2017 IEEE 37th International Conference on Distributed Computing Systems

On the Feasibility of Inter-domain Routing via a Small Broker Set

Dong Lin*, David Hui*, Weijie Wu*, Tingwei Liu†, Yating Yang†, Yi Wang*, John C.S. Lui†, Gong Zhang*, Yingtao Li* *Huawei Technology Co., †The Chinese University of Hong Kong
Email: *{lin.dong, huis.david, wuweijie2, yang.yating, wangyi18, nicholas.zhang, liyingtao}@huawei.com, †tingweiliu2013@gmail.com, cslui@cuhk.edu.hk

***Abstract*—The current inter-domain routing protocol, namely, the Border Gateway Protocol (BGP), cannot provide end-to-end (E2E) quality-of-service (QoS) guarantees. The main reason is that an autonomous system (AS) can only receive guarantees from its first hop ASes via service level agreements (SLAs). But beyond the first hop, QoS along the path from source to destination AS is not within the source AS's control regime. In this paper, we investigate the feasibility of providing high QoS-guaranteed E2E transit services by utilizing a (small) set of ASes/IXPs to serve as "*brokers*" to provide supervision, control and resource negotiation. Finding an optimal set of ASes as brokers can be formulated as a Maximum Coverage with $B-$dominating path Guarantee (MCBG) problem, which we prove to be NP-hard. To address this problem, we design a $(\frac{1-e^{-1}}{4})-$approximation algorithm and also an efficient heuristic algorithm when considering additional constraints (e.g., path length). Based on the current Internet topology, we discover a "3540-alliance" subset (accounting only 6.8%) of 52,079 ASes/IXPs, which can provide high QoS guarantees for 99.29% E2E connections.**

I. Introduction

E2E QoS guarantees, which impose a stringent inter-AS QoS support, are becoming more and more important with the explosion of Internet video traffic. By 2020, global IP traffic will reach 1.3 ZB per year, in which 82% is IP video traffic [1], and E2E QoS guarantees for such applications are urgently needed. However, E2E QoS cannot be guaranteed by the current inter-domain routing protocol, namely, the Border Gateway Protocol (BGP). The main reason is that an AS can only receive guarantees from its first hop ASes via service level agreements (SLAs). But beyond the first hop, QoS guarantee along the path from a source AS to a destination AS is beyond the source AS's control regime. From our collected data of 52,079 ASes or Internet eXchange Points (IXPs), it reveals that more than 90% of E2E AS connections are more than one-hop.

To address this issue, content providers typically use content delivery networks (CDNs) to distribute their contents around the world so that most requests can be served by nearby copies stored by CDNs. However, the CDN technology is not effective for realtime and delay sensitive services, such as VoIP and video conferencing, because for most of these applications the E2E AS hop count is usually larger than one and unfortunately, there is no inter-AS QoS support in the current Internet.

For decades, an increasing number of proposals coming from diverse angles advocate inter-domain routing mediator [2]–[5] as an approach to enable ISPs to cooperate and provide E2E guarantees. In these schemes, QoS enabled pathlets (i.e., fragments of paths represented as sequences of a virtual node [6]) provided by ISPs are stitched together by an inter-domain routing mediator (e.g., a bandwidth broker [2] or IXPs [5]) to construct global paths. These initiatives are currently being explored in the industry and also in standardization bodies (e.g., in the context of the Path Computation Element (PCE) architecture [7]).

Pushing the idea of "stitching pathlet" a step further, we consider how the Internet could be improved from the perspective of "centralized inter-domain routing brokers." We show that a "small broker set" can be utilized to stitch each inter-AS hop along the AS routing path, centralize routing control for mission-critical traffic across domains, working in parallel to BGP. Such broker set is formed by a small subset of ASes or IXPs, which are selected to serve as inter-AS routing brokerage agencies so as to take up the responsibilities of network performance measurement, control, resource negotiation, as well as providing transit services. When every hop of AS-path can be covered by the broker set (i.e., for every AS hop, at least one of its source or destination belongs to the broker set), this AS path is said to be *dominated* by the broker set. Note that in this paper, we will not focus on how exactly the E2E QoS will be guaranteed by constructing the brokers set, but we assume that the broker set's monitoring and controlling of the (almost) whole Internet can provide a possible way to achieve it.

The technical challenge is how to efficiently find such a broker set that provides dominating paths for most inter-AS connections in the current Internet. To address such challenge, we have to consider the following issues:

- Which AS/IXP should be in the broker set, which we denote as B?
- How small can B be such that it can provide $B-$dominating paths for most, if not all, inter-AS connections?
- Is there any economic incentive to form and maintain such kind of broker set?

Contributions: In this paper, we introduce an inter-AS routing framework where a broker set is selected to improve the ASes' E2E QoS by dominating the associated AS paths. We model the broker set selection problem as the Maximum Coverage with $B-$dominating path Guarantee

1063-6927/17 $31.00 © 2017 IEEE
DOI 10.1109/ICDCS.2017.18

IEEE computer society

(MCBG) problem and prove its NP-hardness. We further propose an approximation algorithm which can provide at least $(\frac{1-e^{-1}}{4})-$guarantee as compared to the best E2E connectivity with the dominating AS paths, given the size constraint of B. We also propose a heuristic algorithm when additional factors (e.g., path length constraints) are considered. The algorithm is not only computationally efficient, but also offers a minimal reduction of the QoS guarantees of no more than 0.5% E2E AS connections and applies well in dynamic scenario where AS-level Internet evolves continuously. By studying our collected data from 52,079 ASes/IXPs, we demonstrate that it is indeed feasible to provide QoS guarantees for 99.29% E2E AS connections with only around 6.8% ASes and IXPs serving as brokers. The broker set size can be further reduced if we focus on providing QoS guarantees for the majority E2E AS connections: 1,000 brokers for 85.41% saturated connectivity and 100 brokers for 53.14% saturated connectivity. We also provide an economic model in the technical report [8] as a possible way to incentivize ASes/IXPs to form and maintain such brokerage coalition.

The rest of the paper is organized as follows. In Section II, we present the related work. In Section III, we describe our collected datasets. In Section IV, we present our problem statement and the approximation algorithm. In Section V, we further consider two practical issues, i.e., lower computational complexity and the path length constraint, and propose an efficient heuristic algorithm. In Section VI, we present experimental results and findings to demonstrate the feasibility of an inter-domain routing using a small broker set. Finally, Section VII concludes.

II. Related Work

Previous proposals which attempt to provide E2E QoS can be classified in the following three categories:

Computing QoS-constrained path. To provide E2E QoS, many works focus on finding paths satisfying the QoS constraints. In [9], authors consider one pre-determined route and construct a subgraph containing the known route and all its neighbors. It uses existing routing algorithms to find a QoS-constrained path by assuming the graph link weights are known. Authors in [10] propose a distributed solution for calculating a QoS-constrained path over multiple pre-determined routes. In these two works, the performances depend on the predetermined routes, and the assumption that the link weights are known is not realistic.

Stitching QoS-enable pathlets. Authors in [2] utilize bandwidth mediators for mediating the concatenation of multiple guaranteed bandwidth pathlets. Authors in [3] propose outsourcing routing control to inter-domain SDN controllers. Such controllers can deal with E2E pathlet stitching using their bird's eye view over the participating domains. Recently, authors in [5] propose to use central control points (CXPs) to stitch the QoS enabled pathlets provided by ISPs to construct global paths. A CXP is external to an ISP entity and it applies centralized inter-domain control over the fractions of Internet traffic are routed. These schemes seriously increase the burden of CXPs since they need to exchange Internet traffic for ASes and also calculate optimal QoS E2E paths for all the routing requests.

Economic method. Authors in [11] seek to develop an economic plane solution for E2E QoS. They introduce marketplaces that providers and users can meet and supply the minimum necessary semantics for them to exchange information. Yet, the actual holder of a marketplace is still unclear, and providers do not often have the detailed quality information of different AS routes for all end users. Authors in [12] introduce "route bazaars", a contractual system where ASes and end users agree on QoS-aware routes in the absence of preexisting trust relations between the networks. These works are still essentially based on stitching QoS-enable pathlets.

Finally, all works listed above encounter the scalability problem when facing a large scale network such as the Internet. In the dataset we collected, there are tens of thousands of ASes. In this work, we aim to improve the E2E QoS by finding a small set of ASes to stitch each two AS hops along the AS routing path to provide supervision, control and resource negotiation for each AS hops. One attractive property of our approach is that, considering all connections are bidirectional, we can achieve QoS guarantees for 99.29% E2E connectivity with only 6.8% ASes/IXPs as brokers; considering the directional connections, we just need a minor change to the current AS peering relationships to serve 72.5% E2E connectivity with quality assurance with only 2% ASes/IXPs as brokers.

III. Topology and Datasets

Since the Internet topology heavily influences how one should model and design an effective inter-domain routing strategy with E2E QoS support, here, we first present our collected data, which includes AS sources and their routes, and then we describe our data processing method.

In this study, we consider the AS-level topology. The AS-level Internet ecosystem can be considered as a logical fabric of the Internet. AS-level topology, which is composed of different ASes and their interconnections, has been widely used to characterize the Internet traffic. There are basically two mechanisms to connect ASes. One is via dedicated links, which relies on the business agreement between two ASes, e.g., provider-to-customer peering or P2P peering. Another is to make use of a physical interconnection infrastructure called IXP, which provides efficient and cost effective means for traffic exchange between ASes. We collected data for the AS topology as well as connections to IXPs, and built a network topology to cover both direct and IXP-based connections.

Currently, there are some excellent public Internet AS-level topology datasets. Here we adopt the dataset from [13], which offers the most comprehensive and long-term data. The AS topology is constructed using BGP data of IPv4 collected by Route views, RIPE RIS, PCH and Internet2 [13]. The data are stored on a monthly basis. To make a complete AS-level topology, we use the data of the whole year for 2014. In addition to the traditional AS topology, we also manage to discover those AS connections via IXPs. We obtained the data of IXPs membership and IXPs peering in 2014 using similar approaches described in [14].

It is important to point out that it is inevitable to have an incomplete AS topology. This is due to the limited scope of the BGP data collection method, e.g., some interconnections between ASes may not be discovered. Also, some short-life connections may be falsely presented, originating from unintentional misconfigurations or intentional trials [13]. For the IXP data, there are around 400 IXPs which are providing global traffic switching services in 2014, and we were able to collect around 80% (or 322) of these IXPs based on targeted traceroute and targeted source routing techniques. Note that the large numbers of ASes, IXPs and connections, as illustrated in Table I, show that our dataset is indeed representative.

TABLE I
SUMMARY ON THE COLLECTED DATASET

Description	Numbers
IXPs	322
ASes	51757
Size of the maximum connected sub graph	51,895
# of Connections	347,332
# of Connections among ASes	292,050
# of Connections between IXPs and ASes	55,282

Similar to [5], we also assume that IXPs are independent entities. This is proposed due to the rich connectivity of IXPs and the huge amounts of traffic exchanged at IXPs [5]. This assumption assigns a new role to IXPs which typically provide switching service only instead of routing. Our experiments also show that IXPs play a critical role in the broker set.

IV. Problems and Algorithms: Theoretical Basis

In this section, we formulate our inter-domain routing brokerage problem and develop some theoretical foundations. First, we formulate it as a Maximum Coverage with $B-$dominating path Guarantee (MCBG) problem and analyze its complexity; then we propose an approximation algorithm to solve the MCBG problem.

A. Problem statement

Let $G=(V,E)$ denote an undirected graph consisting of the vertex set V and edge set E. For each vertex $v \in V$, we define the neighborhood $N(v)$ as the set of all vertices in V that are adjacent to v. Similarly, define $N(V')$ as the set of all vertices in V which are adjacent to $\forall v \in V'$, i.e., $N(V')=\cup_{v\in V'} N(v)$. We first define the notion of a "$B-$*dominating path*".

Definition *1: Given a graph $G=(V,E)$, a routing path is called a "$B-$dominating path" if for every hop along the path, at least one of its source or destination vertex belongs to the set B, where $B \subseteq V$.*

In the context of inter-domain routing brokerage, we treat the AS-level topology we mentioned in Section III as the input graph G, where vertex set V is the set of ASes/IXPs, a connection between AS/IXP u and AS/IXP v is represented by an edge (u,v) in G. If we can find a small set $B \subseteq V$ such that for every source-destination AS pair, we can find a $B-$dominating path, then the E2E network performance can be maintained and managed. Furthermore, since B is small, it is easier to create economic incentives to form B such that the E2E QoS can be greatly improved. To this end, we aim to find a broker set B such that $\forall v, v' \in V$, there exists at least one $B-$dominating path between them.

Mathematically, the inter-domain routing brokerage problem can be formulated as a path-dominating set (PDS) problem.

Path-Dominating Set (PDS) Problem: Given an input graph $G=(V,E)$ and an integer $k \geq 1$, determine whether it is feasible to find a set $B \subseteq V$ such that

- $|B| \leq k$, and
- there exists at least one $B-$dominating path between u and v, for $\forall u, v \in V$.

Sometimes, it may not be possible to find a solution to the PDS problem to provide *all* connections with the $B-$dominating path guarantees. Nevertheless, we still want to find a small broker set B so as to provide as many connections with $B-$dominating path guarantees as possible. To this end, we formulate the optimization version for the inter-domain routing brokerage problem in Problem 1.

Problem 1 **M**aximum **C**overage with **B**-dominating path **G**uarantee (MCBG) problem

Input: A connected non-trivial graph $G=(V,E)$, and a positive integer k.
Output: A subset $B \subseteq V$ which guarantees:
1) $|B| \leq k$;
2) for $\forall u, v \in B \cup N(B)$, there exists at least one $B-$dominating path between u and v;
3) $f(B)=|B \cup N(B)|$ is maximized.

Note that for $\forall u, v \in B \cup N(B)$, if there exists a path containing only nodes in $B \cup N(B)$, then the path must be dominated by B. Thus the coverage function f can help to evaluate the satisfiability of the E2E connectivity with $B-$dominating path guarantees. Now, let us state our first result.

Theorem *1: If there exists a solution to the PDS problem, then it is also the solution to the MCBG problem. If there is no solution to the PDS problem, then the solution to the MCBG problem can provide dominating path guarantees to the largest possible source-destination pairs.*

Proof: If there is a solution to the PDS problem, then we denote it as B, which satisfies $|B| \leq k$ and can provide $B-$dominating path guarantee for $\forall u, v \in V$. Thus both u and v must connect to at least one broker, i.e., $B \cup N(B)=V$. Therefore, B is the solution to the MCBG problem. If there is no solution to the PDS problem, then denote the solution to the MCBG problem as B. If there is a set B' which satisfies $|B'| \leq k$ and can provide $B'-$dominating path guarantee for $\forall u, v \in B' \cup V'$ and $|B' \cup V'| > |B \cup N(B)|$. To satisfy the $B'-$dominating path constraint, any vertex in V' must connect to at least one broker in B', i.e., $V' \subseteq B' \cup N(B')$. As $|B' \cup N(B')| \geq |B' \cup V'| > |B \cup N(B)|$, B is not the solution of problem 1. Therefore, there doesn't exist such a set B'. Thus B can provide $B-$dominating path guarantees for as many connections as possible. ∎

B. Computational complexity

Let us now quantify the computational complexity of the PDS problem.

Lemma *1: The PDS problem is NP-complete.*

Proof: One can prove this by reducing the vertex cover problem to the PDS problem in polynomial time. Due to page limit, we leave the detailed proof in the technical report [8]. ∎

To analyze the computational complexity of the MCBG problem 1, we first consider its decision version.

Lemma *2: The decision version of the MCBG problem is NP-complete.*

Proof: One can prove this by reducing the PDS problem to the decision version of MCBG problem in polynomial-time. We leave the detailed definition of the decision version of MCBG problem and the proof in the technical report [8]. ∎

Theorem *2: Problem 1, the MCBG problem, is NP-hard.*

Proof: Since its decision version is NP-complete, the MCBG problem is NP-hard. ∎

C. Approximation algorithm for MCBG

Given that the MCBG problem is NP-hard, we propose an approximation algorithm to solve the MCBG problem. The high level idea is that, we divide the broker set B into two parts: B^*, pre-selected for approximating the optimal coverage, and B', added for guaranteeing the $B-$dominating path constraint.

To find B^*, we define the *Maximum Coverage with broker set B* (MCB) problem, and then present its approximation algorithm Alg. 1. The selection of B^* is realized by Alg. 1.

Problem 2 **M**aximum **C**overage with broker set **B** (MCB) problem

Input: A connected non-trivial graph $G = (V,E)$ and a positive integer k.
Output: A subset $B \subseteq V$ which guarantees:
1) $|B| \leq k$;
2) $f(B) = |B \cup N(B)|$ is maximized.

For convenience, let $MCB(V,k)$ and $MCBG(V,k)$ denote an instance of MCB and MCBG problem respectively. We can use the following approximation algorithm to solve $MCB(V,k)$, or in other words, to find B^*.

Algorithm 1 Approximation algorithm for $MCB(V,k)$ [15]

Input: The vertex set V and an integer k
Output: A set B which satisfies $|B| \leq k$
1: Start with $B_0 = \emptyset$;
2: **for** $i = 1$ to k **do**
3: $s_i \leftarrow \arg\max_s f(B_{i-1} \cup \{s\}) - f(B_{i-1})$;
4: $B_i \leftarrow B_{i-1} \cup \{s_i\}$;
5: **end for**
6: Return $B = B_k$.

Now the remaining issue is to find B'. To achieve this, we take advantage of the special property of our graph. Note that for the AS-level Internet graph we study, it has a special characteristic, that is, more than 99.2% of the source and destination pairs' hop count distances are within four hops. This special characteristic helps us to design efficient brokerage algorithm. Let us first formally define this characteristic.

Definition *2: A graph $G=(V,E)$ is called an $(\alpha,\beta)-$ graph if the following condition is satisfied:*

$$Prob[d(u,v) \leq \beta] \geq \alpha \qquad \forall u,v \in V,$$

where $d(u,v)$ is the shortest hop distance between node u and v, β is an integer which is much smaller than the diameter of G, and $\alpha \in [0.5,1]$.

For example, the AS-level graph we have is a $(0.99,4)-$graph. Note that the property of an $(\alpha,\beta)-$graph can help us to decide the size of B' to satisfy the $B-$dominating path constraint. The details about how to solve the MCBG problem, including finding B^* and B', are shown in the following approximation algorithm Alg. 2.

Algorithm 2 Approximation algorithm for $MCBG(V,k)$ on an $(\alpha,\beta)-$graph G

Input: The vertex set V and an integer k
Output: A set B which satisfies $|B| \leq k$ and guarantees at least one $B-$dominating path between $\forall u,v \in B \cup N(B)$
1: B^* = the solution returned by applying Alg. 1 to $MCB(V,x^*)$ and $B' = V - B^*$, where $x^* = \left\lfloor \frac{k-1}{\lceil \frac{\beta}{2} \rceil} + 1 \right\rfloor$;
2: **for all** $r \in B^*$ **do**
3: $B'_r = \emptyset$;
4: **for all** $v \in B^* - \{r\}$ **do**
5: Find the shortest path from v to r on $G(V,E)$;
6: Add at most $\lceil \frac{\beta}{2} \rceil - 1$ members along the path to B'_r to guarantee this path is a $(B^* \cup B'_r)-$dominating path (every two adjacent brokers are one-hop neighbor or 2 hop neighbor connected by a non-broker);
7: **end for**
8: **if** $|B'_r| < |B'|$ **then**
9: $B' = B'_r$;
10: **end if**
11: **end for**
12: Return $B = B^* \cup B'$.

In Alg. 2, the computational complexities for the selection of B^* and B' are $O(k(|V|+|E|))$ and $O(k^2(|V|\log|V|+|E|))$, when adopting the Fibonacci heap implementation of the Dijkstra's algorithm for calculating the shortest path in line 5 of Alg. 2, respectively.

Now we can prove how Alg. 2 can achieve an approximation with the pre-selected B^*. Let us first present the following lemmas to aid the proof.

Lemma *3: The coverage function f is a submodular and nondecreasing set function [16].*

Lemma *4: Alg. 1 provides $(1-e^{-1})-$approximation for the Maximum Coverage with broker set B (MCB) problem [15].*

Lemma *5: A tree with k vertices can be divided into no more than $p = \left\lfloor \frac{2(k-1)}{m} \right\rfloor + 1$ subtrees in which each subtree has no more than m vertices.*

Proof: We design a tree partition algorithm to realize this partitioning. Please refer to the technical report [8] for the detailed proof and the tree partition algorithm. ∎

Now we are in the position to state the following theorem.

Theorem *3: When a graph is an $(\alpha,\beta)-$graph, we can obtain an approximation algorithm for the MCBG problem with the approximation ratio of $\frac{1-e^{-1}}{\theta}$ where:*

$$\theta = p = \left\lfloor 2\left\lceil \frac{\beta}{2} \right\rceil \right\rfloor = \begin{cases} \beta, \beta \text{ is even;} \\ \beta+1, \beta \text{ is odd.} \end{cases} \quad (1)$$

Proof: Let OPT_{x^*} and OPT_k denote the optimal solutions for $MCBG(V,x^*)$ and $MCBG(V,k)$ respectively, and let B denote the solution to $MCBG(V,k)$ obtained through Alg. 2.

We are trying to prove that:

$$\left(\frac{\theta e}{e-1}\right) f(B) \geq \theta f(OPT_{x^*}) \geq f(OPT_k). \quad (2)$$

First we will prove $\theta f(OPT_{x^*}) \geq f(OPT_k)$.

As $\forall u, w \in OPT_k \cup N(OPT_k)$, there exists an OPT_k-dominating path between u, w. Now redefine the connectivity in OPT_k: two vertices in the broker set are considered to be connected, if they are one-hop neighbors or they are connected by a non-broker vertex,. Thus we have a newly defined connected graph of OPT_k. Since every connected graph has a spanning tree, we can construct a tree with size k according to the connected graph of OPT_k.

Based on lemma 5, the tree T constructed from the connected graph of OPT_k can be divided into $p=\left\lfloor \frac{2(k-1)}{m} \right\rfloor+1$ subtrees with sizes no more than m. Here $m=x^*=\left\lfloor \frac{k-1}{\lceil \frac{\beta}{2} \rceil}+1 \right\rfloor > \frac{k-1}{\lceil \frac{\beta}{2} \rceil}$ such that after the operation in line 6 of Alg. 2, the broker set size will not exceed the size constraint k, and $p=\left\lfloor 2\left\lceil \frac{\beta}{2} \right\rceil \right\rfloor$.

Denote those p subtrees as $T_i, 1 \leq i \leq p$ and the vertices in T_i as N_i. Based on the property of f in lemma 3, we have:

$$\begin{aligned} & f(OPT_k) = f\left(\bigcup_{i=1}^{p} N_i\right) \leq \bigcup_{i=1}^{p} f(N_i) \\ & \leq p f(OPT_m) = \theta f(OPT_{x^*}) \end{aligned} \quad (3)$$

Next, we will prove $\left(\frac{e}{e-1}\right) f(B) \geq f(OPT_{x^*})$.

Denote the optimal solution of $MCB(V, x^*)$ as OPT'. We have $f(OPT') \geq f(OPT_{x^*})$ for OPT' does not have the $OPT'-$dominating path constraint. From lemma 4, we have $\left(\frac{e}{e-1}\right) f(B) \geq f(OPT') \geq f(OPT_{x^*})$. ∎

As for the AS-level Internet topology, 99.2% E2E connections are within four hops.

Corollary *1: Given that our AS-level topology is a* $(0.99, 4)-$*graph, Alg. 2 is a* $\frac{1-e^{-1}}{4}-$*approximation algorithm for the MCBG problem.*

V. Problems and Algorithms: Practical Considerations

The previous section provides the theoretical foundation of the broker set selection problem. To address the needs of the inter-domain E2E QoS guarantee, we have to consider several engineering and practical issues. First, to further improve the computation efficiency of the approximation algorithm, we propose a heuristic algorithm with lower computational complexity while maintaining a good $B-$dominating path coverage with broker set B. Second, we consider a more general version of the MCBG problem in practice by taking the path length constraint into consideration.

A. Efficient heuristic algorithm and baseline algorithms

The MaxSubGraph-Greedy algorithm, as depicted in Alg. 3, is a pseudocode of an effective algorithm for broker set selection. It has a computational complexity of $O(k(|V|+|E|))$ while maintaining a good $B-$dominating path coverage with broker set B.

Note that Alg. 3 aims to maximize the connected graph size in each iteration. As we will show, our experiment results indicate that Alg. 3 is capable of finding a broker set with a very high coverage in only few thousand iterations. Furthermore, this algorithm also applies well in dynamic scenarios where AS-level Internet continuously evolves as new ASes are born, new connections are established, new IXPs are formed, and ASes peered at new IXPs. For a newly added AS or IXP $i \in N(B)$, there is no need to add new broker. When the number of newly added uncovered ASes and IXPs $\{i | i \notin N(B)\}$ increases to a certain threshold, to guarantee the E2E connectivity with QoS guarantee, we can use Alg. 3 to add a new brokers that maximize the connected graph size in each iteration.

Algorithm 3 MaxSubGraph-Greedy

Input: A connected non-trivial graph $G=(V,E)$ and a positive integer k
Output: A set B which satisfies $|B| \leq k$
1: Select a vertex $v \in V$, and let $B=\{v\}$;
2: If $|B|=k$ or $V-(B \cup N(B))=\emptyset$, then Stop;
3: Select a vertex $w \in V-B$, and assign $B \leftarrow B \cup \{w\}$ if the size of maximum sub graph in $B \cup \{w\}$ is maximized. Go to Step 2.

We want to emphasize that, an AS-path just contains members in $B \cup N(B)$ does not mean that it is dominated by the broker set. The latter one requires for every AS hop, at least one of its source or destination belongs to B. A special case is that, an AS-path containing only members in $N(B) \setminus B$ is not dominated by B. Alg. 2 can output a broker set B_2 that every AS hop along the AS-path is dominated by the broker set, but the outputted broker set B_3 of Alg. 3 has no such guarantee. However the following experiment results show that Alg. 3 can achieve an E2E connectivity (i.e., percentage of AS-paths dominated by broker set) which is as good as Alg. 2.

To evaluate the performance gain of our proposed algorithm, we compare it with four baseline algorithms, whose detailed pseudocodes are listed in the technical report [8]. The **Set Cover** (SC) algorithm is proposed in [17] to find some but not necessarily the smallest dominating sets. We compare with this algorithm to help us to gain some understanding on the importance of a broker set selection process. The **IXP-Based** (IXPB) algorithm returns a set of IXPs with a degree higher than a given threshold. Since IXP is often treated as an ideal node for inter-domain control [5], it is important for us to understand the influence of an IXP if it is used as a broker. The **Degree-Based** (DB) and **PageRank-Based** (PRB) algorithms are greedy algorithms widely used in identifying important vertices in a graph. In each round, the node with the largest degree or page rank value will be added to the broker set.

B. Path length constraint and its evaluation method

Note that in the MCBG problem, for each source and destination pair (u, v), the $B-$dominating path can be of an arbitrary length. For some ISPs, they may want to restrict the number of AS hop counts on the E2E path, e.g., the number of AS hop counts should follow some ISPs' specified distribution. Therefore, during the search of the broker set B, we introduce an extra requirement on the path length l_{uv}. As a result, we further refine the MCBG problem, and we call it MCBG with Path Length Constraint, which can be stated as follows.

For the AS hop-count to follow a specified probability distribution, one can use the following probability interpretation. If the choice of a source u and destination v pair can be viewed

Problem 3 MCBG problem with Path Length Constraint

Input: A connected non-trivial graph $G=(V,E)$, a positive integer k and positive integers l_{uv}, representing the path length parameter for any pair $u, v \in V$ $(u \neq v)$.
Output: A subset $B \subseteq V$ which guarantees:
1) $|B| \leq k$;
2) for $\forall u, v \in B \cup N(B)$, there exists at least one $B-$dominating path of length l_{uv} between u and v;
3) $f(B) = |B \cup N(B)|$ is maximized.

as a random event with every possible source-destination pair as the sample space, then the corresponding path length l_{uv} can be viewed as a random variable l. If every event of the sample space is equally probable, then the normalized histogram of l_{uv} would give the probability mass function of the random variable l and hence the distribution function of l can also be deduced.

Here, we say a selection strategy is feasible if it gives a candidate solution that satisfies the specified path length distribution up to ϵ fraction of error for each value of l. Mathematically, for a given distribution based path characterization $F(l)$ (where the distribution $F(l)$ is written in terms of the cumulative path-length histogram, i.e., the number of admissible path with path-length less than or equal to l), an algorithm A of producing a broker set $B_{\mathcal{A}}$ is called feasible if the produced set $B_{\mathcal{A}}$ by the algorithm $\mathcal{A}$ gives a distribution $F_{B_{\mathcal{A}}}(l)$ which deviates from the required $F(l)$ by at most ϵ, for all values of l, i.e.,

$$|F_{B_{\mathcal{A}}}(l) - F(l)| \leq \epsilon, \quad \forall l. \tag{4}$$

Using Equation (4), to verify the feasibility of a candidate algorithm $\mathcal{A}$, we need to be able to compute the cumulative distribution $F_{B_{\mathcal{A}}}(l_{uv})$ for a broker set $B_{\mathcal{A}}$ produced by the candidate algorithm $\mathcal{A}$. Hence, let us define a $B_{\mathcal{A}}$ operator on the adjacency matrix A of a graph G as $B_{\mathcal{A}} \cdot A$. This operation will erase an entry of A whenever neither of its row nor column indices belongs to $B_{\mathcal{A}}$. Let us call the output of $B_{\mathcal{A}} \cdot A$ as $\tilde{A}$. This matrix can give the desired cumulative $B_{\mathcal{A}}-$dominating path length distribution $F_{B_{\mathcal{A}}}(l)$ in the follow manner: the number of nonzero entries in $\tilde{A}^l$ gives the number of $B_{\mathcal{A}}-$dominating paths with length less than or equal to l. Therefore, we call this as the "$l-$hop E2E connectivity."

Tab. II depicts the $l-$hop connectivity of different topologies (e.g., *ER-Random, WS-Small-World, BA-Scale-free, ASes with/without IXPs*) under such evaluation metric. Here *ASes with/without IXPs* are the AS level topologies used in this paper with/without considering IXPs as independent entities. The other topologies, *ER-Random, WS-Small-World and BA-Scale-free*, have the same vertex sets (including 52,079 ASes/IXPs) with *ASes with IXPs*, but the edge sets are generated according to the topologies' features accordingly. Note that for *ASes with IXPs*, if we set $l = 4$, we have 99.21% E2E connectivity.

TABLE II
$l-$HOP CONNECTIVITY OF DIFFERENT TOPOLOGIES.

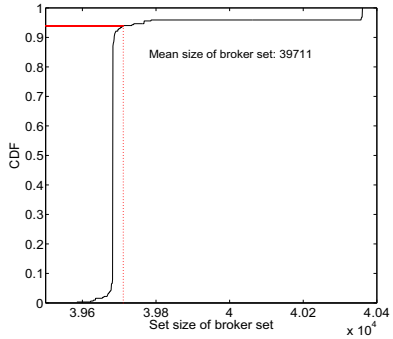
hop count	1	2	3	4	5	6	7
ES	0.37	4.91	47.47	99.30	99.69	99.69	99.69
WS	0.24	2.28	18.76	83.23	99.69	99.69	99.69
BA	1.11	26.17	95.50	99.69	99.69	99.69	99.69
ASes w/ IXPs	10.00	65.74	96.65	99.21	99.29	99.29	99.29
ASes w/o IXPs	5.39	47.98	90.02	97.35	98.00	98.06	98.06

VI. Structural Feasibility and Broker Set's Properties

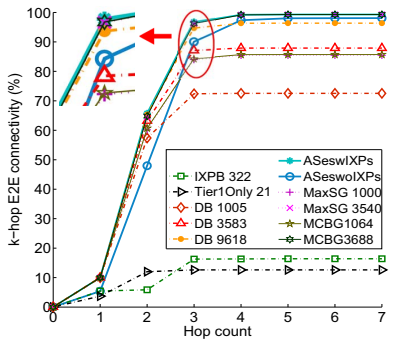
In this section, we describe some experimental results showing the possible composition of a broker set in the current Internet. We also compare our proposed algorithm on broker selection with other state-of-the-art algorithms. Due to page limit, we leave some detailed results in the technical report [8].

A. Evaluation for $l-$hop E2E connectivity

The selection of a broker set is non-trivial. Improper selection may lead to a large size broker set, making it more difficult to incentivize ASes to join the broker set, or with very poor $l-$hop E2E connectivity. Fig. 1(a) illustrates the cumulative distribution function (CDF) of the AS E2E connectivity by running the SC algorithm over 300 iterations. Although 100% E2E connectivity is guaranteed, the SC algorithm has to select around 40,000 nodes in the broker set, accounting for more than 76% of the overall network vertices. Fig. 1(b) illustrates the achieved $l-$hop E2E connectivity by varying hop count requirement l of the other algorithms. The results of IXPB and Tier1Only algorithms, which were considered in previous works, show that it is not appropriate to merely rely on the IXPs or the tier 1 ISPs to act as brokers. Both of them suffer from low E2E connectivity: IXPB algorithm could only reach at most 15.70% E2E connectivity with 322 brokers, and it is far worse for the Tier1Only algorithm.



(a) CDF of broker set size by SC algorithm



(b) $l-$hop E2E connectivity of other algorithms

Fig. 1. Comparison for $l-$hop E2E connectivity of different algorithms.

The DB and PRB algorithms can lead to serious marginal effect: the marginal increase of the $l-$hop E2E connectivity decreases with the increasing broker set size. This can be caused

by the decreasing correlation between degree/PageRank value and saturated E2E connectivity with the increasing broker set size. The broker set selected by the DB algorithm, which consists of high degree ASes and IXPs, can achieve around 72.53% E2E connectivity with 1,005 brokers. However, the DB algorithm requires a large size broker set to guarantee a high (e.g., 99%) E2E connectivity for the serious marginal effect when $|B| > 1,000$: the DB algorithm can only achieve 96.35% E2E connectivity even with 9,618 brokers. By taking a more detailed view of the selected broker set by DB algorithm, we also find that most selected brokers are located at the center network core, leaving network edge mostly uncovered. The PRB algorithm has similar problem, for the PageRank distribution of the undirected graph is statistically close to its degree distribution [18].

Our approximation algorithm for the MCBG problem can achieve 85.71% saturated connectivity with 1,064 brokers and 99.29% saturated connectivity with 3,688 brokers, making it the best algorithms among all we considered. Compared with the approximation algorithm, our MaxSG algorithm achieves equivalent performance (i.e., sacrificing less than 0.5% connectivity) while greatly reduces the computational complexity. Also, MaxSG algorithm outputs a broker set consisted with 3,540 members which totally dominate the maximum connected sub graph of the given Internet topology, i.e., 51,895 out of 52,079 ASes/IXPs, leading to a saturated E2E connectivity as high as 99.29%. Unlike the DB algorithm, our MaxSG algorithm does not have an overcrowded network core and the network outer ring can be well covered.

Remark: Note that the broker set with 3,540 members, which only accounting 6.7% of 52,079 ASes/IXPs, is proposed to achieve 99.29% saturated E2E connectivity. Due to the marginal effect, the broker set's size can be greatly reduced if we mainly focus on the majority part of the E2E AS connections, e.g., 1,000 brokers for 85.41% saturated connectivity and 100 brokers for 53.14% saturated connectivity.

B. Attractive Properties of the 3540-alliance broker set B

We name the broker set with 3,540 brokers output by the MaxSG algorithm as the "3540-alliance", and discuss some of its attractive properties.

Minimal Path Inflations: Path length inflations (i.e., previously l hops reachable pairs now require l' hops, where $l' > l$) are observed in Fig. 1(b). Consider the DB algorithm. With 1,005 brokers, only 72.40% E2E connection can be satisfied within four hops, in contrast to that of 90.02% in a free-path selection scheme (i.e., denoted as "ASeswithIXPs"). As illustrated in Tab. III, if the internal connections inside such broker set are bidirectional (i.e., there exist peering connections), minimal path inflations via this broker set can be achieved (i.e., the E2E connectivity curve of 3540-alliance almost overlaps the one of "ASesWithIXPs").

Diversified Compositions: As illustrated in Fig. 2(a), the 3540-alliance consists of different types of ASes and IXPs. This avoids monopoly by some tier 1 ISPs. Here, we use the same definition and data in [19] to divide the brokers into different categories according to the services they offered. Table IV lists some brokers and their rankings as well, which also illustrates the importance of IXPs for $B-$dominating path routing with broker set B.

TABLE III
THE 3540-ALLIANCE CAN GUARANTEE MINIMAL PATH INFLATIONS.

hop count	1	2	3	4	5	6	7
ASes w/o IXPs	5.39	47.98	90.02	97.35	98.00	98.06	98.06
ASes w/ IXPs	10.00	65.74	96.65	99.21	99.29	99.29	99.29
MaxSG 3540	9.96	64.53	96.09	99.17	99.29	99.29	99.29

90% of the E2E Connections Only Used Nodes in the Broker Set: As illustrated in Fig. 2(a), although for some connections a re-route through non-brokers is still necessary, more than 90% E2E connections can be carried out the 3540-alliance solely without the aid of non-brokers, which implies that the broker set does not need to pay any non-broker node (AS or IXP) to complete the traffic transmission. For the remaining 10% E2E connections, we show how to incorporate non-broker nodes in the technical report [8].

TABLE IV
BROKER LIST

Rank	Type	Name	Rank	Type	Name
1	IXP	Equinix Palo Alto	8	T/A	TWTC
2	T/A	LVLT-3549	9	IXP	Equinix Chicago
3	T/A	COGENT-174	232	C	YAHOO-1
4	IXP	LINX	260	C	ViaWest
5	T/A	ATT-INTERNET4	380	C	Host Virtual, Inc
6	T/A	HURRICANE	438	E	PE Voronov Evgen Sergi
7	IXP	DE-CIX Frankfurt	470	E	Serverius Holding

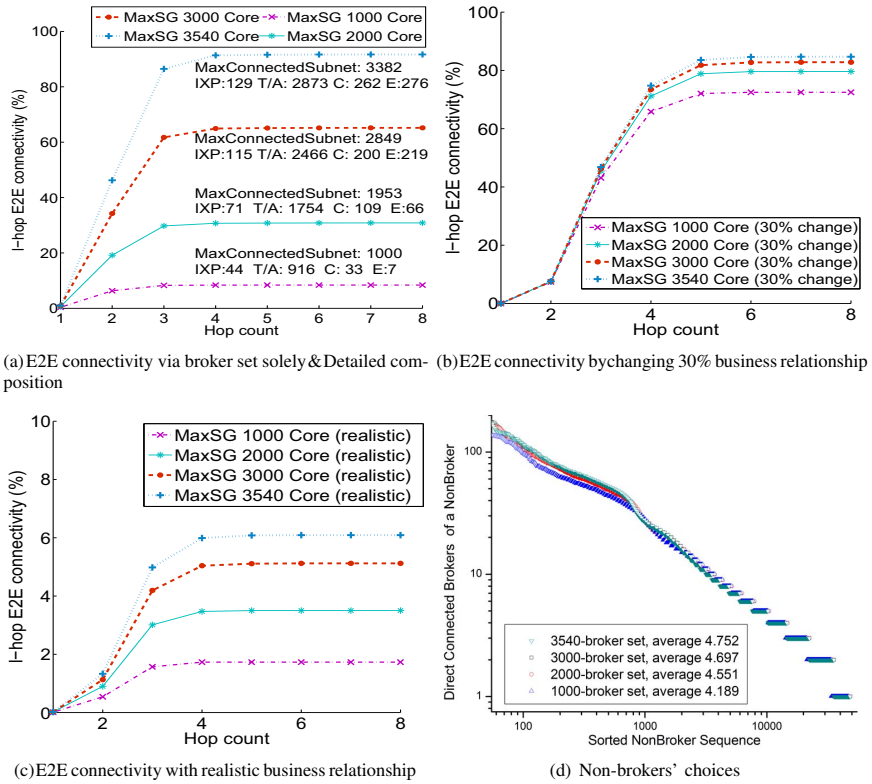
IXP: Internet Exchange Point.
Transit/Access(T/A): ASes which serve as either transit and/or access provider.
Content: ASes which provide content hosting and distribution systems.
Enterprise: Various organizations, universities and companies at the network edge that are mostly users, rather than providers of Internet access, transit or content.

Minimal Changes in the Business Relationships: While for real-life inter-domain routing, the business relationships (e.g. high-tier and low-tier, or peering) among ASes/IXPs has a significant influence and must be taken into consideration. Fig. 2(c) shows the actual performance of a broker set in the current Internet by forcing them to obey existing business relationships probed, i.e., previously assumed bidirectional routing policy becomes directional. A sharply decreased E2E connectivity over different sizes of broker sets has been observed. However, we also notice in Fig. 2(b) that by randomly changing only 30% inter-broker connections to bidirectional (e.g., peering), such degradation can be greatly suppressed. Even a 1000-broker set with 30% random change at its inter-broker connections can achieve a 72.5% E2E connectivity, and the 3540-alliance with 30% random change can achieve 84.68% E2E connectivity.

Potential for Multi-path Routing: As shown in Fig. 2(d), given specific broker sets, we find that a non-broker node typically connects to more than four brokers on average. This means that one can consider the multi-path routing framework in AS information delivery to further improve the E2E QoS.

VII. Conclusion

In this paper, we propose an inter-domain routing brokerage framework, and show that an inter-AS routing path can be totally dominated by a small set of ASes and IXPs to provide E2E QoS guarantee. We model the problem as a MCBG problem and prove it to be NP-hard. To address the MCBG problem, we propose a $(\frac{1-e^{-1}}{4})-$approximation algorithm. To further improve the computation efficiency, we design an



(a) E2E connectivity via broker set solely & Detailed composition

(b) E2E connectivity bychanging 30% business relationship

(c) E2E connectivity with realistic business relationship

(d) Non-brokers' choices

Fig. 2. Findings for the 3540-alliance broker set found by the MaxSG algorithm.

algorithm which, compared with the approximation algorithm, has equivalent good performance while greatly reduces the computational complexity from $O(k^2(|V|\log|V|+|E|))$ to $O(k(|V|+|E|))$ and also applies well in dynamic scenarios where AS-level Internet evolves continuously. We further investigate the feasibility of deploying the broker set in the current Internet from structural perspectives. We also show that with little change to the current AS peering relationships, 72.5% E2E connectivity can be served with high quality assurance by selecting only 2% ASes/IXPs as brokers. The broker set size can be further reduced when the providing QoS guarantees for the majority (e.g., 50%) E2E AS connections.

REFERENCES

[1] "Cisco visual networking index: Forecast & methodology, 2015-2020," http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/complete-white-paper-c11-481360.html.

[2] V. Valancius, N. Feamster *et al.*, "Mint: A market for internet transit," in *Proceedings of the 2008 ACM CoNEXT Conference*. ACM, 2008, p. 70.

[3] V. Kotronis, X. Dimitropoulos, and B. Ager, "Outsourcing the routing control logic: better internet routing based on sdn principles," in *Proceedings of the 11th ACM Workshop on Hot Topics in Networks*. ACM, 2012, pp. 55–60.

[4] V. K. X. Dimitropoulos, R. Kloti *et al.*, "Control exchange points: Providing qos-enabled end-to-end services via sdn-based inter-domain routing orchestration," *Open Networking Summit*, Mar. 2014.

[5] V. Kotronis, M. Rost *et al.*, "Stitching inter-domain paths over ixps," *ACM SOSR*, 2016.

[6] P. Godfrey, I. Ganichev *et al.*, "Pathlet routing," *Proceedings of ACM SIGCOMM*, Oct. 2009.

[7] J. Vasseur and J. Le Roux, "Path computation element (pce) communication protocol (pcep)," 2009.

[8] "Technical report," https://ia601506.us.archive.org/11/items/technical_report/technical_report.pdf.

[9] R. Jacquet, G. Texier, and A. Blanc, "Sanp: An algorithm for selecting end-to-end paths with qos guarantees," in *Proceedings of IEEE Future Network and Mobile Summit*, Jul. 2013.

[10] N. B. Djarallah, N. L. Sauze *et al.*, "Distributed e2e qos-based path computation algorithm over multiple inter-domain routes," in *Proceedings of IEEE 3PGCIC*, Oct 2011.

[11] T. Wolf, J. Griffioen *et al.*, "Choicenet: toward an economy plane for the internet," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 44, no. 3, pp. 58–65, Jul. 2014.

[12] I. Castro, A. Panda *et al.*, "Route bazaar: automatic interdomain contract negotiation," in *15th Workshop on Hot Topics in Operating Systems (HotOS XV)*, 2015.

[13] "Data source," http://irl.cs.ucla.edu/topology/.

[14] B. Augustin, B. Krishnamurthy, and W. Willinger, "Ixps: mapped?" in *Proceedings of ACM SIGCOMM*, Aug. 2009.

[15] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher, "An analysis of approximations for maximizing submodular set functionsłi," *Mathematical Programming*, vol. 14, no. 1, pp. 265–294, Dec. 1978.

[16] "Submodular set function," https://en.wikipedia.org/wiki/Submodular_set_function.

[17] A.-H. Esfahanian, "Connectivity algorithms," http://www.cse.msu.edu/~cse835/Papers/Graph_connectivity_revised.pdf.

[18] F. S. Perra Nicola, "Spectral centrality measures in complex networks," *Physical Review E*, Sept. 2008.

[19] "As ranking," http://as-rank.caida.org/.