On the scalability of BGP: the roles of topology growth and update rate-limiting

Ahmed Elmokashfi
Simula

Amund Kvalbein
Simula

Constantine Dovrolis
Georgia Tech

ABSTRACT

The scalability of BGP routing is a major concern for the Internet community. Scalability is an issue in two different aspects: increasing routing table size, and increasing rate of BGP updates. In this paper, we focus on the latter. Our objective is to characterize the churn increase experienced by ASes in different levels of the Internet hierarchy as the network grows. We look at several "what-if" growth scenarios that are either plausible directions in the evolution of the Internet or educational corner cases, and investigate their scalability implications. In addition, we examine the effect of the BGP update rate-limiting timer (MRAI), considering both major variations with which it has been deployed. Our findings explain the dramatically different impact of multihoming and peering on BGP scalability, identify which topological growth scenarios will lead to faster churn increase, and emphasize the importance of not rate-limiting explicit withdrawals (despite what RFC-4271 recently required).

1. INTRODUCTION

Recently, there is a significant concern among both Internet operators and researchers about the scalability of interdomain routing with BGP. A workshop organized by the Internet Architecture Board concluded that *"routing scalability is the most important problem facing the Internet today"* [24]. The concern is that we are soon approaching the point where the global routing system, and the core routers in particular, will no longer be able to keep up with routing dynamics. BGP scalability is an issue in two different aspects: *increasing routing table size, and increasing rate of BGP updates (churn)*. Note that, in general, an increase in the routing table size (number of routable prefixes) also in-
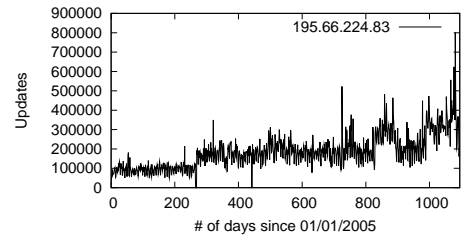

Figure 1: Growth in churn from a monitor in France Telecom's network.

creases churn, since the number of networks that can fail or trigger a route change increases. In this paper, *we focus on the issue of increasing churn.*

A recent measurement study [16] showed that BGP churn increases at a much faster pace than the routing table size. During 2005, *the daily rate of update messages almost doubled, while the size of the routing table grew by only 18%.* Projections of this growth trend indicate that the processing load on core routers demands expensive router upgrades. This problem is exacerbated by the burstiness of BGP update traffic: routers should be able to process peak update rates that are up to 1000 times higher than the daily averages [15]. To illustrate the growth and variability in churn, we plot the rate of BGP updates received from a RIPE routing monitor located in France Telecom's backbone network [29]. Figure 1 shows how the number of updates sent each day increased through 2005-2007. Due to the high variability, we used the Mann-Kendall test to estimate the trend in churn growth. Using this estimation technique, the number of updates received from this monitor grew approximately by a total of 200% over these three years.

The goal of this study is to improve our understanding of the underlying reasons for the experienced growth in churn. Churn is a result of a complex interplay of 1) the *routing protocol*, including policy annotations and various BGP mechanisms like update rate limiting, route flap dampening etc. 2) *events* like prefix announcements, link failures, session resets, traffic engineering operations that generate routing updates, and

3) the characteristics of the Internet *topology*. The last factor, in particular, is the primary focus of this paper. We aim to understand how topological characteristics of the AS-level graph influence the scalability of BGP churn.

Describing the AS-level Internet topology and how it evolves has been the subject of much research (and heated debate) in the last decade. We refer the reader to the following representative references [7, 21, 23, 31, 34]. In this paper, we do not use an existing topology generation model because we want to explore a wide range of "what-if" possibilities that none of the existing models captures in a parsimonious and intuitive manner. For the same reasons, we do not base our investigations on inferred historical internet topologies. Instead, we first identify four basic but fundamental characteristics of the Internet graph that have persisted over the last decade. Then, we design a simple and controllable topology generator that satisfies the previous properties, and at the same time allows us to easily navigate the topological space. The "knobs" of this generator are parameters with operational relevance in practice, such as the multihoming degree (MHD) of stubs versus transit providers, instead of abstract measures such as betweenness or assortativity.

Using our topology generator, we establish the factors that determine churn at different locations in the Internet hierarchy, and investigate the importance of each factor in a growth model that resembles the evolution of the Internet over the last decade. We then examine several deviations from this growth model, and investigate how the number of routing updates generated by a specific routing event grows with the size of the topology in each case. We ask questions such as: "What if the MHD of stub ASes increases with the network size instead of staying constant?" "What if the Internet becomes denser mostly due to peering links?" "What if tier-1 providers dominate the transit market, reducing the number of tier-2 providers?" Beyond the topological characteristics, however, we cannot ignore the importance of the BGP update rate-limiting timer MRAI (Minimum Route Advertisement Interval). A recent change in the BGP specification (RFC-4271) required that explicit withdrawals are subject to MRAI rate-limiting, just as any other update. We consider the dramatic impact of this specification change on the scalability of BGP churn.

The rest of the paper is organized as follows. In the next section, we explain our overall approach and describe the model that we base our investigation on. In Sec. 3, we describe our topology generator and present our Baseline growth model. In Sec. 4, we present a model for the churn experienced at different locations in the Internet, and discuss the importance of different factors as the network grows. In Sec. 5, we examine
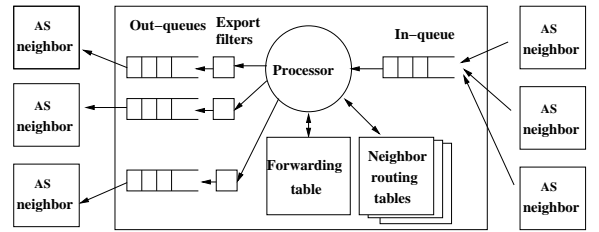

Figure 2: Model for a node representing an AS.

several topology growth scenarios and investigate how they affect BGP churn. In Sec. 6, we show how different implementations of the MRAI timer have a major impact on the resulting churn. We review related work in Sec. 7, and draw our conclusions in Sec. 8.

2. APPROACH AND MODEL

We can only study the problems described above using simulations. Since our goal is to look at scalability under different hypothetical topology growth models, our investigation cannot be performed by doing measurements in the current Internet. Also, the complexity of BGP and large Internet-like topologies makes it difficult to create a tractable and useful mathematical model. Such modeling has been attempted before, but only for regular topologies, and without taking the effects of MRAI into account [33]. Simulations of any system of the size and complexity of interdomain routing requires to make several simplifying assumptions in our model. In this section we describe the choices and assumptions we make, and argue why the resulting model captures the effects we want to investigate in our study.

We study different growth models of the AS-level topology of the Internet, using our topology generator described in Sec. 3. In order to do this in a scalable way, *we model each AS as a single node*, and *connections between two neighboring ASes as a single logical link*. This implies that we do not capture routing effects within an AS, introduced by iBGP or interactions with IGP routing protocols (e.g., hot-potato routing). However, while such effects do have an impact on how routing updates are generated, they are orthogonal to the effects we want to study.

We focus on events where individual destination prefixes are withdrawn and then re-announced by the owner. This is the most basic routing event that can take place in the Internet, and at the same time the most radical; these changes must be communicated all over the network. For different topology growth scenarios, *we measure the number of routing updates received by nodes at different locations in the network.*

Figure 2 shows the structure of a node in our simulation model. A node exchanges routing messages with

its neighbors. Incoming messages are placed in a FIFO queue and processed sequentially by a single processor. The time it takes to process an update message is uniformly distributed between 0 and 100 ms. Each node maintains a table with the routes learned from each neighbor. Upon receiving an update from a neighbor, a node will update this table, and re-run its decision process to select a new best route. The new preferred route is then installed in the forwarding table and announced to its neighbors. For each neighbor, we maintain an export filter that blocks the propagation of some updates according to the policies installed in the network. Outgoing messages are stored in an output queue until the MRAI timer for that queue expires. If a queued update becomes invalid by a new update, the former is removed from the output queue.

For our study, we need a simulator that is capable of capturing the exchange of routing updates described above, and that scales to network sizes of thousands of nodes. Existing interdomain routing simulators fall into two broad categories. Either they only calculate steady state routes, and do not capture routing dynamics [26], or they include a detailed model of each eBGP session, and hence do not scale to network sizes in the order of today's AS-level Internet topology of about 27000 nodes [2, 10]. Because of this, we have chosen to develop a new simulation model that suits our requirements[1]. Using this simulator, we are able to efficiently simulate networks up to about 10000 nodes.

We consider policy-based routing, with the use of MRAI timers to limit the frequency with which a node sends updates to a neighbor. By "policies", we refer to a configuration where relationships between neighboring ASes are either peer-to-peer or customer-provider. We use normal "no-valley" and "prefer-customer" policies. Routes learned from customers are announced to all neighbors, while routes learned from peers or providers are only announced to customers. A node prefers a route learned from a customer over a route learned from a peer, over a route learned from a provider. Ties among routes with the same local preference are broken by selecting the route with the shortest AS path, then based on a hashed value of the node IDs.

By "MRAI" or "rate-limiting", we refer to a configuration where two route announcements from an AS to the same neighbor must be separated in time by at least one MRAI timer interval. We use a default MRAI timer value of 30 seconds. To avoid synchronization, we jitter the timer as specified in the BGP-4 standard. According to the BGP-4 standard [28], the MRAI timer should be implemented on a per-prefix basis. However, for efficiency reasons, router vendors typically implement it

---

[1]The simulator code and the scripts used to generate the results in this paper is available at `http://simula.no/research/networks/software`

on a per-interface basis. We adopt this approach in our model.

We look at two different implementation choices for the MRAI timer, with respect to whether explicit route withdrawals are subject to rate limiting. According to an outdated BGP specification (RFC1771) [27], explicit withdrawals are *not* subject to the MRAI timer. This approach is still used by some router implementations, including the open-source software router project Quagga [1]. In the most recent RFC (RFC4271) [28] however, it is stated that explicit withdrawals *should* be rate-limited just like other updates. The reasons for making this important change are not clear to us. Discussions on the IDR mailing list (see [30] and the related thread) indicate that some people opposed treating withdrawals differently than other updates, since there is no general way to distinguish "good news" from "bad news" based on the update type. While this might be true for some corner cases, we do not think this argument is a sufficient reason for a change that as we will show has a major impact on routing scalability. We let WRATE denote an implementation where explicit withdrawals are rate limited just like other updates, while NO-WRATE denotes an implementation where explicit withdrawals are sent immediately when they are generated.

3. CONTROLLABLE TOPOLOGIES

In this section, we first describe some key properties that characterize the AS-level Internet topology. We believe that these properties will remain valid in the future. We then describe a model that allows us to construct topologies with different configurable properties while still capturing these key properties.

Most existing topology generators are not capable of producing topologies annotated with business relations, which are essential in our study. Those who are [9, 14], do not have the flexibility we need for controlling different topological characteristics. It is possible to infer historical Internet topologies from routing update traces [12], but it is well known that such inference tends to underestimate the number of peering links, and it is difficult to infer topologies of a tractable size that are representative of todays Internet. We therefore implement our own topology generator.

The input parameters to our generator have "operational" semantics. Instead of specifying abstract graph properties like the clustering coefficient, the betweenness or the assortativity of the topology, we define our topology in a more hands-on, real-world related manner by specifying parameters like how many providers an AS has, how likely it is to peer with other types of ASes etc.

Stable topological properties The AS-level Internet topology is far from a random graph. Over the past

decade it has experienced tremendous growth, but the following key characteristics have remained constant:

Hierarchical structure. On a large scale, the nodes in the Internet graph form a hierarchical structure. By hierarchical we mean that customer-provider relationships are formed so that there are normally no provider loops, where A is the provider of B who is the provider of C who again is the provider of A.

Power-law degree distribution. The degree distribution in the Internet topology has been shown to follow a truncated power-law, with few very well-connected nodes, while the majority of nodes have only few connections [11]. The well connected nodes typically reside at the top of the hierarchy.

Strong clustering. The nodes in the Internet are grouped together in clusters, with nodes in the same cluster more likely to be connected to each other. One reason for this clustering is that networks operate in different geographical areas.

Constant average path length. Recent measurements show that in spite of a tremendous growth in the number of nodes, the AS-level path length has stayed virtually constant at about 4 hops for the last 10 years [8].

Topology generator Next, we describe a flexible model for generating topologies that captures the above properties about the AS-level graph. Several design choices and parameters in our topology generator were guided by a recent measurement study [8].

We use four types of nodes in our model. At the top of the hierarchy are the tier-1 (T) nodes. T nodes do not have any providers, and all T nodes are connected in a clique using peering links. Below the T nodes, we have the mid-level (M) nodes. All M nodes have one or more providers, which can be either T nodes or other M nodes. In addition, M nodes can have peering links with other M nodes. At the bottom of the hierarchy, we have two different types of stub nodes. We distinguish between customer networks (C) and content providers (CP). In this context, CP nodes would include content provider networks, but also networks providing Internet access or hosting services to non-BGP speaking customers. In our model, the difference between C and CP nodes is that only CP nodes can enter peering agreements with M nodes or CP nodes, while C nodes do not have peering links.

To capture clustering in our model, we introduce the notion of *regions*. The purpose of regions is to model geographical constraints; networks that are only present in one region are not allowed to connect with networks that are not present in the same region. In our model T nodes are present in all regions. 20% of M nodes and 5% of CP nodes are present in two regions, the rest are present in only one region. C nodes are only present in one region.

|	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 1: Topology parameters

We generate topologies top-down in two steps. First we add nodes and transit links, then we add peering links. The input parameters n_T, n_M, n_{CP} and n_C decide how many of the n nodes belong to each node type, respectively. First, we create the clique of T nodes. Next, we add M nodes one at a time. Each M node connects to an average of d_M providers, uniformly distributed between one and twice the specified average. M nodes can have providers among both T and M nodes, and we use a parameter t_M to decide the fraction T node providers. M nodes can only select providers that are present in the same region. M nodes select their providers using preferential attachment, which gives a power-law degree distribution [4].

We then add the CP and C nodes, which have an average number of providers d_{CP} or d_C, respectively. CP and C nodes can select T nodes as providers with a probability t_{CP} and t_C, respectively. Just like the M nodes, C and CP nodes select their providers using preferential attachment.

When all nodes have been added to the topology, we add peering links. We start by adding p_M peering links to each M node. As for the provider links, p_M is uniformly distributed between zero and twice the specified average. M nodes select their peers using preferential attachment, considering only the peering degree of each potential peer. Each CP node adds p_{CP-M} peering links terminating at M nodes, and p_{CP-CP} peering links terminating at other CP nodes. CP nodes select their peers among nodes in the same region with uniform probability. Importantly, we enforce the invariant that a node not peer with another node in its customer tree. Such peering would prey on the revenue the node gets from its customer traffic, and hence such peering agreements are not likely in practice.

Baseline growth scenario Next, we define a Baseline growth model that is similar to the growth seen in the Internet over the last decade [8]. The Baseline growth model is characterized by a slow increase in the MHD of stub nodes, and a faster growth in the MHD of
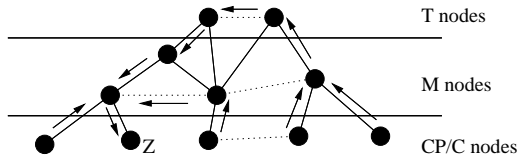
Figure 3: Illustration of network based on our topology model.

middle nodes and the number of peering links. In the Baseline topology we use 5 regions, containing one fifth of all nodes each. Table 1 gives the parameter values for the Baseline growth model.

The topologies generated using this growth model capture the four mentioned properties of the AS-level Internet graph. The hierarchical structure is trivially fulfilled through the way we construct the topologies. The use of preferential attachment when selecting which nodes to connect to gives a power-law degree distribution [3]. We have measured the clustering coefficient of our Baseline topologies to about 0.15, which is significantly higher than for random graphs of the same size. Finally, the average path lengths in our Baseline topologies is constant at around 4 hops as the network grows.

4. EXPLAINING CHURN IN A GROWING NETWORK

In this section, we first present our analytical model describing the number of updates received at a node. Then we use the Baseline growth model to show how this model can be simplified for the different node types, and to determine the most important factors driving the churn growth.

Our main metric is the number of updates received at a node after withdrawing a prefix from a C-type node, letting the network converge, and then re-announcing the prefix again. The experiment is repeated for 100 different C nodes (increasing this number does not change the results), and the number of received updates is measured at every node in the network. We then average over all nodes of a given type, and report this average. In the following, we refer to this procedure as a "C-event". Note that due to the heavy-tailed node degree distribution, we expect a significant variation in the churn experienced across nodes of the same type.

4.1 A framework for update analysis

We give a formulation for the number of updates received at a node, and discuss how the different churn increase factors depend on the use of policies, the topological properties of the network, and the convergence properties of the routing protocol used.

Figure 3 shows a generic network of the type de-

scribed in Sec. 3. Transit links are represented as solid lines, while peer-to-peer links are dotted. For each node, we have indicated the preferred path to the *event originator Z*, which is the node announcing the active prefix. The routing updates that establish these paths flow in the opposite direction. We observe that due to the use of policies, updates (and the resulting paths) will follow a particular pattern: a node N will only announce a route to its providers and peers after an event at node Z if N has Z in its customer tree. On the other hand, N will always send an update to its customers, unless its preferred path to Z goes through the customer itself.

Let $U(X)$ denote the number of updates a node of type X receives after a C-event. X can be either of the four node types in our model; T, M, CP or C. We distinguish between the number of updates received from customers $U_c(X)$, peers $U_p(X)$ and providers $U_d(X)$ respectively. The total number of updates will be the sum of these: $U(X) = U_c(X) + U_p(X) + U_d(X)$. Each of these values will depend on three factors - the number $\mathbf{m}_{y,X}$ of direct neighbors of a given business relation y, the fraction $\mathbf{q}_{y,X}$ of these neighbors that sends updates during convergence, and the number of updates $\mathbf{e}_{y,X}$ each of these neighbors contribute. The expected number of updates from a certain class of neighbors will be the product of these three factors, and we can write

$$U(X) = \mathbf{m}_{c,X}\mathbf{q}_{c,X}\mathbf{e}_{c,X} + \mathbf{m}_{p,X}\mathbf{q}_{p,X}\mathbf{e}_{p,X} + \mathbf{m}_{d,X}\mathbf{q}_{d,X}\mathbf{e}_{d,X}$$
(1)

Note that for some node types, some of these terms will be 0, e.g., T nodes have no providers, and stub nodes have no customers. In the sequel, we will discuss how each of these factors depend on various topological characteristics and their interactions with properties of the routing protocol.

In this section, we will assume the NO-WRATE implementation of the rate-limiting MRAI timer. When explicit withdrawals are not rate limited, their propagation through the network will only be delayed by the queuing and processing delay at each node. This delay is short (in the order of 100 ms in our model) compared to the MRAI timer interval (order of 30 s). Hence, the explicit withdrawals will reach all nodes in the network fast, usually before any announcements about alternative paths (which are subject to rate limiting) are sent out. This way, almost no *path exploration* [18] takes place with NO-WRATE. We return to the WRATE case in Sec. 6.

4.2 Churn at different node types

We focus our discussion on the churn experienced by transit providers (T and M nodes), and content providers (CP nodes). These are the AS types that are most likely to be affected by increasing churn rates, since they must maintain larger routing tables with few or no default routes. Also, as seen in Fig. 4, these are the nodes
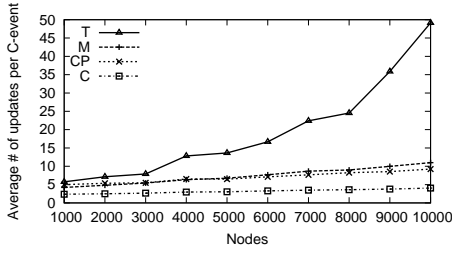
**Figure 4: Number of updates received at T, M, CP and C nodes.**
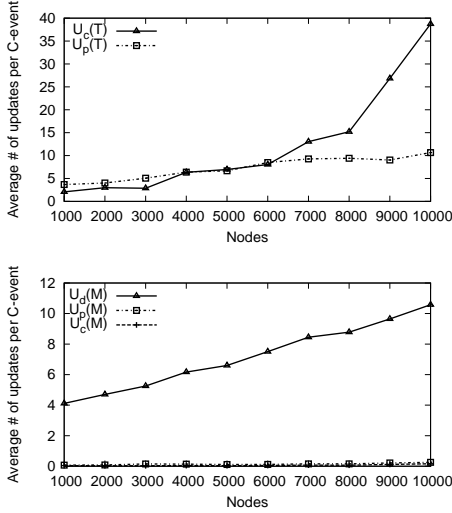
Figure 5: Number of updates received from peers and customers for T nodes (top), and from providers, peers and customers for M nodes (bottom).

that experience the stronger growth in the number of updates received after a C-event.

We have calculated 95% confidence intervals for the values shown in Fig. 4, and they are too narrow to be shown in the graph. This tells us that increasing the number of event originators beyond the 100 used in this experiment will not reduce the observed variance. This variance is a result of the often significant differences between topology instances of different size, caused by the heavy-tailed node degree distribution.

T nodes have no providers, so we have $U(T) = U_p(T) + U_c(T) = \mathbf{m}_{p,T}\mathbf{q}_{p,T}\mathbf{e}_{p,T} + \mathbf{m}_{c,T}\mathbf{p}_{c,T}\mathbf{e}_{c,T}$. The top panel in Fig. 5 shows $U_c(T)$ and $U_p(T)$, for topologies of increasing size created with our Baseline topology model. We observe that both $U_c(T)$ and $U_p(T)$ increase with network size, and that both these factors contribute significantly to the total number of updates. As the network grows, the increased multihoming increases the number of routes that a T node learns from both its
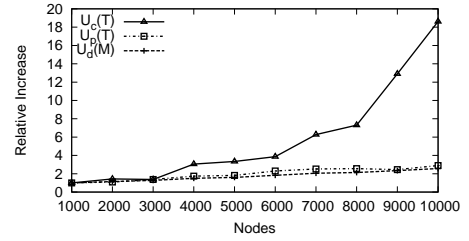
Figure 6: Relative increase in $U_c(T)$, $U_p(T)$ and $U_d(M)$.

customers and peers. $U_p(T)$ is the larger factor for small network sizes, and it grows approximately linearly with network size, with a coefficient of determination $R^2 = 0.95$. The strongest growth is seen in $U_c(T)$, which dominates for larger network sizes. Regression analysis shows that the growth of $U_c(T)$ is *quadratic*, with a coefficient of determination $R^2 = 0.92$.

While routes are only exported to peers and providers if they are received from a customer, routes are always exported to customers. As we can see in the bottom panel of Fig. 5, M nodes receive the large majority of their updates from their providers. Hence, a good estimate for the number of updates at M nodes is $U(M) = U_d(M) = \mathbf{m}_{d,M}\mathbf{q}_{d,M}\mathbf{e}_{d,M}$. The intuition behind this is that M nodes reach the "main part of the Internet" through their providers, and hence also receive the majority of routing updates from them. This is a major simplification, that makes our analysis much simpler. The same is true for CP nodes, so we limit our discussion to M nodes in the following.

Figure 6 shows the increase ratio in $U_c(T)$, $U_p(T)$ and $U_d(M)$. Each term is normalized so that the number of updates is 1 for $n = 1000$. To explain the observed trends for these terms, we look at the different factors described in Eq. 1 to find out how much of the growth is caused by each of them.

First, we look at the increase in the number of neighbors of different types. Figure 7 (top) shows the relative increase in the $\mathbf{m}_{c,T}$, $\mathbf{m}_{p,T}$ and $\mathbf{m}_{d,M}$ factors as the network grows. $\mathbf{m}_{c,T}$ grows much faster than the other factors. With our Baseline topology growth model, $\mathbf{m}_{c,T}$ grows approximately linearly with n in the range of network sizes we consider. The number of peers $\mathbf{m}_{p,T}$ is given directly by $n_T - 1$, which grows very slowly with n. Similarly, $\mathbf{m}_{d,M}$ is determined by the MHD of M nodes $d_M = 2 + 2.5n/10000$, which also grows linearly with n.

The middle panel in Fig. 7 shows the relative increase in $\mathbf{e}_{c,T}$, $\mathbf{e}_{p,T}$ and $\mathbf{e}_{d,M}$, representing the average number of updates received from each neighbor of a given type that exports a route. Note that the increase in these factors is much smaller than for the $\mathbf{m}$ factors, and that they are therefore less important for the overall increase

in churn. The increase in the **e** factors we see here is caused by random timer interactions when a prefix is announced. The increase is stronger for $\mathbf{e}_{d,M}$, since this represents links that are further away the event originator, giving more cases where a secondary path is received before the preferred one. We return to how this changes with WRATE in Sec. 6.

The bottom panel in Fig. 7 shows the fraction of neighbors of a given type that annouces a route after a C-event, represented by the $\mathbf{q}_{c,T}$, $\mathbf{q}_{p,T}$ and $\mathbf{q}_{d,M}$ factors. A provider will always announce a route to its customer, unless it prefers the path through the customer itself. Hence $\mathbf{q}_{d,M}$ is almost constant, and always larger than 0.99. $\mathbf{q}_{c,T}$ and $\mathbf{q}_{p,T}$ are both generally increasing with network size. This illustrates how *increased multihoming makes it increasingly likely that the event originator is in the customer tree of a given customer or peer of a T node.* This probability is much higher for peers than for customers of T nodes, since the peers, which are T nodes themselves, have a much larger number of nodes in their customer tree.

To sum up our discussion, we have that the churn at M nodes is dominated by the updates received from providers. The number of updates $U_d(M)$ grows with network size, since both the number of providers $\mathbf{m}_{d,M}$ and the average number of updates $\mathbf{e}_{d,M}$ received from each active provider grows, while the probability that a provider will announce a path is constant. The growth in $U_d(M)$ (a factor 2.6 in our range of $n = 1000$ to $n = 10000$, as seen in Fig. 6) is dominated by the (linear) growth in the MHD (a factor 2.2), which makes the total growth seem linear. $\mathbf{e}_{d,M}$ also grows, but at a much slower rate (factor 1.2), and does not produce a strong super-linearity.

For T nodes, both the updates $U_c(T)$ received from customers and the updates $U_p(T)$ received from peers are important, and both grow with network size. The strongest growth is contributed by $U_c(T)$, with a factor 18.5. Much of this growth can be attributed to a strong linear growth in the number of customers (a factor 9.5). Combined with the generally increasing trend for $\mathbf{q}_{c,T}$ (a factor 1.85), this gives a clearly superlinear growth in $U_c(T)$. The factor $\mathbf{e}_{c,T}$ also shows a small increase (a factor 1.06), but this is less important for the overall trend.

The number of updates $U_p(T)$ received from peers also grows, but at a slower rate. This is mainly because of the much slower growth in the number of peers - while the number of customers $\mathbf{m}_{c,T}$ increases with a factor of 9.5 over our range of topology sizes, the number of peers $\mathbf{m}_{p,T}$ grows only by a factor 1.7. The factor $\mathbf{q}_{p,T}$ also contributes to the growth in $U_p(T)$ by a factor of 1.5, while $\mathbf{e}_{p,T}$ is of less importance (a factor of 1.1).

This section has shown how the T nodes experience the highest growth in churn as the network grows with
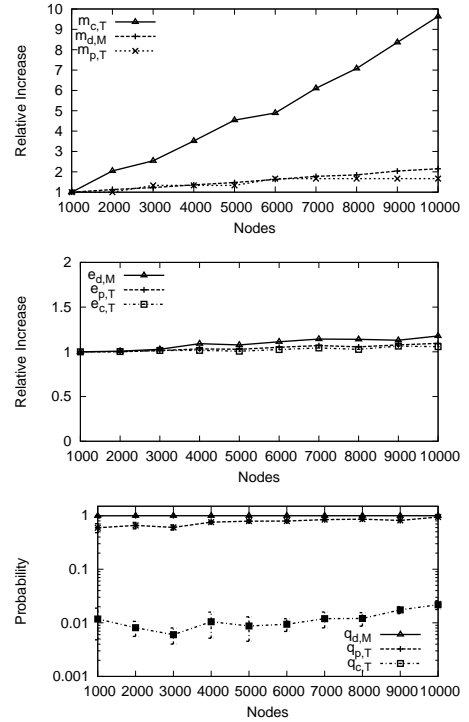

Figure 7: Relative increase in the factors that determine the update growth rate.

our Baseline growth model. This increase is driven mainly by an increased number of updates from customers. M and CP nodes also see increased churn, driven mainly by their increased MHD. In the next section, we will see how changes in the topology growth model affect the various churn factors.

5. TOPOLOGY GROWTH SCENARIOS

In this section, we look at several single-dimensional deviations from the Baseline model presented above. By looking at how BGP churn increases at various hypothetical growth models, we are able to answer different "what-if" questions about Internet growth. For example, what if multihoming to several providers becomes much more common than today for stub networks? Or what if buying transit services from tier-1 nodes becomes so cheap that they drive regional providers out of business? Our goal is not always to create realistic growth scenarios, but also to highlight the effect of altering different topological properties. Hence, we sometimes look at the effect of large changes to a single property at a time. We stick to the NO-WRATE implementation of the MRAI timer. This minimizes the effects of path exploration, and allows us to focus only on the effects of topological characteristics.

As seen in Sec. 4, T nodes experience both the strongest churn in absolute terms, and the strongest increase as
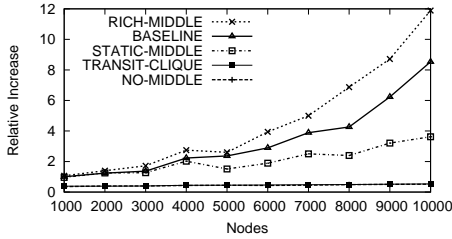
Figure 8: The effect of the AS population mix on T nodes.

the network grows. Hence, we focus mainly on the number of updates received at T nodes.

5.1 The effect of the AS population mix

First, we look at how the mix of different node types affects churn, by considering four different deviations from the Baseline model with respect to the mix of T, M, CP and C nodes. These deviations illustrate how economic factors can create a very different fauna of networks than what we see today. To implement these scenarios in our model, we change the parameters n_T, n_M, n_{CP} and n_C, while keeping all other parameters fixed.

NO-MIDDLE In the first deviation, we look at a network without M nodes, by setting $n_M = 0$. This illustrates a scenario where the price for transit services from the globally present tier-1 nodes is so low that they have driven regional transit providers out of business.

RICH-MIDDLE In the second deviation, we focus on the opposite scenario, where the ISP market is booming and there is room for a plethora of M nodes. We implement this by multiplying n_M by 3 ($n_M = 0.45n$), and reducing n_{CP} and n_C accordingly (while keeping their ratio constant).

STATIC-MIDDLE In the third deviation, we look at a situation where all network growth happens at the edges of the network. The number of transit providers (T and M nodes) is kept fixed, and the network grows only by adding P and C nodes. This could be a plausible scenario for the future, if the ISP population becomes stable.

TRANSIT-CLIQUE In the fourth and final deviation, we let all transit nodes be part of the top-level clique. This scenario may seem far-fetched, but it is important because it shows what would happen if the transit provider hierarchy collapses to a clique of "equals" connected by peering links. We implement this by setting $n_T = 0.15n$ and $n_M = 0$.

Figure 8 shows the increase in the number of updates seen after a C-event at T nodes for each deviation as the network grows. The increase is normalized with the number of updates in the Baseline model at $n = 1000$.

A first observation from the graphs is that the node

mix has a substantial influence on churn. In particular, the comparison of RICH-MIDDLE, Baseline, and STATIC-MIDDLE shows that the number of M nodes is crucial. There are two ways in which M nodes increase churn at T nodes. First, an increasing number of M nodes increases the customers $m_{c,T}$ of T nodes. For instance, in the RICH-MIDDLE deviation $m_{c,T}$ increases by a factor of 10.2 when n increases from 1000 to 10000. On the other hand, $m_{c,T}$ increases only by a factor of 5.3 in the STATIC-MIDDLE deviation. Second, an increasing number of M nodes, when they are multihomed to other providers (M or T nodes), tends to also increase the factor $q_{c,T}$. The reason is that M nodes create additional valid paths from the source of a C-event (at stub networks) to T nodes, and so it becomes more likely that a T node will receive updates from its peers and customers after a C-event. Regression analysis shows that the growth of $U(T)$ in the RICH-MIDDLE, Baseline and STATIC-MIDDLE deviations can be modeled as *quadratic*, with different scaling factors.

We also observe that the number of T nodes in the network does not have any impact on the number of updates by itself. The only difference between deviations NO-MIDDLE and TRANSIT-CLIQUE is in the number of T nodes, and we see that the number of updates is the same in these two scenarios. In the absence of M nodes, T nodes will receive one update for each provider the event originator has - either directly from the event originator, or from a peer. This number only increases as a function of the multihoming degree of the event originator, and is not influenced by the network size per se.

An important conclusion from the above observations is that *the increased number of updates does not primarily come from an increased number of transit nodes, but from the hierarchical structure in which they are organized.* An Internet with several tiers of providers buying transit services from other providers gives a much higher update rate than a more flat topology where most stub networks connect directly to tier-1 providers. Whether the Internet will move towards a more hierarchical or flat topology in the future is hard to tell. We do know however, that the average path length, measured in AS-level hops, has remained roughly constant, at around 4 hops, during the last ten years [8]. This implies that the Internet retains some hierarchical structure, and that the depth of that structure does not seem to vary with the size of the network.

5.2 The effect of the multihoming degree

Next, we look at the effect of varying the number of transit links each node brings to the network. Both stub and mid-tier nodes have an incentive to connect to several providers to increase their reliability and load balancing capability. We implement these scenarios by
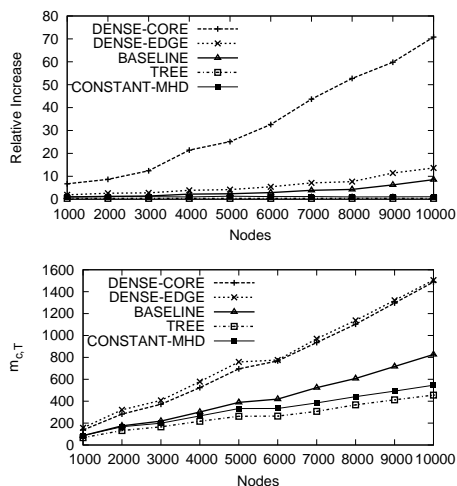
Figure 9: The effect of the multihoming degree at T nodes.

Figure 10: The effect of peering relations.

varying the d_M, d_{CP} and d_C parameters, while keeping all other parameters fixed.

DENSE-CORE We look at the effect of much stronger multihoming in the core of the network (M nodes). We implement this deviation by multiplying d_M by 3.
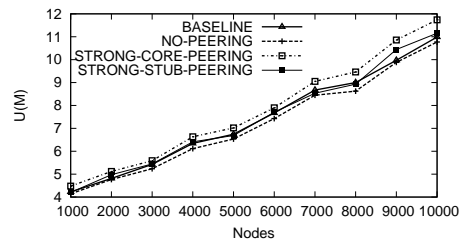
DENSE-EDGE We look at the effect of densification at the edges of the network. In this deviation, stub nodes increase their multihoming degree. We implement this by multiplying d_C and d_{CP} by 3.

TREE We look at a tree-like graph, where all nodes have only a single provider. Here, d_M, d_{CP} and d_C are all set to 1. This is clearly not a realistic scenario, but helps us explore the extreme version of a trend.

CONSTANT-MHD Finally, we look at a scenario where the multihoming degree of all nodes stays constant. We implement this by removing the component of d_M, d_{CP} and d_C that depends on n.

Figure 9 shows the number of received updates (top) and the number of customers m_c (bottom) for T nodes in the different scenarios. First, note that there is a clear connection between the MHD and the number of updates seen at a T node - for the same network size, a higher MHD causes larger churn. Second, even though the number of customers $m_{c,T}$ is about the same in DENSE-CORE and DENSE-EDGE, the churn is significantly higher in the former. This fact illustrates how the meshed connectivity of multihomed M nodes increases the likelihood that a T node will receive updates from a peer or customer. In other words, increased multihoming at the core of the network causes a larger growth in the factor $q_{c,T}$ than increased multihoming at the edges of the network. Specifically, we measured that $q_{c,T}$ increased by a factor of 1.6 in DENSE-CORE, while it increased by a factor of 1.3 in DENSE-EDGE.

When the MHD degree stays constant (in TREE and CONSTANT-MHD), the churn at T nodes is much less. In the extreme case of the TREE model, the churn at T nodes remains constant at two updates per C-event, because the T node learns about the event from exactly one peer or customer (once for the DOWN event and once for the UP event). In the CONSTANT-MHD model, the number of updates is also roughly constant because the increase in the number of customers $m_{c,T}$ as the network grows is offset by a corresponding decrease in the probability $q_{c,T}$ that any given customer of the T node will have the source of that C-event in its customer tree.

According to a recent measurement study [8], the average MHD of both stub nodes and providers has been increasing during the last decade (from 1.4 to 1.8 for stub nodes and from 1.8 to 3.2 for providers). The fact that the MHD has been increasing more rapidly in the core of the network implies that the Internet is closer to the DENSE-CORE model than to the DENSE-EDGE or the CONSTANT-MHD deviations. This can be viewed as bad news, at least in terms of BGP churn.

5.3 The effect of peering relations

In this subsection, we look at the impact of varying the peering degree between different types of nodes. The fraction of peering links in the Internet has increased over the last decade [8]. However, various difficulties in detecting such links do not allow us to know which peering model is most realistic.

NO-PEERING There are no peering interconnections, except in the clique of T nodes. This is clearly not realistic, but it serves as a reference point.

STRONG-CORE-PEERING We look at densification through more peering links in the core of the network. We model this deviation by doubling p_M.

STRONG-EDGE-PEERING Another possibility is densification through more peering links at the network edges. We model this deviation by multiplying p_{P-M} and p_{P-C} by 3.

Since the peering degree is only changed at M and CP nodes, we show the number of updates received at M nodes rather than T nodes. Figure 10 shows the number of updates received at M nodes as a function of network size for the Baseline and each deviation. The main

conclusion is that the peering degree does *not* cause a significant change in the generated churn. Adding or removing a significant number of peering links at the edge or at the core of the network does not give major differences in the number of updates. This conclusion also holds for other node types. To explain this observation recall that updates are propagated over peering links only for customer routes. Hence, the fraction of peering links that are active during a C-event is low. Moreover, such updates have limited export-scope (only to customers), compared to routes received from customers.

5.4 The effect of provider preference

Next, we look at the effect of provider preferences, i.e., the probability that a node chooses to buy transit services from a T or an M node. This choice has implications for how the network will grow; a higher preference for T nodes gives a more "flat" structure, while a higher preference for M nodes diverts more paths through several layers of hierarchy. We define two deviations of the Baseline model:

PREFER-MIDDLE In the first deviation, nodes prefer to buy transit services from M nodes rather than T nodes. We implement this by setting $t_P = t_C = 0$, and limiting the number of T providers for M nodes to one at most.

PREFER-TOP In this deviation, nodes prefer to buy transit services directly from T nodes. We implement this by limiting the number of M providers for M, CP and C nodes to be at most one.

The top panel in figure 11 shows that a scenario where most nodes buy transit from M nodes results in a higher churn at T nodes, while more direct connections to T nodes decreases churn. *If we are moving towards an Internet in which customers and content providers at the edges prefer to connect to mid-tier ISPs, the number of BGP updates at T nodes will be much higher than if they prefer to connect to tier-1 ISPs.* Looking at the different factors that determine $U(T)$, we observe that the PREFER-TOP deviation gives a much higher $\mathbf{m}_{c,T}$ than PREFER-MIDDLE, but that this is more than offset by a strong decrease in $\mathbf{q}_{c,T}$, as shown in the middle, and bottom panels in figure 11. An M node is more likely to notify its provider about a C-event than a stub node, because an M node has several potential event sources in its customer tree.

A recent study [8] observed that content providers and regional transit providers tend to buy transit service from either tier-1 or tier-2 providers with almost equal probability. The equivalent of C nodes ("Enterprise Customers") however, show a preference for tier-2 providers during the last 3-4 years, justifying the selection of the corresponding probabilities in the Baseline model.
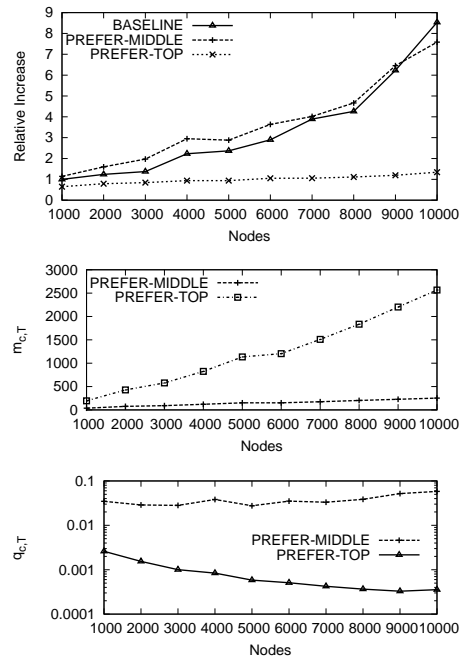

Figure 11: Updates increase/Factors for **PREFER-TOP** and **PREFER-MIDDLE** deviations

6. THE EFFECT OF WRATE

We have seen how topological properties like the MHD affect the number of neighbors and the probability that a neighbor will be active (send BGP updates). The topology, however, has little effect on the number of updates received from a single neighbor after a C-event. As explained in Sec. 4, the NO-WRATE implementation gives virtually no path exploration after the withdrawal of a prefix, and the **u** factors stay close to the minimum 2 updates (one after the withdrawal and one after the re-announcement). In this section, we investigate the effect of implementing the MRAI timer with the WRATE option, so that explicit withdrawals are rate limited like any other update. With WRATE, it takes more time for the route withdrawals to reach all nodes. During that time a node will announce other alternate paths, and hence we expect increased churn.

The upper panel in Fig. 12 shows the increase in updates received at T, M, CP and C nodes when WRATE is applied. We use the Baseline topology growth model. The plot shows the number of updates $U(X)$ received with WRATE, divided by $U(X)$ in the NO-WRATE case. We observe that WRATE causes a significant increase relative to NO-WRATE for all node types. Further, this increase factor grows with network size. For T nodes, the churn doubles when $n = 10000$. The relative increase is larger for nodes at the periphery of the network, since they generally have longer paths to the
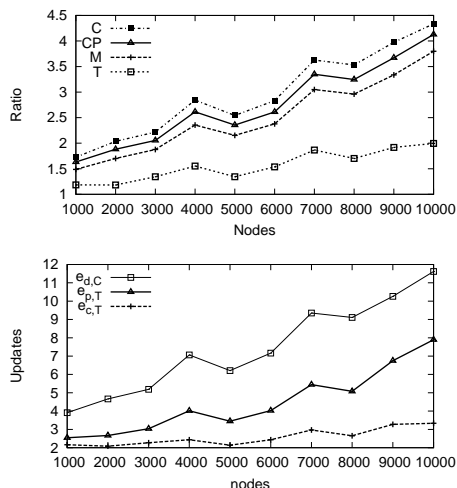
Figure 12: Number of updates (top) and e-factors for the WRATE case.

event originator, and hence a larger potential for path exploration. However, the absolute number of updates is much smaller at those nodes (see Fig. 4).[2]

The lower panel in Fig. 12 shows $\mathbf{e}_{d,C}$, $\mathbf{e}_{p,T}$ and $\mathbf{e}_{c,T}$ with WRATE. This graph confirms that the difference between the two MRAI implementations is caused by an increase in these factors. We also see that the increase in the number of received updates is stronger for neighbors with a larger number of policy-compliant paths that can be explored. In the Baseline growth model, the number of valid paths from a T node to an event originator increases superlinearly, which causes a superlinear growth in $\mathbf{e}_{p,T}$. On the other hand, the slower growth in the number of paths exported by customers also gives a slower growth in $\mathbf{e}_{c,T}$.

We have also explored the effect of WRATE in the topological deviations described in Sec. 5. Those results show that the churn increase with WRATE is stronger in a more well-connected network, especially at its core, where there are more potential paths that can be explored after a prefix withdrawal. In the DENSE-CORE deviation, the number of updates received at T nodes when $n = 10000$ increases by a factor 3.6 with WRATE, compared to 2.0 in the Baseline model.

The results presented here show how the use of WRATE exacerbates path exploration, and hence increases churn. This effect gets stronger as the network grows, and even more so in a network that is densely connected at its core. In summary, the previous observations make us question the wisdom of requiring WRATE in the most recent specification of the BGP protocol [28].

---

[2] Again, the large variance in this graph is due to the natural heterogeneity in the underlying topologies. The confidence interval at each network size is too narrow to show.

7. RELATED WORK

Interdomain routing dynamics and scalability has been a topic in the literature for the last decade or so, after it first was shown that BGP suffers from excessive churn caused by pathological protocol behavior [20]. The phenomenon of *path exploration* was discussed in [18], and upper and lower bounds for the number of updates exchanged during convergence were given. In a follow-up work, it was shown that the duration of path exploration depends on the length of the longest possible backup path to the affected destination [19]. The impact of MRAI timers to limit path exploration was discussed in [13]. In a more recent measurement study, it was shown that path exploration is less severe in the core of the network than on the edges [25], which is in accordance with the results presented in Sec. 6. Another study [5] showed that a small fraction of ASes is responsible for most of the churn seen in the Internet. It has also been shown that events at the edge of the network affect a larger number of nodes than those in the core [33]. A recent measurement study concluded that the state of BGP routing is now "healthier" than it was a decade ago, with less update traffic caused by configuration mistakes or protocol pathologies [22]. While all previous papers deal with measuring and explaining BGP routing dynamics, our work differs in that it focuses on the relation of topology growth and BGP dynamics. Increased churn is also a main motivation for completely new routing architectures, like [32]. Other work has focused on BGP scalability in the context of increasing routing table size. One study finds that address fragmentation is the largest contributor to routing table growth [6]. As a response to the increased routing table sizes, a radically different routing strategy called compact routing has been proposed [17]. This approach can give routing table sizes that scale logarithmically with the number of routable addresses, but performs poorly under dynamic conditions.

8. CONCLUSIONS

We have examined the roles of topology growth and MRAI timer implementation on the scalability of BGP, by looking at the number of updates received at nodes at different locations in the AS hierarchy after a C-event. For different node types, we have identified the most significant sources of churn, and described how different factors contribute to increased churn as the network grows. We have shown that nodes at the top of the AS hierarchy experience both the highest churn in absolute terms, and the strongest increase as the network grows.

Using our flexible topology model, we have explored scalability in several plausible and educational "what-if" scenarios for the growth of the AS-level topology. We have shown that *the most important topological fac-*

tor deciding the number of updates generated is connectivity in the core of the network. In particular, the number mid-tier transit providers and the multihoming degree of these nodes plays a crucial role, since *transit nodes in the mid-level of the Internet hierarchy have a special role in multiplying update messages.* Measurements have shown that the update rate in the Internet increases at a faster rate than the number of prefixes in the routing table [16]. Our study shows that *a main explanation for the reported churn growth can be the increasing connectivity in the core.* Another important finding from this study is that peering links play a very different role than transit links with respect to scalability. *The peering degree in the Internet does not influence churn.* Finally, we have shown that the depth of the hierarchical structure in the Internet plays a significant role. *A relatively flat Internet core is much more scalable than a vertically deep core.*

In the last section of this paper, we investigated the role of a particular aspect of the rate-limiting mechanism - whether explicit route withdrawals are subject to the MRAI timer. We showed that rate-limiting explicit withdrawals leads to a significant increase in churn under our event model, because of the effect of path exploration. This difference grows with network size, and is stronger in well-connected networks.

In our future work we plan to look at more complex events than the C-event, more elaborate topologies that include intradomain protocols and iBGP configurations, and other BGP mechanisms and configurations, such as Route Flap Dampening and BGP multipath extensions.

9. REFERENCES

[1] Quagga project website. http://www.quagga.net/.
[2] SSFNet website. http://www.ssfnet.org/.
[3] R. Albert and A. L. Barabasi. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74:47, 2002.
[4] A. L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–512, October 1999.
[5] A. Broido, E. Nemeth, and k. claffy. Internet expansion, refinement, and churn. *European Transactions on Telecommunications*, January 2002.
[6] T. Bu, L. Gao, and D. Towsley. On characterizing BGP routing table growth. *Computer Networks*, 45(1), May 2004.
[7] H. Chang, S. Jamin, and W. Willinger. Internet Connectivity at the AS-level: An Optimization-Driven Modeling Approach. In *ACM SIGCOMM Workshop on MoMeTools*, 2003.
[8] A. Dhamdhere and C. Dovrolis. Ten years in the evolution of the Internet ecosystem. In *IMC 2008*, 2008.
[9] X. Dimitropoulos, D. Krioukov, A. Vahdat, and G. Riley. Graph annotations in modeling complex network topologies. *arXiv:0708.3879*, 2008.
[10] X. Dimitropoulos and G. Riley. Efficient large-scale BGP simulations. *Elsevier Computer Networks, Special Issue on Network Modeling and Simulation*, 50(12):2013–2027, 2006.
[11] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the Internet topology. In *ACM SIGCOMM*, pages 251–262, 1999.
[12] Lixin Gao. On inferring autonomous system relationships in the Internet. *IEEE/ACM Transactions on Networking (TON)*, 9(6), December 2001.
[13] T. Griffin and B. Premore. An experimental analysis of BGP convergence time. In *ICNP*, 2001.
[14] Y. He, S. V. Krishnamurthy, M. Faloutsos, and M. Chrobak. Policy-aware topologies for efficient inter-domain routing evaluations. In *IEEE INFOCOM 2008 Mini-Conference*, Phoenix, AZ, USA, April 2008.
[15] G. Huston. The BGP instability report. http://bgpupdates.potaroo.net/.
[16] G. Huston and G. Armitage. Projecting future IPv4 router requirements from trends in dynamic BGP behaviour. In *ATNAC*, Australia, December 2006.
[17] D. Krioukov, k. claffy, K. Fall, and A. Brady. On compact routing for the Internet. *Computer Communications Review*, 37(3), July 2007.
[18] C. Labovitz, A. Ahuja, A. Bose, and F. Jahanian. Delayed internet routing convergence. In *ACM SIGCOMM*, pages 175–187, August 2000.
[19] C. Labovitz, A. Ahuja, R. Wattenhofer, and S. Venkatachary. The impact of Internet policy and topology on delayed routing convergence. In *INFOCOM*, Anchorage, AK, USA, April 2001.
[20] C. Labovitz, G. Robert Malan, and F. Jahanian. Internet routing instability. *SIGCOMM Comput. Commun. Rev.*, 27(4):115–126, 1997.
[21] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graph Evolution: Densification and Shrinking Diameters. *ACM Transactions on Knowledge Discovery from Data (ACM TKDD)*, 2007.
[22] J. Li, M. Guidero, Z. Wu, E. Purpus, and T. Ehrenkranz. BGP routing dynamics revisited. *Computer Communications Review*, April 2007.
[23] L. Li, D. Alderson, R. Tanaka, J. C. Doyle, and W. Willinger. Towards a theory of scale-free graphs: Definition, properties, and implications (extended version). *Internet Mathematics*, 2005.
[24] D. Meyer, L. Zhang, and K. Fall. Report from the IAB workshop on routing and addressing. http://tools.ietf.org/id/draft-iab-raws-report-02.txt, April 2007.
[25] R. Oliveira, B. Zhang, D. Pei, R. Izhak-Ratzin, and L. Zhang. Quantifying path exploration in the Internet. In *IMC*, Rio de Janeiro, Brazil, October 2006.
[26] B. Quoitin and S. Uhlig. Modeling the routing of an autonomous system with C-BGP. *IEEE Network*, 19(6), November 2005.
[27] Y. Rekhter and T. Li. A border gateway protocol 4 (BGP-4). RFC1771, March 1995.
[28] Y. Rekhter, T. Li, and S. Hares. A border gateway protocol 4 (BGP-4). RFC4271, January 2006.
[29] RIPE's Routing Information Service. http://www.ripe.net/ris/.
[30] J. G. Scudder. [idr] re: [rrg] re: BGP path hunting, MRAI timer and path length damping. Message to the IDR mailing list, http://www1.ietf.org/mail-archive/web/idr/current/msg02415.html, June 2007.
[31] G. Siganos, M. Faloutsos, and C. Faloutsos. The Evolution of the Internet: Topology and Routing. *University of California, Riverside technical report*, 2002.
[32] L. Subramanian, M. Caesar, C. Tien Ee, M. Handley, Z. Morley Mao, S. Shenker, and I. Stoica. HLP: a next generation inter-domain routing protocol. In *Proceedings SIGCOMM*, pages 13–24, Philadelphia, USA, August 2005. ACM.
[33] X. Zhao, B. Zhang, A. Terzis, D. Massey, and L. Zhang. The impact of link failure location on routing dynamics: A formal analysis. In *ACM SIGCOMM Asia Workshop*, April 2005.
[34] S. Zhou and R. Mondragon. Accurately Modeling the Internet Topology. *Physical Review E, vol. 70*, 2004.