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

Tingwei Liu, John C. S. Lui, *Fellow, IEEE*, Dong Lin, and David Hui

Abstract—The Internet is a gigantic distributed system where the end-to-end (E2E) quality-of-service (QoS) plays an important role. Yet the current inter-domain routing protocol, namely, the Border Gateway Protocol (BGP), cannot provide E2E 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 a source AS to a destination AS is not within the source AS's control regime. This makes it difficult to provide high quality-of-experience services to many Internet users even when many content providers are willing to pay for such high quality E2E guarantees. 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, and we show that it is in fact NP-hard. To address this problem, we design a $(\frac{1-e^{-1}}{2})-$approximation algorithm and also an efficient heuristic algorithm when additional constraints (e.g., the path length) are considered. We further analyze the APX-hardness of the MCBG problem to reveal the existence of the best approximation ratio. Based on the current Internet topology, we demonstrate that it is indeed feasible to provide high QoS guarantees for most E2E connections with only a small broker set: with only 0.19, 1.9 or 6.8 percent ASes/IXPs serving as brokers, 53.13, 85.41 or 99.29 percent of all global E2E connections can receive high QoS-guaranteed services. Finally, we provide an economic model to study the behaviours of ASes when cooperating our brokerage scheme with the BGP protocol, and show that there are incentives to form and maintain such a brokerage coalition.

Index Terms—Inter-domain routing, E2E QoS, dominating-path guarantee

1 Introduction

The Internet, as the world's largest distributed system [1], is an interconnected system of networks which consist of tens of thousands of autonomous systems (ASes) [2]. With the explosion of Internet traffic, the end-to-end (E2E) Quality-of-Service (QoS) guarantees, which impose stringent inter-AS QoS supports, are becoming more and more important [3]. By 2020, global IP traffic will reach 1.3 ZB per year, in which 82 percent is IP video traffic [4], and E2E QoS guarantees for such applications are urgently needed. Yet, the E2E QoS cannot be guaranteed by the current inter-domain routing protocol, i.e., the Border Gateway Protocol (BGP)[5], [6]. The main reason is that an AS can only receive guarantees from its first-hop ASes via service level agreements (SLAs). Beyond the first-hop, QoS guarantees along the path from a source AS to a destination AS is out of the source AS's control regime. However, for time sensitive applications, e.g., multimedia streaming and Voice over IP (VoIP), the states of intermediate AS hops and also the whole AS paths are key factors that influence the E2E QoS. And our collected data of 52,079 ASes and Internet eXchange Points (IXPs) reveals that more than 90 percent of E2E AS connections are more than one-hop.

To address the issue of the E2E QoS provisioning, 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 very effective for realtime and delay sensitive services, e.g., VoIP and video conferencing. This is because for most of these applications, E2E AS hop counts are usually larger than one and unfortunately, there is no inter-AS QoS support in the current Internet.

In this work, we consider the feasibility of improving the E2E QoS for the current Internet from the perspective of "centralized inter-domain routing brokers", on the AS-level topology. We show that *a "small broker set" can be utilized to stitch inter-AS hops along the AS routing path, centralize the routing control for mission-critical traffic across domains, working in parallel to BGP*. Such a 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 responsibilities of network performance measurement, control, resource negotiation, as well as providing transit services. When every hop of an AS-path is 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 we will not focus on how exactly the E2E QoS is guaranteed by constructing a broker set, but we assume that the broker set's capability of

- *T. Liu and J.C.S. Lui are with the Chinese University of Hong Kong, Shatin, Hong Kong. E-mail: twliu@cse.cuhk.edu.hk, cslui@cuhk.edu.hk.*
- *D. Lin and D. Hui are with the Huawei Technology Co., Shenzhen 518129, China. E-mail: {lin.dong, huis.david}@huawei.com.*

Manuscript received 29 Dec. 2017; revised 14 May 2018; accepted 4 Aug. 2018. Date of publication 15 Aug. 2018; date of current version 16 Jan. 2019. (Corresponding author: John C. S. Lui.)
Recommended for acceptance by L. Wang.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TPDS.2018.2865572

monitoring and controlling the (almost) whole Internet can provide a possible way to achieve it. One possibility is that, similar to the case in BGP where ASes sign SLAs with their first-hop ASes and the case in [7] where an alliance of ASes are willing to provide SLAs, ASes can sign agreements with the broker set. By doing so, ASes receive E2E QoS guarantees because the whole paths are dominated by the broker set. Another possibility is, similar to the strategy in [8], the broker set blocks AS connections when it finds QoS requirements are not satisfied.

We like to point out that the realization of such a AS broker set can benefit both CDNs and data center networks (DCNs), because the broker set can enhance the service QoS while reducing the number of needed servers. Besides, the broker set based routing scheme (or the brokerage scheme for short) has a flexible compatibility when cooperating with BGP, i.e., ASes can partially adopt the brokerage scheme and adjust adopting rates so as to maximize their own utilities.

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

- Which AS/IXP should be included in the broker set, which we denote as B?
- How small can B be so 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 sets when cooperating the brokerage scheme with the BGP?

1.1 Summary of Main Contributions and Results

The *key contribution* of this paper is that, by address above challenges, we reveal the potential of improving the E2E QoS from the perspective of "a (small) group of centralized inter-domain routing brokers". We introduce the inter-AS routing framework where a broker set is selected to improve the ASes' E2E QoS by dominating the associated AS paths. Then ASes can enjoy QoS guarantees for the whole routing path by, for instance, signing agreements with the broker set. The main results and contributions of this paper are as follows:

- *The brokerage scheme and the MCBG problem:* We propose a brokerage scheme to provide E2E QoS guarantees. We model the broker set selection problem as the Maximum Coverage with $B-$dominating path Guarantees (MCBG) problem and show it is NP-hard. We further analyze the APX-hardness of the MCBG problem to reveal the existence of the best constant approximation ratio.
- *Efficient algorithms for the broker set selection:* Given the size constraint k of B, we propose an approximation algorithm which provides at least $(\frac{1-e^{-1}}{2})-$guarantee as compared to the best E2E connectivity with dominating AS paths, with a computational complexity of $O(k^2|V|\log|V|+|E|)$, where $|V|$ and $|E|$ are numbers of vertices and edges. To take more practical issues (e.g., computation efficiency and path length constraints) into consideration, we propose a linear-time heuristic algorithm, which achieves a computational complexity of $O(k(|V|+|E|))$ with a minimal reduction of QoS guarantees of no more than 0.5 percent E2E AS connections, compared with the approximation solution.
- *An ideal broker set for the current Internet*: By studying our collected data from 52,079 ASes/IXPs, we demonstrate that it is indeed feasible to provide QoS guarantees for 85.41 or 99.29 percent E2E AS connections with only a small subset of (around 1.9 or 6.8 percent) ASes and IXPs serving as brokers.
- *An economic model for cooperating with BGP*: Finally, we provide an economic model to show that there are incentives in forming and maintaining such brokerage coalition when cooperating the brokerage scheme with BGP. We also study how ASes behave in the cooperation.

1.2 Paper Organization

The remainder of this paper is organized as follows. Section 2 reviews some related work. Section 3 describes our collected datasets and provides interpretations of the topology we study. In Section 4, we present our problem statement, discuss its APX-hardness and propose an approximation solution. In Section 5, we further consider two practical issues such as lower computational complexity and path length constraint, and propose an efficient heuristic algorithm. In Section 6, we present experimental results and findings to demonstrate the feasibility of an inter-domain routing using a small broker set. In Section 7, we analyze the economic feasibility to cooperate our proposed brokerage scheme with the current in-use routing protocol, and study how ASes behave in the cooperation. Finally, Section 8 concludes.

2 Related Work

For decades, there are many excellent studies which attempt to provide E2E QoS guarantees. We classify them into the following three categories and describe their limitations.

Computing QoS-Constrained Path. To provide E2E QoS guarantees, many works focus on finding paths satisfying QoS constraints. Authors in [7] consider one single predetermined route. They construct a subgraph containing the known route and all its neighbours, assume link weights between AS border routers are known, and apply existing routing algorithms to find QoS constrained paths. This work is extended in [9] by considering the existence of AS alliances, in which ASes are willing to offer SLAs. Authors in [10] propose a distributed solution for finding QoS-constrained paths over multiple pre-determined routes. However, the above works [7], [9], [10] are limited by the dependence on predetermined routes and the assumption of pre-known AS link weights, which is not realistic. Authors in [8] provide a connection-oriented switching method where path connections fail if cooperated ASes find they cannot meet QoS requirements. Yet as there is no agreement, it is hard to guarantee quality levels declared by ASes for traffic passing into/through their networks.

Economic Method. Another line of works [11], [12], [13], [14] seek solutions from an economic plane perspective. In [11], [12] market mechanisms are used to automate the deployment of the E2E inter-domain QoS policy among

TABLE 1
Alliance Size (# of Brokers) and QoS Coverage Comparison

Method	Alliance size (# of brokers)	QoS coverage
Our approach	100 (0.19% out of 52,079 ASes/IXPs)	53.14%
	1,000 (1.9% out of 52,079 ASes/IXPs)	85.41%
	3,540 (6.8% out of 52,079 ASes/IXPs)	99.29%
[13], [14]	51,757 (all ASes)	100.00%
[18], [19]	$\geq$51,757 ($\geq$1 brokers per AS)	100.00%
[20], [21], [22]	322 (all IXPs)	15.70%

users and ISPs. iREX [11] facilitates a distributed economic system where resource user domains select and reserve inter-domain QoS resources to construct inter-AS routes, resource provider domains provide available iREX path vector information, including resource quality and price of each path. Authors in [12] further extend iREX into a multi-path routing option. However, the proposed mechanism brings a large overhead for storing and updating enormous iREX path vectors. Authors in [13], [14] consider incentivizing all domains to cooperate and share intra-domain information, e.g., the network topology and resources. Authors in [13] present an alliance paradigm for domain cooperation and revenue sharing, and consider both a centralized alliance model, where a third party exists and ensures the supervision of contracts, and a distributed alliance model, where the responsibility is shared by all domains. And they discuss the detailed implementation that covers network management at different plane levels in [14]. However, their goal of cooperating as many as 52,079 ASes is not realistic.

Stitching QoS-Enable Pathlets. The idea of centralized inter-domain path mediators, which is most similar to our proposal, is well explored in the literature [15], [16], [17] to support QoS requirements across domains. In these setups, ISPs provide QoS enabled *pathlets* (i.e., fragments of paths represented as sequences of virtual nodes), which are stitched together to build global paths by inter-domain routing mediators, such as Path Computation Elements (PCEs) optimally selecting disjoint QoS paths across multiple domains [15], bandwidth brokers [18], [19] providing efficient admission control, or external trusted providers managing routing for multiple ASes [17]. Recently, authors in [20], [21], [22] propose to use central control points (CXPs) as inter-domain routing mediators for the consideration of their rich connectivity, enabling high path diversity and global client reach. A CXP is external to an ISP entity and it applies centralized inter-domain control over how fractions of Internet traffic are routed. Nevertheless, these schemes seriously increase the burden of selected mediators, since they need to calculate optimal QoS E2E paths for all routing requests and, for some mediators, also exchange Internet traffic.

The scalability problem, especially for the Internet with tens of thousands of ASes, is a common problem facing studies attempting to improve the E2E QoS. In this work, we consider utilizing a *"small"* set of ASes to stitch each two AS hops along the AS routing path. As the whole routing path is dominated by the broker set, brokers are able to provide supervision, control and resource negotiation for every AS hop. The broker set's ability of monitoring and controlling AS paths makes it possible to achieve E2E QoS guarantees: for instance, similar to the case in [7] where an AS alliance is willing to provide SLAs, ASes can receive guarantees for the whole routing paths by agreements with the broker set; or similar to the strategy in [8], the broker set block AS connections when it finds QoS requirements are not satisfied. To summarize, our approach has the following attractive properties:

- We provide at least $(\frac{1-e^{-1}}{2})-$guarantee as compared to the best E2E connectivity with dominated AS paths. In particular, we achieve QoS guarantees for 99.29 percent E2E connectivity with only 6.8 percent ASes/IXPs as brokers. If focusing on providing QoS guarantees for the majority E2E AS connections, the broker set size can be further reduced: 1,000 brokers for 85.41 percent saturated connectivity and 100 brokers for 53.14 percent saturated connectivity.
- We provide a flexible compatibility when cooperating with the current BGP protocol. ASes can partially adopt the brokerage scheme and adjust adopting rates so as to maximize their own utilities. We utilize an economic model to show the existence of incentives to form and maintain the brokerage coalition in the cooperation.

To better illustrate gains from the above properties, we compare our approach with previous works from perspectives of both AS alliance size and coverage of E2E connections with QoS guarantees, and summarize in Table 1.

3 Topology and Datasets

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

In this study, we consider the AS-level topology, which is composed of different ASes and their interconnections, and 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 the *Internet eXchange Point* (IXP), which provides efficient and cost effective means for traffic exchange between multiple ASes. If an AS wants to enjoy this cheap traffic exchange, it has to register as a member of the corresponding IXP. We collected data for the AS topology as well as connections to IXPs. We then 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 [23], 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 [23]. 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

TABLE 2
Summary on the Collected Dataset

Description	Numbers
IXPs	322
ASes	51,757
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

obtained the data of IXP memberships and peerings in 2014[1] using similar approaches described in [24].

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 [23]. 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 percent (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 2, show that our dataset is indeed representative.

Similar to [20], [22], we also assume that IXPs are independent entities. This is proposed due to the rich connectivity of IXPs [20], [22]. This assumption assigns a new role to IXPs which typically provide only switching service instead of routing. In our experiment, we will show the importance of IXPs, in particular, they play a critical role in the broker set.

Fig. 1 depicts the visualization of our derived topology using the k-core decomposition method [25] to review the hierarchical structure of the network. Nodes in the figure represent either ASes (in color) or IXPs (in black). Edges in the figure represent either AS-IXP or AS-AS connections. $k-core$ of a graph is defined as the maximal subgraph in which each vertex in this subgraph has at least degree k, then the "*coreness*" of a vertex is k if it is a vertex in a $k-$core subgraph but not a vertex of a $(k+1)-$core subgraph. We assign colors according to the coreness of nodes, as illustrated by the right side of the figure, to reveal the hierarchical structure of the network. Our AS topology consists of more than 120 shells of $k-$cores (each shell is assigned a unique color). Furthermore, the figure also shows the degree of nodes. The scale of node degree is displayed using the *size* of a node, as illustrated in the left side, a node with the degree of 8,741 is the largest node in the graph.

4 Problems and Algorithms: Theoretical Basis

In this section, we propose the inter-domain routing brokerage problem and develop some theoretical foundations. First, we formulate it as a Maximum Coverage with $B-$dominating path Guarantee problem and analyze the complexity; then we present an approximation algorithm to solve the MCBG problem. Finally, we show the best constant approximation ratio of the MCBG problem.

1. The most recent data is collected for 2015 but only includes the data for the first two months. To obtain more reliable conclusions, we utilize the data of 2014 which is the most recent full-year data.

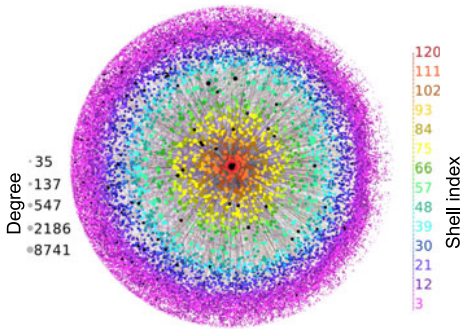


Fig. 1. Visualization of an AS-level Internet topology, which is scale-free and layered network consists of IXPs at both its core and edge.

4.1 Problem Statement

Let $G=(V,E)$ denote an undirected graph with vertex set V and edge set E. For each vertex $v\in V$, define the neighbourhood $N(v)$ 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 that are adjacent to $\forall v\in V'$, i.e., $N(V')=\cup_{v\in V'}N(v)$, and $f(V')$ as the number of vertices covered by V', i.e., $f(V')=|V'\cup N(V')|$. We first define the notion of a "$B-$*dominating path*".

Definition 1. *Given a non-trivial 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 B, where $B\subseteq V$.*

Remark. Note that the definition of $B-$dominating path also implies that along such a path, every two adjacent brokers are 1 hop neighbour or 2 hop neighbour connected by a non-broker.

In the context of inter-domain routing brokerage, we treat the AS-level topology we mentioned in Section 3 as the input graph G, where the vertex set V is the collection of ASes/IXPs, a connection between AS/IXP u and 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, there exists one $B-$dominating path, then the E2E network's 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.

Problem 1 (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 for the PDS problem to provide *all* connections with $B-$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 2.

Problem 2 (Maximum Coverage with B-dominating path Guarantees problem). *Given a connected non-trivial graph $G = (V,E)$ and a positive integer k. Find a subset $B \subseteq V$ which guarantees:*

- $|B| \leq k$;
- there exists at least one $B-$dominating path between u and v, for $\forall u, v \in B \cup N(B)$;
- $f(B) = |B \cup N(B)|$ is maximized.

Note that for $\forall u, v \in B \cup N(B)$, if there exists a path containing only nodes in $B \cup N(B)$, 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 for the PDS problem, then it is also the solution for the MCBG problem. If there is no solution for the PDS problem, the solution for 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, 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., $B \cup N(B) = V$. Hence, B is the solution for the MCBG problem. If there is no solution for the PDS problem, denote the solution to the MCBG problem as B. If there exists a set B' which not only satisfies $|B'| \leq k$ but also provides the $B'-$dominating path guarantees for $\forall u, v \in B' \cup V'$ and if $|B' \cup V'| > |B \cup N(B)|$, 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 2. Therefore, there doesn't exist such a set B'. So B can provide $B-$dominating path guarantees for as many connections as possible. □

4.2 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 appendix, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPDS.2018.2865572. □

To analyze the computational complexity of the MCBG Problem 2, we first consider its decision version.

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

Proof. The decision version of the MCBG problem will output an indicator "YES" if it is feasible to find a subset $B \subseteq V$ which satisfies three constraints in Problem 2, and "NO" otherwise. To prove the NP-completeness for the decision version of the MCBG problem, it is suffice to show that there is a polynomial-time reduction from the NP-complete PDS problem (which is proven to be NP-complete in Lemma 1), to the decision version of the MCBG problem. To conduct such reduction, for each instance in the path-dominating set B, we can construct a corresponding instance of the decision version of the MCBG problem based on their definitions, by setting $p = |V|$. Therefore, the decision version of the MCBG problem is also NP-complete. □

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

Proof. As the maximization problem is NP-hard if its decision version is NP-complete, the MCBG problem is NP-hard. □

4.3 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 to 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 Algorithm 1. The selection of B^* is realized by Algorithm 1.

Algorithm 1. Approximation Algorithm for $MCB(V,k)$ [26]

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: $\quad s_i \leftarrow \text{argmax}_s f(B_{i-1} \cup \{s\}) - f(B_{i-1})$;
4: $\quad B_i \leftarrow B_{i-1} \cup \{s_i\}$;
5: **end for**
6: Return $B = B_k$.

Problem 3 (Maximum Coverage with the broker set B (MCB) problem). *Given a connected non-trivial graph $G = (V,E)$ and a positive integer k. Find a set $B \subseteq V$ which satisfies:*

1) $|B| \leq k$, and
2) $f(B) = |B \cup N(B)|$ is maximized.

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

Now the remaining issue is to find B'. To achieve this, we take advantage of some special properties of the graph that we study, e.g., the graph depicted in Fig. 1. Note that in the AS-level Internet graph, for more than 99.2 percent of source and destination pairs, their hop count distances are within 4 hops. This special characteristic helps us to design an 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*

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

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 $(\alpha,\beta)-$graph can help to decide the size of B' to satisfy the $B-$dominating path constraint. The details of solving the MCBG problem, including finding B^* and B', are shown in the approximation algorithm Algorithm 2.

Algorithm 2. Approximation Algorithm for $MCBG(V,k)$ on a $(\alpha,\beta)-$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: Let B^* be the solution returned by applying Algorithm 1 to $MCB(V,x^*)$ and let $B'=V-B^*$, 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;
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 Algorithm 2, the computational complexities for selecting 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 Algorithm 2, respectively.

Now we can prove how Algorithm 2 can achieve the approximation of the pre-selected B^*. We first present the following lemmas to aid the proof.

Lemma 3. *The coverage function f is a submodular and nondecreasing set function [27].*

Lemma 4. *Algorithm 1 provides a $(1-e^{-1})-$approximation for the MCB problem [26].*

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

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

$$\theta=\left\lceil\frac{\beta}{2}\right\rceil=\begin{cases}\beta, & \beta \text{ is even;}\\ \beta+1, & \beta \text{ is odd.}\end{cases} \tag{1}$$

Proof. We start by providing the sketch of the proof. Let $OPT_{MCBG(k)}$ be the optimal solution for $MCBG(V,k)$ and $APX_{MCBG(k)}$ be the corresponding approximation solution obtained through the Algorithm 2. Let $OPT_{MCB(x^*)}$ be the optimal solution for $MCB(V,x^*)$ and $APX_{MCB(x^*)}$ be the corresponding approximation solution obtained through the Algorithm 1. Furthermore, denote the optimal solutions for $MCB(V,\theta x^*)$ and $MCB(V,k)$ as $OPT_{MCB(\theta x^*)}$ and $OPT_{MCB(k)}$, respectively. To prove the approximation ratio, it is suffice to show

$$\frac{\theta}{1-e^{-1}}f\big(APX_{MCBG(k)}\big)\geq\frac{\theta}{1-e^{-1}}f\big(APX_{MCB(x^*)}\big), \tag{2a}$$

$$\geq\theta f\big(OPT_{MCB(x^*)}\big), \tag{2b}$$

$$\geq f\big(OPT_{MCB(\theta x^*)}\big), \tag{2c}$$

$$\geq f\big(OPT_{MCB(k)}\big), \tag{2d}$$

$$\geq f\big(OPT_{MCBG(k)}\big). \tag{2e}$$

The inequality (2a), $f(APX_{MCBG(k)})\geq f(APX_{MCB(x^*)})$, can be derived using $APX_{MCB(x^*)}\subseteq APX_{MCBG(k)}$ and the monotonicity of the coverage function f. Here $APX_{MCB(x^*)}$ is the pre-selected brokers in $APX_{MCBG(k)}$.

The inequality (2b), $\frac{1}{1-e^{-1}}f(APX_{MCB(x^*)})\geq f(OPT_{MCB(x^*)})$, is the direct application of Lemma 5.

The inequality (2c), $\theta f(OPT_{MCB(x^*)})\geq f(OPT_{MCB(\theta x^*)})$, is obtained based on the property of function f mentioned in Lemma 3. If we divide $OPT_{MCB(\theta x^*)}$ into θ disjoint subsets $S_1,S_2,\ldots,S_\theta$ with equal size x^*, then

$$\begin{aligned}\theta f\big(OPT_{MCB(x^*)}\big)&\geq\theta\max_{1\leq i\leq\theta}f(S_i)\\ &\geq\sum_{1\leq i\leq\theta}f(S_i)\geq f(\cup_{1\leq i\leq\theta}S_i)=f\big(OPT_{MCB(\theta x^*)}\big).\end{aligned} \tag{3}$$

The inequality (2d), $f(OPT_{MCB(\theta x^*)})\geq f(OPT_{MCB(k)})$, can be derived according to $\theta x^*\leq k$ and the monotonicity of the coverage function f. Here, x^* is the number of pre-selected brokers (i.e., B^* in line 1 of Algorithm 2) for approximating the optimal coverage. We choose one of those x^* pre-selected brokers as the root. As for each of the other x^*-1 pre-selected brokers, we need to add $\lceil\frac{\beta}{2}\rceil-1$ extra brokers in the worst case to satisfy $B-$dominating path constraints (i. e., construct B' in line 6 of Algorithm 2), because path lengths among those x^* brokers are no more than β hops with a high probability α. Here, to achieve the best approximation ratio: x^* is selected as the biggest integer satisfying $x^*+(x^*-1)(\lceil\frac{\beta}{2}\rceil-1)\leq k$, such that the broker set size will not exceed the size constraint k; also, θ is selected as the smallest one satisfying $\theta x^*\geq k$. Thus, $x^*=\left\lfloor\frac{k-1}{\lceil\beta/2\rceil}+1\right\rfloor$ and $\theta=\lceil\frac{\beta}{2}\rceil$.

The inequality (2e), $f(OPT_{MCB(k)})\geq f(OPT_{MCBG(k)})$, is straightforward: compared with $OPT_{MCBG(k)}$, $OPT_{MCB(k)}$ is obtained without the $OPT_{MCB(k)}-$dominating path constraints. Thus, we conclude that $f(APX_{MCBG(k)})\geq f(OPT_{MCBG(k)})$. □

As for the AS-level Internet topology, 99.2 percent E2E connections are within 4 hops. By applying Theorem 3, we have the following corollary.

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

4.4 Approximation Class Analysis for MCBG

Now, we make further discussion about the approximation class of the MCBG problem to reveal the existence of the best approximation ratio for the MCBG problem. We will first prove the MCB problem to be APX-hard, which means there is a constant c such that it is NP-hard to find an approximation algorithm with an approximation ratio better than c. And we prove the APX-hardness of the MCBG problem by utilizing the proposed approximation algorithm Algorithm 2 to construct a PTAS reduction from MCB problem.

Lemma 5. *The MCB problem belongs to the APX class, i.e., MCB $\in$ APX.*

Proof. A NP-hard optimization (NPO) problem belongs to the class APX if it is approximable within a constant [28]. As the greedy algorithm Algorithm 1 provides a *best* approximation ratio of $1-e^{-1}$ for the MCB problem, it is NP-hard to find an approximation algorithm with an approximation ratio better than $1-e^{-1}$ [29]. Thus MCB $\in$ APX. □

Lemma 6. *The MCB problem is PTAS-reducible to the MCBG problem, i.e., MCB$\leq_{PTAS}$MCBG, on the $(\alpha,\beta)-$graph.*

Proof. To construct a PTAS-reduction from the MCB problem to the MCBG problem, we need not only a function h to map from instances of MCB into instances of MCBG, but also a function g to map from solutions of MCBG into solutions of MCB preserving the performance ratio [30]. The detailed PTAS-reduction process is presented in the appendix, available in the online supplemental material. □

Theorem 4. *The MCBG problem on the $(\alpha,\beta)-$graph is APX-hard.*

Proof. An NPO problem P is APX-hard if for any problem $P' \in APX$, $P' \leq_{PTAS} P$ [30]. As Theorem 2 has proved that MCBG$\in NPO$, and Lemma 6 has shown that MCB$\in APX$ is PTAS-reducible to MCBG on the $(\alpha,\beta)-$graph, the MCBG problem on the $(\alpha,\beta)-$graph is APX-hard. □

Remark. Note that the APX-hardness of the MCBG problem on the $(\alpha,\beta)-$graph reveals the existence of the best constant approximation ratio, which leaves the research potential for developing approximation algorithms with "tighter" and even "tight" approximation ratios. We leave this in our future work.

5 Problems and Algorithms: Practical Considerations

The previous section provides the theoretical foundation for 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, we propose a heuristic algorithm with a lower computational complexity while maintaining a good $B-$dominating path coverage with the broker set B. Second, we generalize the MCBG problem by taking the path length constraint into consideration.

5.1 Efficient Heuristic Algorithm and Baseline Algorithms

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

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 the maximum sub graph in $B \cup \{w\}$ is maximized. Go to step 2.

Note that Algorithm 3 aims to maximize the connected graph size in each iteration. As we will show, experiment results indicate that Algorithm 3 is capable of finding a broker set with a very high coverage in only few thousand iterations.

To evaluate the performance gain of our proposed algorithm, we compare it with four baseline algorithms, whose detailed pseudocodes are listed in the appendix, available in the online supplemental material. The *Set Cover* (SC) algorithm is an algorithm proposed in [31] to find some but not necessarily the smallest dominating sets. Comparisons with the SC algorithm are helpful to gain some understanding on the importance of a broker set selection process. The *IXP-Based* (IXPB) algorithm returns a collection of IXPs whose degrees are higher than some given threshold. Since IXPs are often treated as ideal nodes for inter-domain control [20], [22], 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 of a graph. At each round, the node with the largest degree or page rank value will be added to the broker set. In later experiments, we shall further explore those above algorithms' differences in selecting a broker set and examine their performances against each others.

5.2 Path Length Constraint and Its Evaluation Method

Note that in the MCBG problem, the $B-$dominating path between any source-destination pair (u,v) can be of arbitrary length. However, in practice, some ISPs may want to restrict AS hop counts of E2E paths, for instance, require AS hop counts to follow some distribution specified by ISPs. Therefore, during searching the broker set B, we introduce an extra requirement on the path length l_{uv} and present the following MCBG problem with path length constraints.

Problem 4 (MCBG problem with path length constraints). *Given 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 $u, v \in V$ $(u \neq v)$. Find a subset $B \subseteq V$ which guarantees:*

TABLE 3
$l-$Hop Connectivities 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

1) $|B| \leq k$;
2) there exists at least one $B-$dominating path of length l_{uv} between u and v, for $\forall u, v \in B \cup N(B)$;
3) $f(B) = |B \cup N(B)|$ is maximized.

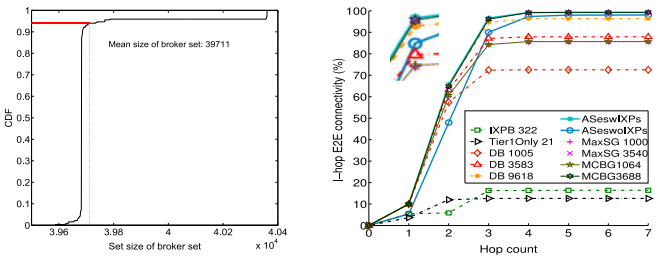
We consider evaluating a candidate broker set's satisfaction of path length constraints from a stochastic perspective. We start by providing a probability distribution based interpretation for the path length characterization in the AS level topology. By treating the choice of a source-destination pair (u,v) as a random event whose sample space contains all possible source-destination pairs, the corresponding path length l_{uv} can be viewed as a random variable l. We denote $F(l)$ as the cumulative histogram of l_{uv}, i.e., the number of admissible paths with path-lengths no more than l, which gives the path length distribution. We define "*a broker selection strategy $\mathcal{A}$ is feasible*" if it gives a candidate broker set $B_{\mathcal{A}}$ satisfying $F(l)$ up to an ϵ fraction of errors for all values of l. More specifically, $B_{\mathcal{A}}$ gives a distribution $F_{B_{\mathcal{A}}}(l)$ such that

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

Eq. (4) provides us *a stochastic way to verify the feasibility of a candidate broker selection algorithm $\mathcal{A}$*. Yet, we still need to compute the cumulative distribution $F_{B_{\mathcal{A}}}(l_{uv})$ for a broker set $B_{\mathcal{A}}$ produced by $\mathcal{A}$. Here we provide an efficient way to compute $F_{B_{\mathcal{A}}}(l_{uv})$. Given a graph G with an adjacent matrix A, we define a operate * such that $B_{\mathcal{A}}{}^\star A$ will erase all entries of A when neither of its row nor column indices belongs to $B_{\mathcal{A}}$. The output matrix of $B_{\mathcal{A}}{}^\star A$, denoted as $\tilde{A}$, 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 no more than l. We also call this as the "*$l-$hop E2E connectivity*".

Towards a better understanding of the above definition, we depict $l-$hop connectivities of different topologies (e.g., *ER-Random, WS-Small-World, BA-Scale-free, ASes with/without IXPs*) in Table 3. 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 a 99.21 percent E2E connectivity.

Remark. Note that when the hop count threshold l increases, the E2E connectivity also increases and eventually stabilizes. This saturated value is defined as the "*saturated E2E connectivity*". Also, we need to point out that only connections detected in our dataset are considered when constructing the adjacent matrix A, which means that *there is no connection if in reality an AS is unwilling to directly connected to IXPs*.



(a) CDF of broker set size by the SC algorithm (b) $l-$hop E2E connectivities of the other algorithms

Fig. 2. Comparisons of different algorithms' l-hop E2E conductivities.

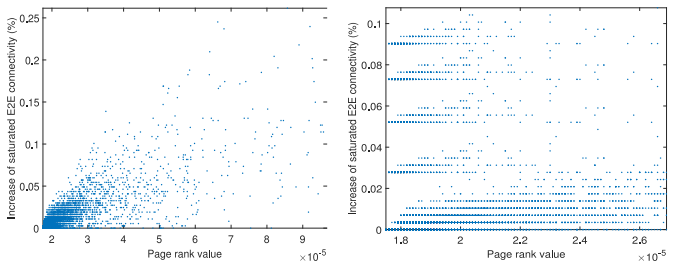
6 Structural Feasibility and Broker Set's Properties

In this section, we study possibilities of the broker set composition in the current Internet according to some experimental results. We also compare our proposed algorithms with other state-of-the-art algorithms on the broker selection.

6.1 Evaluation for the $l-$Hop E2E Connectivity

The selection of a broker set is non-trivial. *An improper selection may lead to a large size broker set, making it more difficult to incentivize ASes to join the broker set, or with a very poor l-hop E2E connectivity*. Fig. 2a shows the cumulative distribution function (CDF) of the broker set size by running the SC algorithm 300 iterations. Although a 100 percent E2E connectivity is guaranteed, the SC algorithm takes around 40,000 nodes into the broker set, accounting for more than 76 percent of the overall network vertices. No doubt, incentivizing such a large population of ASes to join the broker set and maintain it is unrealistic. Fig. 2b shows the achieved $l-$hop E2E connectivities of the other algorithms when varying hop count requirement l. For IXPB and Tier1Only algorithms, which were considered in previous works, their $l-$hop E2E connectivity results imply that *it is not appropriate to merely rely on IXPs or tier 1 ISPs to act as brokers*. Both of them suffer from low E2E connectivities: the IXPB algorithm could reach at most a 15.70 percent E2E connectivity with 322 brokers, and it is far worse for the Tier1-Only algorithm. Given the fact that only 40.2 percent ASes are directly connected to IXPs, it's not hard to foresee the low E2E connectivity by choosing only IXPs as brokers: with a limited network coverage, only a very small amount of routing paths can be served by the broker set. The case is similar when merely selecting tier 1 ISPs as brokers.

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 the degree/PageRank value and the saturated E2E connectivity with an increasing broker set size. The broker set selected by the DB algorithm, which consists of high degree ASes and IXPs, can achieve an around 72.53 percent E2E connectivity with 1,005 brokers. However, the DB algorithm requires a large size broker set to guarantee a high (e.g., 99 percent) E2E connectivity for the serious marginal effect when $|B| > 1{,}000$: the DB algorithm can only achieve a 96.35 percent E2E



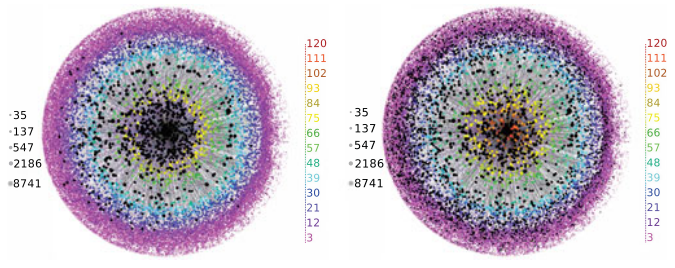
(a) The 101th candidate (b) The 1,001th candidate

Fig. 3. The correlation between the PageRank value and the increase of the connectivity when adding a new broker.

connectivity even with 9,618 brokers. By taking a more detailed view of the selected broker set by DB algorithm in Fig. 4a, we also find that most selected brokers are located at the center network core, leaving the network edge mostly uncovered. The PRB algorithm has the similar problem, for the undirected graph's PageRank distribution of is statistically close to its degree distribution [32].

Towards a better understanding of the possible cause of such a serious marginal effect, we take the PRB algorithm as an example for the further analysis. In Fig. 3, we use the PRB algorithm to find broker sets of size 100 and 1,000, then take different ASes as the 101th and 1,001th brokers. When the broker set is small, as shown in Fig. 3a, ASes with larger PageRank values are more likely to bring higher saturated E2E connectivity increases. Thus it is reasonable to select ASes with lager PageRank values. However, the correlation between the PageRank value and the connectivity contribution decreases from 0.818 to 0.227 when the broker set size increases from 100 to 1,000. Thus the PRB algorithm, which picks ASes with lager PageRank values does not work, as illustrated in Fig. 3b.

Our approximation algorithm for the MCBG problem can achieve a 85.71 percent saturated connectivity with 1,064 brokers and a 99.29 percent saturated connectivity with 3,688 brokers, making it the best algorithm among all we considered. Compared with the approximation algorithm, our MaxSG algorithm achieves an equivalent performance (i.e., sacrifices less than 0.5 percent connectivity) while greatly reduces the computational complexity. Also, MaxSG algorithm outputs a broker set consisted of 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 percent. And Fig. 4b shows a close look of the selected broker set. Unlike the DB algorithm, the MaxSG



(a) Degree-based (b) MaxSG

Fig. 4. A close look at brokers selected by PRB and MaxSG algorithms. Vertices in black are brokers.

TABLE 4
The 3,540-Alliance Guarantees Almost Zero Path Inflation

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

algorithm does not have such an overcrowded network core and the network outer ring can be well covered. More detailed findings about the 3,540-alliance broker set will be discussed in the following sections.

Remark. Note that the broker set with 3,540 members, which only accounts 6.7 percent of 52,079 ASes/IXPs, is proposed to achieve a 99.29 percent 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 E2E AS connections, e.g., 1,000 brokers for a 85.41 percent saturated connectivity and 100 brokers for a 53.14 percent saturated connectivity.

6.2 Attractive Properties of the 3,540-Alliance Broker Set B

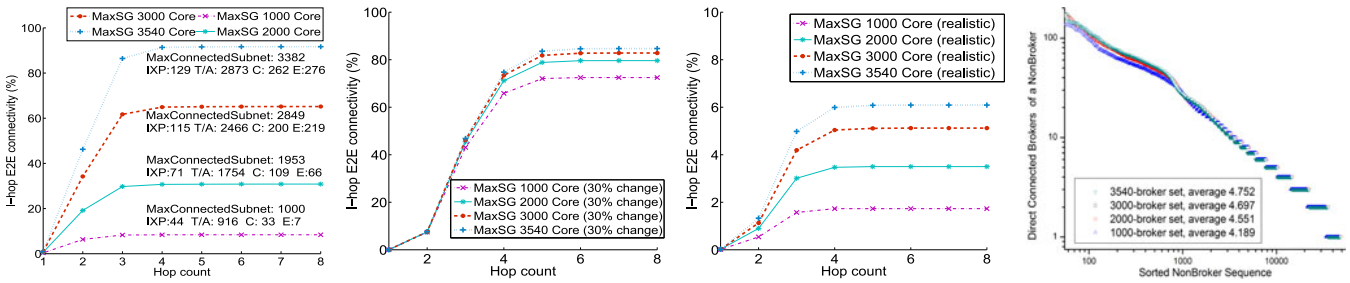
We name the broker set with 3,540 brokers output by the MaxSG algorithm as the "3,540-alliance", and discuss some of its attractive properties.

Minimal Path Inflations. Path length inflations (i.e., previously a l-hop reachable pair now requires l' hops, where $l' > l$) are observed in Fig. 2b. Consider the DB algorithm. With 1,005 brokers, only 72.40 percent E2E connections can be satisfied within 4 hops, in contrast to that of 90.02 percent in the free-path selection scheme denoted as "ASeswithIXPs". While as illustrated in Table 4, if 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 3,540-alliance almost overlaps the one of "ASesWithIXPs").

Diversified Compositions. As illustrated in Fig. 5a, the 3,540-alliance consists of different types of ASes/IXPs, rather than being monopolized by tier 1 ISPs. This avoids the monopoly of some tier 1 ISPs. Here, we use the same definition and data in [33] to divide brokers into different categories according to their offered services. Table 5 lists some brokers and their rankings as well, which illustrates the importance of IXPs for the $B-$dominating path routing with the broker set B.

90 Percent of E2E Connections Only Use Nodes in the Broker Set. As illustrated in Fig. 5a, although for some connections a re-route through non-brokers is still necessary, more than 90 percent E2E connections can be carried out by the 3,540-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. As for the remaining 10 percent E2E connections, we show how to incorporate non-broker nodes in the later discussion.

Minimal Changes in Business Relationships. While for the real-life inter-domain routing issue, business relationships (e.g., high-tier and low-tier, or peering) among ASes/IXPs have significant influences and must be taken into consideration. Fig. 5c shows a broker set's performance in the



(a) The l-hop E2E connectivity via the broker set solely and the detailed composition (b) The l-hop E2E connectivity by changing 30% business relationships (c) The l-hop E2E connectivity with realistic business relationships (d) Non-brokers' choices

Fig. 5. Findings for the 3,540-alliance broker set found by the MaxSG algorithm.

current Internet if forcing ASes/IXPs to obey existing business relationships, i.e., the previously assumed bidirectional routing policy becomes directional. A sharply decreased E2E connectivity over different broker set sizes has been observed. However, we also notice in Fig. 5b that, by randomly changing only 30 percent inter-broker connections to bidirectional (e.g., peering), such degradation can be greatly suppressed. Even a 1,000-broker set with 30 percent changes at its inter-broker connections still achieve a 72.5 percent E2E connectivity, and the 3,540-alliance with 30 percent random changes can achieve a 84.68 percent E2E connectivity.

Potentials for the Multi-Path Routing: As shown in Fig. 5d, given specific broker sets, a non-broker node typically connects to more than four brokers on average. This implies that the multi-path routing framework can be cosidered in the AS information delivery to further improve the E2E QoS.

7 Economic Incentives

Now we discuss the economic feasibility for realizing such a brokerage scheme on the existing Internet. We will analyze the possibility of cooperating the brokerage scheme with the current inter-domain routing protocol and study ASes' behaviors in the cooperation. We provide a flexible compatibility when cooperating with the current inter-domain routing protocol:

- We do *not* assume all ASes agree to join the broker set, especially at the early stage. Hence, *the broker set B mentioned in this section needs not to be the 3540-alliance*, it can be a much smaller one, e.g., 100 brokers to provide a 53.41 percent saturated connectivity or 1,000 brokers to provide a 85.41 percent saturated connectivity.
- We allow ASes to flexibly adjust adopting rates of brokerage scheme and BGP protocol to maximize their utilities.

TABLE 5
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.

We formulate a game-theoretic framework to understand the interactions among ASes. More specifically, we first treat all ASes in B as one identity, and use Stackelberg game model and Nash bargaining solution to study the game between any AS in $\bar{B}$ and B. Then we state the rationale for treating all ASes in B as one identity, and discuss the revenue distribution among ASes in B. The analysis can reveal that:

- ASes in both the broker set B and the non-broker set $\bar{B}$ are willing to follow this new routing rule.
- There exists incentives for ASes to form and maintain the brokerage coalition B in the cooperation.

7.1 Interactions of Non-Broker & Broker Sets

We start with an example of the business model illustrated in Fig. 6. All brokers, i.e., ASes in B are treated as one identity, and the rationale for doing so will be explained later. We first discuss behaviors of non-broker ASes, i.e., ASes in $\bar{B}$. Let i denote a specific AS in $\bar{B}$. AS i plays a double role in the game with B: (1) the *customer* of B, e.g., AS 1, which has a *routing strategy* α_i to determine the fraction of AS i's traffic routed to B, and need to pay B for the routing service; (2) the *employee* of B, e.g., AS 5, which is hired by B to transit traffic from one broker to another broker and can receive payment for the transition service it provided. Fig. 6 offers an instruction of the payment flow. Our goal is to guarantee the existence of a steady state of economic relationships among all ASes, and make $\alpha_i \rightarrow 1$, which means that our new routing scheme will be fully adopted, under the steady state.

Interactions Between the Employee AS in $\bar{B}$ and B. We first consider when an AS $\in \bar{B}$ acts as an *employee* of B. Let p_B denote the routing price for per unit volume traffic charged

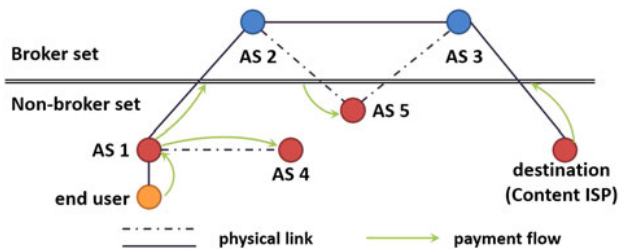


Fig. 6. A typical example for the business model.

by B. As shown in Fig. 6, B can charge from both the customer AS and the destination. Sometimes, the E2E connectivity can be achieved via B solely, e.g., AS 2 and 3 have a direct connection. When there is no direct connection, B needs to hire a non-broker to form the $B-$dominating path, e.g., B hires AS 5 to achieve the E2E connectivity, and pays the employee AS p_e for routing per unit volume traffic. The cost for each AS to route per unit volume traffic is denoted by c. We apply the *Nash bargaining solution* to achieve an agreement on the price p_e between the employee AS j and B.

Let u_j^* denote the utility function of the employee AS j for per unit volume traffic

$$u_j^* = p_e - c. \tag{5}$$

Note that AS j has no global knowledge, i.e., it does not know the exact hop count of the whole $B-$dominating path. On the other hand, j knows the hop count distance between any two vertices is no more than β, it can therefore assume that for this $B-$dominating path, B needs to hire at most $\lfloor \frac{\beta}{2} \rfloor$ employees, which also means B's utility for per unit volume traffic is at least

$$u_b^* = 2p_B - \left\lfloor \frac{\beta}{2} \right\rfloor p_e - \left(\beta - \left\lfloor \frac{\beta}{2} \right\rfloor\right) c. \tag{6}$$

Thus the Nash bargaining solution can be obtained by solving the following optimization problem:

$$\begin{aligned} \max_{p_e} \quad & u_j^* \cdot u_b^*, \\ \text{s.t.} \quad & p_e > c. \end{aligned} \tag{7}$$

Theorem 5. *For the bargaining proposed above, there exists a Nash bargaining solution.*

Proof. As the above optimization problem is maximizing a continuous function over a compact set, it must have at least one solution. □

Interactions Between the Customer AS in $\bar{B}$ and B. Now we consider when an AS $\in \bar{B}$ acts as a *customer* of B. For any AS i, when acting as a customer, we normalize its total traffic volume as 1. Let α_0 denote the fraction of traffic routed to B in the traditional routing mechanism. Thus $\alpha_i = 1$ ($\alpha_i = \alpha_0$) means the new routing scheme is fully adopted (rejected). Let u_i denote the utility function of AS i. Then

$$u_i = V_i(\alpha_i) + P_i(\alpha_i) - p_B \alpha_i, \tag{8}$$

where $V_i(\alpha_i)$ denotes the income received from end-users; $P_i(\alpha_i)$ and $p_B\alpha_i$ denote costs for routing through $\bar{B}$ and B, respectively.

To better understand non-broker ASes' behaviors, we do a further analysis on the above AS utility function. Assume $V_i(\alpha_i)$ is proportional to the user satisfaction level on the QoS. When α_i increases, more traffic is routed through B, so the QoS is improved. The user satisfaction increases with the improved QoS, but the rate of the user satisfaction increase decreases due to the law of diminishing returns. Therefore, $V_i(\alpha_i)$ is a continuous, concave and strictly increasing function. As for $P_i(\alpha_i)$, its value can be positive, negative, or zero if i is a high-tier, a lower-tier, or peering with other ASes in $\bar{B}$. Therefore, we can roughly classify the traffic routed by AS i into five classes, i.e., high paid, low paid, peering, low charged, high charged, with the $P(\alpha_i)$ value increasing from negative to positive. When α_i increases, AS i will first consider transferring its high paid traffic to B so as to increase u_i, and then it may also consider transferring traffic to B in the order of low paid, peering, low charged traffic. Hence, we assume that $P_i(\alpha_i)$ is a continuous and concave function which is non-decreasing in $[\alpha_0, \alpha_*]$ and non-increasing in $[\alpha_*, 1]$, $P_i(1) = 0$.

The utility of B is

$$u_b = 2p_B\alpha - C(\alpha, p_e), \tag{9}$$

where $\alpha = \sum_{i \in \bar{B}} \alpha_i v_i$, and $C(\alpha, p_e)$ is the cost function for routing data and hiring employees, assumed to be concavely increasing in α and p_e.

Therefore, the interaction between AS $i \in \bar{B}$ and B, where both AS i and B make decisions *sequentially* to maximize their own utilities, can be formulated as a *Stackelberg game* [34]. In particular, we have:

Players. All ASes in both $\bar{B}$ and B.

Strategies. Each non-broker AS $i \in \bar{B}$ determines its routing strategy α_i; B determines the routing price p_B.

Roles. B is the first mover and decides p_B. Non-broker ASes are the second movers and decide $\alpha_i, i \in \bar{B}$. Each AS $i \in \bar{B}$ makes its decision individually.

Outcome. The outcome can be determined by backward induction, i.e., for any given p_B, each AS i decides $\alpha_i = \alpha_i(p_B)$ to maximize its own utility; based on this knowledge, B decides p_B to maximize its utility.

Theorem 6. *For the Stackelberg game proposed above, there exists the Stackelberg equilibrium.*

Proof. We first focus on the second stage, e.g., AS i determines α_i for given p_B. We argue that for any given p_B, the optimization problem

$$\begin{aligned} \max_{\alpha_i} \quad & u_i(\alpha_i) = V_i(\alpha_i) + P_i(\alpha_i) - p_B\alpha_i, \\ \text{s.t.} \quad & \alpha_0 \le \alpha_i \le 1, \end{aligned} \tag{10}$$

has a unique solution.

The reason is straightforward: $u_i(\alpha_i)$ is a strictly concave function of α_i, and $[\alpha_0, 1]$ is a convex set. Thus, the solution $\alpha_i(p_B) = \arg\max u_i(\alpha_i)$ is unique. As $\alpha_i(p_B)$ and $p_e(p_B)$ are continuous, so is $\alpha(p_B) = \sum_{i \in \bar{B}} \alpha_i(p_B)$. Thus, we have

$$\begin{aligned} \max_{p_B} \quad & u_b(p_B) = 2p_B\alpha(p_B) - C(\alpha(p_B), p_e(p_B)), \\ \text{s.t.} \quad & 0 \le p_B \le p_{B0}, \end{aligned} \tag{11}$$

where p_{B0} is the maximum price B can set. As the above optimization is maximizing a continuous function over a compact set, it must have at least one solution. This guarantees the existence of the Stackelberg equilibrium. □

Theorems 5 and 6 guarantee the existence of steady state economic relationships among all ASes. To further explore how to achieve a large value of α_i, we consider a simple example assuming homogeneous ASes $\forall i \in \bar{B}$. The result shows that, *by including high-tier ISPs into the broker set, lower-tier ISPs become more willing to follow the new rule*.

7.2 Cooperation Among ASes in the Broker Set

So far, we regard all ASes in B as one identity and assumed they cooperate so as to provide most connections to guarantee the E2E connectivity and reduce the path length. Now let us discuss how to achieve such a cooperation. As the cooperation among ASes in B can greatly improve the E2E QoS, B can charge more from other ASes and end users. Our goal is to design a *fair revenue distribution mechanism* to guarantee that no AS in B has an incentive to leave the broker set.

For the cooperative game among ASes in B, we apply the *Shapley value approach* [35] to capture the revenue distribution. Let $U(B)$ denote B's profit under the Stackelberg equilibrium. $U(B) = u_b(p_B^*)$ if p_B^* is the price at the Stackelberg equilibrium. $U(K)$ denotes set K's profit under the equilibrium if $K \subseteq B$ is selected as the broker set. The *marginal contribution* of AS $j \notin K$ to K is defined as

$$\Delta_j(K) = U(K \cup \{j\}) - U(K). \tag{12}$$

Then, AS j's Shapley value is

$$\varphi_j(B) = \frac{1}{|B|!} \sum_{\pi \in \Pi} \Delta_j(\tilde{B}(\pi, j)), \tag{13}$$

where Π is the set of all $|B|!$ orderings of B, and $\tilde{B}(\pi, j)$ is the set of ASes preceding j in the ordering of π.

We assume the revenue distributed to AS $j \in B$ is equal to its Shapley value, for Shapley value satisfies *efficiency*, *symmetry* and *fairness* conditions [36]. Here, efficiency means the sum of Shapley values equals B's total profit. Symmetry means if two ASes contribute equally to any subset of B excluding themselves, their Shapley values are the same. Fairness means for any two ASes in B, their mutual contributions are equal. Authors in [35], [37] provide methods to approximate the Shapley value in Eq. (13).

Based on the above properties of the Shapley value, we have the following theorems to achieve a stable cooperation in B.

Theorem 7. *If $\forall K, L \subseteq B$ and $K \cap L = \emptyset$, $U(K \cup L) \geq U(K) + U(L)$ (superadditivity) holds, then the Shapley value is individually rational, i.e., $\varphi_j(B) \geq U(\{j\}), \forall j \in B$.*

Obviously, the superadditivity condition satisfies in B, since only a full cooperation over B can guarantee the E2E connectivity for the whole network. This condition guarantees the *stability* of the cooperation in B: no AS in B has an incentive to leave for it will not achieve a higher revenue by doing so.

Theorem 8. *If $\forall j \in B$ and $\forall K \subseteq L \subseteq B \backslash \{j\}$, $\Delta_j(K) \leq \Delta_j(L)$ (supermodularity) holds, then the Shapley value satisfies $\sum_{j \in M} \varphi_j(B) \geq U(M), \forall M \subseteq B$.*

If the supermodularity condition holds, the *strong stability* of B can be guaranteed: no subset of ASes has an incentive to leave B and form another coalition, for their small coalition cannot bring more them revenues. This finding gives insights on the *proper size of B*. At the beginning of the formation of B, supermodularity satisfies easily not only because that the first added ASes are super ASes, but also due to the "network externality" effect caused by the cooperation. Yet, when the set size gradually reaches some threshold, most important ASes are already included in B, and new joiners have only marginal contributions, so the supermodularity condition does not hold any more. That's the time to stop increasing the set size.

8 Conclusion

In this paper, we propose an inter-domain routing brokerage framework, in which an inter-AS routing path can be totally dominated by a small set of ASes and IXPs to provide the E2E QoS guarantee. We model the problem as the MCBG problem and prove it to be NP-hard. To address the MCBG problem, we propose a $(\frac{1+e^{-1}}{2})-$approximation algorithm and prove the APX-hardness of the MCBG problem on the $(\alpha, \beta)-$graph. To further improve the computation efficiency and deal with the path length constraint, we design a heuristic algorithm which, compared with the approximation algorithm, has the equivalent good performance while greatly reduces the computational complexity from $O(k^2(|V|\log|V|+|E|))$ to $O(k(|V|+|E|))$. We further investigate the feasibility of deploying the broker set in the current Internet from both structural and economic perspectives, and show that our proposed brokerage framework is capable of providing enough incentive to persuade ASes to follow when cooperating with BGP. We also take the realistic business relationships into consideration and show that with little change to the current AS peering relationships, 72.5 percent E2E connectivity can be served with a high quality assurance by selecting only 2 percent ASes/IXPs as brokers.

Acknowledgments

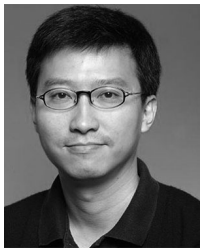
John C.S. Lui was supported in part by the GRF 14200117 and the Huawei Research Fund.

References

[1] D. A. Grier, "Keeping the Internet global," *IEEE Comput.*, vol. 50, 2017, Art. no. 108.
[2] L. Gao and F. Wang, "The extent of AS path inflation by routing policies," in *Proc. IEEE Global Telecommun. Conf.*, 2002, pp. 2180–2184.
[3] R. Martinez-Morais, F. J. Alfaro-Cortes, and J. L. Sanchez, "Providing QoS with the deficit table scheduler," *IEEE Trans. Parallel Distrib. Syst.*, vol. 21, no. 3, pp. 327–341, Mar. 2010.
[4] Cisco visual networking index: Forecast & methodology, 2015–2020. [Online]. Available: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/complete-white-paper-c11-481360.html
[5] T.-I. Kim, H.-W. Jung, et al., "Inter-domain routing based on link state information for end-to-end QoS guarantee," in *Proc. IEEE Int. Conf. Adv. Commun. Technol.*, 2009, pp. 1729–1732.
[6] M. Shirazipour, S. Pierre, and Y. Lemieux, "Inter-domain traffic engineering using MPLS," in *Intelligence in Communication Systems*. Berlin, Germany: Springer, 2005, pp. 43–52.
[7] R. Jacquet, G. Texier, and A. Blanc, "SANP: An algorithm for selecting end-to-end paths with QoS guarantees," in *Proc. IEEE Future Netw. Mobile Summit*, 2013, pp. 1–10.
[8] V. A. Danciu, D. Kranzlmüller, et al., "A border-friendly, non-overlay mechanism for inter-domain QoS support in the internet," *J. Cases Inf. Technol.*, vol. 19, pp. 223–229, 2011.
[9] R. Jacquet, G. Texier, and A. Blanc, "Computing end-to-end QoS paths in the Internet considering multiple alliances," in *Proc. IEEE 16th Int. Telecommun. Netw. Strategy Planning Symp.*, 2014, pp. 1–6.
[10] N. B. Djarallah, N. L. Sauze, et al., "Distributed E2E QoS-based path computation algorithm over multiple inter-domain routes," in *Proc. IEEE Int. Conf. P2P Parallel Grid Cloud Internet Comput.*, 2011, pp. 169–176.

[11] A. D. Yahaya and T. Suda, "iREX: Inter-domain resource exchange architecture," in *Proc. IEEE INFOCOM*, 2006, pp. 1–12.
[12] A. D. Yahaya and T. Suda, "iREX MPO: A multi-path option for the iREX inter-domain QoS policy architecture," in *Proc. IEEE Int. Conf. Commun.*, 2008, pp. 5815–5822.
[13] H. Pouyllau, R. Douville, et al., "Economic and technical propositions for inter-domain services," *Bell Labs Tech. J.*, vol. 14, pp. 185–201, 2009.
[14] H. Pouyllau, R. Douville, et al., "Architecture for inter-domain service delivery," in *Proc. IEEE Netw. Operations Manage. Symp.*, 2010, pp. 748–762.
[15] A. Sprintson, M. Yannuzzi, et al., "Reliable routing with QoS guarantees for multi-domain IP/MPLS networks," in *Proc. IEEE INFOCOM*, 2007, pp. 1820–1828.
[16] F. Racaru, M. Diaz, and C. Chassot, "Quality of service management in heterogeneous networks," in *Proc. IEEE Int. Conf. Commun. Theory Rel. Quality Service*, 2008, pp. 83–88.
[17] V. Kotronis, X. Dimitropoulos, and B. Ager, "Outsourcing the routing control logic: Better internet routing based on SDN principles," in *Proc. ACM Workshop Hot Topics Netw.*, 2012, pp. 55–60.
[18] Z.-L. Zhang, Z. Duan, et al., "Decoupling QoS control from core routers: A novel bandwidth broker architecture for scalable support of guaranteed services," in *Proc. ACM Conf. Appl. Technol. Archit. Protocols Comput. Commun.*, 2000, pp. 71–83.
[19] Z. Duan, Z.-L. Zhang, et al., "A core stateless bandwidth broker architecture for scalable support of guaranteed services," *IEEE Trans. Parallel Distrib. Syst.*, vol. 15, no. 2, pp. 167–182, Feb. 2004.
[20] V. Kotronis, X. Dimitropoulos, R. Klöti, B. Ager, P. Georgopoulos, and S. Schmid, "Control exchange points: Providing QoS-enabled end-to-end services via SDN-based inter-domain routing orchestration," *LINX*, vol. 2429, no. 1093, p. 2443, 2014.
[21] V. Kotronis, R. Klöti, et al., "Investigating the potential of the Inter-IXP multigraph for the provisioning of guaranteed end-to-end services," in *Proc. ACM SIGMETRICS Int. Conf. Meas. Model. Comput. Syst.*, 2015, pp. 429–430.
[22] V. Kotronis, M. Rost, et al., "Stitching inter-domain paths over IXPs," in *Proc. ACM Symp. SDN Res.*, 2016, Art. no. 17.
[23] Data source. 2015. [Online]. Available: http://irl.cs.ucla.edu/topology/
[24] B. Augustin, B. Krishnamurthy, and W. Willinger, "IXPs: Mapped?" in *Proc. ACM SIGCOMM Conf. Internet Meas.*, 2009, pp. 336–349.
[25] J. I. Alvarez-Hamelin, L. Dall'Asta, et al., "Large scale networks fingerprinting and visualization using the k-core decomposition," in *Proc. 18th Int. Conf. Neural Inf. Process. Syst.*, 2005, pp. 41–50.
[26] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher, "An analysis of approximations for maximizing submodular set functions," *Math. Program.*, vol. 8, pp. 73–87, 1978.
[27] Submodular set function. 2017. [Online]. Available: https://en.wikipedia.org/wiki/Submodular_set_function
[28] G. Ausiello, P. Crescenzi, et al., *Complexity and Approximation: Combinatorial Optimization Problems and Their Approximability Properties*. Berlin, Germany: Springer, 2012.
[29] G. L. Nemhauser and L. A. Wolsey, "Best algorithms for approximating the maximum of a submodular set function," *Math. Operations Res.*, vol. 3, pp. 177–188, 1978.
[30] P. Crescenzi and L. Trevisan, "On approximation scheme preserving reductibility and its applications," in *Proc. Int. Conf. Found. Softw. Technol. Theoretical Comput. Sci.*, 1994, pp. 330–341.
[31] A.-H. Esfahanian, "Connectivity algorithms." 2013. [Online]. Available: http://cse.msu.edu/~cse835/Papers/Graph_connectivity_revised.pdf
[32] F. S. Perra Nicola, "Spectral centrality measures in complex networks," *Phys. Rev. E*, vol. 78, 2008, Art. no. 036107.
[33] AS ranking. 2018. [Online]. Available: http://as-rank.caida.org/
[34] M. J. Osborne, *An Introduction to Game Theory*. London, U.K.: Oxford Univ. Press, 2004.
[35] A. E. Roth, *The Shapley Value: Essays in Honor of Lloyd S. Shapley*. Cambridge, U.K.: Cambridge Univ. Press, 1988.
[36] R. B. Myerson, "Graphs and cooperation in games," *Math. Operations Res.*, vol. 2, pp. 225–229, 1977.
[37] S. Maleki, L. Tran-Thanh, G. Hines, T. Rahwan, andA. Rogers, "Bounding the estimation error of sampling-based Shapley value approximation," arXiv preprint arXiv:1306.4265, 2013.

Tingwei Liu received the BE and ME degrees in electronics and information engineering from the Huazhong University of Science and Technology, China, in 2013 and 2016, respectively. She is currently working toward the PhD degree in the Department of Computer Science and Engineering, Chinese University of Hong Kong (CUHK). Her research interests include networking algorithm and protocol design, network availability measurement, and network economics.



John C. S. Lui received the PhD degree in computer science from the University of California, Los Angeles, 1992. He is currently a professor with the Department of Computer Science and Engineering, Chinese University of Hong Kong (CUHK), Hong Kong. He was the chairman of the Department from 2005 to 2011. His current research interests include communication networks, network/system security (e.g., cloud security, mobile security, etc.), network economics, network sciences (e.g., online social networks, information spreading, etc.), cloud computing, large-scale distributed systems, and performance evaluation theory. He serves on the editorial board of the *IEEE/ACM Transactions on Networking*, the *IEEE Transactions on Computers*, the *IEEE Transactions on Parallel and Distributed Systems*, the *Journal of Performance Evaluation* and the *International Journal of Network Security*. He received various departmental teaching awards and the CUHK Vice-Chancellor's Exemplary Teaching Award. He is also a co-recipient of the IFIP WG 7.3 Performance 2005 and IEEE/IFIP NOMS 2006 Best Student Paper Awards. He is a fellow of the Association for Computing Machinery (ACM), a fellow of the IEEE, a croucher senior research fellow, and an elected member of the IFIP WG 7.3.

Dong Lin received the bachelor's degree in computer science from the Beijing University of Aeronautics & Astronautics, in 2005, the master's degree in computer science from Tsinghua University, in 2008, and the PhD degree in computer science and engineering from the Hong Kong University of Science and Technology, in 2012. He was a research group member of Software Development Division, ClusterTech during 2012 and 2014, managed to develop cloud management platforms for both public and private cloud infrastructures. He joined Huawei as a researcher in 2014, focusing on the research areas that related to computer networks, including data center networks, information-centric networks, content delivery networks, complex networks, etc.

David Hui received the BEng (with 1st Class Honor) and MPhil degrees in electrical and electronic engineering from the University of Hong Kong (HKU), in 2004 and 2007, respectively, and the PhD degree in electronic and computer engineering from the Hong Kong University of Science and Technology (HKUST), in 2014. He is currently a researcher of Huawei Technologies. His research interests include stochastic modelling, learning and control of communication networks, with current focus on queueing performance analysis and network slicing design.