

On Route Selection for Interdomain Traffic Engineering

Yang Richard Yang Haiyong Xie Hao Wang Avi Silberschatz
Yanbin Liu Li Erran Li Arvind Krishnamurthy *

Abstract

In this paper, we investigate a model of route selection for interdomain traffic engineering where the routing to multiple destinations can be coordinated. We identify potential routing instability and inefficiency problems, and derive a set of practical guidelines to guarantee stability without global coordination. Using a realistic Internet topology, we show that route oscillations can happen even when a small number of ASes coordinate route selection for just a small number of destinations, if the coordination does not follow our guidelines. We further extend our model so that 1) ASes can adopt any route selection algorithms in a class of algorithms which we call *rational route selection algorithms*; and 2) the local ranking of routes of an AS can depend on ingress traffic patterns. We show that persistent route oscillations can happen in certain network settings even if the ASes strictly follow the constraints imposed by business considerations, and adopt any rational route selection algorithms.

Keywords: Interdomain routing, traffic engineering, routing instability

1 Introduction

The global Internet consists of a large number of interconnected autonomous systems (AS), where each AS is administrated autonomously. Neighboring ASes exchange their routes using the Border Gateway Protocol (BGP), where each route consists of a path vector of ASes from a source to a destination. Upon learning routes from its neighbors, an AS chooses the best routes from all of its available routes, according to its local route selection policies. Recently, Internet Service Providers (ISPs) are increasingly adopting traffic-engineering-based local route selection policies for the ASes under their control (*e.g.*, [10]). Several vendors also provide traffic engineering blackboxes which are deployed [11]. We have recently conducted an email survey of ISPs in [12], and the results indicate that many ISPs choose routes to achieve their interdomain traffic engineering objectives, such as satisfying the capacity constraints of links between neighboring ASes, load-balancing interdomain traffic, and/or minimizing cost.

Despite this emerging trend, so far there are few systematic studies on the stability and efficiency of the global Internet with interdomain route selection for interdomain traffic engineering. A major breakthrough was made recently when Griffin *et al.* [6, 7] proposed systematic models to study the stability of path-vector interdomain routing. In particular, they identified the existence of policy disputes as a potential reason for routing instability. By routing instability, they mean persistent route oscillations even though the network topology is stable. Although these models capture a wide range of interdomain traffic engineering objectives, several aspects of route selection for interdomain traffic engineering still have not been analyzed. First, the previous models assume that the routing decisions for different destinations can be separated. Thus the models apply only to networks where

*Y. Richard Yang, Haiyong Xie, Hao Wang, Avi Silberschatz and Arvind Krishnamurthy are with Yale University. Yanbin Liu is with the University of Texas at Austin. Li Erran Li is with Bell-labs. The authors are listed in reverse alphabetical order. The authors are supported in part by NSF grants ANI-0207399, ANI-0238038 and CNS-0435201. We thank Joan Feigenbaum, Eric Friedman, Aaron Jaggar, Tim Griffin, Vijay Ramachandran, and Jennifer Rexford for valuable comments. We are also grateful to the anonymous reviewers and the shepherd Olivier Bonaventure for their valuable comments.

there is no AS whose routing policies require it to coordinate its route selection among multiple destinations. On the other hand, a fundamental feature of route selection for interdomain traffic engineering in particular and traffic engineering in general is that route selection constraints (*e.g.*, traffic assigned to a link remains within link capacity) and/or objectives (*e.g.*, balancing the load) involve the route selection of multiple destinations. Thus, in route selection for interdomain traffic engineering, whether a route will be chosen for a given destination will depend on what routes are available or chosen for other destinations. For example, if an AS selects routes for each destination independently without considering the chosen/available routes of other destinations, in the worst case it may choose the same access link for all destinations, violating link capacity constraints and/or causing load imbalance. Second, the previous studies focus on the stability of a homogeneous network where each AS runs the same specific interdomain route selection algorithm (*i.e.*, the BGP-based greedy route selection algorithm). However, with increasing usage of route selection for interdomain traffic engineering, route selection algorithms with more sophisticated strategies are likely to be designed and deployed in the Internet. Thus it is necessary to analyze the stability of a *heterogeneous* network where ASes may adopt a larger class of route selection algorithms beyond the greedy strategy. Third, the previous studies focus on local policies which rank only the egress routes; that is, they assume that the local ranking of egress routes at each AS is independent of the inbound traffic pattern of that AS. However, in practice, the local policies of ASes may involve both the egress routes and the pattern of inbound traffic. In the last few years, several traffic-demand-matrix-based traffic engineering algorithms have been proposed. Although such route selection algorithms have been shown to be effective, the evaluations often assume that the route selection of each AS does not affect the inbound traffic, whereas the inbound traffic is likely to change with the chosen egress routes, introducing unexpected interactions. Thus it is necessary to analyze the stability of route selection algorithms implementing local policies that take into account inbound traffic patterns.

In this article, we summarize some of our recent results on analyzing route selection for interdomain traffic engineering; for formal models and detailed analysis, we refer interested readers to [12]. In Section 2, we analyze a general model to capture route selection for interdomain egress traffic engineering where route selection among multiple destinations is coordinated. We first identify that there exist networks where the interaction between route selection among multiple destinations can cause routing instability, even though the route selection of the networks is guaranteed to converge when each destination is considered alone. We then propose a set of practical guidelines, and show that if these guidelines are followed by the ASes, route selection for interdomain traffic engineering will be stable without explicit global coordination. In this section, we also conduct simulations using a realistic Internet AS topology to show that even with a small number of ASes coordinating route selection for just a small number of destinations, if the coordinated route selection does not follow our guidelines, we can observe instability. In Section 3, we study a more general route selection model for interdomain traffic engineering. There are two extensions in this general model. First, instead of studying a specific route selection algorithm, we allow ASes to choose any algorithm from a class of algorithms. Specifically, since we are modeling the route selection behaviors of self-optimizing ASes, our only requirement is that it should be “unjustified” or “irrational” for a self-optimizing AS to choose an inferior route infinitely often when there are better routes available; thus we refer to the class of algorithms we study as *rational route selection algorithms*. Second, in this more general model, we allow the local ranking of routes of an AS to depend on not only its routes to the destinations but also its ingress traffic patterns. We show that there are networks which will be unstable when the ASes strictly follow AS business guidelines, and adopt any rational route selection algorithms. Our conclusion and future work are in Section 4.

2 Route Selection for Egress Interdomain Traffic Engineering

2.1 Motivation

We start with a very simple illustrative example as shown in Figure 1. Suppose the majority of the traffic of S goes to two destinations D_1 and D_2 . Assume S wants to balance its outgoing traffic to its two neighbors A and B . Thus, it wants to choose a combination of routes for destinations D_1 and D_2 such that they use different neighbors, if possible, in order to have low utilization on the two links SA and SB . We refer to a combination of routes for D_1 and D_2 as a *route profile*. Since S may not know in advance the routes it will learn from its neighbors, and the routes that A and B export to S can change due to network dynamics, S needs an automatic method to pick the best route profile, according to currently available routes. One method by which S can specify its local ranking of interdomain routes is to define an interdomain traffic engineering objective function (e.g., minimize the maximum of the utilization of the two links for this case). An advantage of using an objective function is its compact representation. Given the objective function, link capacities, and traffic demands, a traffic engineering program searches for the best route profile automatically and dynamically, according to currently available routes. The local ranking of interdomain routes can also be specified by a policy language. An example policy can be: if D_1 and D_2 use different links, assign a base local preference value of 100; otherwise, a base local preference value of 0. If D_1 uses link SA , add 10 to local preference value. If D_2 uses link SB , add 5 to local preference value. The program picks the available route profile with the highest local preference value. For generality, we assume a ranking table at each AS, which lists, in decreasing order, all of the potential route profiles. An example route ranking table for S is shown in Figure 1, where each row is a route profile, i.e., a combination of routes for D_1 and D_2 . For example, the best route profile for S is (SAD_1, SAD_2) ; i.e., S uses SAD_1 for destination D_1 , and SAD_2 for destination D_2 . The worst route profile is (SBD_1, SBD_2) . Thus, if the route profile (SAD_1, SAD_2) is available, S will choose it. On the other hand, if the only available route profile is (SBD_1, SBD_2) , S has no choice but to use it.

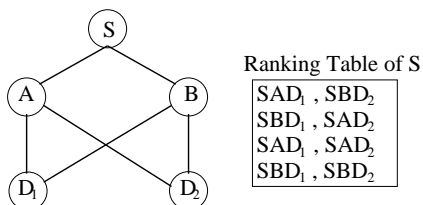


Figure 1: Egress load balancing: an example motivating the need for coordinated route selection.

2.2 Problem Definition

To simplify our exposition, we make the following assumptions. We assume a connected network with a set \mathcal{S} of source ASes and a set \mathcal{D} of destination ASes. We assume an AS-level topology, and leave an investigation of the interaction of intradomain routing and interdomain routing as future work (e.g., [3]). We assume that the underlying network infrastructure is stable (i.e., no links up and down). We focus on route selection and thus assume that the export policies (i.e., which routes to export to which neighbors) are static. We assume that the local ranking of route profiles of the ASes are fixed during our analysis so that we can focus on the interactions among local routing policies. In practice, traffic patterns may change due to factors such as diurnal trends; as a result, ASes may adjust their rankings. We also leave such dynamics as future work.

We first consider the case in which the ranking of interdomain routes of an AS depends only on the routes from the AS itself to the destinations. In other words, the ASes are conducting *egress interdomain traffic engineering*, which is one of the major tasks of ISP interdomain traffic engineering [1]. In Section 3 we will

extend this model and study route selection for general interdomain traffic engineering, where the route from each source to the AS itself also matters.

We now define the *stable route selection for egress interdomain traffic engineering problem*. We define for AS i the set of all potential routes to all destinations as $R_i = \prod_{d \in \mathcal{D}} R_{i \rightarrow d}$, where $R_{i \rightarrow d}$ is the set of all possible routes from i to destination d . We allow empty routes. We refer to an element $r_i \in R_i$ as a *route profile* of i , since r_i completely specifies the routes from i to each destination. The ranking of interdomain routes of i on routes in R_i is represented by a ranking table of route profiles. Specifically, there is a ranking function \mathcal{R}_i which maps a set of routes R_i to a totally ordered set Λ ; i.e., $\mathcal{R}_i : R_i \rightarrow \Lambda$. The ranking function \mathcal{R}_i can be i 's traffic engineering objective function. Hereafter, we assume that i 's ranking of interdomain routes is given in the form of a ranking table. We emphasize that our introduction of ranking tables is conceptual and solely for the purpose of analysis. Ranking tables are just a general representation of some more compact representations of route selection methods such as objective functions or policy languages. If the route selection behavior of an AS is consistent when faced with different sets of available routes, a ranking table can be constructed accordingly.

An AS uses its route ranking table to select the best available routes. Figure 2 shows the standard, greedy BGP protocol/process model of interdomain route selection [6, 7], naturally extended to multiple destinations. Each AS maintains a routing cache of currently available routes for each destination, exported by its neighbors. AS i selects routes from its routing cache, one route $r_{i \rightarrow d}$ for each destination d , so that the chosen route profile r_i has the highest rank; i.e., $\mathcal{R}_i(r_i) > \mathcal{R}_i(r'_i)$, for any other route profile r'_i available from the routing cache. This chosen route profile r_i will then be used by i to route packets. If $r_{i \rightarrow d}$ is different from the previously selected route to d , i then withdraws the previous route, and exports the new route to the neighbors that are allowed to receive this route under i 's export policy. We assume that BGP route update messages between neighboring ASes are reliably delivered in FIFO order. This is reasonable as the messages are sent via TCP. We also assume that each message will be processed with bounded delay.

A *network route selection* is a combination of route profiles, one for each AS. A network route selection is *stable* if no AS can choose a higher ranked route profile from the exported routes of its neighbors. We also call a stable network route selection a *stable route solution* or *solution* for short.

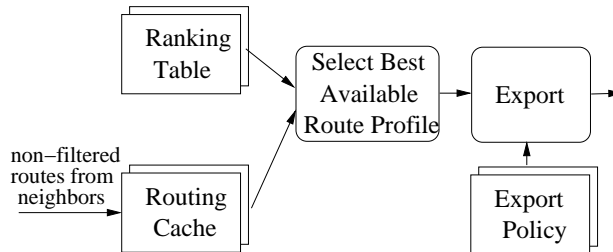


Figure 2: The BGP-based greedy protocol/process model of route selection for interdomain egress traffic engineering.

2.3 Multi-Destination Egress Traffic Engineering Can Cause Instability

A somewhat unexpected result is that the interaction of the routing to multiple destinations due to interdomain traffic engineering can cause routing instability. The network shown in Figure 3(a) is one such interesting example. For clarity, we show only the highest-ranked three route profiles of A and B . To make this example more realistic, the ASes export their routes according to their business relationship. There are two major types of business relationship in the current Internet. The first type is the provider-and-customer relationship, where a provider provides transit service to its customers. We refer to the connection from a provider to a customer as a provider-to-customer link; such a link is represented by a directed edge from the provider to the customer. The

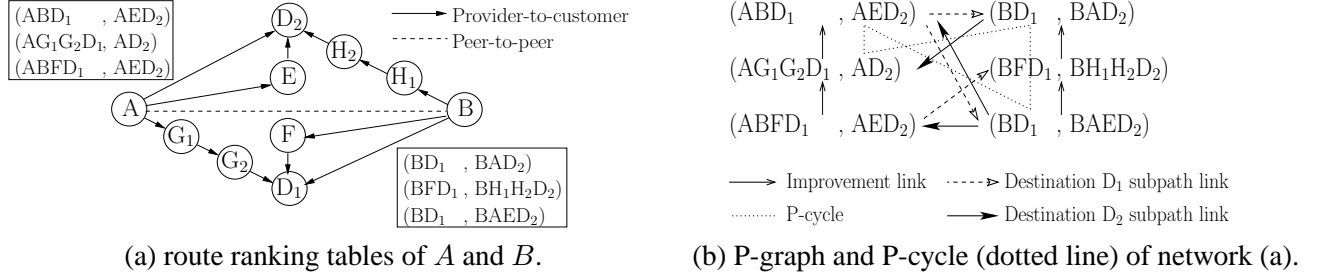


Figure 3: A network which has no stable route selection.

second major type of business relationship is the peer-to-peer relationship, where a pair of ASes provide transit services to each other's customers. We refer to the connection between a pair of peers as a peer link; such a link is represented by a dashed edge between the two peers. These business relationships imply that the export policies of ASes in the Internet follow the *typical export policies* [5]: 1) each AS exports to its providers its own routes and those learned from its customers, but not those learned from its peers or other providers; 2) each AS exports to its customers its own routes and any routes learned from others; 3) each AS exports to its peers its own routes and those learned from its customers, but not those learned from its providers or other peers.

We first consider each destination separately. For destination D_1 , A has ABD_1 and $AG_1G_2D_1$, and B has BD_1 and BFD_1 , respectively, as the two highest ranked route profiles. Given this combination of route preference for D_1 , the network has a stable route solution of ABD_1 and BD_1 for A and B , respectively. One can also verify that if we consider D_2 alone, the network has a stable route solution of AED_2 and $BH_1H_2D_2$ for A and B , respectively. Thus, if there were no interaction among destinations, A and B would settle on the stable solutions of (ABD_1, AED_2) and $(BD_1, BH_1H_2D_2)$, respectively.

Next we consider coordinated route selection for both destinations. The above solutions obtained by considering each destination alone are no longer stable. For example, B will not choose $(BD_1, BH_1H_2D_2)$ since this route profile has a lower rank. Note that for clarity, we show only the highest three route profiles of A and B in Figure 3. One can verify that the network has no stable solution at all. Specifically, we observe that the export policies of the ASes make the route profile $(AG_1G_2D_1, AD_2)$ always available to A . Thus to see that the network has no stable solutions, we just need to verify that there is no stable route solution when A chooses $(AG_1G_2D_1, AD_2)$ or (ABD_1, AED_2) . Clearly, there is no stable solution for $(AG_1G_2D_1, AD_2)$ since if A chooses $(AG_1G_2D_1, AD_2)$, B will choose (BD_1, BAD_2) ; this causes A to change to (ABD_1, AED_2) . However, there will be no stable route selection for (ABD_1, AED_2) , either. To make (ABD_1, AED_2) available to A , B must choose BD_1 for D_1 . Since $(BFD_1, BH_1H_2D_2)$ is always available to B , it must be the case that B chooses (BD_1, BAD_2) . However, this requires A to choose AD_2 , which is inconsistent with (ABD_1, AED_2) . Thus, the network has no stable route selections due to destination interaction! Furthermore, note that the instability above is with only two destinations and a few ASes. Traffic engineering techniques would in practice play with tens or hundreds of destinations over the Internet composed of thousands of ASes. Thus, the potential for instability is likely to be much higher. As verification, in Section 2.7, we will show that under realistic Internet topologies, the coordination of route selection by a small number of ASes for a small number of destinations can lead to instability.

2.4 Stable, Robust Egress Traffic Engineering and Protocol Convergence

Given that multi-destination interaction due to interdomain traffic engineering can result in no stable route selection, in this section, we derive a sufficient condition that can guarantee stable, robust route selection and protocol convergence.

We first introduce the notion of a *P-graph* to capture the interaction between the interdomain traffic engineer-

ing policies of multiple ASes. The notion of a P-graph is motivated by the partial order graph of Griffin *et al.* [6], but generalized to interdomain traffic engineering. The nodes of a P-graph are from the union $\bigcup_i R_i$, namely, all route profiles of all ASes. We consider only route profiles that are allowed by export policies. There are two types of directed edges in a P-graph. The first type of edges are improvement edges. There is an improvement edge from node r_i to r'_i if i prefers route profile r'_i to r_i . The second type of edges are sub-path edges. There is a destination D_j sub-path edge from a node r_i to another node r_j if the path in r_j for destination D_j is a sub path of that in r_i . A *P-cycle* is a loop in the P-graph of the following special format: one or more improvement edges, followed by one or more sub-path edges of the same destination, followed by one or more improvement edges, and so on. For example, Figure 3(b) shows the P-graph and the P-cycle for the network of Figure 3(a). Note that there may be trivial loops in a P-graph which are not of the format of a P-cycle. For example, the loop consisting of (BD_1, BAD_2) , $(AG_1G_2D_1, AD_2)$ and (ABD_1, AED_2) is not a P-cycle, since there are two consecutive sub-path edges of different destinations.

As our following theorem shows, if there is no P-cycle, the BGP protocol will converge. We refer interested readers to [12] for a proof of this theorem.

Theorem 1 *If the P-graph has no P-cycle, then the BGP protocol converges.*

2.5 Multi-Destination Egress Traffic Engineering Can Cause non-Pareto Optimal Solution

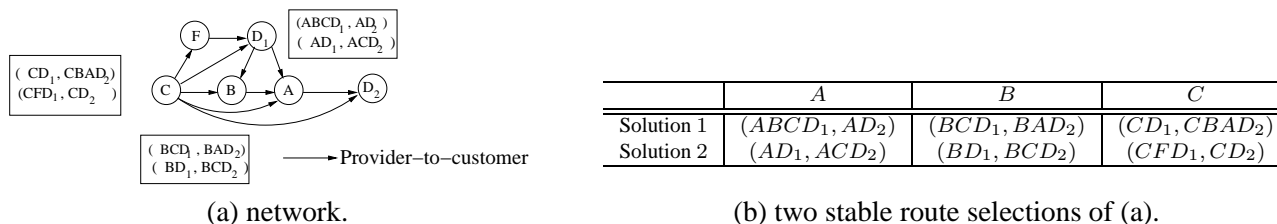


Figure 4: An example with two solutions but one of them is not Pareto optimal.

A network with stable solutions can have multiple solutions. It is important that a stable route selection for interdomain traffic engineering be Pareto optimal; namely, that there does not exist another stable route solution where each AS has a higher ranked route profile. The requirement of Pareto optimality is a fundamental requirement in economics and social welfare theory. One can show that BGP satisfies the Pareto optimality property in the domain of *strict route preferences*. In other words, when the routing to multiple destinations is not coordinated, any stable route selection computed by BGP protocol is Pareto optimal.

However, when route selection among multiple destinations is coordinated, a stable route selection for interdomain traffic engineering can be non-Pareto optimal. The example in Figure 4(a) is one such interesting example. This example has two stable route solutions, as shown in Figure 4(b), and the solution at the second row is not even Pareto optimal. This example clearly demonstrates that to be effective, explicit coordination of route selection (*e.g.*, negotiation [8]) may involve more than two parties.

2.6 Stable Egress Route Selection for Interdomain Traffic Engineering without Global Coordination

In Section 2.4, we give a sufficient condition to guarantee the existence of and convergence to a stable route selection. The condition depends on checking for P-cycle. In practice, it is difficult to obtain the P-graph and check whether it contains a P-cycle. This is due to the fact that BGP is a distributed protocol, and generally ASes do not share their traffic engineering policies. Furthermore, the preceding section considers general networks,

while in the current Internet, route selection of ASes is constrained by their business relationships. In this section, we seek rigorous, practical interdomain traffic engineering guidelines that are reasonable according to current AS business relationships, can be checked locally, and can guarantee convergence. In other words, if each AS follows these practical interdomain traffic engineering guidelines, route selection is stable and the converged route selection is unique. We are particularly motivated by the study of Gao and Rexford [5], which proposed guidelines to guarantee route convergence without global coordination in the practical setting of the Internet if each destination is considered separately. When ASes coordinate route selection among multiple destinations, we find that to guarantee the existence and uniqueness of stable route selection for interdomain traffic engineering, we need stronger conditions.

We assume that ASes follow the typical export policies (please see Section 2.3). Such export policies imply that all valid routes have the following patterns [5]: a provider-to-customer link can be followed by only provider-to-customer links, and a peer link can be followed by only provider-to-customer links. Accordingly, we divide the routes from an AS i to a destination d into three categories:

- *Customer route*: each link along a customer route is a provider-to-customer link.
- *Peer route*: the first link along a peer route is a peer link, and the remaining links are all provider-to-customer links.
- *Provider route*: the first link is a customer-to-provider link, and the remaining route consists of zero or multiple customer-to-provider links, followed by zero or one peer link, and then zero or multiple provider-to-customer links.

We also divide the set of destinations of an AS i into two categories, according to the given network topology and export policies:

- *Customer-reachable destinations*: these destinations are direct or transitive customers of AS i .
- *Peer-provider-reachable destinations*: these destinations are direct or transitive customers of one of AS i 's peers, or direct or transitive providers, but they are not direct or transitive customers of AS i .

Given the above definitions of different types of routes, Gao and Rexford [5] observe that business considerations imply that ASes prefer customer routes over peer/provider routes. We call such route preference, namely, customer routes \succ peer/provider routes, the *standard individual-route preference policy*. Assuming the typical export policies, the standard individual-route preference policy, together with the assumption that there are no provider-customer loops in the business relationships formed by ASes, Gao and Rexford prove that these conditions guarantee convergence in the global Internet. However, a potential issue in their analysis is that their route selection model assumes that there is no coordination among destinations, while in the current Internet, ISPs are increasingly adopting coordinated route selection policies to achieve their interdomain traffic engineering objectives. Therefore, we need to re-evaluate how the standard individual-route preference policy will change if an AS coordinates between its routes to multiple destinations.

Our key guideline is that routing decisions for different categories of destinations be *decomposed*. Theorem 1 motivates us to find practical conditions to prevent the formation of P-cycles. On the other hand, Figure 3 indicates that when different types of routes are coordinated, there may exist P-cycles. Therefore, we require some guideline as a constraint on local routing policies to prevent P-cycles. Specifically, we say that the routing decisions of an AS are *decomposed* if the following condition is satisfied: AS i 's routing decision for customer-reachable destinations depend only on the routing decisions for its other customer-reachable destinations, and are independent of the routing decisions for its peer-provider-reachable destinations; similarly, AS i 's routing decisions for its peer-provider-reachable destinations are independent of those of its customer-reachable destinations.

When the routing decisions of AS i are decomposed for customer- and peer-provider-reachable destinations, we say that it follows *the standard joint-route preference policy*.

We now have the following theorem if the above guidelines are followed by all ASes:

Theorem 2 *The network has a unique stable route selection which BGP is guaranteed to converge to, if the following conditions hold:*

1. *there is no provider-customer loop in the network;*
2. *all ASes have fixed typical export policies;*
3. *the routing decisions for customer-reachable and peer-provider-reachable destinations follow the standard joint-route preference policy.*

A proof by contradiction can be constructed to prove the above theorem. Specifically, we can prove that a network does not have P-cycle in its P-graph when all conditions in Theorem 2 are satisfied. We refer interested readers to [12] for a detailed proof.

2.7 Simulation Studies of Instability of Egress Traffic Engineering

The preceding sections analyze the stability of route selection for interdomain traffic engineering. In this section, we use simulations to study the likelihood of routing instability when the conditions of Theorem 2 are not satisfied.

Methodology

We construct the AS topology of the Internet using the BGP table of University of Oregon Routeviews and the BGP tables of 18 Looking Glass servers. In order to make the simulations more efficient, we iteratively remove 6157 leaf ASes (degree 1 nodes) and their links from the topology. The remaining network has 13,048 ASes and 37,999 links. We infer business relationships among the ASes to produce the *AS business relationship graph*. We refer interested readers to [12] for details.

An important component of our simulation studies is route ranking tables. For AS i who does not coordinate the route selection among multiple destinations, we use the subjective routing framework to construct its route ranking table [2]. The subjective routing framework is motivated by the observation that different ASes often use different performance metrics in comparing routes. Thus, in this framework, there is a set M of performance metrics assigned to each link. Each AS computes the cost of a route using its own set of weights. Specifically, AS i has a set of weights, $W_i = \{w_{i,m} | m \in M\}$, where $w_{i,m}$ is the weight associated with the performance metric m . Note that $w_{i,m} = 0$ if i is not concerned with the metric m . Let $C_l^{(m)}$ be the value of metric m at link l . Given a route $r_{i \rightarrow d}$ from AS i to destination d , AS i computes the cost of this route as $c(r_{i \rightarrow d}) = \sum_{m \in M} w_{i,m} \sum_{l \in r_{i \rightarrow d}} C_l^{(m)}$. For each destination, AS i chooses the route with the lowest subjective cost as its best route for that destination.

For an AS i who coordinates its route selection of multiple destinations, we construct its ranking table as follows. First, for each destination d , we compute the set $R_{i \rightarrow d}$ of all feasible routes from i to d , assuming all ASes have typical export policies. Then we construct the set of all possible route profiles $R_i = \prod_{d \in \mathcal{D}} R_{i \rightarrow d}$. For efficiency, we do not explicitly store R_i ; instead, we store just the set of all feasible routes to all destinations (i.e., $\cup_{d \in \mathcal{D}} R_{i \rightarrow d}$), and assign a unique ID to each route in this set; therefore, we represent a route profile using a set of IDs corresponding to the routes in the route profile. Finally, we construct the ranking table of AS i by randomly permuting the entries of R_i .

We implement our own event-driven simulator to study the stable route selection problem for interdomain traffic engineering. We refer interested readers to [12] for more details. The simulator simulates the BGP

protocol process including route import/export, route announcement/withdrawal, and so on. Each AS selects its routes as described above. We also add random delays (in simulation time units) to route import/export events in order to simulate network asynchronicity.

To detect instability, our simulator keeps a history of its selected route profiles for each AS. Specifically, according to its route selection history, each AS constructs a directed stability graph with each node representing a unique route profile and each directed edge representing a temporal transition between two route profiles. An AS has no stable route selection if all nodes in the stability graph are in one single strongly connected component. Hereafter, we refer to such ASes as *unstable* ASes. Since this condition is a sufficient condition, we may underestimate the extent of instability. In order to avoid mistaking initial route exchanges for unstable route selection, we wait for a long enough time before checking instability. Specifically, we start to keep a history of previous best route profiles for each AS after 500 simulation time units when all ASes have routes to all destinations. We start to check the instability condition for each AS every 20 simulation time units after the routing history starts. We run the simulation for 7,000 simulation time units so that the number of ASes identified as unstable does not change any more, and take this number as the number of unstable ASes.

Routing Instability Caused by Route Coordination

We investigate routing instability caused by coordinated route selection among multiple destinations. The experiment is set up as follows. We start with a candidate set consisting of a randomly chosen Tier-2 AS, and grow the candidate set by randomly choosing the neighboring ASes of the candidates with probability 0.5. This process continues until the set consists of a sufficient number of ASes. We also randomly choose a small set of ASes as the destinations. All ASes in the candidate set coordinate their route selections for this common set of destinations. We choose the candidate ASes in this way so as to model a scenario where ASes are more likely to coordinate route selections when their neighbors are doing so. To investigate the potential seriousness of the problem, we set up the experiments so that only 40 ASes coordinate route selection for only 2 destinations and violate the standard joint-route preference policy. All remaining ASes select routes for each destination separately. We also repeat the experiment with different random seeds and obtain the cumulative fraction of unstable ASes.

We study the following two cases: (a) the remaining ASes strictly follow the standard individual-route preference; and (b) the remaining ASes violate the standard individual-route preference with probability 0.03. Figure 5 shows the empirical distribution of the number of unstable candidate ASes for both cases. We conduct a distribution fitting and find that the negative binomial distribution best fits the empirical distributions, as shown in the figures. We observe in case (a) that in worst cases, almost all 40 candidate ASes are unstable in the network. This result is surprising in that even when a small percentage (40 out of 13,048) of ASes coordinate their routes to just two destinations, the network might be unstable. Furthermore, although Figure 5(a) shows that the probability of instability due to route coordination is low, the number of ASes that might be affected by instability can still be very high, given the fact that the Internet consists of a much larger number of ASes. We also vary the number of ASes who coordinate route selection and the number of destinations. We observe that the number of unstable ASes further increases as the number of ASes who coordinate route selection but do not follow the joint-route preference policy increases. We also observe in case (b) that the number of unstable ASes strictly increases when the remaining ASes violate the standard individual-route preference.

3 Route Selection for General Interdomain Traffic Engineering

3.1 Motivation

In the preceding section, each AS runs the standard, greedy BGP route selection algorithm; that is, an AS always picks the best currently available routes. Also, the ranking of egress route profiles of an AS depends only on the egress route profiles and is independent of ingress traffic demand patterns. However, as the two examples below

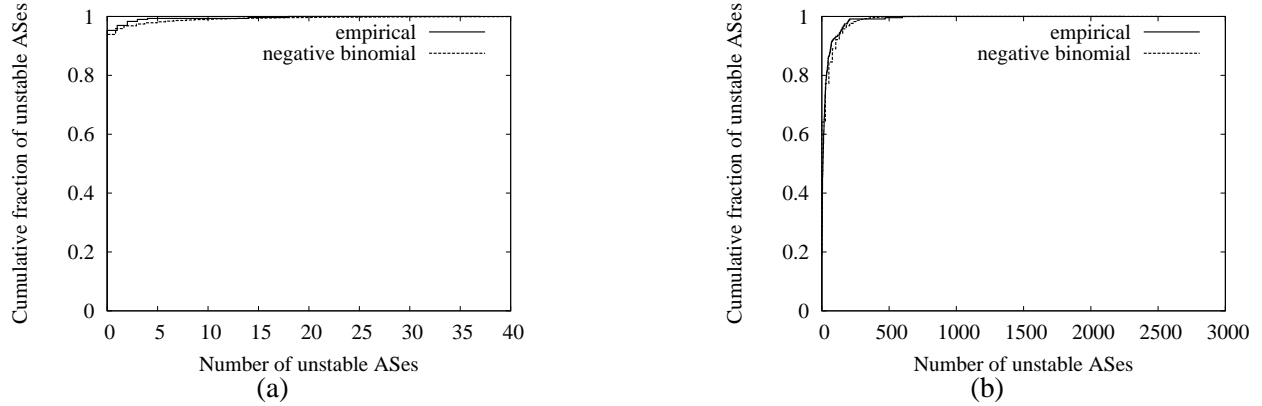


Figure 5: Distributions of total number of unstable ASes due to violation of the standard joint-route preference policy.

illustrate, ASes may adopt more sophisticated route selection algorithms.

A Non-Greedy Strategy Can Perform Better

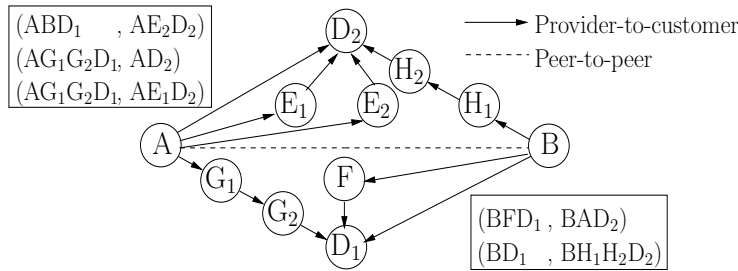


Figure 6: Illustration of a non-greedy route selection strategy.

We first give an example where an AS may achieve a better outcome by adopting a non-greedy route selection algorithm. In Figure 6, to meet the traffic engineering needs, both A and B coordinate the ranking of routes to two destinations D_1 and D_2 . The ranking tables are shown in the two boxes. Let's assume that B uses the greedy strategy. Suppose both A and B start with empty routes, and A announces its routes first. If A uses the greedy strategy, it will select and announce $(AG_1G_2D_1, AD_2)$. This will result in B selecting and announcing (BFD_1, BAD_2) , and the network becomes stable. However, if A selects and announces inferior routes $(AG_1G_2D_1, AE_1D_2)$ to B , B will select and announce $(BD_1, BH_1H_2D_2)$ to A . This provides an opportunity for A to select the highest ranked routes (ABD_1, AE_2D_2) as its stable route selection. This non-greedy route selection strategy is better for A than the greedy strategy.

Inbound-Dependent Route Ranking and Selection

We next give an example to show that it can be reasonable for an AS to rank routes according to inbound traffic patterns. The example also shows that a trivial extension of the greedy route selection algorithm to the scenario of inbound-dependency can lead to instability.

Figure 7(a) is constructed in such a way that the export policies and the route ranking tables of the ASes follow standard AS business assumptions: ASes follow the typical export policies, and prefer customer routes over provider routes. Note that the example avoids peer links in order to have a clean setup. The special feature of this example is that the ranking of AS B , who is one of the two competing providers of S , now depends on *outcomes*, instead of route profiles. An outcome consists of both route selection and ingress traffic pattern for an

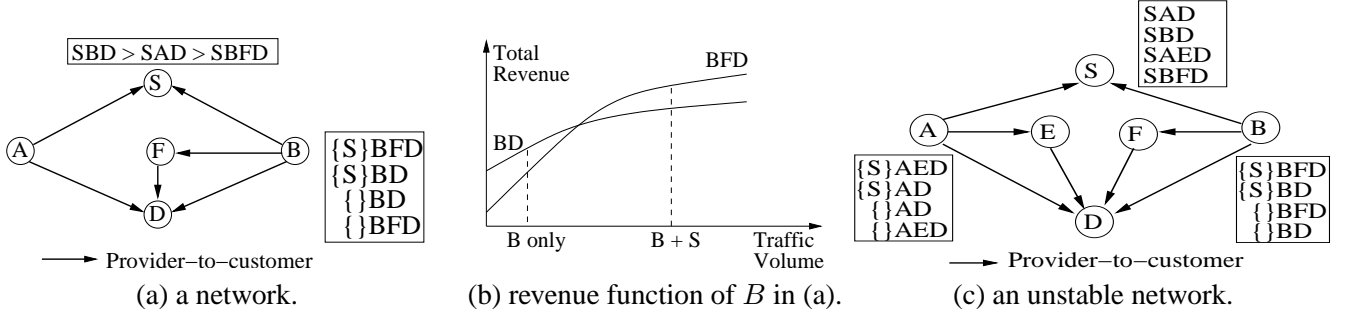


Figure 7: Ingress-dependent traffic engineering.

AS. Specifically, $\{S\}BFD$ denotes the outcome that B uses the route BFD and S sends traffic for destination D through B ; $\{\}BD$ denotes the outcome that B uses the route BD and S does not send any traffic through B . This example can well happen in practice. The ranking table of S is constructed according to the standard BGP decision process: S prefers routes with small hop counts; and for routes with the same hop count, it uses the next-hop ID to break the tie. As for B , note that B prefers traffic from a customer than no traffic. Thus it is a typical ISP behavior. Also note that when S sends traffic through B , the route BFD is preferred than the route BD ; otherwise, the route BD is preferred. A potential revenue function that may cause this scenario to happen is shown in Figure 7(b); that is, BFD is more profitable for B when the traffic volume is high, while BD is more profitable for B when the traffic volume is low.

Given this example, we consider how a BGP-like greedy route selection algorithm will perform. Specifically, B could follow a simple greedy algorithm: assume the current ingress traffic pattern, pick the available route such that the current ingress pattern and the chosen route have the highest rank. Using this algorithm, assume initially S does not use B . Then B first picks BD . Since B chooses BD , S chooses the route SBD , and the traffic from S arrives at B . Since B likes to use BFD when it has high traffic volume, it switches to BFD . Then S chooses SAD , and S no longer uses B . Thus B switches back to BD and we have a loop. This loop is an instance of what we call trivial instability. The above trivial instability is due to the fact that B uses a route selection algorithm which does not keep state to learn the outcome of choosing BD or BFD .

3.2 Rational Route Selection Model

The above two examples motivate the need to study the stability of route selection where ASes may adopt route selection algorithms other than the BGP greedy algorithm. A major challenge, however, is to identify the class of algorithms to investigate. Inspired by previous work on adaptive learning [9] and learning on the Internet [4], in this article, we study a class of route selection algorithms which we call *rational route selection algorithms*. The only condition we impose on this class of route selection algorithm is that, asymptotically, an algorithm in this class will not choose route profiles that are known to be inferior to some other available route profiles. Since we are modeling self-optimizing ASes, we feel that this is a very generic characterization of such ASes. To capture such generic behavior, we avoid any detailed specification of how the ASes actually select route profiles. Instead, we focus on the sequence of network route selections over time, and define the class of algorithms we consider by identifying the general properties of the sequences generated by the route selection algorithms.

We now describe rational route selection algorithms in more detail. We first define the notion of overwhelmed route profiles. Suppose that AS i has observed a history H of network route selections. If this history is long enough for AS i to believe that it has observed all possible route profiles that will be used by each other AS in the future. Suppose AS i has two route profiles r_i and r'_i . If, whenever r_i is available, r'_i is also available and choosing r'_i always yields strictly higher payoff than r_i , then it would be “unjustified” or “irrational” for i to choose r_i . In this case, r_i is said to be *overwhelmed* by r'_i with respect to H , and is called an overwhelmed route

profile. A rational route selection algorithm is one where recursively, overwhelmed route profiles are no longer chosen. In particular, we show in Theorem 3 that the BGP protocol defined in Section 2 is an instance of rational route selection algorithms given that some mild conditions are satisfied. We refer interested readers to [12] for a detailed proof.

Theorem 3 *The BGP protocol defined in Section 2 is an instance of rational route selection, if the following conditions are satisfied:*

1. *BGP update messages between neighboring ASes are delivered reliably in FIFO order, and have bounded delay;*
2. *Each AS sends out BGP update messages in bounded time after it updates its route profile;*
3. *Each BGP update message is processed immediately.*

3.3 Optimal and Stable Inbound-dependent Rational Route Selection by a Single AS

We now return to our motivating example in Figure 7(a). There is a simple rational route selection algorithm that can choose the optimal route and maintain stability for B , if B does not restrict its route selection algorithm to always use the greedy strategy. This algorithm consists of an experimentation phase and a selection phase. At the beginning, B does not know the outcome associated with choosing BD or BFD , thus it will first experiment with these two actions, one at a time. In this phase, B will fix its chosen action for a sufficient amount of time so that it can observe the outcome associated with this chosen egress route (we assume that S will respond to B 's chosen egress route in bounded time). Note that this simple algorithm conforms to the definition of rational route selection. On the other hand, the greedy algorithm does not since it chooses BFD infinitely often which is overwhelmed by BD . One can generalize this example to more general networks.

3.4 Instability of a Network Under Any Rational Route Selection

After introducing the notion of rational route selection algorithms, we now study whether a network consisting of ASes running rational route selection algorithms for general interdomain traffic engineering (*i.e.*, ranking depends on both an AS's route profile and the route profiles of other ASes) has stable route selections.

Definition 1 *A network consisting of ASes each of which is running an rational route selection algorithm has a stable route selection if the route selection of each AS is a single route profile, as time goes to infinity.*

In the above definition, we require that, in a stable route selection, the route selection of each AS be a "pure" routing decision. We do not allow mixed strategies (*i.e.*, a random combination of routes), since mixed strategies involve frequent route fluctuations, and are thus not desirable as "stable" solutions for interdomain routing.

Under the general rational route selection scheme, there are well-behaved network setups with no stable route selection. Figure 7(c) shows an example where no rational route selection algorithm can converge to a stable route selection. It is motivated by the wide spread usage of multihoming. The setup is constructed to satisfy all standard ISP business relationship constraints so that under previous route selection models [5] there is a unique stable route selection. We observe the following instability when ASes use rational algorithms. When AS A and B choose AD and BFD . The outcome is SAD since S ranks SAD higher than $SBFD$. A has an incentive to change from AD to AED since A ranks $\{S\}AED$ higher than $\{S\}AD$. However, AS B realizes that, it can achieve a better outcome by changing BFD to BD since S will choose SBD over $SAED$. This in turn triggers A to switch from AED back to AD . Thus we end up with A choosing AD and B choosing BFD again, and the process continues.

4 Conclusions and Future Work

In this article, we report the results of our study on the stability and efficiency of using route selection to achieve interdomain traffic engineering objectives. We show that interdomain traffic engineering requires that route selection be coordinated over multiple destinations. We show that the interaction among the routing to multiple destinations can cause routing instability even if the routing to each destination individually does indeed have a unique solution. Taking into account business relationships among ISPs in the current Internet, we analyze a set of practical interdomain traffic engineering guidelines and show that if every AS follows them, the existence and uniqueness of stable route solutions in interdomain egress traffic engineering are guaranteed. Using a realistic Internet AS topology, we show that if the guidelines are violated, even when a small number of ASes coordinate their routes for just two destinations, instability could happen.

Despite the success of the analysis and the guidelines, we also show that route selection for interdomain traffic engineering is an extremely important but challenging subject. In a more general model where the local ranking of routes of an AS depends on both egress routes and ingress traffic patterns, we derive an important negative result: there are networks which will be unstable under any rational route selection algorithms where inferior routes are iteratively eliminated. There are many avenues for future work. In particular, although we propose a set of practical guidelines to guarantee convergence, ISPs may still have no incentives to follow these guidelines. How to design incentive-compatible interdomain routing protocols which can guarantee convergence in the most generic setting is a major remaining challenge. The negative result is particularly troubling in that it suggests a fundamental trade-off between local optimality of each AS and global stability. Thus to have a stable, incentive-compatible route selection protocol, the ASes must be willing to look into the future and sacrifice some of the short-term benefits of the current BGP model.

References

- [1] N. Feamster, H. Balakrishnan, and J. Rexford. Some foundational problems in interdomain routing. In *Proceedings of Third Workshop on Hot Topics in Networks (HotNets-III)*, San Diego, CA, Nov. 2004.
- [2] J. Feigenbaum, D. Karger, V. Mirrokni, and R. Sami. Subjective-cost policy routing. Technical Report YALEU/DCS/TR-1302, Yale University, Sept. 2004.
- [3] A. Feldmann, O. Maennel, Z. M. Mao, A. Berger, and B. Maggs. Locating Internet routing instabilities. In *Proceedings of ACM SIGCOMM '04*, Portland, OR, Aug. 2004.
- [4] E. Friedman and S. Shenker. Learning and implementation on the Internet. Working paper. Available at <http://www.orie.cornell.edu/~friedman/pfiles/decent.ps>, 1997.
- [5] L. Gao and J. Rexford. Stable Internet routing without global coordination. *IEEE/ACM Transactions on Networking*, 9(6):681–692, Dec. 2001.
- [6] T. G. Griffin, A. D. Jaggard, and V. Ramachandran. Design principles of policy languages for path vector protocols. In *Proceedings of ACM SIGCOMM '03*, Karlsruhe, Germany, Aug. 2003.
- [7] T. G. Griffin, F. B. Shepherd, and G. Wilfong. The stable paths problem and interdomain routing. *IEEE/ACM Transactions on Networking*, 10(22):232–243, Apr. 2002.
- [8] R. Mahajan, D. Wetherall, and T. Anderson. Negotiation-based routing between neighboring domains. In *Proceedings of USENIX/ACM Symposium on Networked Systems Design and Implementation (NSDI '05)*, San Francisco, CA, May 2005.

- [9] P. Milgrom and J. Roberts. Adaptive and sophisticated learning in normal form games. *Games and Economic Behaviors*, 3:82–100, 1991.
- [10] B. Quoitin, S. Uhlig, C. Pelsser, L. Swinnen, and O. Bonaventure. Interdomain traffic engineering with BGP. *IEEE Communications Magazine*, 41(5):122–128, May 2002.
- [11] S. Uhlig and O. Bonaventure. Designing BGP-based outbound traffic engineering techniques for stub ASes. *ACM SIGCOMM Computer Communication Review*, 34(5), 2004.
- [12] H. Wang, H. Xie, Y. R. Yang, L. E. Li, Y. Liu, and A. Silberschatz. On stable route selection for interdomain traffic engineering: Models, analysis, and guidelines. Technical Report YALEU/DCS/TR-1316, Yale University, Feb. 2005.