitude? More precisely, the research is an attempt to comprehend the influence of topology dynamics, in terms of its scale and Internet's operational topological characteristics, upon BGP performance.

1.2.Our Approach

We intend to use a BGP simulator [19] that implements the BGP protocol with certain level of abstraction. The abstraction enables the simulator to model at the order of magnitude of current Internet and higher, letting up to 60,000 AS's to interact with each other and maintain sessions of BGP. In the simulator, each AS is implemented as single entity similar to a vertex of a graph.

We are not interested in exact behavior of BGP in its current deployment environment, rather we want to understand and comprehend the dynamics of BGP.

Moreover we do not intend to analyze BGP fully and comprehensively, rather we want to study its sensitivity to underlying topology only. Our purpose is to find out different *trends* of behavior BGP exhibits, if it is run with different kinds of topologies and varying parameters. We want to understand the way BGP expresses, its strong and week points that appear due to certain kind of underlying topology.

It is also important to note that we analyze BGP behavior with respect to its convergence delay and signal duration [23, 19] across a topology. In simple terms, convergence delay is the amount of time in which a network information change is propagated over Internet. Signal duration also has a similar definition. It is not our intention to measure the amount of churn (number of BGP update messages) [34] faced by the routers across Internet. Nor do we intend to analyze variation in routing table sizes as the years pass.

Furthermore, by topological parameters we refer to parameters describing specific characteristics of real world Internet. We consider size, shape discussed in [32] for BGP convergence, as well as parameters like average node degree, node degree distribution, joint node degree distribution mentioned in literature [9, 10, 46, 47, 57] as mere graph theory aspects of topology. Actual Internet topology constitutes of relationships along with hierarchy, with loose boundaries between Tier 1, Tier 2 and Tier 3 AS's [2, 20, 25]. Moreover inter-AS routing in Internet is policy based, in which each AS has its own priorities for path selection and path preference for specific prefixes.

Therefore, our focus is to use topologies which keep the actual Internet structure and annotations in consideration. The annotations include AS link relationships in addition to connectivity information and different types of AS belonging to different Tiers in Internet. ASes have some specific attributes because of their position in the network hierarchy. The hierarchy constitutes a few large sized Tier-1 nodes at the top level and a customer AS at the bottom, with Transit relationship to a service provider. The kind of topology generator, we selected is further described in Chapter 3.

It is also important to note that with respect to BGP, we are not and should not be concerned with bandwidth utilization or availability, traffic volumes, link capacities at exchange points, etc.; such aspects of topology or network are not in our focus. They represent a separate study related to amount of pathological messages. Our prime focus is only the logical updates and amount of time they take to converge.

Chapter 2

Background

For a better comprehension of influence of topology on BGP, it will be more appropriate to first illustrate the related concepts and their relevance with our work. Starting from some basic concepts of Internet on autonomous system level, we will continue to explain related work for BGP, followed by giving some background for topology of Internet.

2.1.AS level Internet

Internet is divided into a large number of distinct regions of administrative control, commonly called Autonomous Systems (AS) [20]. An AS, also known as routing domain, typically consists of a network service provider or a large organizational unit, such as a college campus or a corporate network. In turn, each AS interconnects a number of sub-networks, such as remote corporate ones or customer networks. An AS has a single set and clearly defined routing policies [24] and connects to one or more remote ASes at neutral private or public exchange points

The routers in Internet are responsible for receiving and forwarding packets through this interconnected maze of sub-networks and ASes. Each router makes routing decisions based on its knowledge of the topology, the conditions on the network, and complex routing policies specified by network administrators within their domain. In order to make such dynamic decisions, routers exchange path and topology information using special purpose routing protocols. Broadly, there are two classes of such routing protocols:

1.An inter-domain (or exterior) routing protocol is used to exchange information between peer routers in different ASes.
2.An intra-domain (or interior) routing protocol, is used to pass information between routers within an AS.

Internally within an AS, routers use a variety of intra-domain (interior) protocols to distribute local routing information, including Open Shortest Path First (OSPF), ISIS, and IGRP [20]. Usually interior protocols build their own reliability on top of a datagram service [20].

Our focus is towards exterior routing protocols, more specifically Border Gateway Protocol (BGP) which is the most common, rather de facto inter-domain (exterior) routing protocol used by ASes in Internet. BGP uses TCP as its underlying transport layer protocol to exchange routing information about how to reach the destination prefixes. Routers exchange information of a route when there is a change in old information, such as an old route disappearing or a new route becoming available. The BGP update message includes list of

ASes with reachability information, along with other attributes such as next-hop IP address. This enables BGP to hide the topological details and routing inside each network domain. The routers that communicate each other using BGP protocol across a network domain are called BGP speakers. Routing information is propagated according to complex policies configured in BGP speakers by administrator.

BGP speakers within a domain synchronize using intra-domain routing protocols. Synchronization means routers exchange reachability information in such a way that all speakers have consistent information. Consequently, the BGP information collected from any border router should reflect the routing behavior of the AS depending upon local router policies, and local hardware or software failures.

2.1.1.Transit and Peering Relationships

An Internet topology model, for studying protocols like BGP, requires not only considering the inter-connectivity information of network but also the type of link or relationship [25]. Simulations of these protocols without relationship information may result in misleading inferences. The relationships between AS are generally described by one of the following two categories:

- Transit relationship: One AS pays money (or *settlement*) to another AS network for Internet access (or *transit*). It is also known as provider-customer relationship.
- Peer (or *swap*) relationship: Two networks exchange traffic between each other's customers without cost, and for mutual benefit.

Usually ASes of bigger size and administration provide transit services to ASes of smaller sizes. AS of more or less similar sizes have peering relationships with each other. Unlike transit, peering traffic always has one network as source and the other network as its destination.

2.1.1.BGP as a Routing Protocol

As stated earlier, BGP is an exterior routing protocol. The position of BGP amongst the other routing protocols can be understood by considering Figure 2.1. The AS65101, AS65202, AS65404, AS65303 represent independent ASes having routers R1, R2 and R3, R4 and R6, R7 and R9 respectively as BGP routers, having inter-connectivity as shown in Figure 2.1. As we can see that a single AS has more than one BGP speaking routers, but typically they reflect identical behavior based on how they are configured by administration.

The routers belonging to same AS interact with each other using interior routing protocols. Since they belong to same administration domain, they help the AS to maintain a stable behavior for the rest of Internet. This allows us to abstract the whole AS by a single node, as done in our used simulator [19], based on consistent similar routing role of different routers within an AS.

After a policy change or a network failure affects the availability of a path to a set of destinations, the routers topologically closest to the failure will detect the fault, withdraw the route and make a new local decision on the preferred alternative route, if any, to the set of destinations. For instance in Figure 1, in case of link failure between R1 and R2, R2 will withdraw the route information in which the path includes R1. These routers thus propagate the new topological information to each router within the AS. The network's border routers will in turn propagate the updated information to each external peer router, pending local policy decisions. Routing policies on an AS's border routers may result in different update information being transmitted to each external peer.

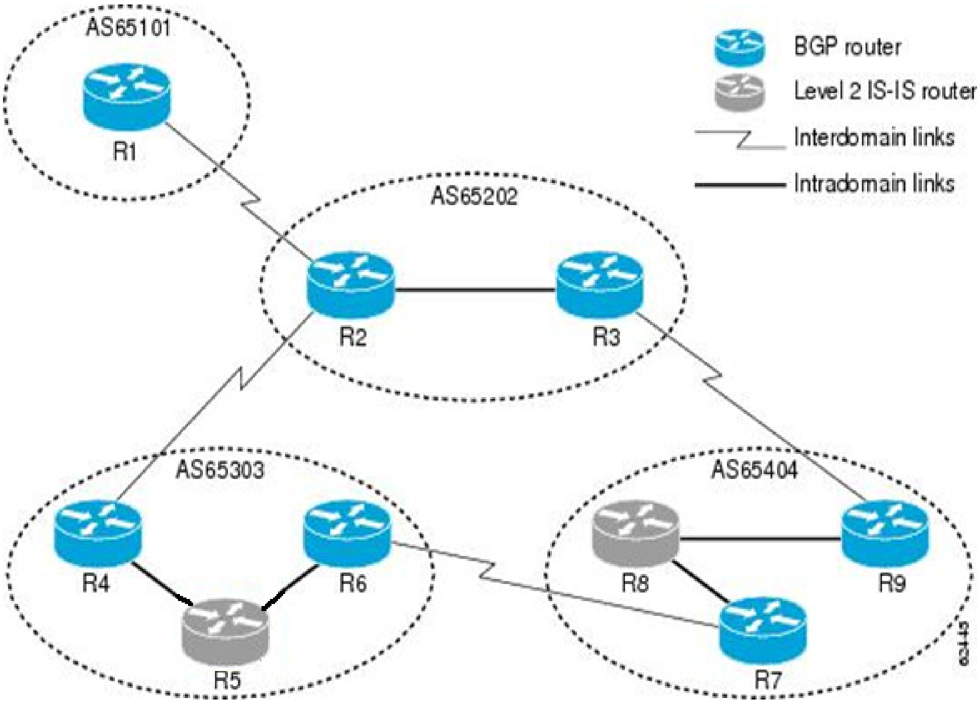


Figure 2.1: *BGP protocol illustration*

2.1.Current State of Related Research

In this section, we will go through related literature that comes under the domain of this project.

2.2.1.BGP Protocol

As BGP and all its enhancements are designed by network researchers and engineers, its specification itself, as a peer to peer protocol is very well understood. However, recent works by researchers [3, 4, 5, 6, 7, 12, 14, 15, 26, 27, and 29] have shown that BGP's dynamics are poorly comprehended. For instance, as mentioned in [26, 27, 29], there is considerable doubt about BGP route-flap damping [28], MRAI timer value configuration [1] and continuous recommendations about optimum values are presented by research community [30, 31, 36].

Studies like Labovitz et al [6] shows that establishment of stable routes after a node failure can take on the order of 3 to 15 minutes, but it is not very well understood when and why the convergence is delayed. Furthermore the reason and causes of high amount of unexpected messages are also not clear [3, 4]. Different problems like forward loops [32] and stable paths problem [35], oscillations [51] or even misconfiguration [12] have also been reported in an attempt to apprehend the protocol. In [51] it is stated that although BGP itself is made self-stabilizing, inconsistent policies my give rise to oscillations problem. Some tools like [8, 53] are also presented to use the BGP data available at the Internet for decision making at various levels and anomaly detection [52].

Routing instability, informally defined as the rapid change of network reachability and topology information, has a number of origins including router configuration errors, transient physical and data link problems, and software bugs. High levels of network instability can lead to packet loss, increased network latency and time to convergence. At the extreme, high levels of routing instability has led to the loss of internal connectivity in wide-area, national networks.

The community is eager to understand the dynamics and interactions of the protocol. The knowledge of the causes of BGP's behavior could lead to improvement of the reliability of routing in Internet. Additionally, understanding the interactions would enable us to influence the convergence processes.

2.2.2.BGP Simulation

Simulations should be able to mimic the behavior of BGP because whatever complicated dynamics it exhibits, fundamentally it is a peer to peer Gossip-based protocol. In this perspective, it should be possible to comprehend BGP and justify how and why it behaves the way it behaves. It is important to find a simulator by utilizing which we may have maximum resemblance of protocol behavior between the deployment environment (the Internet) and the simulated environment. Many attempts of evaluating BGP's operation have used a variety of approaches including networking test beds, simulation, and study of monitored network measurement data. These attempts were at small scale [7, 14, 15, 29, 32, 36], with only few exceptions going beyond a thousand ASes [34].

Modeling an original scenario is critical and requires including conditions representative of the deployment environment. Modeling Internet, from any aspect, is a challenging task [39]. This is mainly due to its enormous size, involving tens of thousands of Ases, which is continually increasing not only in number of ASes but also in inter-AS links [40]. These characteristics are not negligible for modeling and evaluating an AS-level routing protocol like BGP. Sufficient scale is required to understand the effect of local decisions of individual ASes upon global properties of robustness and convergence. However, large-scale Internet models tend to exhaust the resources of the test platform [41].

2.2.3.Internet Topology and BGP

The topology of Internet at the AS-level has evolved rapidly and its evolution pattern is changing due to network usage, development and deployment. New ASes arise daily and others disappear, and the connections between these systems also change. Recent studies [45] reveal that new ASes arise at the rate of 10.3 per day, while rate of disappearance is 2.87 per day. The links in this topology arise at a rate of 67.3 per day and disappear at the rate of 45.7 per day. To keep pace with this evolution, the topology categorized in terms of its properties. Knowledge acquired in characterizing this evolution is important in many areas of network research including topology analysis. It will also help in the implementation of topology generator for producing synthetic topologies which are used in network simulators and in laboratory tests. The construction of Internet-like topological graphs contribute to the effectiveness of tests and experiments of new protocols, updates of existing protocols and Internet applications.

The Internet research community has spent a decade to investigate the topology of Internet [9, 10, 46 and 47]. Before the appearance of research by Faloutsos et al. [44] hierarchal topology generators were considered best for topology generation because of their emphasis on maintaining structure but later, after the power laws discovery [44], different topology generators focusing on local properties like node degree distribution were proposed. For instance, Claffy and Krioukov [48] focuses on metric of Joint Degree Distribution along with other global and local metrics of graphs like average degree, degree distribution, clustering, rich club connectivity, spectrum, etc.

In [49], the authors claim that three metrics, expansion, resilience and distortion are the smallest set of parameters that can qualitatively distinguish the topologies into well defined categories. They also stated that graphs generated by considering local properties, without focusing on structural properties tend to automatically capture some Internet like hierarchal structure and thus generate topologies near to Internet.

Different types of topology generators are discussed by Mahadevan et. al. [50]; also the authors present a topology generator with some advanced concepts. But most of the above mentioned AS-level topology generators do not consider AS relationships, which is a very important characteristic for routing protocols. Dimitropoulos et al [25] and Elmokashfi et. al. [34] give one of the first approach to consider the AS relationships while generating topologies.

One of the initial work that looks into BGP from topological perspective is presented by Govindan and Reddy [54]. The authors focus on topology and its impact on BGP by varying the topology parameters of domain degree distribution, diameter and connectivity. Analysis was done on real-time data of a single backbone provider. The same authors latter claimed [55] for the first time to discover the Internet Map using a path probing tool, but it doesn't characterize the Internet through graph theory parameters.

The authors in [7] describe how Internet Topology adds in routing convergence delay of BGP. The work of some others [45, 55] uses BGP monitors, which are dedicated routers that analyze BGP messages at different spots across internet. These authors gave an insight to Topology discovery and characterization using BGP data available at several kind of BGP monitors i.e. looking glasses, Route Views [42] and route servers.

Chapter 3

Approach and Techniques

This chapter will define the adapted methodology of this research project. We will begin with describing several BGP and topology parameters that were used in the past for studying them, followed by stating available tools for generating topologies. Then we will illustrate our approach in a step-by-step manner.

3.1.Considerable Characteristics of BGP and Topology

BGP as a core routing protocol is one of the widely studied routing protocol. Researchers have investigated its performance and accuracy by following different ways and approaches.

3.1.1.BGP Parameters

BGP is analyzed in different ways by the research community using different metrics, depending on the research goals in their focus. Its performance is evaluated from various parameters using different methodologies. Some of these are listed below:

No. of Update Messages: Certain pathological behavior has been shown in [3, 4], where authors experimentally show that BGP propagate almost 99 percent of the messages which are not legitimate. They measured number of update messages by dividing messages into different types. In a recent paper [34], the amount of update messages were examined for different hypothetical topologies. We consider this work very relevant to ours.

Packet Delivery Performance: This is very rarely studied parameter of BGP. It is about the packet delivery performance when the route is not converged yet. In general, if the connectivity increases performance also increases, this might be because number of alternate paths increases and the probability that alternate path will go through the same link decreases. Both these factors increase delivery performance [56]. It is an important parameter but its significance is reduced if convergence time is minimal.

Stability or Instability: Some of the researchers tried to figure out in different cases if BGP is overall all a stable protocol, and will it converge at every possible scenario of topology and policy based routing? For instance, Ahronovitz et. al [51] states that the inconsistent policies of ASes may create oscillation problem. Through data received from various Internet backbones, Lavovitz et. al. [5] also lists stability problem. Elmokashfi [35] also focuses on stability of BGP by examining a small test bed.

Convergence: one of the widely studied properties of BGP is Convergence delay. It can be defined as time it takes that a route becomes stable over internet. Different researchers [2, 3, 4, 6, 7 14, 15, 23, 32] focus on convergence times from various perspectives. Convergence of BGP has also being analyzed with respect to router configuration parameters of MRAI and route flap damping [26, 27, 28 and 29]. Recommendations are given for the optimum values of MRAI or route flap damping for optimum convergence [30, 31 and 36]. We will also analyze performance of BGP by considering this parameter for our studies.

3.1.2.Topology Metrics

Since we want to study the performance of BGP for different kinds of topologies, it is important to know how topologies are characterized by network research community. There are many different topology generators that focus on different characteristics and dimensions of Internet while generating topologies. In broader terms there are two different ways in which we can categorize topologies:

Graph Theory Metrics: These metrics have their origin from the graph theory. They consider Internet topology as a graph consisting of only nodes and edges, and try to investigate it in different dimensions. Such parameters include average node degree, average distance, average clustering, resilience, spectrum as well as degree and distance distributions of a graph [9, 10, 46 and 47]. In most cases this kind of categorization does not take AS relationships into account, which is a major drawback and makes them far from real Internet graphs.

Operational Metrics: Internet topology is in fact a graph but with some additional features and characteristics that normal graphs do not have. It has AS relationships, policy-based connection utilizations, and different AS-types [2]. By operational parameters we mean metrics having operational semantics like real Internet has, i.e. number of Tier 1, middle or edge nodes, multi-homing degree, levels of hierarchy, peering intensity at different levels in the hierarchy, etc. In [46] we get one of the earliest such characterization of Internet which also presents a tool for generating topologies using such parameters. Later [34] investigated impact of different topologies on the number of BGP update messages by varying these parameters.

We wanted to investigate the impact of topology on BGP, and operational metrics being closer to real Internet gives us better approach.

3.1.3.Topology Generating Tools

For studying the influence of topology on BGP, we wanted to generate real Internet-like topologies. Random topology generators were inconsiderable, since Internet is far from a random graph. We had the following other options for generating Internet like topologies.

RealNet: RealNet tool [58] claims to generate real Internet like topologies. It takes the original core of Internet and then adds edge nodes according to the required scale. Although it takes into account AS relationships, yet we were not confident about its validation experiments. Further it was able to generate topologies with varying size only.

Orbis: Oribis [57] by CAIDA is a very powerful tool for generating as well as rewiring a topology. It is also capable of measuring certain properties like number of edges, average node degree, clustering etc. Although it claims to be an AS-level topology generator, an important drawback, was that it doesn't consider AS- relationships while rewiring or topology generation. Due to this, we were unable to use it for generating topologies. It captures most of the Internet properties and can be considered one of the best tools for manipulating and generating AS-level topologies without AS relations.

Tool by Dimitropoulos: Similar like Orbis, in [25] a tool is presented which also takes AS relationships into account. It is capable of scaling the topologies while maintaining most of its features. The reason why it was not selected for our topology generation was that it does not give us enough opportunity to generate different kinds of topologies. We can only generate topologies with varying sizes and no other operational metric could be varied. This would limit of our scope of researchable topologies.

ILTG: Internet-Like Topology Generator (ILTG) [34] was most appropriate for our purpose. It not only generates the topologies considering different AS types, and their relationships, but most importantly it gives us *knobs* to vary the topologies by changing operational semantics of Internet. Further details of tool are given in next section.

3.1.Our Approach

In this section we will briefly state the methodology of our work. We will illustrate work that inspired us and used by us as a reference work. We will mention the tools and the steps that we followed to achieve the research goal.

3.2.1.Reference Work

The most important aspect to validate a research is that its results can be compared and associated with reference to older work of the research domain. We were very alert and motivated for selecting reference points for our research work. Mao [23] presented the real world beacon's experiment which was used as reference point by [19]. Since we are using the same tool as developed in [19], we consider both of them closely associated with our work. In addition while we are using ILTG topology generator [34] as well as a subset of topologies types mentioned by them (using their tool), we consider our work as somewhat related to theirs. Since they have investigated BGP performance with respect to number of update messages, we consider our work complementary to theirs.

3.2.2.Tools Used

We used two major tools, one for topology generation and other for simulating BGP. The topology generated from topology generator is given as input to later. Their description is stated below:

a. BGP Simulator

In 2008, a highly scalable BGP simulator [19] is developed, capable to run on Grid infrastructure of the DAS-3 cluster [41] using 32, 48, 64 or 79 compute nodes. The simulator is validated by comparing the BGP beacons signal duration observed in real BGP network [23]. We used this simulator for our experiments and gave diverse topologies as input.

One of its noticeable features is that instead of focusing on intra-domain communication, network and protocol are highly abstracted. It mimics the behavior of BGP with a certain level of generalization in which each AS is treated as a single node without considering intra-AS relationships. This enables to have high scalability, but with loss of some accuracy. As a result one cannot analyze and make conclusive statements about a single AS behavior, however different trends of global BGP behavior can be inferred.

Furthermore BGP Simulator can take any kind of topology as input, whether it is a completely unrealistic topology like grid, torus or a real Internet topology taken from topology resources like CAIDA [43]. In [19] it is also validated by real world-topology (of year 2008) experiments.

One may argue about the benefit of inquiring the impact of topology on BGP behavior, when the actual AS topology with respect to BGP is not based on the way BGP peers have established BGP sessions with each other. Rather actually topology faced by BGP traffic is influenced by configurable policies and complex interactions with intra-domain routing protocols [8]. In response, there are two reasons that make studying the influence of topology on BGP dynamics worthwhile. First, it is not possible to obtain the policy information of an AS that it configured on each of its routers. Even for Internet's AS-level topology, the networking research community has spent decades [9, 10] for coming up with partial topology graphs [43]. It is irrational to wait for such information and then start the research about routing protocols. Second, policy based routing (and intra-domain protocol influenced topology) would be an *instance* of original AS-level topology, which implies that original topology is fully influential on policy-based routing. Moreover, as mentioned earlier, we are interested in global generic behavior and trends, not the exact behavior that is seen in the current Internet.

b. ILTG Topology Generator

It is very important to note that in our study, mere connection information between two ASes is not sufficient knowledge of a topology. We needed a topology generator that considers the network annotations i.e. AS relationships, during the generation process.

Internet-Like Topology Generator ILTG [34] is one of the few tools that generates Internet topologies with AS relationships, a very essential aspect for our study. It is capable of generating the topology with a high number of ASes, but if we increase the number of Ases,

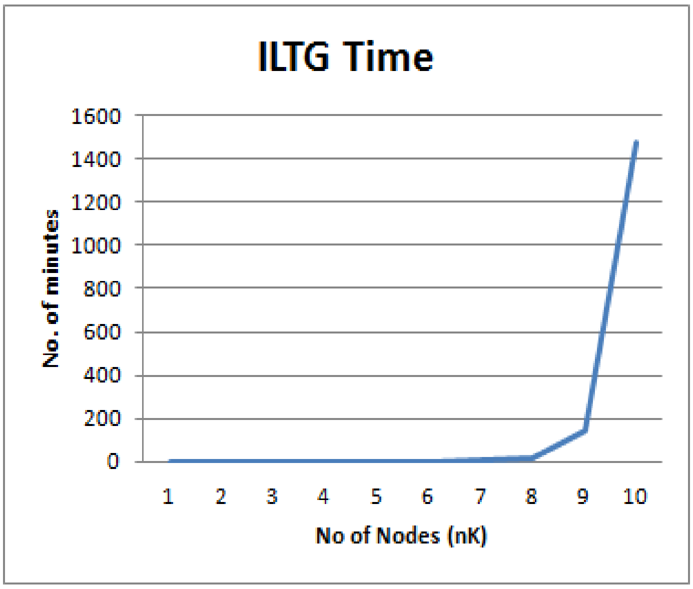


Figure 3.1: ILTG Time Complexity

the time complexity increases exponentially as shown in Figure 3.1. This time complexity limits us to generate topologies of not more than 10K nodes. Although simulator [19] is capable of simulating BGP with more than 60K AS nodes, we are limited by the selected topology generator.

Though we can create topologies with realistic growth scenarios, ILTG allows alteration of different topological properties. It has very interesting *knobs*, by which we can generate topologies with very different characteristics through parameter tuning.

The input parameters have operational semantics. Instead of taking graph theory parameters like clustering coefficient, assortativity of topology, average node degree etc., ILTG enables us to specify topology parameters with real world related aspects, i.e. number of Tier 1 ASes, multi-parenting degree between middle layer or at the edge of network etc. Thus it gives us the opportunity to generate annotated topologies with realistic knobs that makes it easier to map the network with real world semantics.

In ILTG as well as in our simulator, an AS is abstracted as a single node. If the term node is used with reference to topology then it means a node in an AS level Internet, which is basically an AS. ILTG assures that it captures four AS-level properties in its generated topologies. These characteristics despite tremendous growth of Internet in the last decade have remained constant [40]. These are:

Hierarchical Structure: There exists a notion of hierarchy in Internet topology, with a parent giving transit services to a child node. Normally the customer-provider relationships are formed in such a way that there are no provider loops where A is provider of B, who is provider of C, who again is provider of A. Yet there is multi-parenting i.e. one child having two transit providers.

Power Law Degree Distribution: Internet degree distribution is discovered to follow a power law, with only few nodes of higher degrees and a majority of the nodes with minimal degrees [44]. Usually the most well connected nodes reside at the top of hierarchy as Tier 1 nodes.

Strong Clustering: The nodes in Internet are grouped together in clusters, with nodes in same cluster more likely to be connected to each other. Internet has these natural clusters mostly because of different geographical regions.

Constant Average Path Length: Although the number of nodes in Internet have increased tremendously, average AS-level path length has remained approximately constant at about 4 hops for the last 10 years [40].

ILTG generates topology in top down fashion and differentiates the nodes in four types. Tier-1 nodes (T) reside at the top of hierarchy and form a peering clique. Attached to T nodes are Mid-level nodes (M), which have one or more providers that can be either T nodes or M nodes. M nodes may also have peering links with each other. Two different types of nodes reside at the bottom of hierarchy, customer networks (C) and content providers (CP). Only CP nodes can have peering relationships with M nodes or CP nodes, while C nodes do not have peering links.

Regions are modeled in ILTG to capture the clustering concept in Internet. Networks in one region are generally not allowed to connect with networks of other regions.

It is important to note that the unrealistic topological scenarios, i.e. without peering relationships or a topology with collapse of hierarchy, will not ensure the above mentioned four properties. However if the topologies are generated Internet-like, the mentioned properties are guaranteed to be found in them.

3.2.1.Methodology

After explaining the tools that we used, we are in the position to briefly state the step by step approach followed by us, Figure 3.2 . First of all, a topology was generated using parameter values according to the kind of topology we wanted to generate. This topology is passed to the BGP simulator as a parameter along with details of selected locations to serve as monitoring points and beacons. In addition, the simulator also takes a sequence of beacon events [23] in the form of input. As a last step, the output files from monitoring points are processed to obtain the require plots for analysis.

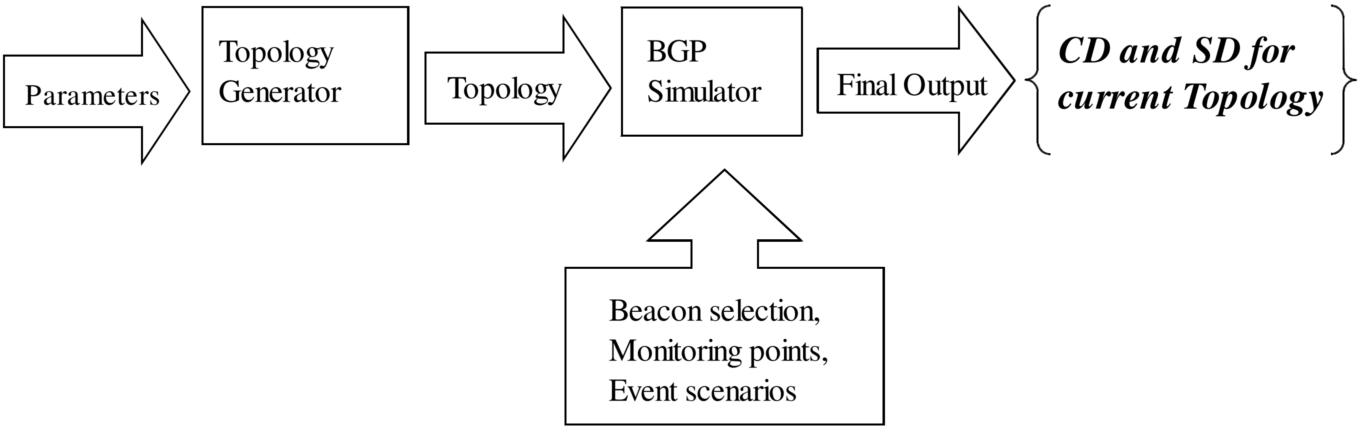

Parameters
Topology Generator
Topology
BGP Simulator
Final Output
CD and SD for current Topology
Beacon selection, Monitoring points, Event scenarios

Figure 3.2: Our Approach

In the next paragraphs, we will explain some terms, concepts, and ways that are used in our methodology.

Multi-parenting: Multi-homing [46, 47] concept in Internet literature means that Customers at the edge of network buy transit services from multiple transit providers. We will use a term multi-parenting to explain the concept that an AS at any layer in hierarchy is taking transit services from more than one parent. The AS can be from Tier 1 nodes, or from any other level in hierarchy.

Propagation Pattern of Input Signal: BGP is a distance vector routing protocol, which requires that each node maintains the distance from itself to each possible destination, and the vector or neighbor to use to reach that destination. Whenever this connectivity information changes, the node propagates its new distance vector to each of its neighbors, allowing each to recalculate its routing table [2].

We use the term *propagation pattern* to explain the routes followed by an information data to propagate over the network. The propagation pattern followed by BGP announcements and withdrawals are different. The difference in propagation pattern is the following. When an AS receives an announcement, it decides to whether to propagate it further or not, according to its policies. An AS can receive the announcements of a route from multiple neighbors and as a result it may or may not have to reconsider the propagation decision. Normally the propagation pattern for announcements resembles a tree graph with extending branches. Due to multi-parenting and peering, the nodes receive multiple messages from their neighbors, thus increasing the convergence delay and signal duration.

For withdrawals the situation is a bit different and more complex. When a particular route is withdrawn, a path vector algorithm like BGP attempts to find an alternate path by iterating over the available paths of equal or decreased priority according to local policy. This route is then propagated to the neighbors as currently available route. In case the destination has disconnected from Internet, none of the alternate routes would be valid. But the algorithm will explore all possible paths and then will end with no path going towards that destination. This path exploration [6] increases the complexity of understanding of withdrawal propagation pattern. Also, if an announcement of a correct path is received later than announcement of another invalid path, still that invalid alternate path will be explored. The reason is that BGP processes the messages received in order [6], which adds to the message complexity which in turn results in increasing convergence delay and signal duration.

Beacons and Monitoring Points: As mentioned earlier we will measure BGP performance by measuring convergence delay and signal duration for a given topology. There is a need for reference time by which for a given announcement or withdrawal, it can be observed that how much time it takes to converge. Wojciechowski [19] repeated the real-world *BGP beacons* experiment by Mao [23] using the simulator, we have followed the similar approach for our experiments.

In [23], a beacon is a publicly known prefix having global visibility and a published schedule for announcements and withdrawals. A beacon AS is the AS at which beacon the daemon resides. The announcements and withdrawals by specific beacons can be considered as input signals. We will use the term input signal, update message, BGP messages in similar context. The idea behind this is that since the schedule for input signal is known, and the prefixes are publically visible, one can monitor the prefixes at the other end of the network, and calculate the time by which the prefix routes become stable across Internet. Several observation or monitoring points can be installed in the network for calculating the convergence of a known input event. Similar strategy is adapted in all the experiments of different topologies.

It is important to note that monitoring points are silent BGP listeners. They receive the BGP update messages from their neighbors but are not responsible to send out the signals. This means they have no role in further propagation of the received or seen messages. In real-world environments the BGP monitors like Oregon's Route Views [42] are placed at core of network for this purpose.

We have tried to apply similar approach to select beacons and monitoring point as adopted in the real world experiments by Mao [23] and in simulator experiments by Wojciechowski [19]. For our experiments, two beacons were selected with node degree of 1 and 25. We will call them EDGE beacon and MIDDLE beacon respectively. Two BGP monitoring points, each with 25 highest degree nodes, were selected at the core of each network topology.

It should be added at this point that the strategy to select beacons at the relative edge of the network and monitoring points at the core, is a rational decision with two benefits. First we are able to compare our result interpretations with [34], in which BGP performance is shown with respect to number of update messages using similar strategy. Secondly, it makes the task of analyzing the propagation of update messages relatively simple. Since Internet has somewhat hierarchal structure, we can assume that if beacon is at the edge, the input messages travel up from the edge to the core of the network. If the monitoring points are not at core, we will have to consider that the messages received by the monitoring points may include not only messages coming from the edge of the network as well as those who met the core and then propagated down to the monitoring point. This makes the task of understanding the signal propagation pattern difficult and complex.

Signal Duration and Relative Convergence Delay: Mao [23] and later Wojciechowski [19] studied BGP beacons convergence delay by means of two different metrics, namely relative convergence time (shortly say convergence delay), and signal duration time. We will study both parameters and compare the results with the ones from the real BGP network of the current Internet which will make us understand how much variation is found in the current Internet.

Signal Duration (SD) is the amount of time between the first signal received by a monitor from one of its neighbors, and the last signal received from same neighbor. For instance Table 3.1 shows monitoring point M1 received one of its neighbors AS1 at time 1 and the last signal from AS1 is received at time 12 second. Therefore the signal duration is 11 seconds for AS1.

Neighbor AS	Beacon Number	Signal Type	AS Path	Reception Time
AS1	B1	Announcement	AS1, AS8, AS10	1
AS1	B1	Announcement	AS1, AS3, AS7, AS10	4
AS1	B1	Announcement	AS1, AS6, AS7, AS10	6
AS1	B1	Announcement	AS1, AS6, AS14, AS 9, AS10	12

Table 3.1: *Signals received by monitor M1 from its neighbor AS1*

Relative convergence time or in short *Convergence Delay* (CD) for an AS is the amount of time elapsed between first signal received from any neighbor AS and the last signal received from the AS from which perspective it is measured. For instance, if measure the CD for AS1, Table 3.2 shows that first signal received by M1 is at time 0 (from AS4) and last signal received from AS1 is at time 12. This means that CD for AS1 is 12 seconds.

Neighbor AS	Beacon Number	Signal Type	AS Path	Reception Time
AS4	B1	Announcement	AS4, AS9, AS10	0
AS1	B1	Announcement	AS1, AS8, AS10	1
AS1	B1	Announcement	AS1, AS3, AS7, AS10	4
AS4	B1	Announcement	AS4, AS6, AS7, AS10	5
AS4	B1	Announcement	AS4, AS3, AS13, AS10	6
AS1	B1	Announcement	AS1, AS6, AS7, AS10	6
AS1	B1	Announcement	AS1, AS6, AS14, AS 9, AS10	12

Table 3.1: *Signals received by monitor M1 from its neighbor AS1and AS4*

In this chapter we explained major steps of our adapted methodology. We also explained the reasons behind selection of various choices at different steps of research. During our research we also used some specific terms that were not used in Internet literature and we explained them as well.

In the next chapter we will give explanation of experiments carried by us. We will also define how we categorized those experiments so that it becomes simple to understand and drive conclusions from them.

Chapter 4

Experiments and Evaluations

In this chapter we will first describe the procedure in which various experiments were carried out, then we will discuss about the process we followed to accumulate our results. Finally, we will report what the outputs of different experiments are and what information can be deduced and comprehended from them.

4.1.Experimental Setup

We conducted a series of experiments with diverse topologies and evaluated BGP performance on them. We tried to keep the rest of BGP simulator parameters, other than input topology, constant and similar to the values used by Wojciechowski [19]. For each topology, we run three experiments, without changing any parameter. We have not found notable differences in all the three runs of a single instance and only included one of the runs of an experiment in this report.

4.1.1.Output Processing and Result Presentation

In this section we will illustrate the procedure to process the simulator output. It is very important to know how outputs of a simulator are handled in order to understand validity of the results. The output of the simulator is in the form of files generated by each monitoring point. The information in the file, which are basically the signals received by the monitors from each of its neighbors, is further processed to calculate the required BGP performance. The format of file with each line corresponding to a single signal is:

NeighbourAs ; MonitorName ; Prefix ; TimeSeen ; FullAsPath ; Announcement/Withdrawal

The Prefix is used to identify the originating beacon of a signal. Announcement and Withdrawal signals are processed separately in order to analyze difference in their behaviors. MonitorName is constant in a file and is used to identify the name of monitoring point which generated the output file. The signals received from different neighbors are distinguished by NeighbourAs. The time at which a signal is received by monitoring point is determined through TimeSeen. FullAsPath discloses all the ASes involved in propagating a route and consequently they *make* a route. Following the definitions of SD and CD described in Chapter 3, the output files are processed for each monitor. A signal is uniquely identified by a combination of Prefix and NeighbourAs, i.e. all the lines of an output file that have similar NeighbourAs and Prefix will be considered as a unit signal. For each such signal, it is figured out that what is the time when it was received first as well as the time it was received at last. Therefore information about each signal's first and last reception from a neighbor AS is extracted. The data is arranged in the following format for additional processing:

NeighbourAs ; Prefix ; FirstSeen ; LastSeen ; CD ; SD ; Announcement/Withdrawal

Where CD corresponds to convergence delay and SD means signal duration.

As a last step, the data is organized in such a way that histograms for announcements and withdrawals can be plotted. Similarly, cumulative distribution function (CDF) curve showing the trend of CD and SD are plotted for both cases. Figure 4.1 (a) shows the announcement histogram of an experiment and (b) is the plot of corresponding CDF curve.

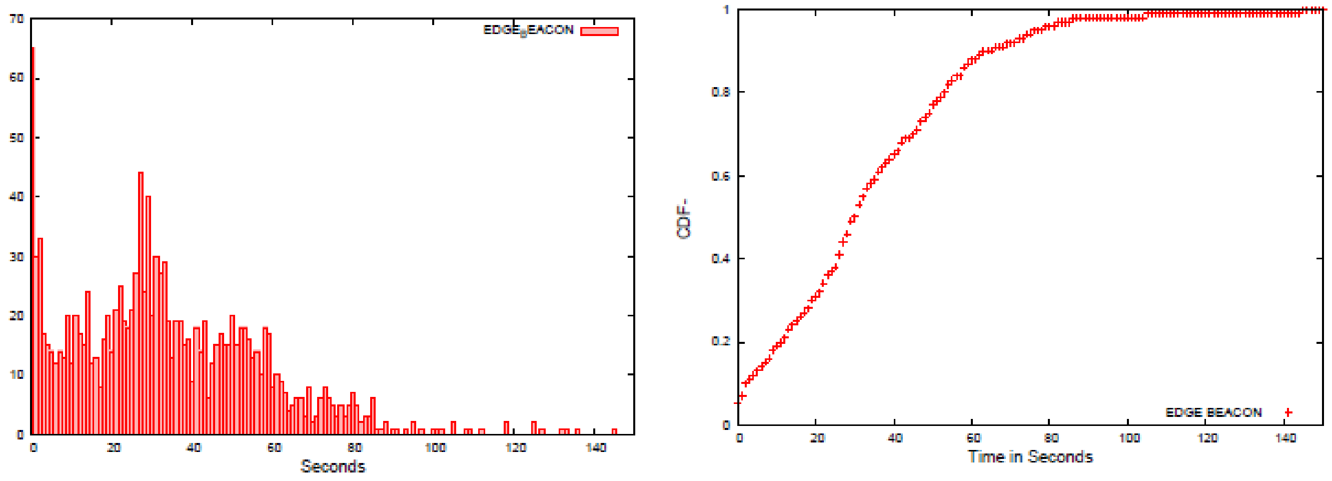


Figure 4.1: *(a) announcement Histogram, (b) CDF curve*

Majority of observed SD and CD values fall into a 0 to 180 seconds interval. In our topology experiments of 10K As nodes, most values are in 0 to 150 seconds interval. In [23] and particularly in [19], the authors have used CDF curves in the range of 0 to 180 seconds. We will use the interval of 0 to 150 seconds for our plots. For the rest of results we will only show the CDF curve, which gives a clearer picture of CD and SD.
In short the one experiment process can be illustrated by Figure 4.2:

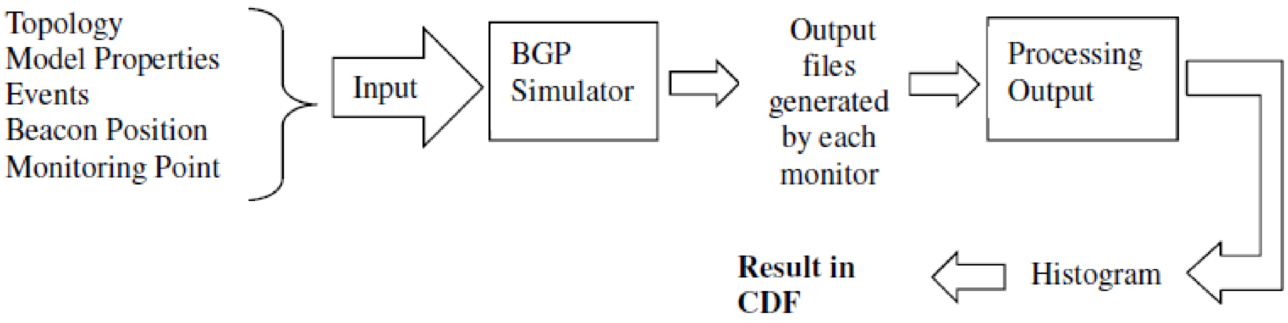


Figure 4.2: An Experiment Instance

4.1.2.Effect of Position of Beacons and Monitoring Points

We have observed that positioning of beacon and monitoring points is a delicate decision with concrete impact on the results of experiments. It determines the perspective by which we look to a particular topology.

For further illustration we will show results of an experiment in which we follow equivalent criteria for beacon selection and monitoring point placement but yet when two different sets of beacons were selected, completely different results were obtained. Figure 4.3 (a) and (b) show CD in two scenarios on TREE topology. The sensitivity level for beacon selection can be observed. The only difference between the two experiment instances is that the beacons are different, although the selection criteria are alike for both MIDDLE beacon and EDGE beacon. From Figure 4.1 (a), one can infer that edge is taking higher time than MIDDLE beacon, but at the same time 4.1(b) shows that there is not a considerable difference in CD for both beacon input signals.

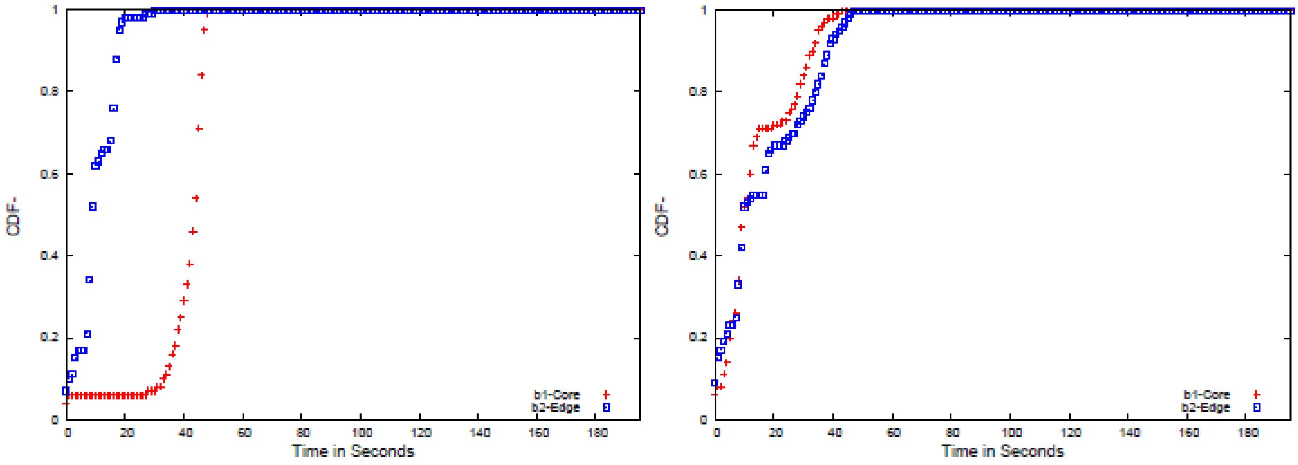


Figure 4.3: *CDF curve for TREE topology with different beacon selection (a) CD, (b) SD*

Despite its complexity, we tried to include the results that show the majority behavior in a certain configuration of an experiment. Although we were limited by number of attempts, yet we were successful to find the rational of different patterns and included those results in our descriptions.

4.1.3.Runtime Environment

The simulator is designed to run on a homogeneous cluster. The experiments were performed on the DAS-3 cluster at Vrije University which is consisted of 85 dual-processor dual –core nodes, i.e 4 CPU cores per node. We used 32 or 64 nodes for our experiments. In general the ASes are randomly distributed across all nodes, and maintain TCP channels for communication with each other.

The BGP simulator is capable to run in a time-scaled environment, i.e it can run about 200 times faster than the real-time. To avoid congestion on compute nodes, we used time scaling of 60 to100. For further details of various features of the simulator we refer to [19].

4.1.4.A Single Experiment Instance

The BGP simulator is a relatively complex simulator not only with respect to resource consumption, but also it requires significant preprocessing for preparing its input files. An experiment instance means one execution of the simulator in which a single experiment is performed. We run the simulator for almost 40 hours of real-time in an instance. It was feasible since the simulator runs in scaled time. It takes the following input files:

a. Network Topology

It takes the network topology along with full relationship data. The information of positions of monitoring points and their neighbors is also included in the topology information.

b. Input Events

It is the list of update events that will serve as the input in the form of BGP update messages. It also defines which beacon will send out a prefix at which time. Following the approach of [23] and [19], the time between two input events is fixed to two hours. For each event, its type is also specified i.e. announcement or withdrawal.

c. Simulator Properties

There are two files that are used to specify some simulation and model parameters. Simulator properties are adjustable according to the experimental or environmental requirements. For instance timeScaler defines how much faster the simulator should run with respect to the real-time. Similarly total number of nodes on which simulator will run. Some of the parameters have default values as well. In general these properties do not have an impact on simulator results however they may result in increased or decreased performance of simulator. Also increase in load on the executing nodes is also possible by inappropriate setting of such parameters.

d. Model Properties

These values determine the behavior of BGP simulation. They may have dramatic impact on simulation results. For instance, how many percent of ASes should have MRAI timers or route flap damping technique turned on, parameter effecting calculation of delay caused by iBGP?

4.1.Observing Current Internet Topology

In this section we will illustrate an experiment carried out with the simulator 2010 current CAIDA network topology. Internet is assuring its several properties from the last decade, due to which despite extravagant growth in size, the convergence behavior has not changed dramatically [40]. We found out in our experiments that CD has slightly decreased which is consistent with observations by [19]. The reason behind this is that, as the size of Internet increases with addition of new ASes, the density of interconnections between them also increased. This density reduced the CD since it reduces average path length between two nodes [56].

In case of BGP announcements, the results are completely consistent with [19] and [23]. The CDF curve at Figure 4.4 (a) shows 90 percent of the announcements have CD of less than 80 seconds. There is slight step in the curve around 30 seconds and 60 seconds. This is probably due to setting of MRAI timer values for simulation. Almost 40 percent of the total ASe's have MRAI timers set to 30 seconds due to which we observed the steps of 30 seconds. Furthermore there is no significant difference between MIDDLE beacon and EDGE beacon for CD. In different runs of the same instance these curves of both beacons were sometimes swapped.

Observing the SD at Figure 4.4 (b), we see that near 50 percent of signals have SD of zero, thus one single update/announcement is received by the monitoring points. Steps of 30 and 60 seconds are more obvious since it is calculated with respect to same neighbor.

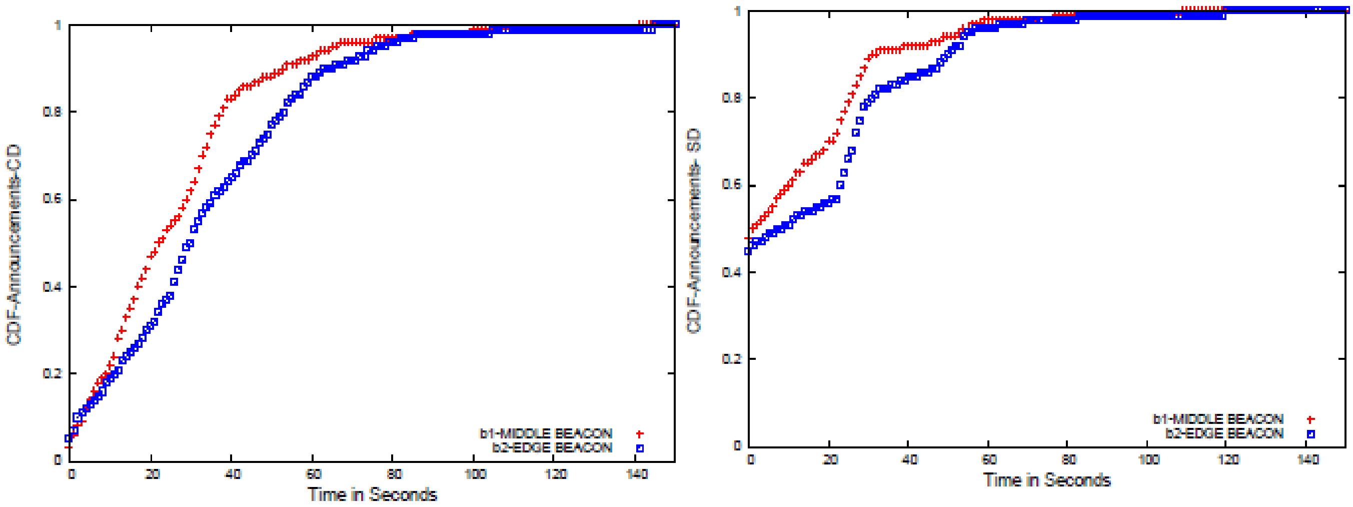


***Figure 4.4:** Current Internet Topology (a) Convergence Delay, (b) Signal Duration*

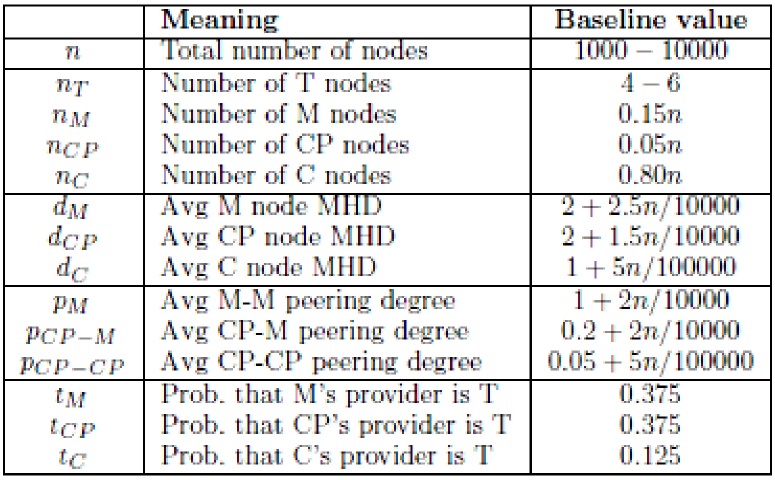
Withdrawals for the current Internet topology on the simulator take a substantially larger amount of time with only less than 20 percent taking 160 seconds. According to [19], it is possible that simulator is maintaining more alternate routes than real world routers in Internet. Due to this reason the path exploration immensely increases resulting in a high CD.

4.2.Topology Scenarios

We will now describe different types of topological scenarios that we have considered for understanding BGP behavior. All these topologies are generated through ILTG described in Chapter 3. To compare results of topologies with each other, we needed a reference topology. It was not possible to use CAIDA network topology for this purpose, since it has about 32K nodes while we are able to generate topologies of only 10K nodes with ILTG. Therefore we generated a simple topology using ILTG, for serving as reference point to other topologies and call it *baseline topology*.

baseline Topology:

For baseline topology, the parameter tuning is done in such a way that it is similar to the growth seen in Internet over the last decade [40]. It is characterized by a slow increase in multi-parenting degree (MPD) of stub nodes, and a faster growth in the MPD of nodes at middle as well as the number of peering links. Five regions are used containing one fifth of all nodes each. The Table 4.1 gives the parameter values for baseline topology.



	Meaning	Baseline value
n	Total number of nodes	$1000 - 10000$
n_T	Number of T nodes	$4 - 6$
n_M	Number of M nodes	$0.15n$
n_{CP}	Number of CP nodes	$0.05n$
n_C	Number of C nodes	$0.80n$
d_M	Avg M node MHD	$2 + 2.5n/10000$
d_{CP}	Avg CP node MHD	$2 + 1.5n/10000$
d_C	Avg C node MHD	$1 + 5n/100000$
p_M	Avg M-M peering degree	$1 + 2n/10000$
p_{CP-M}	Avg CP-M peering degree	$0.2 + 2n/10000$
p_{CP-CP}	Avg CP-CP peering degree	$0.05 + 5n/100000$
t_M	Prob. that M's provider is T	0.375
t_{CP}	Prob. that CP's provider is T	0.375
t_C	Prob. that C's provider is T	0.125

***Table 4.1:** ILTG- baseline topology parameter values*

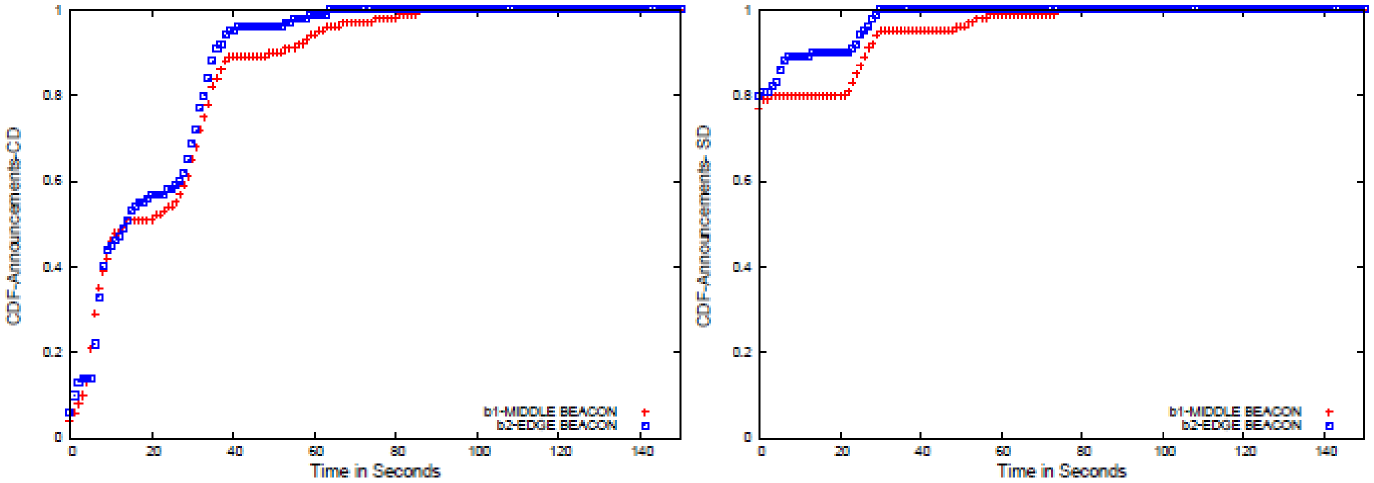


***Figure 4.5:** CDF for ILTG- baseline Topology- Announcements (a) CD, (b) SD*

The performance of BGP for the baseline topology is show in Figure 4.5. Announcements and withdrawals have completely different patterns from each other. In both cases we observe longer duration for CD and SD than in the real Internet graph. One of the reasons is probably because of reduced topology size. It also certifies that generated topologies are very close to the real topology, but can never be considered as the actual topology.

Figure 4.5 (a) shows plot of the CD, with 30 second interval steps clearly. From both Figure 4.5 (a) and (b), it seems apparently that MIDDLE beacon takes more time to converge than the EDGE beacon, but in fact the curves swap positions when the same experiment instance is run multiple times. In (b) the CDF curve appears with result that 80 percent of announcements have SD of 0 second. All SDs are below 60 seconds which is regarded as good performance with respect to BGP.

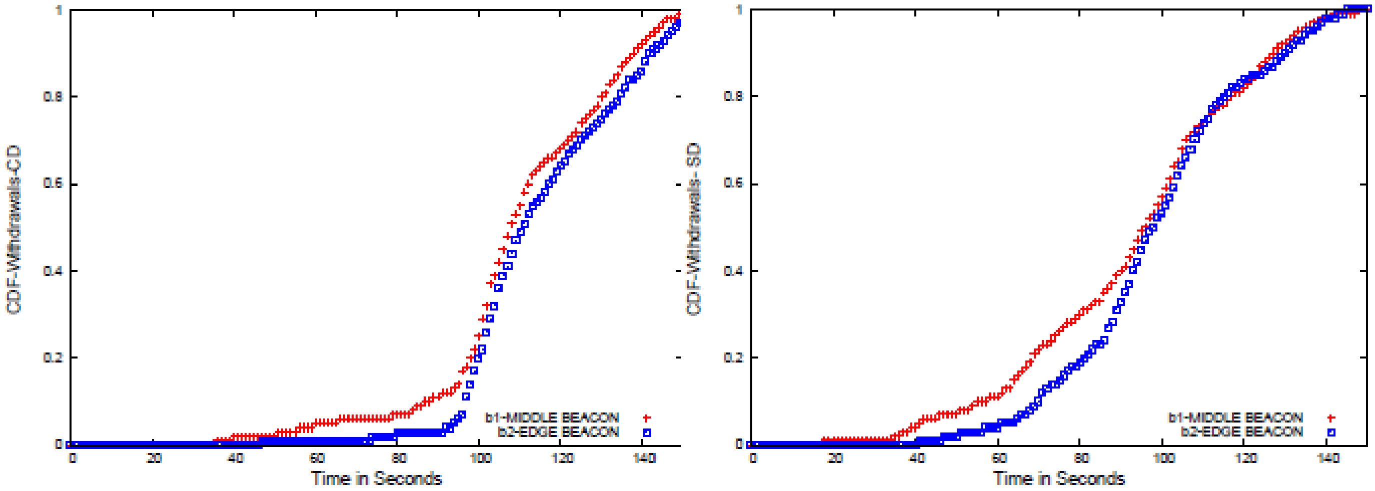


***Figure 4.6:** CDF for ILTG- baseline Topology- Withdrawals (a) CD, (b) SD*

Figure 4.6 illustrates the plot for withdrawals for the baseline topology. Relatively fast convergence is observed with respect to real Internet graph, but yet the time the time it takes to converge is much higher than announcements. Most of the withdrawals have CD less than 120 seconds. (a) shows a slight bending near 120 seconds, which is again a multiple of 30 seconds. Similar steps can be seen in (b) as well. SD plot at (b) and CD at (a) also reveals that most of the withdrawals have values of 60 to 130 seconds respectively.

The shorter convergence for withdrawals is mostly because of path exploration issues faced by routing algorithm. The position of beacon in the network doesn't seem to affect the convergence of withdrawals since both beacons have identical CDF curves.

Now that we have described the reference topology, we will continue to explain the rest of topological scenarios and later present performance of BGP over them. We divide our scenarios in two groups:

1.Variation in Peering links:

In Internet as the year pass, more and more ASes with approximately similar size undergo peering relationship with each other [40]. Although it is difficult to capture all peering links between ASes, yet it can be said that currently at least 20 to 25 percent of the total links are peering links, and the proportion is continually increasing [40]. We were particularly interested in examining if the peering links are a threat to Internet or what kind of implications this trend can influence on BGP. It is difficult to say which kind of scenario actually matches the growth of peering relationships in Internet, but we will examine different possibilities for understanding the impact of peering.

NO PEERING: In this scenario the hierarchy may have multi-parenting, different depth levels but no peering relationship exists. Only to ensure network connectivity, the tier-1 nodes form a peering clique. Although it is a superficial scenario, yet we want to better understand different models of peering relationships, and this is one of it. This scenario may also serve as reference point for the other peering scenarios.

STRONG CORE PEERING: In this case the peering relationships among the M nodes have been increased. This results in a dense core. The peering relationships are doubled to the relationships in the baseline scenario. The rest of the structure at the edge of network remains similar to baseline topology. It is possible that the real Internet, the network might evolves in this direction.

STRONG EDGE PEERING: This model has increased peering relationships towards the edge of network, i.e for stub nodes. We multiply the peering between stub nodes and M nodes as well as between stub nodes and stub nodes by three. It might be the case that Internet is also facing the similar growth. Nevertheless we examine it to enhance our understanding about peering relationships.

2.Variations in Hierarchal Structure:

Internet has maintained a hierarchy, far from a random graph, right from the beginning of its existence [20]. The structure of Internet maintains a power law with respect to degree distribution [44]. We wanted to examine what influence this hierarchal structure has on BGP performance. For this we modeled the topology with some superficial variations which may seem far-fetched, but they help us to understand the impact of increasing or decreasing certain structural properties like multi-parenting and level of hierarchy. The models are explained below:

TRANSIT CLIQUE: In this variation all the nodes, other than stub nodes are made part of a tier-1 clique. It is an interesting case since it makes the top as clique of “equals” connected by peering links. Basically it is a collapse of the provider-customer hierarchy. During topology generation, fifteen percent of the total nodes were made tier-1 nodes and all of the rest as direct stub nodes of them. In such case we have a degree distribution with two extremes, either very high in case of a T node or very low in case of stub node. One important thing to note is that in this topology, because of this two extremes of node degree, we were unable to position MIDDLE beacons according to our standard criteria of beacon positioning, i.e. a random 25 degree node. The nearest to 25 that we found was 5 degree nodes, and we selected one of them as our MIDDLE beacon.

TREE: There is a recent trend of multi-parenting [40], possibly for the sake of guaranteed connectivity. Although it increases the load on BGP routers by increasing the size of routing tables [21], we were interested on its impact on BGP performance. We used a Tree Topology, with no multi-parenting at all, to evaluate BGP performance over it.

In the following sections, we will illustrate performance of BGP on the above mentioned topological scenarios. For BGP, since the nature of announcements and withdrawals is different with respect to propagation pattern, we will discuss them separately for each topology.

4.4. Effect on BGP Announcements

First, the impact on BGP announcements will be described because of simplicity of their propagation pattern. An AS can receive various messages for the same announcement from multiple neighbors. On receiving an announcement message, the AS decides to propagate it depending on its local policies.

4.4.1. Impact of Peering Links

We will first demonstrate the impact that peering links have on BGP announcements. Peers do not propagate the route of one neighbor to all of its other neighbors. However they advertise only their own routes to the AS with which they have peering relationship. If this difference is not maintained, peering will

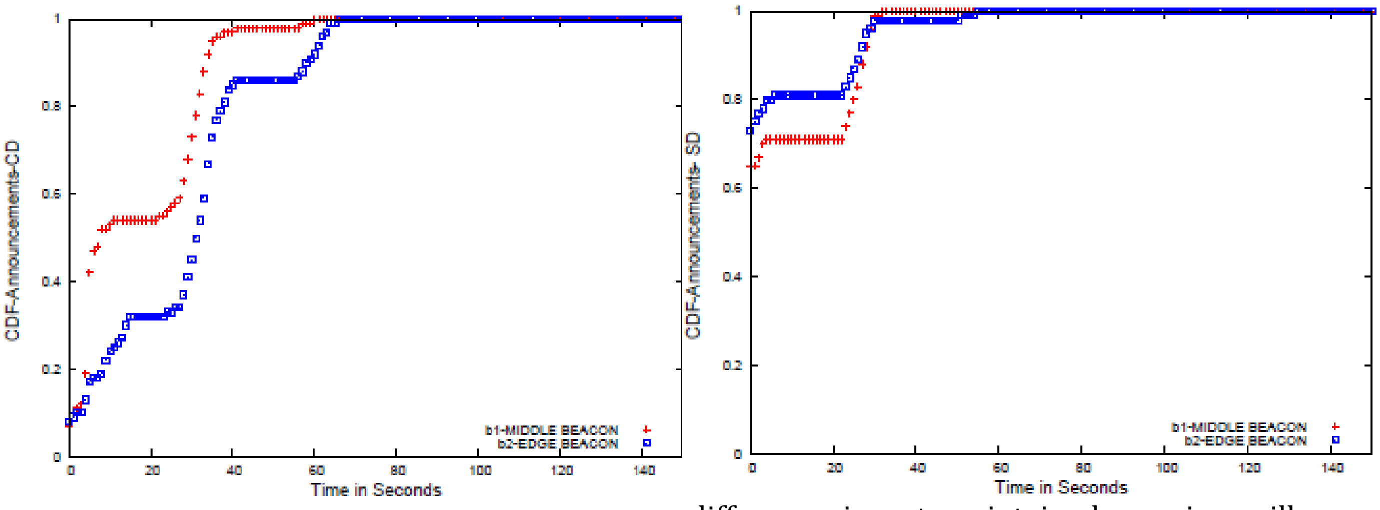


Figure 4.7: *CDF for ILTG- NOPEERING- Announcements (a) CD, (b) SD*

become similar to transit provisioning. ASes advertise only addresses of their peers to their customers, and not the routes announced by that peer. So if a route is advertised via a peering link, it will only go one hop and then downward towards customers. Since our monitoring points are at the core, we expect that the addition of peering links shouldn't have an enormous impact on BGP convergence.

NOPEERING: In the absence of peering, the only way of signal propagation is through transit-customer relationships. Figure 4.7 shows that for announcements, the plots are almost identical with the baseline topology. This is mainly due to the fact that in case of peering links, the route is not propagated upward to the core (i.e. Tier 1) beyond one hop, and since our monitoring points lie at the core, they remain mostly unaffected by the addition of peering links.

STRONG CORE PEERING: From BGP's perspective, a denser core will result in that announcement messages will propagate only one hope above, but at more locations of the network. The number of locations will increase because of increase of density of the CORE due to additional peering in this special scenario. The SD and CD plots at Figure 4.8 shows that they are identical baseline topology. Since the monitoring points are at the core, peering links doesn't seem to have any noticeable impact on convergence

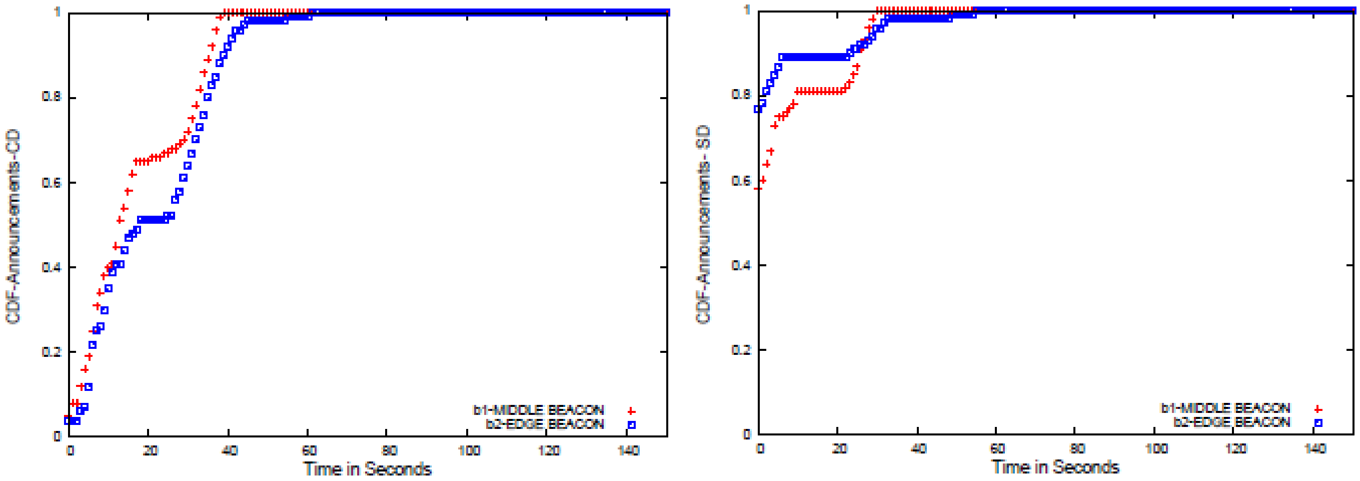


Figure 4.8: *CDF for ILTG- STRONG CORE PEERING- Announcements (a) CD, (b) SD*

STRONG EDGE PEERING: The announcements plots for the strong edge peering, similar to results of the strong core peering, are completely identical to baseline topology. In Figure 4.9 similar step of 30 seconds is visible. Furthermore overall percentage of announcements less than a specific value is also approximately alike.

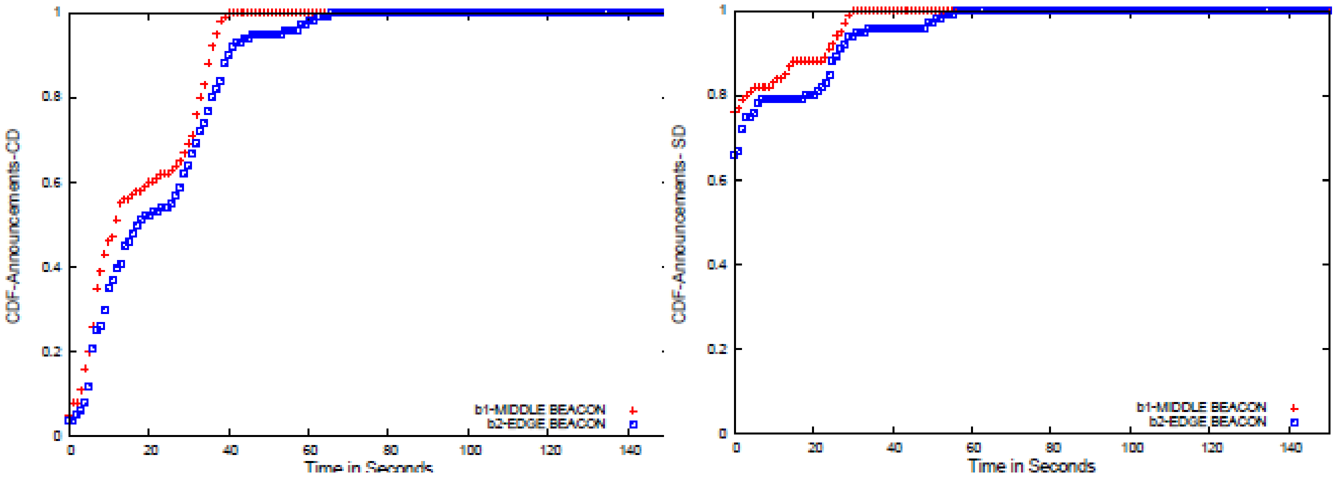


***Figure 4.9:** CDF for ILTG- STRONG EDGE PEERING- Announcements (a) CD, (b) SD*

4.4.2. Impact of Hierarchal Structure

TRANSIT CLIQUE:
We were expecting that hierarchy is an important property of Internet network structure, which might be helpful in BGP convergence. It is interesting to note that the collapse of hierarchy has tremendously increased the BGP performance. The Figure 4.10 shows that in the absence of middle nodes, the announcement signal directly goes into the core, and reaches instantly to the monitoring point at core due to high peering degree at the core consisting or Tier-1 nodes. Figure 4.10 (b) shows almost 100 percent of nodes with SD of zero seconds. Althoug announcements h the collapse of hierarchy results in high number of multiple paths, due to their propagation pattern remain unaffected.

The Figure 4.10 (a) and (b) also shows that because of shorter paths, only a few MRAI timers are triggered with only one slight turn round 30 seconds.

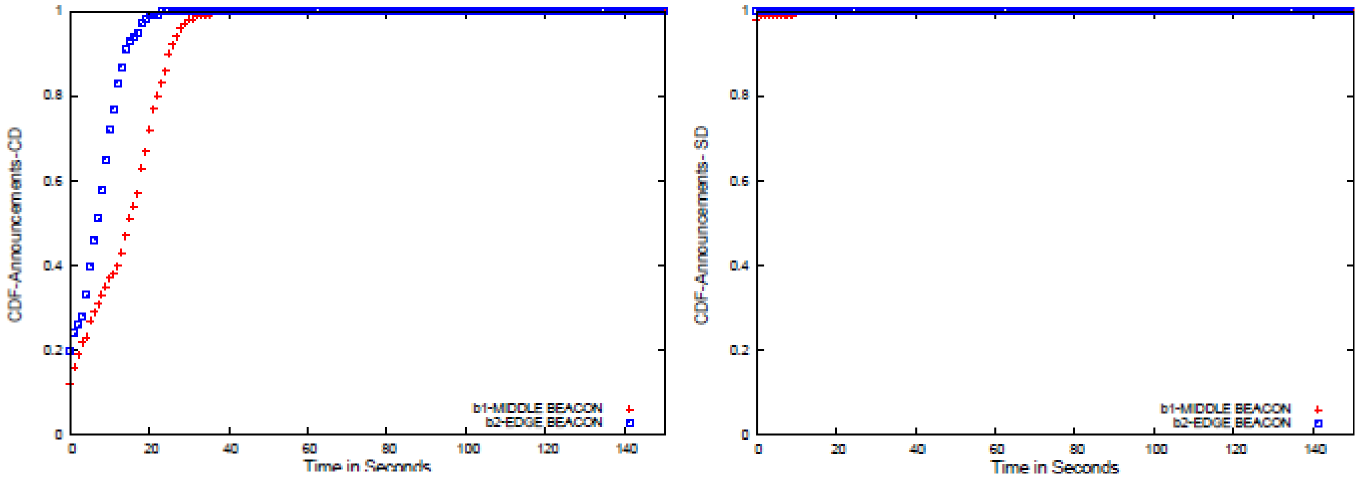


***Figure 4.10:** CDF for ILTG- TRANSIT CLIQUE- Announcements (a) CD, (b) SD*

TREE: The Figure 4.11 shows that multi-parenting degree also has significant influence on BGP convergence. In Tree topology, with each AS having exactly one provider, the announcements go straight up towards the core. The existence to peering links let the signal propagate one hop farther towards the core. The Figure 4.11 (b) clearly shows that signal duration is exactly zero for 100 percent of signals. This is because the monitoring point receives a signal exactly once by each of its neighbors in the absence of multiple transit provider across the network.

There is a step as well in Figure 4.11 (a), which is a special case for these beacons. It might be because of position of beacon. Similar steps are not evident in other positions of beacon in the same topology. We do not consider it as an effect of MRAI timer as the TREE topology should trigger less MRAI timers than baseline topology, since it reduces the number of paths available. The reduction in multi-parenting makes the network less robust but more efficient in terms of BGP performance.

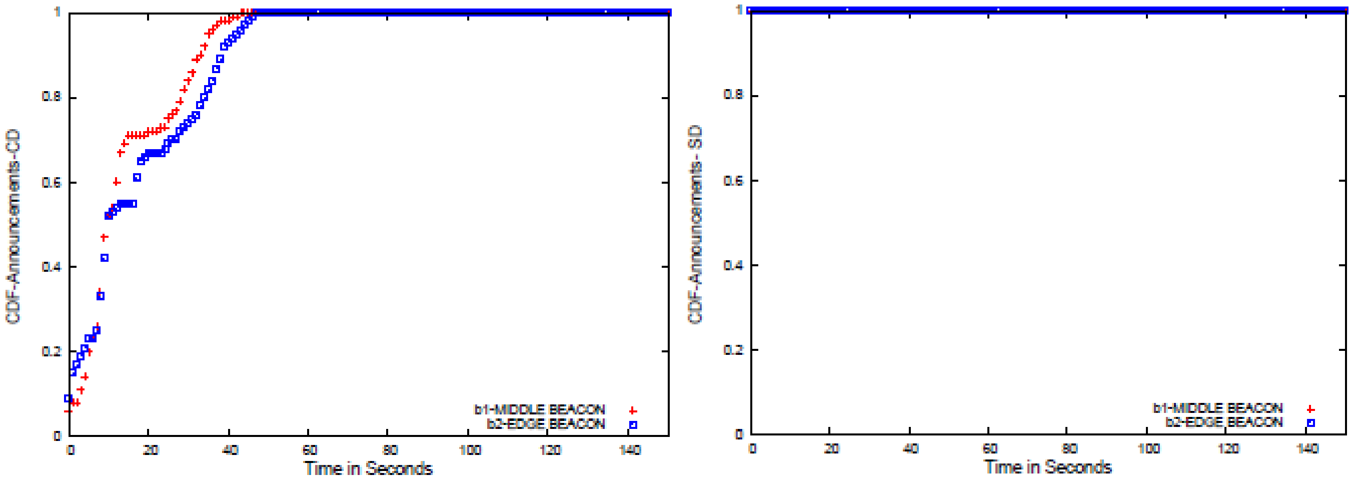


Figure 4.11: *CDF for ILTG- TREE- Announcements (a) CD, (b) SD*

4.5. Effect on BGP Withdrawals

As second step, the impact of various topologies on BGP withdrawals will be illustrated. As described earlier, withdrawals have a complex propagation pattern due to path explorations. In an extreme case they might overload the BGP routers, resulting in partial disconnectivity as well [3, 4].

4.5.1.Impact of Peering Links

We will illustrate impact of peering links on withdrawal convergence. In case of withdrawals, when an AS notifies a withdrawal to its peer, the peer will look for alternate routes.

NOPEERING: The withdrawals in this scenario have a little bit different pattern than in the baseline topology. The main difference as seen in Figure 4.12 is that the two beacons have converged differently. It is discussed earlier that positions of beacons and monitoring points have great impact on convergence. We suspect that one of the beacons, mainly the MIDDLE beacon is selected very near to the core. During path explorations, such paths may be selected in which the beacon may receive the signals that encounter the core, and then get back to it, ultimately we find an increase in CD and SD. This increase is due to the fact that first signals might be received by the peer link, and the last signals received from the core residing at upper level in hierarchy. We also notice that almost none of the signals have zero SD, which also strengthens our hypothesis.

Nevertheless, if our suspicion is correct, the results are not affected by removal of peering relationships from the network. It is the matter of position of MIDDLE beacons that influenced the results. If we are wrong in our suspicion, then further granular examination is required to investigate if some outliers are driving the results.

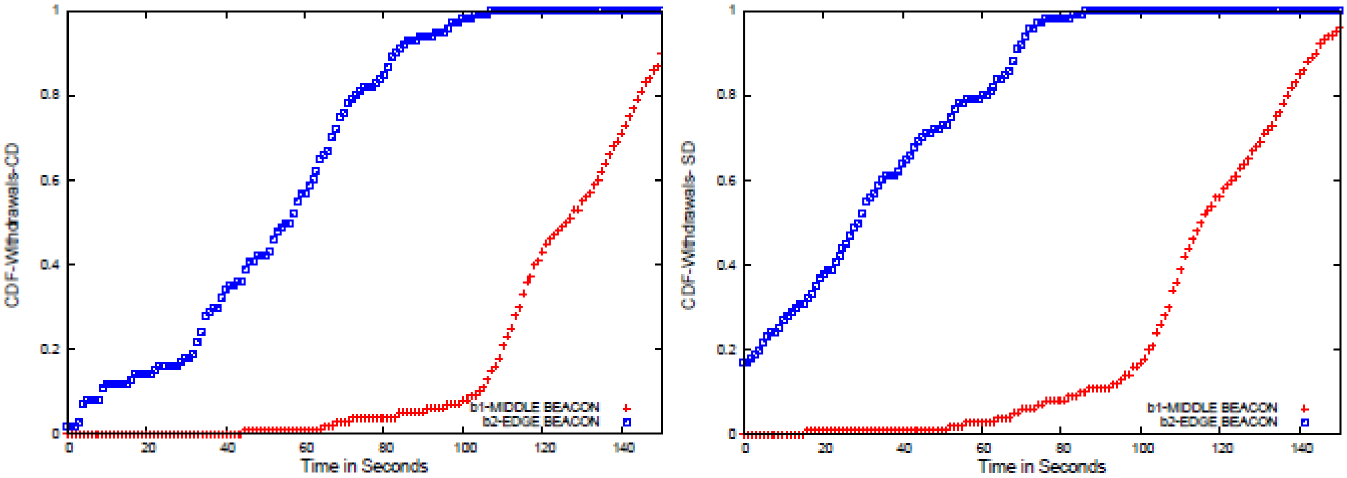


Figure 4.12: CDF for ILTG- NOPEERING- Withdrawals (a) CD, (b) SD

STRONG CORE PEERING: The Figure 4.13 shows that like in case of announcements, the withdrawals also remain unaffected by dense core. Apparently the beacons are not very near to core, so that they might have different behaviors. The graphs are completely matching with withdrawals of baseline topology.

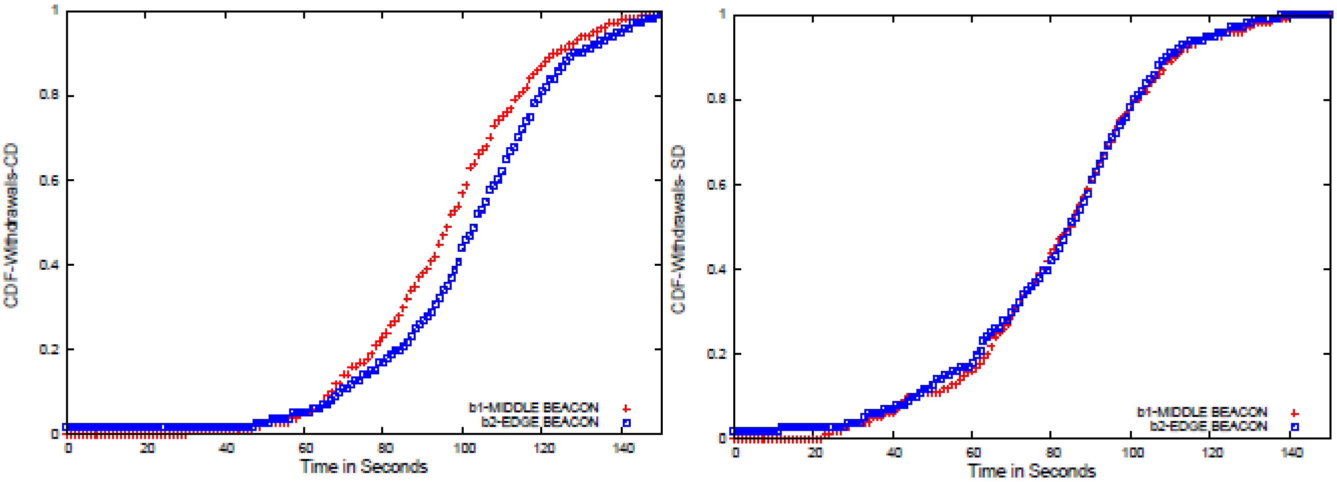


Figure 4.13: CDF for ILTG- STRONG CORE PEERING- Withdrawals (a) CD, (b) SD

STRONG EDGE PEERING: Figure 4. 14 shows that if peering links are increased at the edge of network, the convergence of withdrawals remains almost unaffected. The plots are completely consistent with the baseline topology withdrawals. .

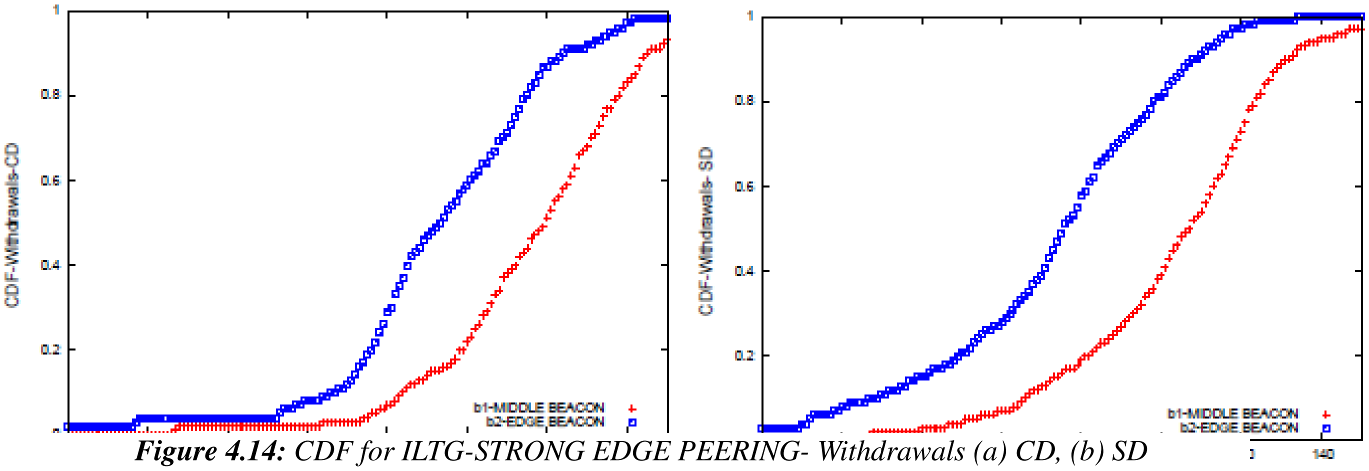


Figure 4.14: CDF for ILTG-STRONG EDGE PEERING- Withdrawals (a) CD, (b) SD

4.5.2.Impact of Hierarchal Structure

TRANSIT CLIQUE: The CD and SD of withdrawals have also increased when there are only two levels of hierarchy. Figure 4.15 shows no steps at 30 seconds interval. Figure 4.15 (b) shows that because of existence of alternate paths, withdrawals faced path explorations. Yet SD is several times less than the baseline scenario. Withdrawals usually do not trigger MRAI timers. This shows that increased number of transit providers will not decline the BGP convergence rather the steps or levels of transit provider and customer links cause delayed convergence.

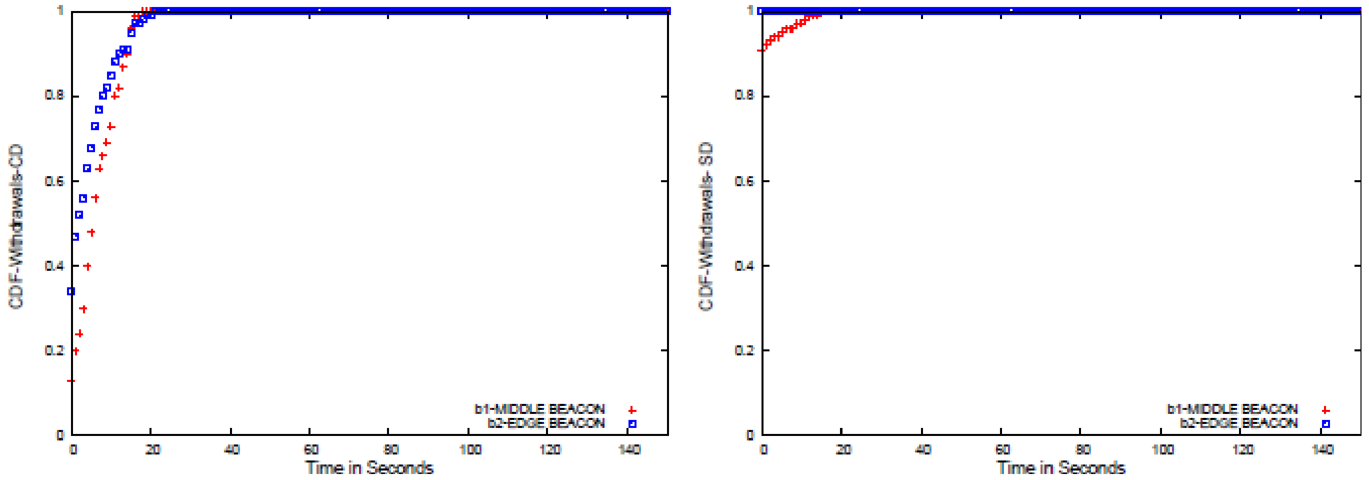


Figure 4.15: CDF for ILTG-TRANSIT CLIQUE- Withdrawals (a) CD, (b) SD

TREE: We observe a decrease in CD for withdrawals in TREE topology, as compared to the baseline topology, as shown in Figure 4.16. The absence of multi-parenting reduces the number of available alternative to only paths available through peering links. This automatically limits the number of available alternate paths which in turn abstain from excessive path explorations. The very small step in Figure 4.16 (a) is as mentioned in announcements, a special case of beacon position. It is difficult to make a concrete statement about this specific position of beacons.

In contrary to TRANSIT CLIQUE scenario which is completely unrealistic, the TREE topology is relatively near to real world. Multi- homing can be reduced to a certain extent through appropriate policy making.

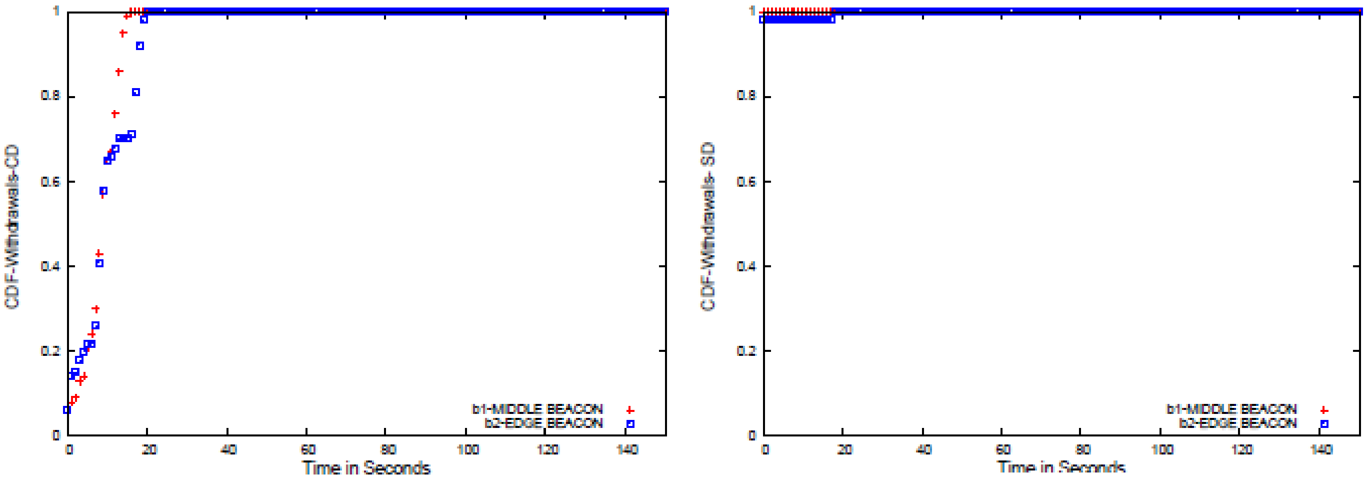


Figure 4.16: *CDF for ILTG-TREE- Withdrawals (a) CD, (b) SD*

In this chapter we illustrated the experiments that we carried out in this project using the BGP simulator and ILTG topology generator. We briefly stated our explanations, reasoning of different findings from the experiment outputs. In next chapter we will state conclusions derived from this research effort.

Chapter 5

Conclusion

We were interested in this research to understand what implications topological variations can have on BGP performance. It is a very small contribution in understanding the behavior of BGP, which can be considered as first step in the direction where we are able to suggest that what should be the trend of growth of future Internet topology. It is also possible that such results and conclusions may lead to a policy document suggesting the AS level operational community to formulate the relationships among ASes for the better of all.

The work can be considered contemporary to [34] in which the authors investigated the impact of different kinds of topologies on number of BGP messages. We investigated the impact of different topologies, with relatively small set of topologies, on BGP Convergence Time. Quite surprisingly the results are somewhat similar. This implies that the number of BGP update messages and BGP Convergence is almost directly proportional to each other. We consider it a very significant finding of our work.

5.1.Impact of Work

Overall for BGP we figured out by our experiments that the amount of peering relationships does not influence the BGP performance considerably. Only in one case for NO PEERING withdrawals, we faced a deviated performance. We explained that the possible reason of this deviation is not absence of peering links. If ASes increase peering relationships with each other for saving the transit costs, we suggest that BGP performance in Internet will remain unaffected.

Contrary to the conclusions in case of peering links, the hierarchal structure of internet has a significant influence on BGP performance. The performance decline because of increasing number of multi-homed links can be bad news for BGP routers. Multi-homing not only ensures connectivity, but as a side affect, results in high growth of routing table size [21] and increase in number of update messages [34]. In [7], it was stated that multi-homed sites face experience of degraded performance. Our results show that they serve as a case of decreasing overall BGP convergence. Furthermore our experiments show that more specific prefixes seem to trigger more MRAI timers across the internet, causing delay in convergence. Similarly, withdrawals suffer from extra path explorations and therefore the signal is propagated across the network with significant delay. The internet operators can make policies for multi-parenting since the current trend in its increase [40] can be a threat for BGP performance.

If the depth of hierarchy is decreased, then according to our experiments, BGP convergence will increase. The special case of TRANSIT CLIQUE provided us a deep insight into impact of levels of hierarchy. Surprisingly we found that if, hypothetically, the regional providers are thrown out of the market, BGP will perform better. However, in such case there could be more problems like enormous increase in routing table size which makes it impractical. It gives another important conclusion that if number of transit nodes is increased, convergence is not delayed but the hierarchical structure in which tiers are organized causes decreased performance of BGP. An Internet with multiple service providers, providing each other transit services, has more delayed convergence than the Internet with relatively flat structure. The fact that average path length remained the same during last decade [40] shows that despite enormous growth of Internet, its depth size remained almost constant. This lead BGP performance to remain consistent as well.

We also notice the role of MRAI timers in CD. Whenever there is delayed convergence, with only 20 to 40 percent of nodes having CD below 40 seconds and SD of less than 30 seconds, we always see the steps at round 30 seconds interval. The steps are not steep since time is not fully synchronized across the ASes. Similar steps at 30 seconds interval are not evident in case of faster convergence with almost 90 percent of nodes having CD less than 40 seconds and SD less than 2 seconds.

5.2.Future Suggestions

In future it is possible to investigate the sensitivity level of positions of BGP beacons and the monitoring points. It might be possible to find out the appropriate positioning criteria, by which we are more confident in our BGP behavior analysis.

There is also a possibility, as with every experiment, that any of our experiment results are driven by some outliers, i.e. a particular input, phenomena, or model parameter is affecting a result in such a way that actual impact of topology remains hidden. This can be analyzed by careful analysis of the output at finer granularity, and by co-relating different model input parameters with each other as well as with outputs.

More topological scenarios can be generated and it will be interesting to find out their impact on BGP. For instance, what if the customers try to prefer to buy transit from Tier 1 nodes, resulting in relatively flat hierarchical structure. What if we make the number of middle nodes (M) constant and continue to increase the customers (C). These are only a few of analyzable scenarios.

The time limitations of ILTG did not allow us to generate networks beyond 10k size, even though the BGP simulator is capable of simulating network sizes several times higher than 10k. It might be possible to optimize ILTG or to use some other topology generator for getting topologies of e.g. 60k. Although the tool by Dimitropoulos [25] was capable of generating topologies of 100k with relationships, no knobs for parameter tuning were available. These knobs were very significant for generating different *kinds* of topologies.

Another valid research area could be to analyze whether operational parameters or graph-oriented parameters are more beneficial and important to study. It will be very interesting to find which kind of analysis would benefit the BGP community more?

We are aware that there are various solutions being proposed for the future routing of Internet. Some of them suggest making BGP disappear. For instance, Clark et. al. [33] suggests that current Internet just evolved by many external factors with several discrepancies and there is need to revisit Internet design and architecture. O. Bonaventure [38] also suggests reconsidering inter-domain routing architecture. Similarly suggestions Massey et. al. [37] may require substantial modifications in BGP in order to give customer and transit networks different address spaces for increasing scalability. Despite these proposals, to realize the dream of replacement of BGP are far ahead and may require at least a decade. The current studies and future similar behavioral analysis of BGP makes us understand strengths and weaknesses of the routing system. We suggest that similar studies of BGP behavior will not only help us in making valid improvements for BGP, but will also help in better designs of new routing architectures.

References:

[1] Y. Rekhter, T. Li, S. Hares, ***A Border Gateway Protocol*, RFC 4271** (BGP-4), January 2006.

[2] S. Halabi and D. McPherson, ***Internet Routing Architectures***, Cisco Press, 2000

[3] C. Labovitz, R. Malan, and F. Jahanian, ***Internet Routing Stability,*** in Proceedings of ACM SIGCOMM 1997

[4] C. Labovitz, R. Malan, and F. Jahanian, ***Origins of Internet Routing Instability***, in Proceedings of INFOCOM 1999.

[5] C. Labovitz, A. Ahuja, and F. Jahanian, ***Experimental Study of Internet Stability and Wide-Area Network Failures***, in Proceedings of FTCS 1999.

[6] C. Labovitz, A. Ahuja, A. Bose, and F. Jahanian, ***Delayed Internet Routing Convergence***, in Proceedings of ACM SIGCOMM 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 INFOCOM 2001

[8] N. Feamster, J Winick, J Rexford, ***A model of BGP Routing for Network Engineering***, ACM SIGMETRICS Performance Evaluation Review, 2004

[9] B. Donnet, T. Friedman, ***Internet Topology Discovery***: A Survey in IEEE communications Surveys, Volume 9, 2007

[10] V. Krishnamurthy, M. Faloutsos, M. Chrobak, J. Cui, L. Lao, A. G. Percus, ***Sampling Large Internet Topologies for Simulation purposes***. Computer Networks , June 2007.

[11] D. Meyer, L. Zhang, and K. Fall. ***Report from the IAB workshop on Routing and Addressing***, February 2007.

[12] R. Mahajan, D. wetherall, T. Anderson, ***Understanding BGP Misconfiguration***, In ACM SIGCOMM 2002.

[13] Geoff Huston - ***An Update on IPv6 Deployment (RIPE 56)*** link: http://www.ripe.net/ripe/meetings/ripe-56/presentations/Huston-Measuring_IPv6_Deployment.pdf

[14] T. Griffin, B. J. Premore, ***An experimental analysis of BGP convergence time***. In In proceeding of ICNP 2001.

[15] D. Pei, M. Azuma, D. Massey and L. Zhang, **BGP-RCN*: Improving BGP convergence through Root Cause Notification***. Comput. Network, ISDN, 2005

[16] A. Greenberg, G. Hjalmtysson, D. A. Maltz, A. Myers, J. Resford, G. Xie, H. Yahn, J. Zhan and H. Zhang. ***A clean state 4D approach to network control and management,*** SIGCOMM, 2005.

[17] Cisco and Juniper, www.cisco.com, www.juniper.net

[18] K. Lougheed, Y. Rekhter, ***A Border Gateway Protocol, RFC 1105*** *(BGP-1)*, June 1989.

[19] M. Wojciechowski, ***Border Gateway Protocol Modeling and Simulation,*** a student project of NlNetlabs under supervision of Benno Overiender for Project BGP Simulator, 2008.

[20] C. Huitema, ***Routing In Internet***, Prentice Hall, New Jersey, 1995

[21] G. Huston and G. Armitage. ***Projecting future Ipv4 router requirements from trends in dynamic BGP behavior***. In ATNAC, Australia, December 2006

[22] J.C. Mogul, ***Emergent (Mis)behavior vs. Complex Software Systems,*** Proceedings of the 1st ACM SIGOPS/EuroSys European Conference on Computer Systems 2006

[23] Z. M. Mao, R. Bush, T. G. Griffin, and M. Roughan. ***BGP Beacons***. In Proceedings of 3rd ACMCOMM conference on Internet measurement , 2003.

[24] J. Hawkinson, T. Bates, ***Guidelines for creation, selection, and registration of an Autonomous System (AS),*** RFC 1930, 1996

[25] X. Dimitropoulos, D. Krioukov, A. Vahdat, and G. Riley. ***Graph annotations in modeling complex network topologies,*** ACM Transactions on Modeling and Computer Simulation. 2009

[26] Randy Bush, Tim Griffin, and Zhuoqing Morley Mao. ***Route flap damping: harmful?*** Talk at RIPE 43, September 2002.

[27] Zhuoqing MorleyMao, Ramesh Govindan, George Varghese, and Randy Katz. ***Route Flap Damping Exacerbates Internet Routing Convergence***. In Proceedings of ACM SIGCOMM, 2002

[28] C. Villamizar, R. Chandra, and R. Govindan. **BGP route flap damping**, 1998. RFC 2439.

[29] S. Deshpande , B. Sikdar, ***On the impact of route processing and MRAI timers on BGP convergence times***, In Global Telecommunications Conference, 2004.

[30] P. Jakma, ***Revised Default Values for the BGP 'Minimum Route Advertisement Interval,*** Internet Draft by Inter-domain routing. 2008

[31] Abdelshakour Abuzneid and B. J. Stark, ***Improving BGP Convergence Time via MRAI Timer***, Book Chapter in Novel Algorithms and Techniques in Telecommunications and Networking published by Springer Netherlands, 2010

[32] TG Griffin, G. Wilfong , ***An Analysis of BGP Convergence Properties***, ACM SIGCOMM Computer Communication, 1999

[33] DD Clark, K Sollins, J Wroclawski, T Fabe, ***Addressing Reality: An Architectural Response to Real World Demands on Evolving Internet***, ACM SIGCOMM 2003

[34] A. Elmokashfi, A. Kvalbein, C. Dovrolis, ***On the scalability of BGP: the roles of topology growth and update rate-limiting***, ACM CoNEXT 2008

[35] T.G Griffin, F. B. Shepherd, G. Wilfong, ***Stable Paths Problem and Interdomain Routing***, IEEE/ACM Transactions on Networking, 2002

[36] Z Duan, J Chandrashekar, J Krasky, K Xu, ZL Zhang, ***Damping BGP Route Flaps***, in Proceedings of IEEE IPCCC, 2004 .

[37] D Massey, L Wang, B Zhang, L Zhang, ***A scalable Routing System Design for Future Internet,*** In proceedings of ACM SIGCOMM, 2007

[38] O. Bonaventure. ***Reconsidering Internet Routing Architecture***. *Internet Draft*, March 2007

[39] J. H. Cowie, D. M. Nicol, A. T. Ogielski, ***Modeling the Global Internet***, Computing in Science and Engineering, vol. 1, no. 1, pp. 42-50, Jan./Feb. 1999

[40] A Dhamdhere, C Dovrolis, ***Ten Years of the Evolution of Internet Ecosystem***, Proceedings of the 8th ACM SIGCOMM conference on Internet measurement, 2008

[41] D. M. Nicol, M. Liljenstam, J. Liu, ***Advanced concepts in large-scale network simulations***. In Proceedings of the Winter Simulation Conference, 2005

[42] University of Oregon Route Views Project. **http://www.routeviews.org/**

[43] www.caida.org

[44] M. Faloutsos, P. Faloutsos, and C. Faloutsos***. On Power –law relationships of the Internet Topology***, In ACM SIGCOMM 1999.

[45] R.V. Oliveira, B. Zhang, L. Zhang, ***Observing the Evolution of Internet AS Topology***. SIGCOMM , 2007

[46] K.L Calvert, E.W. ZEgura, ***Modeling Internet Topology***, IEEE, 1997

[47] B. Zhang, ***Collectig Internet AS level Topology***, ACM SIGCOMM, 2005

[48] K. C. Claffy, D. Krioukov, ***The Internet AS-level Topology: Three data sources and one Definitive Metric***, ACM SIGCOMM 2006

[49] R. Govindon, W. Willinger, ***Network Topology Generators; Degree vs Structural***, ACM, 2002

[50] P. Mahadevan, D. Krioukov, K. Fall, A. Vahdat, ***Systematic Topology Analysis and Generation using degree correlations***, ACM SIGCOM 2006

[51] E. Ahronovitz, JC Konig, C Saad, ***A Distributed Method for dynamic Resolution of BGP Oscillations***, IEEE International Parallel & Distributed Processing Symposium, 2006.

[52] T. Wong, J. Van, C. Alaettinoglu, ***Internet Routing Anomaly Detection and Visualization,*** IEEE Dependable Systems and Networks, 2005

[53] L. Colliti, G Di Battista, M Patrignani, M Pizzonia ***Investigating Prefix Propagation through Active BGP Probing***, Microprocessors and Microsystems, 2007.

[54] R. Govindan, A Reddy, ***An Analysis of Internet Inter-Domain Topology and Route Stability***, IEEE INFOCOM 1997.

[55] W. Muhlbaur, O. Maennel, S. Uhilig, A. Feldmann, M. Roughan, ***Building an AS-topology model that captures Route Diversity***, ACM SIGCOMM, 2006

[56] D. Pei, L.Zhang, ***A study of Packet Delivery Performance during Routing Convergence***, IEEE, 2003

[57] P Mahadevan, C Hubble, D Krioukov, B Huffaker, ***Orbis: rescaling degree correlations to generate annotated internet topologies***, ACM SIGCOMM, 2007

[58] L Cheng, NC Hutchinson, MR Ito, ***RealNet: A Topology Generator Based on Real Internet Topology***, International Conference on Advanced Information Networking and Applications, 2008