Can SDN Accelerate BGP Convergence?

A Performance Analysis of Inter-domain Routing Centralization

Pavlos Sermpezis
FORTH, Greece
sermpezis@ics.forth.gr

Xenofontas Dimitropoulos
FORTH / University of Crete, Greece
fontas@ics.forth.gr

***Abstract*—The Internet is composed of Autonomous Systems (ASes) or domains, i.e., networks belonging to different administrative entities. Routing between domains/ASes is realised in a distributed way, over the Border Gateway Protocol (BGP). Despite its global adoption, BGP has several shortcomings, like slow convergence after routing changes, which can cause packet losses and interrupt communication even for several minutes. To accelerate convergence, inter-domain routing centralization approaches, based on Software Defined Networking (SDN), have been recently proposed. Initial studies show that these approaches can significantly improve performance and routing control over BGP. In this paper, we complement existing system-oriented works, by analytically studying the gains of inter-domain SDN. We propose a probabilistic framework to analyse the effects of centralization on the inter-domain routing performance. We derive bounds for the time needed to establish data plane connectivity between ASes after a routing change, as well as predictions for the control-plane convergence time. Our results provide useful insights (e.g., related to the penetration of SDN in the Internet) that can facilitate future research. We discuss applications of our results, and demonstrate the gains through simulations on the Internet AS-topology.**

I. Introduction

The Border Gateway Protocol (BGP) is globally used, since the early days of the Internet, to route traffic between *Autonomous Systems* (ASes) or *domains*, i.e., networks belonging to different administrative entities. BGP is a distributed, shortest path vector protocol, over which ASes exchange routing information with their neighbors, and establish route paths.

Although BGP is known to suffer from a number of issues related to security [1], [2], or slow convergence [3], [4], [5], deployment of other protocols or modified versions of BGP is difficult, due to its widespread use, and the entailed political, technical, and economic challenges. Hence, any advances and proposed solutions, should be seamless to BGP.

Taking this into account, it has been proposed recently that Software Defined Networking (SDN) principles could be applied to improve BGP and inter-domain routing [6], [7], [8], [9], [10], [11]. The SDN paradigm has been successfully applied in enterprise (i.e., *intra*-AS) networks, like LANs, data centers, or WANs (e.g., Google). However, its application to inter-domain routing (i.e., between different ASes) has to overcome many challenges, like the potential unwillingness of some ASes to participate in the routing centralization. For instance, a small ISP might not have incentives (due to the high investment costs) to change its network configuration. This led previous works on inter-domain SDN to consider (a) partial deployment, only by a fraction of ASes, and (b) interoperability with BGP.

The proposed solutions have demonstrated that bringing SDN to inter-domain routing can indeed improve the convergence performance of BGP [12], offer new routing capabilities [6], or lay the groundwork for new services and markets [13], [7]. However, most of previous works are system-oriented: they propose new systems or architectures, and focus on design or implementation aspects. Hence, despite some initial evaluations (e.g., experiments, emulations, simulations) we still lack a clear understanding about the interplay between inter-domain centralization and routing performance.

To this end, in this paper, we study *in an analytic way* the effects of centralization on the performance of inter-domain routing. We focus on the potential improvements on the (slow) BGP convergence, a long-standing issue that keeps on concerning industry and researchers [14]. Our goal is to complement previous (system-oriented) works, obtain an analytic understanding, and answer questions such as: *"To what extent can inter-domain centralization accelerate BGP convergence? How many ASes need to cooperate (partial deployment) for a significant performance improvement? Is the participation of certain ASes more crucial? Will all ASes experience equal performance gains?"* Specifically, our contributions are:

- We propose a model (Section II) and methodology (Sections III and IV) for the performance analysis of inter-domain routing centralization. To our best knowledge, we are the first to employ a probabilistic approach to study the performance of inter-domain SDN.
- We analyse the time that the network needs to establish connectivity after a routing change. In particular, we derive upper and lower bounds for the time needed to achieve data-plane connectivity between two ASes (Section III), and exact expressions and approximations for the time till control-plane convergence over the entire network (Section IV). Our results are given by closed-form expressions, as a function of network parameters, like network size, path lengths, and number of SDN nodes.
- Based on the theoretical expressions, as well as on extensive simulation results, we provide insights for potential gains of centralization, inter-domain SDN deployment strategies, network economics, etc.

We believe that our study can be useful in a number of

ISBN 978-3-901882-94-4

directions. Research in inter-domain SDN can be accelerated and facilitated, since a fast performance evaluation with our results can precede and limit the volume of required emulations/simulations. The probabilistic framework we propose can be used as the basis (and be extended and/or modified) to study other problems or aspects relating to inter-domain routing, e.g., BGP prefix hijacking, or anycast. Finally, the provided insights can be taken into account in the design of protocols, systems, architectures, pricing policies, etc.

II. Model

A. Network

We consider a network, e.g., the whole Internet or a part of it, that consists of N autonomous systems (ASes). We represent each AS as a *single node* that operates as a BGP router; this abstraction that is common in related literature [3], [12], allows to hide the details of the intra-AS structure and functionality, and focus on inter-domain routing. When two ASes are connected (transit, peering, etc., relation), we consider that a link exists between the corresponding routers, over which data traffic and BGP messages can be exchanged.

B. SDN Cluster

ASes can be ISPs, enterprises, CDNs, IXPs, etc., belong to different administrative entities, and span a wide range of topological, operational, economic, etc., characteristics. As a result, not all ASes should be expected to be willing to cooperate for and/or participate in an inter-domain centralization effort. Routing centralization is envisioned to begin from a group of a few ASes cooperating with each other, e.g., at an IXP location [6], [7]; then, more ASes could be attracted (performance or economics related incentives) to join the group, or form another group.

To this end, we assume that $k \in [1, N]$ ASes, i.e., a fraction of the entire network, cooperate in order to centralize their inter-domain routing. In the remainder, we refer to the set of these k ASes, as the *SDN cluster*[1]. To avoid delving into system-specific issues of the centralization implementation, we assume the following setup, which captures main characteristics of several proposed solutions(e.g., [12], [9], [15]), and is generic enough to accommodate future solutions: ASes in the SDN cluster exchange routing information with a central entity, which we call *multi-domain SDN controller*. The multi-domain SDN controller might be an SDN controller that directly controls the BGP routers of the ASes (e.g., as in [12]), or a central server that only provides information or sends BGP messages to the ASes (e.g., similar to [15]).

C. BGP Updates

Each node has a routing table (Routing Information Base, RIB), in which each entry contains an IP Prefix, and the corresponding AS-path (i.e., sequence of ASes) through which this prefix can be reached. RIBs are built from the information received by the neighbor ASes: upon a routing change, the "source" AS (e.g., the AS that originates a prefix) sends BGP updates to its neighbors to notify them about the change; when an AS receives a BGP update, it calculates the needed updates (if any) for its RIB, and sends BGP updates to its own neighbors. In this way, BGP updates propagate over the entire network, and paths to prefixes are built in a distributed way.

Let us assume that an AS receives a BGP update at time t_1 and forwards it to a neighbor AS at time t_2. We call *BGP update time*, and denote T_{bgp}, the time between the reception of a BGP update in an AS and its forwarding to a neighbor AS, i.e., $T_{bgp} = t_2 - t_1$. The BGP update times may vary a lot among different ASes and/or connections, since they depend on a number of parameters: routers' hardware/software (e.g., time to process BGP data and update RIB) and/or configuration (e.g., MRAI timers), intra-AS network structure (e.g., number of routers, topology) and/or operation (e.g,. iBGP configuration, intra-AS SDN), etc.

Knowing all these parameters for every AS is not possible, and using (upper) bounds for T_{bgp} would not lead to practical conclusions [3]. Thus, to be able to perform a useful analysis, we follow a probabilistic approach, and model the BGP update times as follows.

Assumption 1 (BGP updates - renewal process)**.** *The BGP update times T_{bgp} are independent and identically distributed random variables, drawn from an arbitrary distribution $f_{bgp}(t)$, with $E[T_{bgp}] = \mu_{bgp}$.*

Under Assumption 1, BGP update times are given by a renewal process. The model is very generic, since it allows to use any valid function $f_{bgp}(t)$, and thus describe a wide range of scenarios with different parameters. Real measurements can be used to make a realistic selection for the distribution $f_{bgp}(t)$, as we show in Appendix A; however, a detailed study for fitting the $f_{bgp}(t)$ is beyond the scope of this paper.

D. Inter-domain SDN Routing

Routing information in the SDN cluster propagates in a centralized way, through the multi-domain SDN controller. When an AS in the SDN cluster receives a BGP update from a neighbor AS (not in the SDN cluster), it forwards the update to the SDN controller. The SDN controller, which is aware of the topology in the SDN cluster and the connections/paths to external ASes, informs every AS in the SDN cluster about the needed changes in the routing paths. The ASes that receive the updated routes from the controller, notify their non-SDN neighbors using the standard BGP mechanism.

Let t_1 be the time that the first AS belonging to the SDN cluster receives a BGP update from a non-SDN neighbor, and t_2 the time till *all* ASes in the SDN cluster have been informed (by the controller) for the BGP updates. We denote as T_{sdn} the time needed for all the SDN cluster to be informed after a member has received a BGP update, i.e., $T_{sdn} = t_2 - t_1$. The times T_{sdn} would depend on the system implementation. However, it was shown that system designs can achieve $T_{sdn} \ll T_{bgp}$ [16]. Hence, in the remainder -for the sake of presentation- we assume that $T_{sdn} \to 0$. Nevertheless, our

[1]Although we use the term *SDN*, our framework does not require necessarily that routing centralization is implemented on an SDN architecture.

TABLE I: Important Notation

N	network size (total # of nodes)	
k	SDN cluster size (total # of SDN-nodes)	
T_{bgp}	BGP update time	
$f_{bgp}(t)$	distribution of BGP update times	Assumption 1
d	path length	
$k^{'}$	# of SDN-nodes *on a path*	
T_{SD}	data-plane connectivity time in a SD-path	Theorem 1
T_c	BGP convergence time	Theorem 2
T_ℓ	ℓ-partial BGP convergence time	Corollary 1

results can be easily modified for arbitrary T_{sdn} (even for cases with $E[T_{sdn}] > E[T_{bgp}]$), without this affecting the main conclusions of the study.

E. Preliminaries and Problem Statement

In our analysis, we consider the following setup:

Every node in the network knows at least one (BGP) path to every other node.

A node initiates a routing change that affects the inter-domain routing (e.g., node n_0 in Fig. 1). This could be an announcement or withdrawal of an IP prefix, an interruption of an AS connection (e.g., a link is down), etc. Here, we consider that a node, which we call the "source node", announces a new IP prefix; this routing change affects the entire network, every node will install a path for this prefix in its RIB upon the reception of the BGP update.

Nodes in the SDN cluster, receive route information from the SDN controller, and add an entry in their RIB for the prefix to the source node; even if the path is not established in the node preceding in this path (e.g., in Fig. 1 node n_j might receive the update before node n_{j-1}). In this case only the node in the SDN cluster knows how to route traffic to the new prefix, therefore, if the SDN node sends traffic to the new prefix, this would not necessarily reach the source-node. The connectivity will be established when every AS in the path has been informed about the BGP update.

BGP updates do not propagate backwards in the path; this would create loops or longer paths, which are discarded or not preferred by BGP.

We call "SD-path" the final path, i.e., the shortest conforming to the routing policies, between the source node ("S") and another node ("destination", or "D").

In the remainder of the paper we investigate the effects of routing centralization on: (a) the data-plane connectivity between the source node ("S") and any node ("D") in the network, i.e., the time needed till all nodes in an SD-path have installed the updated BGP paths after a routing change (Section III); and (b) the control-plane convergence, i.e, the time needed till the entire network has established the final paths corresponding to the routing change (Section IV)[2].

For ease of reference, we summarize the notation in Table I.

[2]In the following sections we discuss the main points of our analysis, state the theoretical results, and focus on their implications and practical insights. Due to space limitations, the detailed proofs of all our results are given in [17].

Fig. 1: *SD path* of size d. The node n_0 initiates the routing change; nodes n_i and n_j belong to the SDN cluster.

III. Data-Plane Connectivity

A. Analysis

A source node "S" announces a new IP prefix, and SD-path is the final path from S to a "destination" node D; see, e.g., Fig. 1. Theorem 1 bounds the expectation of the time T_{SD} needed to establish (or, guarantee) data-plane connectivity in the path.

Theorem 1. *The expectation of the time T_{SD} in a path of length d with $k^{'} \in [0, d+1]$ nodes in the SDN cluster, is bounded as follows*

$$LB(d,k^{'})\cdot E[T_{bgp}] \leq E[T_{SD}|d,k^{'}] \leq UB(d,k^{'})\cdot E[T_{bgp}] \quad (1)$$

where

$$LB(d,k^{'}) = \begin{cases} 0 & , \ d \leq k^{'} \leq d+1 \\ \frac{d}{k^{'}+1} & , \ 0 \leq k^{'} < d \end{cases} \quad (2)$$

and

$$UB(d,k^{'}) = \begin{cases} d-k^{'}+1 & , \ 2 \leq k^{'} \leq d+1 \\ d & , \ 0 \leq k^{'} < 2 \end{cases} \quad (3)$$

Proof. The proof is given in [17]. □

We provide the intuition behind the proof of Lemma 1 in relation to Fig. 1. When a node in the SD-path that belongs to the SDN cluster receives the BGP update (e.g., node n_i), then every other node in the SDN cluster (e.g., node n_j) is informed about the update, sometimes even before its preceding node(s) (e.g., n_{j-1}). Hence, the BGP update can propagate on *different sections* of the SD-path *simultaneously* (e.g,. from n_i up to n_{j-1}, and -at the same time- from n_j to n_d). The length of these SD-path sections (which determine the BGP update propagation time) depend on the positions of the SDN nodes on the path. The bounds are derived based on the "best" and "worst" possible positions of the SDN nodes on the SD-path.

B. Network Topology and Routing Centralization

Based on Theorem 1 we can calculate the average time $E[T_{SD}|d]$ over all paths of the same size d (or, equivalently, for an average path of size d), using the property of conditional expectation:

$$E[T_{SD}|d] = \sum_{i=0}^{d+1} E[T_{SD}|d,k^{'}=i]\cdot P\{k^{'}=i|d\} \quad (4)$$

where $P\{k^{'}=i|d\}$ denotes the probability that i nodes (out of the total $d+1$ nodes on the path) belong to the SDN cluster.

Topology-independent SDN cluster. If the SDN cluster is formed independently of the network topology, the quantity $k^{'}$ follows an *hypergeometric distribution* with parameters N

(population size), k (number of successes in the population), and $d+1$ (number of draws), and probability mass function

$$P\{k^{'} = i|d\} = \frac{\binom{k}{i} \cdot \binom{N-k}{d+1-i}}{\binom{N}{d+1}} \quad (5)$$

Topology-related SDN cluster. On the other hand, if the participation of ASes in the SDN cluster is related to the topology, e.g., because ASes are explicitly selected based on topological characteristics (e.g., centrality), or the incentives of cooperation are inherently related to their connectivity (e.g., SDN deployment on tier-1 ISPs, or IXPs [6], [7]), then $k^{'}$ might not be captured accurately by Eq. (5). Therefore, the actual distribution $P\{d, k^{'}\}$ needs to be calculated; however, this might be a difficult (or infeasible) task.

Alternatively, in certain cases, the distribution $P\{k^{'} = i|d\}$ could be approximated with variations of the standard hypergeometric distribution that are able to take into account the fact that different nodes appear in shortest paths with different probabilities. For instance the *Fisher's noncentral hypergeometric distribution* can be used to consider biased selection of ASes for the SDN cluster: let ω_i be the betweenness centrality [18] of a node n_i, and ω_{sdn} and ω_{bgp} the averages among the nodes in the respective sets, i.e.,

$$\omega_{sdn} = \frac{\sum_{n_i \in SDN} \omega_i}{|\{n_i : n_i \in SDN\}|} \ , \quad \omega_{bgp} = \frac{\sum_{n_i \notin SDN} \omega_i}{|\{n_i : n_i \notin SDN\}|}$$

Denoting $\omega = \frac{\omega_{sdn}}{\omega_{bgp}}$, the probability $P\{k^{'} = i\}$ is approximately given by

$$P\{k^{'} = i|d\} = \frac{\binom{k}{i} \cdot \binom{N-k}{d+1-i} \cdot \omega^i}{\sum_{j=0}^{d+1} \binom{k}{j} \cdot \binom{N-k}{d+1-j} \cdot \omega^j} \quad (6)$$

In the above distribution, the higher the betweenness centrality of the ASes in the SDN cluster, the more skewed towards the higher values of $k^{'}$ the distribution $P\{k^{'}|d\}$ is, and, thus, the lower the delay T_{SD}.

Internet AS-topology vs. SDN cluster. We now focus on the Internet topology, which is of higher interest, and apply our -generic- theoretical results to investigate the effects of routing centralization.

We first build the Internet AS graph from a large experimentally collected dataset [19] (consisting of $N = 55567$ ASes), and infer routing policies over existing links based on the Gao-Rexford conditions [20] (this is the most common approach in related literature; more details can be found in [17]). We consider about 10^6 different SD-paths, from which we calculate the path length distribution $P\{d\}$ (see Fig. 3), and the betweenness centrality for each node.

We consider different scenarios with variable SDN cluster size $k = 1, ..., N$, where the set of nodes in the SDN cluster are selected (a) *randomly*, or (b) based on their *betweenness centrality* (i.e., the top k nodes with the highest betweenness centrality values). From Theorem 1, we calculate the lower and upper bounds for the average T_{SD} time over all path lengths, i.e., $E[T_{SD}] = \sum_d E[T_{SD}|d] \cdot P\{d\}$, where $E[T_{SD}|d]$ is given by Eq. (4), and $P\{k^{'}|d\}$ from Eq. (5) or Eq. (6) for the aforementioned cases (a) and (b), respectively.

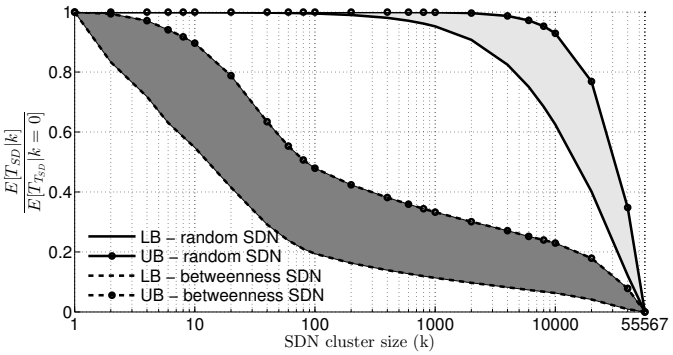


Fig. 2: Bounds for the average data-plane connectivity time, normalized over the no SDN scenario, i.e., $\frac{E[T_{SD}|k]}{E[T_{SD}|k=0]}$, in the Internet AS-graph. Upper (UB) and lower (LB) bounds enclose the colored areas: nodes in the SDN cluster are selected (i) *randomly* (light grey area) and (ii) with decreasing *betweenness centrality* (dark grey area).

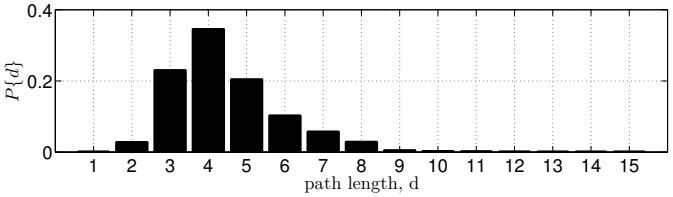


Fig. 3: Path length distribution on the Internet AS-topology.

In Fig. 2, we present the lower (LB) and upper (UB) bounds for $E[T_{SD}]$ for different SDN cluster sizes k, normalized over the case without routing centralization ($k = 0$). When the nodes in the SDN cluster are selected *randomly*, i.e., independently of the topology, a significant decrease in the average connectivity time can be achieved only when at least 20% (around $k = 10000$) of the nodes participate in the SDN cluster (note the log scale of the x-axis). This observation, which is in accordance with previous findings [12], is a rather grim message for the efficiency of (a randomly deployed) inter-domain routing centralization, since even if a few hundreds or thousands of ASes were willing to cooperate, the gains would be marginal.

On the contrary, as is shown in Fig. 2, when the SDN cluster consists of ASes with high *betweenness centrality*, with only a few tens of nodes the average delay can decrease up to 50%. This new insight (compared to previous understanding of the effects of routing centralization) brings optimism for the feasibility of inter-domain centralization: even if deployed incrementally, e.g., starting from a few tier-1 ISPs[3], the Internet can immediately see significant performance improvements.

[3]Large ISPs are central in the Internet topology, with high betweenness centrality. For example, the top-10 ASes with the highest betweenness centrality values belong to the list of the top-50 ASes with the largest number of ASes in customer cone [21]

C. Simulation Results and Implications

To validate our theoretical results, we conduct simulations on scenarios with varying (a) *network topologies*: synthetic graphs such as full-mesh, Poisson graph, Barabasi-Albert (power low graph), Newman-Watts-Strogatz (small world graph), as well as, the real Internet AS-graph; (b) *SDN cluster sizes*: $k = 0, ..., N$; and (c) *distributions* $f_{bgp}(t)$: exponential with rate $\lambda = 1$ and uniform in $[0, 2]$, both with $\mu_{bgp} = 1$. In the following we present a subset of representative results, and discuss some important observations.

The average values of T_{SD} in the simulations, are *always* within the bounds of Theorem 1 for all pairs $\{d, k'\}$ in every scenario we tested.

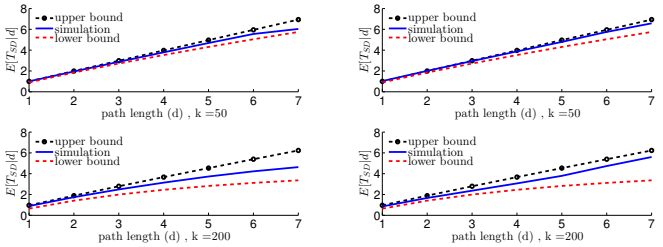
In Fig. 4 we compare the simulation results for $E[T_{SD}|d]$ (average over all k') against the theoretical bounds, which are calculated from Eq. (4) by using the expressions of Eq. (5) (topology-independent SDN cluster) and Theorem 1. For both cases of $f_{bgp}(t)$, the bounds are very tight for $k = 50$, when only a small fraction (5%) of the nodes belong to the SDN cluster (top plots). For larger SDN cluster sizes ($k = 200$, or 20%; bottom plots), the bounds are still very tight for small path lengths (e.g., $d < 4$), while the range *[lower bound, upper bound]* increases with d. In summary, the accuracy of the bounds increases for smaller k or d.

For $k = 200$ and $d = 7$ (rightmost points in bottom plots), while the simulated value lies in the middle of the two bounds in the exponential $f_{bgp}(t)$ case (Fig. 4(a)), it is closer to the upper bound in the uniform $f_{bgp}(t)$ case (Fig. 4(b)). Among all the scenarios we tested, we did not observe any tendency of the values to be closer either to the upper or lower bound. This is an indication that there is probably a limit on how much tighter bounds can be derived.

In Table II, we show how the times T_{SD} change for increasing SDN cluster size k. Comparing the two cases, $d = 2$ and $d = 5$, we can see that the effect of the routing centralization is higher for longer paths. The simulated data-plane connectivity times decrease more and faster for $d = 5$, and this is captured also by the relative changes of the theoretical bounds.

Similar behavior is observed also in Fig. 5, in simulation scenarios on the Internet topology where the SDN cluster comprises nodes with high betweenness centrality. For $k = 10$, paths of length $d = 3$, $d = 6$, and $d = 9$, see a relative decrease on the average connectivity time of about 10%, 20%, and 40%, respectively. The corresponding values for $k = 50$ are about 25%, 40%, and 60% (i.e., almost double than $k = 10$), while for larger SDN cluster sizes ($k > 50$) the extra gain is small.

These findings (Table II and Fig. 5) demonstrate that ASes which have (on average) longer paths to other ASes, e.g., stub networks or small ISPs at the edge of the Internet, would see a higher benefit from routing centralization than central ASes (e.g., tier-1 ISPs) or well connected ASes such as CDNs [22]. Hence, the *node closeness centrality* [18] can be used as a metric to evaluate (or rank) the improvement in the performance of ASes: the lower the closeness centrality, the higher the benefit from routing centralization.



(a) $T_{bgp} \sim exponential(\lambda = 1)$ (b) $T_{bgp} \sim uniform(0, 2)$

Fig. 4: Data-plane connectivity time $E[T_{SD}|d]$ (y-axis), vs. size of network cluster k (x-axis). Simulation scenarios: Poisson graph network topology of size $N = 1000$ and $p = 0.005$, with (a) $T_{bgp} \sim exponential(\lambda = 1)$ and (b) $T_{bgp} \sim uniform(0, 2)$.

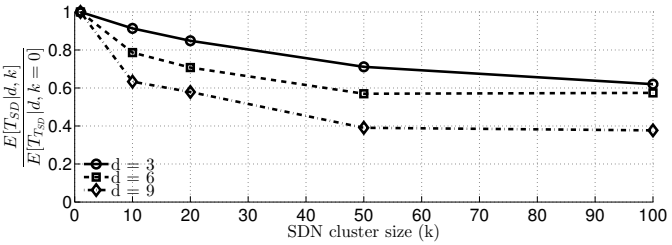


Fig. 5: Data-plane connectivity time normalized over the no SDN scenario $\frac{E[T_{SD}|d,k]}{E[T_{SD}|d,k=0]}$ (y-axis) vs. size of network cluster k (x-axis). Curves correspond to the averages for three different path lengths (i) $d = 3$, (ii) $d = 6$, and (iii) $d = 9$, in simulation scenarios over the Internet AS-graph, with $T_{bgp} \sim exponential(\lambda = 1)$, and nodes in the SDN cluster selected with decreasing *betweenness centrality*.

The above observation sheds light on an interesting trade-off related to which nodes participate to the SDN cluster and which nodes benefit from routing centralization. As shown in Section III-B, nodes with high betweenness centrality improve more the performance if they participate in the SDN cluster (see, e.g., Fig. 2). However, their own gain is smaller since they are central nodes in the network (betweenness and closeness centrality are positively correlated measures). As a result, incentives -other than performance- might be needed for attracting central ASes to cooperate for routing centralization. For instance, tier-1 ISPs could deploy inter-domain centralization in order to offer new services (related to the improved BGP convergence performance) to their customers.

IV. Control-Plane Convergence

In this section we derive results for the control plane convergence time, i.e., the time needed after a routing change till *every* AS in the network has updated and established the final (i.e., shortest, conforming to routing policies) paths.

The control-plane convergence time is equal to the maximum of the T_{SD} times over all the SD-paths. Due to the involved order statistics, proceeding similarly to Section III, would lead to complex computations and loose bounds. Hence, in this section, we proceed to an approximate analysis that allows us to provide useful insights for the effects of routing centralization on the BGP convergence time.

TABLE II: Data-plane connectivity time normalized over the no SDN scenario, $\frac{E[T_{SD}|k]}{E[T_{SD}|k=0]}$.

Upper bound / **Simulation** / Lower bound	$k=20$	$k=50$	$k=100$	$k=200$
$d=2$	99.9% / **99.2%** / 97.0%	99.6% / **97.7%** / 92.5%	98.6% / **92.9%** / 85.1%	94.4% / **85.1%** / 70.4%
$d=5$	99.9% / **97.8%** / 94.2%	99.3% / **93.9%** / 86.2%	97.4% / **86.4%** / 74.5%	90.1% / **75.6%** / 56.4%

Specifically, we first narrow the Assumption 1, by assuming that the renewal process for the BGP update times T_{bgp} is a Poisson process; this allows to study the problem in an analytically tractable way, using a Markovian framework. Our experiments and measurements in the Internet (Appendix A), support the selection of the Poisson assumption for the times T_{bgp}.

Assumption 2 (BGP updates - Poisson process)**.** *The times T_{bgp} are iid random variables, drawn from an exponential distribution with rate $\lambda = \frac{1}{\mu_{bgp}}$, and mean value $E[T_{bgp}] = \mu_{bgp}$.*

Under Assumption 2, we can build a *transient* Markov Chain to model the propagation of BGP updates, where each state denotes the set of nodes that have updated the paths in their RIBs. However, analysing such a Markov chain is still very complex, since the state space contains $2^N - 1$ states, and the transition rates depend on the topology of the network, which cannot be known exactly in most practical cases.

To this end, we first consider the case of a full-mesh network (a common approach in related literature [12], [3], [23]), which can be described by a much simpler Markov chain, and compute the control-plane convergence time as a function of the network size N, SDN cluster size k, and rate λ (Section IV-A). Then, we generalize the results, and derive approximations for sparse topologies, which are of higher practical interest (Section IV-B). Simulation results show that the insights stemming from our analysis are valid also for the (much more complex) Internet AS-graph (Section IV-C).

A. Analysis: Full-Mesh Topology

In a full-mesh network, every pair of nodes has a direct connection, and, thus, the shortest path (i.e., BGP path) to each node is the direct path of size $d = 1$. Hence, every node receives the BGP update from the source node. Moreover, since all nodes in the SDN cluster are informed the time any of them receives the BGP update ($T_{sdn} \ll T_{bgp}$, or $T_{sdn} \to 0$), the SDN cluster can be considered as a single node.

As a result, a Markov Chain as this in Fig. 6 can be used to model the propagation of BGP updates. Each time a node (a single AS or the SDN cluster) receives the BGP update, the Markov chain moves to the next state. We start from the moment/state (time $t = 0$ / state 0) just before the routing change takes place. Control-plane convergence is achieved at state C, when all nodes have the updated paths in their RIBs.

To calculate the transition rates $\lambda_i^{'}$, we first define the following quantities.

Definition 1 (bgp-eligible nodes & bgp-degree)**.**
– A `bgp-eligible` node is a node the (a) has not received the BGP update, and (b) lies on a BGP (shortest) path where the previous node has the updated route in its RIB.

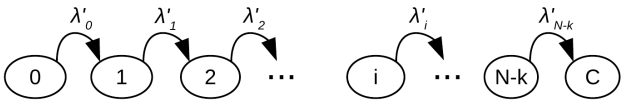


Fig. 6: Markov Chain for the BGP update dissemination process.

– The `bgp-degree` at step i, $D(i)$, is the number of nodes that are bgp-eligible nodes.

Under the above definition, the time to move from a step/state i to the next step/state, is the time needed till the *first* of the bgp-eligible nodes receives the update. Under Assumption 2, it follows that this time is the minimum of $D(i)$ iid random variables exponentially distributed with rate λ. Therefore the transition time is also exponentially distributed with rate (i.e., the transition rate)

$$\lambda_i^{'} = \lambda \cdot D(i) \tag{7}$$

Now, in a full-mesh network, bgp-eligible nodes are all the nodes that have not received the BGP update (since all nodes are directly connected to the source node). We denote as $n(i)$ the number of nodes that have received the BGP update at step i. From the above discussion it follows $n(i)$ depends on which step the SDN cluster received the BGP update. Denoting as x the state/step that the first node in the SDN cluster receives the BGP update, we can write

$$n(i|x) = \begin{cases} i & , i \leq x \\ i + k - 1 & , i > x \end{cases} \tag{8}$$

and the bgp-degree is easily shown to be given by Lemma 1.

Lemma 1. *The `bgp-degree` $D(i|x)$, $i \in [1, N-k], x \in [0, N-k]$, in a full-mesh network topology is given by*

$$D(i|x) = N - n(i|x) \tag{9}$$

Up to this point, we have calculated the transition rates of the Markov chain of Fig. 6 conditionally on x (see, Eq. (7) and Lemma 1). To compute the control-plane convergence time, we need also the probabilities $P_{sdn}(x)$ that the SDN cluster receives the BGP update at step x. In the following lemma, we derive the expression for the probabilities $P_{sdn}(x)$.

Lemma 2. *The probability that the SDN cluster receives the update at step x is given by*

$$P_{sdn}(x) = \frac{k}{N-x} \cdot \prod_{j=0}^{x-1} \left(1 - \frac{k}{N-j}\right) \tag{10}$$

Proof. The proof is given in [17]. □

Now, using Lemmas 1 and 2, we proceed and derive the following result for the distribution of the control-plane

convergence time T_c. Specifically, Lemma 3 gives a closed form expression for the moment generating function (MGF)[4] of the time T_c.

Lemma 3. *The moment generating function (MGF) $M_{T_c}(\theta)$ of the BGP convergence time T_c is given by*

$$M_{T_c}(\theta) = \sum_{x=0}^{N-k} \prod_{i=1}^{N-k} \left(1 - \frac{\theta}{\lambda \cdot D(i|x)}\right)^{-1} \cdot P_{sdn}(x) \quad (11)$$

Proof. The proof is given in [17]. □

Using the above lemma, and applying the property

$$E[X^n] = \left.\frac{d^n M_X(\theta)}{(d\theta)^n}\right|_{\theta=0} \quad (12)$$

we can calculate the moments of T_c. The following theorem gives the mean value (first moment) of T_c as a function of $D(i|x)$ (Lemma 1) and $P_{sdn}(x)$ (Lemma 2), or, equivalently, as a function of the parameters N, k, and λ.

Theorem 2. *The expectation of the BGP convergence time T_c is*

$$E[T_c] = \frac{1}{\lambda} \cdot \sum_{x=0}^{N-k} \sum_{i=1}^{N-k} \frac{1}{D(i|x)} \cdot P_{sdn}(x) \quad (13)$$

Proof. The proof is given in [17]. □

The methodology in the proof of Lemma 3 can be applied to derive useful expressions for other quantities that are of practical interest, and allow us to obtain a better understanding of the effects of routing centralization on control-plane convergence. For example, the following corollary quantifies the speed of the control-plane convergence process.

Corollary 1. *The expectation of the ℓ-Partial BGP Convergence Time, T_ℓ, i.e., the time needed till ℓ ($\ell \leq N$) nodes have the final BGP updates,is given by*

$$E[T_\ell] = \frac{1}{\lambda} \cdot \sum_{x=0}^{N-k} \sum_{i=1}^{M(\ell,x)} \frac{1}{D(i|x)} \cdot P_{sdn}(x) \quad (14)$$

where

$$M(\ell, x) = \begin{cases} \ell - 1 & , \quad 0 < \ell \leq x+1 \\ x & , \; x+1 < \ell \leq x+k \\ \ell - k & , \; x+k < \ell \leq N \end{cases} \quad (15)$$

B. Analysis: Sparse Topologies

As mentioned earlier, computing the control-plane convergence for an arbitrary topology is very complex. For instance, applying the methodology of Section IV-A, the set of bgp-eligible nodes at a step i depends on the exact paths $\mathcal{P}$ that the BGP updates have been propagated. Hence, we need to consider all $S \in \mathcal{P}$ (with $|\mathcal{P}| \sim O\left(2^N\right)$), and we need to keep track of all $D(i|x, S \in \mathcal{P})$ and $P_{sdn}(x|S \in \mathcal{P})$. However, approximating sparse topologies with a Poisson (or, Erdos-Renyi) random graph $G(N,p)$, we derive expressions for the BGP convergence time in the following result. As we show in the validation Section IV-C, our approximations describe well effects of routing centralization also in more generic/realistic topologies, like power-law graphs or the Internet AS-graph.

Result 1. *Lemma 3, Theorem 2, and Corollary 1, with $E[D(i|x)]$ (instead of $D(i|x)$), approximate the control-plane convergence time in a Poisson graph network topology; where $E[D(i|x)]$ is the expectation of the* `bgp-degree` *$D(i|x)$, $i \in [1, N-k], x \in [0, N-k]$, in a Poisson graph*

$$E[D(i|x)] = (N - n(i|x)) \cdot \left(1 - (1-p)^{n(i|x)}\right) \quad (16)$$

Proof. The proof is given in [17]. □

C. Simulation Results and Implications

We evaluate the accuracy of our theoretical results in various simulation scenarios, including also scenarios where the assumptions for (i) exponential $f_{bgp}(t)$, and (ii) full-mesh or Poisson graph networks, do not hold.

In scenarios of full-mesh networks, where the times T_{bgp} are exponentially distributed, our theoretical expressions of Section IV-A predict the simulation results for the expected convergence time $E[T_c]$ with very high accuracy.

For the validation of the theoretical expressions in sparse networks (Section IV-B), we simulate various sparse topologies, like Poisson, Barabasi-Albert (power low), and Newman-Watts-Strogatz (small world) graphs. Although the theoretical results are derived under the Poisson graph assumption, our simulations show that they can predict the performance with similar accuracy in the all the topologies we tested.

In Fig. 7 we present a representative subset of our results that demonstrate how the routing centralization can decrease the BGP convergence time. We plot the partial convergence time, normalized over the scenario without centralization, i.e. $, \frac{E[T_\ell|k]}{E[T_\ell|k=0]}$. We consider three cases, $\ell = 100$ (or, $0.1 \cdot N$) in Fig. 7(a), $\ell = 500$ (or, $0.5 \cdot N$) in Fig. 7(b), and $\ell = N = 1000$ that corresponds to the control-plane convergence in Fig. 7(c).

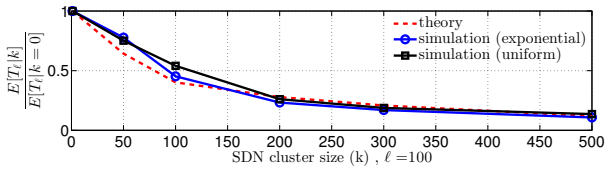
A first observation is that our results can capture well the relative changes[5] in the (partial) convergence time, not only for scenarios with exponential $f_{bgp}(t)$ (as we assume in our analysis), but also for scenarios with uniform $f_{bgp}(t)$.

In Fig. 7(c), we can see that the control-plane convergence time does not significantly improve as the SDN cluster size k increases. For instance, even for $k = 500$ (i.e., 50% of the nodes belong to the SDN cluster), the decrease in the convergence time is less than 30%. This comes to verify the results of [12], which showed that significant gains can be achieved only for high values ($> 50\%$) of SDN penetration.

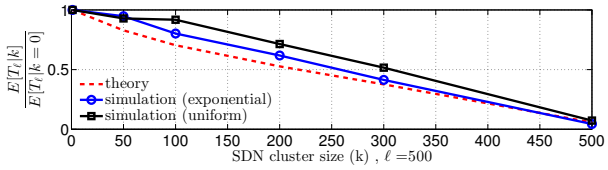
However, when it comes to the partial control-plane convergence (Figs. 7(a) and 7(b)), the effects of routing centralization are higher. The time needed till 10% of the nodes ($\ell = 100$ - Fig. 7(a)) to receive the updated routing information, decreases quickly; e.g., to 0.5 of its no-SDN ($k = 0$) value, only with $k = 100$ nodes (10%) participating in the SDN cluster.

[4] We remind that the MGF of a random variable X is defined as $M_X(\theta) = E[e^{\theta \cdot X}], \theta \in \mathbb{R}$, and completely characterizes a random variable (equivalently to its distribution), since all the moments of X can be calculated from its MGF.

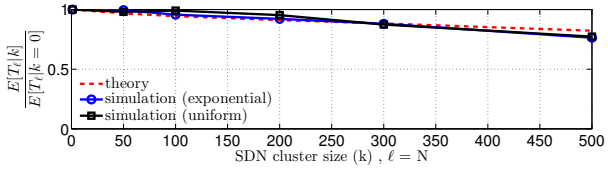
[5] The accuracy of the theoretical results (approximations), when we consider the *actual* -not normalized- values, is lower.



(a) $\ell = 100$



(b) $\ell = 500$



(c) $\ell = 1000 (= N)$

Fig. 7: Partial convergence time, normalized over the no SDN scenario, $\frac{E[T_\ell|k]}{E[T_\ell|k=0]}$ (y-axis), vs. size of SDN cluster k (x-axis). Simulation scenarios: Barabasi-Albert topology with $N = 1000$ and average node degree 10 (i.e., $0.01 \cdot N$); $T_{bgp} \sim exponential(\lambda = 1)$ (black line - squares) and $T_{bgp} \sim uniform(0, 2)$ (blue line - circles).

This reveals an important aspect, relating to the effects of routing centralization, which has not been shown in previous works (e.g., [12]): although the control-plane convergence can significantly improve only if a high percentage ($> 50\%$) of nodes cooperate, we can have very large gains in the *partial convergence* even with small sizes of SDN clusters.

In Fig. 8 we present simulation results on the Internet AS-graph[6], where the top betweenness centrality nodes form the SDN cluster. Despite the fact that the simulated scenario deviates from our assumptions, our main theoretical findings are still valid: centralization can significantly accelerate the connectivity time with a large percentage of ASes (i.e., ℓ-partial convergence, see, e.g., curves for $\ell = 0.1 \cdot N$ and $\ell = 0.5 \cdot N$), while the total convergence $E[T_c]$ improves slower with the SDN cluster size k. Moreover, the efficiency of inter-domain centralization is quite impressing: with only $k = 50$ central nodes in the SDN cluster, the time needed to establish updated paths with half of the Internet nodes ($\ell = 0.5 \cdot N$) is 50% less than in the case without centralization.

V. Related Work

Inter-domain SDN is a new research area that attracts increasing attention [6], [7], [11], [8], [9], [10], [12]. In [6] authors propose and implement SDX, a software-defined component for IXPs, which increases the capabilities on routing control. Another IXP-based system that enables novel services

[6]For scalability issues, we did not consider here stub ASes and ASes with less than 3 neighbors, resulting in a reduced Internet graph with $N = 11527$.

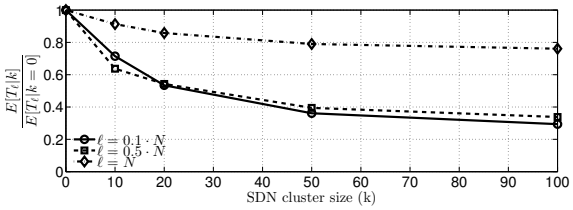


Fig. 8: Partial convergence time, normalized over the no SDN scenario, $\frac{E[T_\ell|k]}{E[T_\ell|k=0]}$ (y-axis), vs. size of SDN cluster k (x-axis). Simulation scenarios on the Internet AS-graph. Nodes in the SDN cluster are selected with decreasing *betweenness centrality*.

for establishing QoS route paths is described in [7]. In [11] a solution for incremental deployment of inter-domain SDN, seamless to traditional IP networks, is proposed, and [8] contributes in this direction by proposing an SDN-based methodology for decoupling BGP policy control from routing. [9] proposes an SDN-based architecture to enhance inter-domain routing, and [10] proposes a component to enable inter-domain SDN. Finally, [12] builds a realistic emulator, and investigates the effects of routing centralization on BGP convergence time.

The slow convergence of BGP has been extensively studied through measurements in [3], [4], [5]. It has been shown that BGP can take several minutes to converge after a routing change, and this can cause severe packet losses [3] and performance degradation [4].

Finally, analytic approaches for the BGP convergence can be found in [23], [3], [24]. In [23], a probabilistic model and automata theory is used to study the BGP convergence (probability of convergence, and convergence time). [3] studies analytically the BGP convergence with respect to the number of exchanged messages, while [24] performs a worst-case analysis of BGP convergence.

VI. Conclusion

In this paper, we analytically studied the effects of inter-domain SDN on the time needed for establishing connectivity and convergence after a routing change. We proposed a probabilistic model, and derived results for the expected data-plane connectivity time (lower/upper bounds) and control-plane convergence time (exact predictions and approximations).

Our results can be used to quickly evaluate the effects of different network parameters, like network size, topology, path lengths, or number of SDN nodes, on the routing performance. Hence, they can complement previous system-oriented studies and facilitate future research. Moreover, our methodology and results can be a useful tool for studying important problems relating to routing changes in the Internet. Finally, they can be applied in practical design problems, like selecting the nodes to participate in the SDN cluster based on performance criteria (i.e., which node can have the highest impact), or for network economics purposes (e.g., detecting the potential incentives for an AS to participate in inter-domain routing centralization).

Acknowledgements. This work has been funded by the European Research Council grant agreement no. 338402.

REFERENCES

[1] S. Kent, C. Lynn, and K. Seo, "Secure border gateway protocol (s-bgp)," *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 4, pp. 582–592, 2000.

[2] L. Subramanian, V. Roth, I. Stoica, S. Shenker, and R. Katz, "Listen and whisper: Security mechanisms for bgp," in *Proc. NSDI*, 2004.

[3] C. Labovitz, A. Ahuja, A. Bose, and F. Jahanian, "Delayed Internet routing convergence," *ACM SIGCOMM Computer Communication Review*, vol. 30, no. 4, pp. 175–187, 2000.

[4] N. Kushman, S. Kandula, and D. Katabi, "Can you hear me now?!: it must be BGP," *ACM SIGCOMM Computer Communication Review*, vol. 37, no. 2, pp. 75–84, 2007.

[5] R. Oliveira, B. Zhang, D. Pei, and L. Zhang, "Quantifying path exploration in the Internet," *Networking, IEEE/ACM Transactions on*, vol. 17, no. 2, pp. 445–458, 2009.

[6] A. Gupta, L. Vanbever, M. Shahbaz, S. Donovan, B. Schlinker, N. Feamster, J. Rexford, S. Shenker, R. Clark, and E. Katz-Bassett, "SDX: A software defined internet exchange," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 44, no. 4, pp. 551–562, 2014.

[7] Kotronis, V. and Klöti, R. and Rost, M. and Georgopoulos, P. and Ager, B. and Schmidt, S. and Dimitropoulos, X., "Stitching inter-domain paths over IXPs," in *ACM SOSR*, 2016.

[8] P. W. Thai and J. C. de Oliveira, "Decoupling bgp policy from routing with programmable reactive policy control," in *Proc. ACM CoNEXT Student*, 2012.

[9] C. Rothenberg, M. Nascimento, M. Salvador, C. Corrêa, S. C. de Lucena, and R. Raszuk, "Revisiting routing control platforms with the eyes and muscles of software-defined networking," in *Proc. ACM HotSDN*, 2012.

[10] R. Bennesby, E. Mota, P. Fonseca, and A. Passito, "Innovating on interdomain routing with an inter-sdn component," in *Proc. IEEE AINA*, 2014.

[11] P. Lin, J. Hart, U. Krishnaswamy, T. Murakami, M. Kobayashi, A. Al-Shabibi, K.-C. Wang, and J. Bi, "Seamless interworking of sdn and ip," in *Proc. ACM SIGCOMM 2013 (demo)*, 2013.

[12] V. Kotronis, A. Gämperli, and X. Dimitropoulos, "Routing centralization across domains via sdn: A model and emulation framework for BGP evolution," *Computer Networks*, pp. –, 2015.

[13] G. Gibb, H. Zeng, and N. McKeown, "Outsourcing network functionality," in *Proc. ACM HotSDN*, 2012.

[14] NANOG mailing list archives, "A survey on BGP convergence." http://seclists.org/nanog/2017/Jan/55, Jan. 2017.

[15] S. Vissicchio, O. Tilmans, L. Vanbever, and J. Rexford, "Central control over distributed routing," *ACM SIGCOMM Computer Communication Review*, vol. 45, no. 4, pp. 43–56, 2015.

[16] M. Alan Chang, T. Holterbach, M. Happe, and L. Vanbever, "Supercharge me: Boost router convergence with sdn," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 45, no. 4, pp. 341–342, 2015.

[17] P. Sermpezis and X. Dimitropoulos, "Can SDN accelerate BGP convergence? a performance analysis of inter-domain routing centralization." Technical Report, arXiv preprint, arxiv.org/abs/1702.00188, 2017.

[18] M. Newman, *Networks: An Introduction*. Oxford University Press, Inc., 2010.

[19] "The CAIDA AS relationships dataset." http://data.caida.org/datasets/as-relationships/, 01 Nov. 2016.

[20] L. Gao and J. Rexford, "Stable internet routing without global coordination," *IEEE/ACM Transactions on Networking (TON)*, vol. 9, no. 6, pp. 681–692, 2001.

[21] "AS-rank, CAIDA." http://as-rank.caida.org.

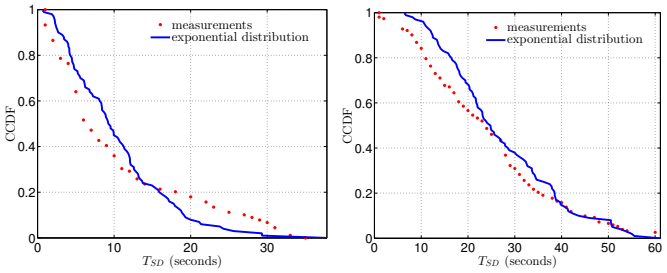
[22] Y.-C. Chiu, B. Schlinker, A. B. Radhakrishnan, E. Katz-Bassett, and R. Govindan, "Are we one hop away from a better internet?," in *Proc. ACM IMC*, 2015.

[23] R. Viswanathan, K. K. Sabnani, R. J. Holt, and A. N. Netravali, "Expected convergence properties of bgp," *Computer Networks*, vol. 55, no. 8, pp. 1957–1981, 2011.

[24] R. Sami, M. Schapira, and A. Zohar, "Searching for stability in inter-domain routing," in *Proc. IEEE INFOCOM*, 2009.

[25] B. Schlinker, K. Z. I., Cunha, N. Feamster, and E. Katz-Bassett, "Peering: An as for us," in *Proc. ACM HotNets 2014*, https://peering.usc.edu.

[26] G. Chaviaras, P. Gigis, P. Sermpezis, and X. Dimitropoulos, "ARTEMIS: Real-time detection and automatic mitigation for bgp prefix hijacking," in *Proc. ACM SIGCOMM Demo*, 2016.



(a) $d = 2$ (b) $d = 5$

Fig. 9: CCDF of the times T_{SD} for SD-paths of length (a) $d = 2$ and (b) $d = 5$. Comparison of times T_{SD} from *measurements* in the Internet (where we found $E[T_{bgp}] = 6.27$), and times T_{SD} generated from our model with $f_{bgp}(t) \sim$ *exponential distribution*$\left(\lambda = \frac{1}{E[T_{bgp}]}\right)$.

APPENDIX A
DISTRIBUTION OF BGP UPDATE TIMES T_{bgp}

To investigate if and how well our modeling assumptions can describe the BGP update times in the Internet, we compare them against real measurement data.

We conducted experiments in the Internet using the PEERING testbed [25], which owns IP prefixes and ASNs, peers with networks in different locations around the world, and allows users to make real BGP announcements. In our experiments/measurements, we follow a similar methodology as in [26]: we (i) announce a /24 prefix from a site of the PEERING testbed, and (ii) use public control-plane monitoring services (route collectors and looking glass servers) to measure the time needed till different ASes receive our announcements.

We collected BGP updates, as seen from the monitors, from $M = 40$ ASes. We repeated the experiments 14 times; each time making a BGP announcement either from the PEERING site at an IXP at Amsterdam (NL), or at an ISP at Los Angeles (US). From each received BGP update i, we consider (a) $T_{SD}(i)$, the time needed till the BGP update i received by the monitor (i.e., timestamp of the BGP update i minus the timestamp of our BGP announcement), and (b) $d(i)$, the length of the AS-path included in the BGP update i.

We group the times $T_{SD}(i)$ by the respective path lengths $d(i)$, and plot the distribution (CCDF) of the measured times T_{SD} in Fig. 9 for two example cases with $d = 2$ and $d = 5$.

Then, we fit the real data with a distribution $f_{bgp}(t)$ (cf. Section II), where we select $f_{bgp}(t) \sim exponential(\lambda)$ in order to test the validity of (the stronger) Assumption 2. We estimate the *average* BGP update time from the measured data as $\hat{E}[T_{bgp}] = \frac{\sum_i T_{SD}(i)}{\sum_i d(i)}$ and set the rate $\lambda = \frac{1}{\hat{E}[T_{bgp}]}$.

We generate from $f_{bgp}(t)$ a large number of times T_{SD} for paths of length $d = 2$ and $d = 5$, calculate their CCDFs, and compare them against the real data in Fig. 9. As we can observe, there is a good match between the generated and real data. This indicates that Assumption 2 is a realistic and reasonable assumption, and, thus, emphasizes the practicality of our theoretical and simulation findings in real settings