Inter-domain SDN: Analysing the Effects of Routing Centralization on BGP Convergence Time

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

Xenofontas Dimitropoulos
FORTH, Greece
fontas@ics.forth.gr

1. INTRODUCTION

Software-defined networking (SDN) has improved the routing functionality in networks like data centers or WANs. Recently, several studies proposed to apply the SDN principles in the Internet's inter-domain routing as well [1–5]. This could offer new routing opportunities [1, 2] and improve the performance of BGP [3–5], which can take minutes to converge to routing changes [6–8].

Previous works have demonstrated that centralization can benefit the functionality of BGP, and improve its slow convergence that causes severe packet losses [6] and performance degradation [7]. However, due to (a) the fact that previous works mainly focus on system design aspects, and (b) the lack of real deployments, it is not clearly understood yet *to what extent* inter-domain SDN can improve performance.

To this end, in this work, we make the first effort towards *analytically* studying the effects of routing centralization on the performance of inter-domain routing, and, in particular, the convergence time of BGP. Specifically, we propose a Markovian model for inter-domain networks, where a subset of nodes (*domains*) coordinate to centralize their inter-domain routing. We then derive analytic results that quantify the BGP convergence time under various network settings (like, SDN penetration, topology, BGP configuration, etc.). Our analysis and results facilitate the performance evaluation of inter-domain SDN networks, which have been studied (till now) only through simulations/emulations that are known to suffer from high time/resource requirements and limited scalability.

2. MODEL

Network Model. We consider a network (e.g., the Internet) composed of N *domains* or *autonomous systems* (ASes). We represent each AS as a single node, i.e., a single BGP router (similarly to [3]). Such an abstraction allows to hide the details of the internal structure of ASes, and focus on inter-domain routing.

We assume that $k \in [1, N]$ ASes cooperate in order to centralize their inter-domain routing: there exists a *multi-domain SDN controller*, which is connected to the BGP routers of these k ASes[1]. In the remainder, we refer to the set of the k ASes, as the *SDN cluster*.

BGP Updates. As in the Internet, ASes use BGP to exchange information and establish routing paths. When a BGP edge router of an AS receives a BGP update, it (i) calculates the updates (if any) for its BGP routing table, (ii) sends updates to the other BGP edge routers within the same AS (e.g., with iBGP), and (iii) sends updates to the BGP routers of the neighboring ASes. The time needed for this process may vary a lot among different connections since it depends on a number of factors, like the employed technology (hardware/software), routers' configuration (e.g., MRAI timers), intra-domain network, etc. To this end, in order to be able to analytically study the BGP updates dissemination (given the uncertainty and complexity), we model the time between the reception and forwarding of a BGP update in a probabilistic way.

Assumption 1. *The time between the reception of a BGP update in an AS/router and its forwarding to a neighbor AS/router, is exponentially distributed with rate λ.*

Despite the simplicity of the above assumption, our results can capture well the behavior of real/emulated networks (see Section 4).

Inter-domain SDN routing. Each AS belonging to the SDN cluster informs the SDN controller upon the reception of a BGP update. The SDN controller, which is aware of the topology of the SDN cluster (neighbors, policies, paths, etc.), calculates the changes in the routing paths and installs the updated routes in each router/AS belonging to the SDN cluster. ASes react to updates from the SDN controller, as in regular BGP updates, and, thus, forward them to their (non SDN) neighbors.

Let T_{sdn} be the time needed for an AS to inform the SDN controller and the controller to install the updated routes in every AS in the SDN cluster. This time can be expected to be in the order of few seconds [3], and much lower than the BGP updating process (cf., the default value for MRAI timers in Cisco routers is $30sec.$), thus, for simplicity, we assume here that $T_{sdn} = 0$.

3. ANALYSIS: BGP CONVERGENCE TIME

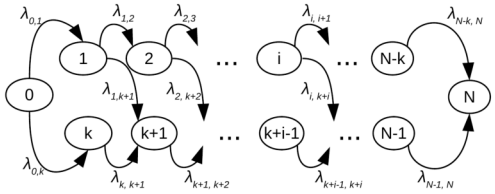
Let us assume a routing change, e.g., an announcement of a new prefix by an AS, in the network at time $t_0 = 0$. Our goal is to calculate the *BGP convergence time*, i.e., the time needed till all ASes/routers in the network have the final (i.e., shortest, conforming to policies) BGP routes for this prefix.

To this end, using Assumption 1, we can model the dissemination of the BGP updates in the network with the Markov Chain (MC) of Fig. 1(a), where each state corresponds to the number of ASes/routers that have the final BGP routes. At time $t_0 = 0$ the system is at state 0, while the state N denotes the BGP convergence. When an AS in the SDN cluster receives the BGP update, all the nodes in the SDN cluster are informed (through the controller); thus, we have a transition, e.g., from state i to state $k + i$. The transition rates in the MC, as we discuss in detail later, depend on the network topology.

The Markov Chain of Fig. 1(a) is transient, and the BGP convergence time is the time needed to move from state 0 to state N.

For notation brevity, in the remainder we use the MC of Fig. 1(b), which is equivalent to the MC of Fig. 1(a). Here, the states repre-

[1]This system abstraction can capture the main functionality of most of the previously proposed approaches.



(a) Markov Chain (number of nodes)

(b) Markov Chain (number of transitions, or *steps*)

Figure 1: Markov Chains where the states correspond to (a) the number of nodes that have updated BGP routes, and (b) the number of transitions, or *steps*, of the BGP update dissemination process.

sent the *number of transitions* in the MC of Fig. 1(a). For example, the state/step 1 corresponds to the state 1 or k of the MC of Fig. 1(a), while the state/step i corresponds to the state i or $k+i-1$ in the MC of Fig. 1(a). The states 0 are equivalent in both MCs, while the state/step C denotes the BGP convergence, and, thus, corresponds to the state N in the MC of Fig. 1(a).

If we denote with x the step at which -for the first time- an AS in the SDN cluster receives the BGP update, then the transitions rates $\lambda_i^{'}$ in the MC of Fig. 1(b) are given by

$$\lambda_i^{'} = \begin{cases} \lambda_{i,i+1} + \lambda_{i,i+k} & , i \leq x \\ \lambda_{k+i-1,k+i} & , i > x \end{cases} \quad (1)$$

We now proceed to calculate the rates $\lambda_i^{'}$. The ASes that have received the BGP updates, will then send BGP updates to some of their neighboring ASes, according to their routing policies. We refer to a neighbor to which the update will be forwarded as a *bgp-eligible* neighbor.

Definition 1. *We define as the* `bgp-degree` *at step i, $D(i)$, the number of (unique) ASes that are bgp-eligible neighbors with any of the ASes that have received the BGP updates at step i.*

Although an AS might receive the same BGP update from more than one neighbors, the final BGP route will correspond to only one of the received updates (i.e., shortest path). Hence, to calculate the transition rate $\lambda_i^{'}$, we take into account only one BGP connection (corresponding to the shortest path) per bgp-eligible neighbor. Since the BGP update times are exponentially distributed with rate λ (Assumption 1), it follows that $\lambda_i^{'}$ will be given by[2] $\lambda_i^{'} = \lambda \cdot D(i)$.

Knowing the rates $\lambda_i^{'}$, we can calculate the transition delays in each step. Adding the delays in each step, we can derive Theorem 1, which gives the BGP convergence time, i.e., the time to move from state 0 to state C. The detailed proof is given in [9].

Theorem 1. *The expectation of the BGP convergence time T in a hybrid SDN/BGP inter-domain topology is given by*

$$E[T] = \frac{1}{\lambda} \cdot \sum_{x=0}^{N-k} \sum_{i=1}^{N-k} \frac{1}{D(i|x)} \cdot P_{sdn}(x) \quad (2)$$

where $D(i|x)$ is the `bgp-degree` *of the network at step i given that the SDN cluster receives the update at step x, and $P_{sdn}(x)$ is the probability that the SDN cluster receives the update at step x.*

[2]The transition time is the minimum of $D(i)$ i.i.d. exponentially distributed times.

In the following sections we calculate the quantities $D(i|x)$ and $P_{sdn}(x)$ for important network topologies.

3.1 Full-Mesh Network Topology

We first consider a basic topology: a full-mesh network, where every AS-pair is connected.

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

$$P_{sdn}(x) = \frac{k}{N-x} \cdot \prod_{j=0}^{x-1} \left(1 - \frac{k}{N-j}\right) \quad (3)$$

Sketch of Proof (detailed proof in [9]). The SDN cluster comprises k (out of the total N) ASes. Thus, the probability that the announcing AS is in the SDN cluster (and thus $x = 0$) is $P_{sdn}(0) = \frac{k}{N}$. If the announcing AS is not in the SDN cluster, which happens w.p. $1 - P_{sdn}(0) = 1 - \frac{k}{N}$, we move to step 1: the remaining ASes without the update are $N-1$, of which k belong to the SDN cluster.Thus, the probability that the next AS that gets the update belongs to the SDN cluster is $\frac{k}{N-1}$. Therefore, it holds that

$$P_{sdn}(1) = \frac{k}{N-1} \cdot (1 - P_{sdn}(0)) = \frac{k}{N-1} \cdot \left(1 - \frac{k}{N}\right) \quad (4)$$

Proceeding similarly for the next steps $i = 2, ..., N-k$, we can derive the expression of Eq. (3). □

Theorem 3 gives the bgp-degrees $D(i|x)$ in a mesh network as a function of $n(i|x)$, which is defined as the number of nodes with updated BGP information at step i, given that the SDN cluster received the update at step x

$$n(i|x) = \begin{cases} i & , i \leq x \\ i + k - 1 & , i > x \end{cases} \quad (5)$$

Theorem 3. *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*

$$D(i|x) = N - n(i|x) \quad (6)$$

Proof. In a mesh network, since every AS-pair is directly connected, only the BGP messages sent by the announcing AS (i.e., shortest path) need to be considered. In step i, the announcing AS has $N - n(i|x)$ neighbors that have not received the BGP updates, and thus, it follows that $D(i|x) = N - n(i|x)$. □

3.2 Random Graph Network Topologies

In networks that are not full-meshes, ASes can be connected in different ways and with diverse policies. Since it is not possible to study every single topology, we use two classes of random graphs to capture the effects of centralization in non full-mesh networks. In this first approach, we consider unconstrained routing policies.

3.2.1 Poisson (Erdos-Renyi) Graph

We first consider the case of a Poisson random graph, where a link between each AS-pair exists with probability p. Varying the value of p we can capture different levels of sparseness.

Using similar arguments as in the full-mesh case, it is easy to show that the probabilities $P_{sdn}(x)$ are given by Theorem 2. The *expected* bgp-degrees, which can be used (as an approximation) instead of $D(i|x)$ in Theorem 1, are given by the following Theorem.

Theorem 4. *The expectation of the* `bgp-degree` *$D(i|x)$, $i \in [1, N-k], x \in [0, N-k]$, in a Poisson graph network topology is*

$$E[D(i|x)] = (N - n(i|x)) \cdot \left(1 - (1-p)^{n(i|x)}\right) \quad (7)$$

Sketch of Proof (detailed proof in [9]). In a non full-mesh network, some ASes are not directly connected to the announcing AS. Thus, in the calculation of the $D(i|x)$ we need to consider the bgp-eligible neighbors of *all* the ASes that have received the update.

Let assume that we are at step i, and $n(i)$ nodes have received the BGP updates. An AS without the update is *not* a bgp-eligible neighbor if it not connected with any of the nodes with the BGP update; this happens with probability $(1-p)^{n(i)}$. Hence, an AS is a bgp-eligible neighbor w.p. $1-(1-p)^{n(i)}$. Since, there are $N-n(i)$ ASes without the update, the expected number of bgp-eligible neighbors is $(N-n(i))\cdot(1-(1-p)^{n(i)})$. □

3.2.2 Arbitrary Degree Sequence Random Graph

The structure of networks where the degrees (i.e., the number of connections) of the ASes are largely heterogeneous, e.g., power-law graphs, can be better described with a Configuration-Model Random Graph (CM-RG) rather than a Poisson graph. In the CM-RG model, a random graph is created by connecting randomly the nodes (i.e., ASes), whose degrees are given. Hence, we can use the CM-RG to model a network with *any arbitrary degree sequence* with mean value μ_d and variance σ_d^2 (and, $CV_d = \frac{\sigma_d}{\mu_d}$).

If the participation of an AS in the SDN cluster is independent of its degree, then the probabilities $P_{sdn}(x)$ are given by Theorem 2, and the bgp-degrees $D(i|x)$ are given by the following Result[3].

Result 1. *The expectation of the* `bgp-degree` $D(i|x)$, $i \in [1, N-k]$, $x \in [0, N-k]$, *in a CM-RG network topology is given by*

$$E[D(i|x)] = D(1|x)\cdot\prod_{j=1}^{i-1} A(j|x) + \sum_{j=1}^{i-1}(\mu_d(j|x)-1)\cdot\prod_{m=j+1}^{i-1} A(m|x)$$

where

$$D(1|x) = \begin{cases} \mu_d & , x > 0 \\ (N-k)\cdot\mu_d\cdot\ln\left(\frac{N}{N-k}\right) & , x = 0 \end{cases} \quad (8)$$

$$\mu_d(j|x) = \mu_d \cdot \prod_{m=1}^{j-1}\left(1 - \frac{CV_d^2}{N-n(m|x)-1}\right) \quad (9)$$

$$A(j|x) = 1 - \frac{\mu_d(j|x)}{N-n(j|x)-1} \quad (10)$$

Proof. The detailed proof is given in [9] □

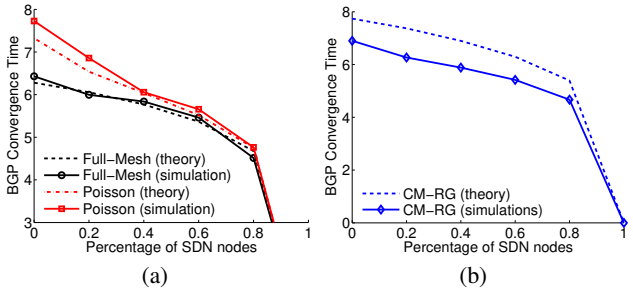
4. VALIDATION AND DISCUSSION

We built a simulator, conforming to our model (i.e., Assumption 1, $T_{sdn}=0$, etc.). In Fig. 2 we compare our theoretical results against simulations (averages over 200 runs). The accuracy is high for the Full-mesh and Poisson graph cases (Fig. 2(a)). In the CM-RG case (Fig. 2(b)) the two curves are similar, and the error of our expression (which is an approximation) is always less than 18%.

Moreover, an initial comparison with the emulation results of [3], where a *real BGP software router* is employed, shows that our theory (despite the assumptions and model simplicity) is in agreement with their observations: e.g., the convergence time (i) decreases faster after $\frac{k}{N} > 50\%$, (ii) has small differences among different topologies when centralization exists, etc.

This highlights one of the contributions of our work: with our results, we can quickly evaluate the effects of routing centralization. The provided expressions are simple (need only to know a few parameters: N and k, and -if needed- p, or μ_d and CV_d) and easy/fast to compute, whereas emulations are time/resource demanding and

[3]We use the notation "Result", instead of "Theorem", because the provided expression is an approximation.



(a) (b)

Figure 2: BGP convergence time vs. SDN penetration (i.e, $\frac{k}{N}$) in networks with $N = 300$. (a) Full-mesh and Poisson graph ($p = \frac{1}{60}$) topologies; and (b) CM-RG topology with degree sequence d_i, $d_i \in [5, 200]$ and power-law distributed with exponent 2.

have limited scalability. Hence, one could use our work to obtain initial insights, and then, e.g., proceed to a few targeted emulations for a more fine-grained or system-specific investigation.

A second contribution is that our methodology and results can be used as the building blocks to model and analyse more complex settings; i.e., model different parts of a large network as full-mesh/Poisson/CM-RG sub-networks, use our expressions for each sub-network, and synthesize them to compute the overall network performance. We consider such an example in [9], where we model the core of the Internet, considering different classes of tier-1 and tier-2 ISPs, while taking into account their routing policies as well.

Finally, we believe that this first analytic study, offers a new useful approach for investigating inter-domain SDN, and can be the basis for analysing further aspects of this field. In particular, we plan to extend our methodology, and consider more settings (e.g., BGP update times), performance metrics, and applications.

Acknowledgements. This work has been funded by the European Research Council Grant Agreement no. 338402.

5. REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] M. Caesar, D. Caldwell, N. Feamster, J. Rexford, A. Shaikh, and J. van der Merwe, "Design and implementation of a routing control platform," in *Proc. NSDI*, 2005.

[6] 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.

[7] 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.

[8] 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.

[9] P. Sermpezis and X. Dimitropoulos, " Analysing the Effects of Routing Centralization on BGP Convergence Time." `http://arxiv.org/abs/1605.08864`, 2016.