Stable Internet Routing Without Global Coordination

Lixin Gao, *Member, IEEE,* and Jennifer Rexford, *Senior Member, IEEE*

***Abstract*—The Border Gateway Protocol (BGP) allows an autonomous system (AS) to apply diverse local policies for selecting routes and propagating reachability information to other domains. However, BGP permits ASs to have conflicting policies that can lead to routing instability. This paper proposes a set of guidelines for an AS to follow in setting its routing policies, without requiring coordination with other ASs. Our approach exploits the Internet's hierarchical structure and the commercial relationships between ASs to impose a partial order on the set of routes to each destination. The guidelines conform to conventional traffic-engineering practices of ISPs, and provide each AS with significant flexibility in selecting its local policies. Furthermore, the guidelines ensure route convergence even under changes in the topology and routing policies. Drawing on a formal model of BGP, we prove that following our proposed policy guidelines guarantees route convergence. We also describe how our methodology can be applied to new types of relationships between ASs, how to verify the hierarchical AS relationships, and how to realize our policy guidelines. Our approach has significant practical value since it preserves the ability of each AS to apply complex local policies without divulging its BGP configurations to others.**

***Index Terms*—Border Gateway Protocol (BGP), convergence, Internet, protocols, routing.**

I. INTRODUCTION

THE INTERNET connects thousands of Autonomous Systems (ASs) operated by different institutions, such as Internet Service Providers (ISPs), companies, and universities. Routing within an AS is controlled by intradomain protocols such as OSPF, IS–IS, and RIP [1]. ASs interconnect via dedicated links and public network access points, and exchange reachability information using the Border Gateway Protocol (BGP) [2], [3]. BGP is an interdomain routing protocol that allows ASs to apply local policies for selecting routes and propagating routing information, without revealing their policies or internal topology to others. However, recent studies have shown that a collection of ASs may have conflicting BGP policies that lead to route divergence [4], [5]. Route divergence can result in route oscillation, which can significantly degrade the end-to-end performance of the Internet. Avoiding these conflicting BGP policies is crucial for the stability of the Internet routing infrastructure. Yet, to be practical, any technique for ensuring convergence should not sacrifice the ability of each AS to apply complex local policies.

A natural approach to the route convergence problem involves the use of the Internet Routing Registry, a repository of routing policies specified in a standard language [6]. A complete and up-to-date registry could check if the set of routing policies has any potential convergence problems. However, this global coordination effort faces several impediments. First, many ISPs may be unwilling to reveal their local policies to others, and may not keep the registry up-to-date. Second, and perhaps more importantly, even if ISPs decide to reveal their local polices, recent work has shown that statically checking for convergence properties is an NP-complete problem [4]. Third, even if the registry could ensure convergent routes under a given topology, BGP still might not converge under router or link failures, or a policy change. Hence, rather than requiring global coordination, we believe that convergence should be achieved by restricting the set of policies that each AS can apply. In this paper, we propose a set of guidelines for an AS to follow in setting its routing policies, without requiring coordination with other ASs [7]. In addition, the guidelines ensure routing convergence even under changes in the underlying topology (e.g., router or link failure) or the routing policies.

Our approach capitalizes on the Internet's hierarchical structure and the commercial relationships between ASs. These relationships include customer–provider, peer-to-peer, and backup [8], [9]. A customer pays its provider for connectivity to the rest of the Internet, whereas peers agree to exchange traffic between their respective customers free of charge; an AS may also provide backup connectivity to the Internet in the event of a failure. To ensure route convergence, we impose a partial order on the set of routes to each destination. Under our guidelines, routing via a peer or a provider is never preferable to routing via a customer link; furthermore, routes via backup links have the lowest preference. An AS is free to apply any local policies to the routes learned from neighbors within each preference class. These guidelines conform to conventional traffic-engineering practices of ISPs, and this might well explain why Internet routing divergence has not occurred yet. However, it is crucial to make these guidelines explicit since BGP itself does not constrain routing policies to ensure convergence. Based on our results, we propose a simple routing registry that stores only the *relationship* between each AS pair, rather than the entire set of routing policies. These relationships can be explicitly registered by the ASs or inferred from the BGP routing tables available throughout the Internet [10], [11].

The remainder of the paper is structured as follows. Section II presents an overview of interdomain routing and discusses previous work on BGP protocol dynamics. Then, Section III presents a formal model of BGP that includes ASs with

Manuscript received September 14, 2000; revised July 1, 2001; recommended by IEEE/ACM TRANSACTIONS ON NETWORKING Editor K. Calvert. The work of L. Gao was supported in part by the National Science Foundation under Grants ANI-9977555 and ANI-0085848, and NSF CAREER Award Grant ANI-9875513.

L. Gao is with the Department of Electrical and Computer Engineering, University of Massachusetts, Amherst, MA 01003 USA (e-mail: lgao@ecs.umass.edu).

J. Rexford is with the Internet and Networking Systems Center, AT&T Labs—Research, Florham Park, NJ 07932 USA (e-mail: jrex@research.att.com).

Publisher Item Identifier S 1063-6692(01)10544-3.

1063–6692/01$10.00 © 2001 IEEE

multiple BGP speakers, both interior BGP (iBGP) and exterior BGP (eBGP), and additional BGP attributes. We define the types of relationships between ASs and describe the hierarchical structure of the AS graph in Section IV. In Section V, we present our policy guidelines and formally prove that adherence to these guidelines guarantees convergence for all possible initial states. We show how to permit additional flexibility in choosing between routes through customers and routes through peers by making realistic assumptions about peer-to-peer relationships. Then, Section VI discusses the robustness of our guidelines to changes in network topology, routing policies, and relationships between ASs. We describe how to apply our methodology to new types of relationships that can arise between ASs, and how an AS pair can transition to a new relationship while preserving BGP stability. Section VII concludes the paper with a discussion of future research directions. The Appendix presents an example of how to configure a Cisco router to obey our policy guidelines.

II. Interdomain Routing

In this section, we present background material on the Internet architecture [12] and the use of BGP for interdomain routing [2], [3]. We also summarize previous work on the protocol dynamics of BGP.

A. Internet Architecture

The Internet consists of a large collection of hosts interconnected by networks of links and routers. The Internet is divided into thousands of distinct regions of administrative control, referred to as *autonomous systems* (ASs). Examples range from college campuses and corporate networks to large ISPs. An AS has its own routers and routing policies, and connects to other ASs to exchange traffic with remote hosts. A router typically has very detailed knowledge of the topology within its AS, and limited reachability information about other ASs. ASs interconnect at public Internet exchange points (IXPs) or dedicated point-to-point links. A public exchange point typically consists of a shared medium, such as a FDDI ring or an ATM switch, that interconnects routers from several different ASs. Physical connectivity at the IXP does not necessarily imply that every pair of ASs exchanges traffic with each other. AS pairs negotiate contractual agreements that control the exchange of traffic. These relationships include customer–provider, peer-to-peer, and backup, and are discussed in more detail in Section IV.

Each AS has responsibility for carrying traffic to and from a set of customer IP addresses. The scalability of the Internet routing infrastructure depends on the aggregation of IP addresses in contiguous blocks, called *prefixes*, each consisting of a 32-bit IP address and a mask length. For example, the prefix 192.0.2.0/24 corresponds to the 256 IP addresses from 192.0.2.0 to 192.0.2.255. An AS employs an *intradomain* routing protocol (such as OSPF or IS–IS) to select paths between routers within the network, and employs an *interdomain* routing protocol (BGP) to advertise the reachability of destination prefixes to neighboring ASs. BGP is a path-vector protocol that constructs paths by successively propagating advertisements between pairs of routers that are configured as *BGP peers* [2], [3]. Each advertisement concerns a particular prefix and includes the list of the ASs along the path (the *AS path*). Upon receiving an advertisement, a BGP speaker must decide whether or not to use this path and, if the path is chosen, whether or not to propagate the advertisement to neighboring ASs (after adding its own AS number to the AS path). A BGP speaker withdraws an advertisement when the prefix is no longer reachable with this route, which leads to a sequence of withdrawals by upstream ASs that are using this path.

The simplest path-vector protocol would employ shortest-path routing. BGP allows a much wider range of policies based on how the routers are configured. An AS can favor one path over another by assigning a *local preference*. BGP also allows an AS to send a hint to a neighbor on the preference that should be given to a route by using the *community* attribute. An AS can control how traffic enters its network by assigning a different *multiple exit discriminator* (MED) value to the advertisements it sends on each link to a neighboring AS. Otherwise, the neighboring AS would select the link based on its own intradomain routing protocol. An AS can also discourage traffic from entering its network by performing *AS prepending*, which inflates the length of the AS path by listing an AS number multiple times. Processing an advertisement involves three steps—*import policies* that decide which routes to consider, *path selection* that decides which route to use, and *export policies* that decide whether (and what) to advertise to a neighboring AS. While the BGP attributes and message formats are defined in standards documents, the BGP decision process and the configuration languages for expressing routing policies have been defined by router vendors. Vendors have settled on de facto standards for the decision process and provide similar support for expressing import and export policies.

B. BGP Protocol Dynamics

The growing importance and complexity of the Internet routing infrastructure has sparked interest in understanding BGP protocol dynamics. Previous work consists of measurement-based studies of BGP protocol traffic and theoretical analysis of BGP convergence properties. Extensive traces of BGP update messages have been used to characterize the structure (and growth) of the Internet topology, as well as the stability of routes to destination prefixes [13]–[17]. In contrast, research on BGP convergence has focused on determining what combination of BGP policies would cause a group of ASs to continually advertise and withdraw routes to a given prefix [18]–[20], [4], [5]. BGP convergence problems would not arise if every AS selects shortest-path routes. However, ASs can have conflicting local policies when they use the local-preference attribute to favor a route with a nonminimal AS path. This can result in route oscillation, where an AS makes a decision and advertises a new route to its neighbors, which causes neighbors to change their decisions; then, these ASs withdraw their previous route and advertise new ones, and the process repeats.

Previous research has studied route convergence under the assumption of global knowledge of the topology and routing policies. The work in [5] analyzes route oscillation in simple ring topologies, and suggests maintaining a global routing registry of interdomain policies that can be checked for potential convergence problems [21], [18], [6], [5]. Expanding on these observations, the work in [4] presents a formal model of BGP that focuses on local-preference and AS-path-length attributes. Since the paper proves *negative* results about BGP convergence properties, it is sufficient to consider a restricted subset of the protocol. In particular, the study establishes that the problem of checking the convergence properties is NP-complete, even with full knowledge of the routing policies of each AS. In addition, the paper presents several examples of conflicting BGP policies, including scenarios when the divergence occurs only after a link failure. A follow-up paper [20] presents a dynamic model that captures the asynchronous processing of updates at each AS. The paper formalizes the notion of a *stable state* where no AS would change its routes, and a *safe BGP system* that is guaranteed to converge to a stable state. The paper presents a sufficient condition for a BGP system to be safe. However, testing adherence to the condition requires full knowledge of the AS graph and the set of routing policies for each AS.

These results suggest that it may be possible to restrict local policies in a way that guarantees BGP convergence, while still allowing greater flexibility than shortest-path routing. Our paper focuses on constructing a set of reasonable policy guidelines that guarantee a safe BGP system, even under changes in network topology and routing policies, without requiring coordination between ASs.

III. Abstract Model of BGP

In this section, we present an abstract model of BGP that we use in establishing the stability properties in Section V. The model extends the work in [20], [4] to include iBGP and eBGP, additional BGP attributes and operations (such as MEDs, community set, and AS prepending), and the possibility that an AS has multiple BGP speakers. This more complete model of BGP is necessary for establishing *positive* results about system stability.

A. BGP Routing

The topology of a BGP system is modeled as a clustered graph $G = (N, V, E)$, where the set N consists of ASs, the vertex set V consists of all BGP-speaking routers, and the edge set E consists of all eBGP peering sessions. Each BGP speaker belongs to one AS and an AS can have one or more BGP speakers. Let $a(i) \in N$ denote the AS that BGP speaker i belongs to. Each eBGP peering session involves a pair of BGP speakers in different ASs. Each BGP-speaker pair in the same AS has an iBGP session and a cost metric that represents the distance between the two BGP speakers based on the intradomain routing protocol. BGP speakers i and j in different ASs [i.e., $a(i) \neq a(j)$] may have an eBGP session, represented as an edge in the graph. Fig. 1 shows an example of the topology in a BGP system. In practice, ASs at a public IXP could exchange routes through a shared route server, rather than having a separate BGP session for each AS pair. The route server applies each ASs' routing policies to create the illusion of a separate BGP session for every pair of ASs [22], consistent with our model.

A route update r includes the destination prefix ($r.prefix$), next-hop interface address ($r.next\_hop$), AS path ($r.as\_path$), local preference ($r.local\_pref$), multiple-exit discriminator ($r.med$), and community set ($r.c\_set$). Each BGP speaker originates updates for one or more prefixes, and can send the updates to the immediate neighbors via an iBGP or eBGP session. BGP-speaker pairs in the same AS use iBGP to exchange routes learned from BGP peers. In practice, ASs may employ route reflectors or confederations [12] to reduce the overhead of exchanging routing updates in a large backbone. These optimizations are intended to reduce the iBGP overhead without affecting the routing decisions and, hence, are not included in our model. Routing updates exchanged via eBGP sessions are transformed according to the BGP policies. We consider an eBGP session $l \in E$ between two BGP speakers, u and v. BGP speaker v receives a set of route updates R on l from u. BGP speaker v applies *import policies* to transform incoming route updates, and applies *export policies* before sending updates to the neighbor u.

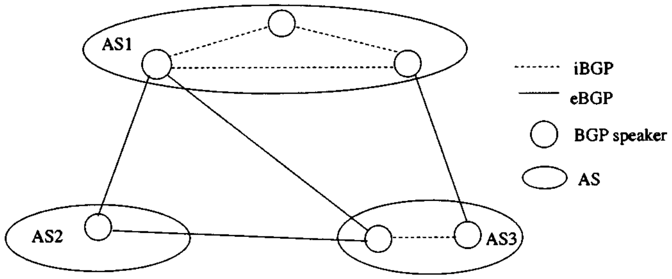


Fig. 1. Example of a BGP system topology.

A BGP speaker applies an implicit import policy defined by the protocol specification and an explicit import policy configured by the network operator. Let $im\_import(l, v)[R]$ denote the set of updates after applying the implicit import policy of v on edge l. Every edge has an implicit import policy that discards a routing update when the receiving BGP speaker's AS already appears in the AS path; this is essential to avoid introducing a loop in the AS path. That is, if $a(v) \in r.as\_path$, then $im\_import(l,v)[\{r\}] = \{\}$ to remove the route; otherwise $im\_import(l,v)[\{r\}] = \{r\}$ to keep the route. Let $ex\_import(l, v)[R]$ represent the set of updates after applying the explicit import policy, such as denying or permitting an update, and assigning a local-preference value. For example, an explicit import policy could assign $r.local\_pref = 100$ if AS 1 appears in $r.as\_path$ or deny any update that includes AS 2 in the path. Ultimately, the import policy transforms the set of updates R as $import(l, v)[R] = ex\_import(l, v)[im\_import(l, v)[R]]$.

After applying the import policies for a route update from an eBGP session, v exchanges the update with all other BGP speakers in the same AS, using iBGP sessions. Each BGP speaker v then follows a route selection process $Select(S)$ that picks the best route for each prefix out of the set S of route updates. The BGP speaker picks the route with the highest $r.local\_pref$, breaking ties by selecting the route with the shortest $r.as\_path$. Note that local preference overrides the AS-path length. Amongst the remaining routes, v picks the one with the smallest $r.med$. In this step, we assume that the operator has configured the router to compare MEDs across all update messages, rather than simply comparing MEDs across routes advertised by the same next-hop AS.[1] Then, the decision process breaks ties by selecting the route with the smallest cost to the BGP speaker that passes the route via an iBGP session. Note that, since the tie-breaking process draws on intradomain cost information, two BGP speakers in the same AS may select different best routes for the same prefix. If a tie still exists, v picks the route with the smallest $r.next\_hop$.

Each BGP speaker sends its best route (one best route for each prefix) via eBGP sessions. The BGP speaker u applies implicit and explicit export policies on each eBGP session l to a neighboring BGP speaker v, defined as $im\_export(l, u)$ and $ex\_export(l, u)$, respectively. Each BGP speaker u applies an implicit policy that sets $r.local\_pref$ and $r.med$ to default values, assigns $r.next\_hop$ to u's interface connecting to l, and prepends u to $r.as\_path$. Explicit export policies include permitting or denying the route, assigning $r.med$, assigning $r.c\_set$, and prepending u one or more times to $r.as\_path$. For example, AS u could decline to advertise routes to AS v that have community 10 in the community set. Also, AS u could prepend u two times to the AS path for prefix 192.0.2.0/24 and for any route that includes AS 2 in the AS path. Ultimately, the export policy transforms the set of updates R as $export(l, u)[R] = ex\_export(l, u)[im\_export(l, u)[R]]$. Then, u transmits these transformed updates to v using eBGP sessions.

B. Distributed Path Selection

The route-selection process proceeds in a distributed and asynchronous fashion, triggered by advertisements and withdrawals of routes. Rather than modeling the exact timing of message transmissions, we focus on the decision-making process of each BGP speaker. For the sake of simplicity, we focus on a single destination prefix d that originates from AS_d; since route aggregation does not affect the convergence properties, it is sufficient to consider the set of routes to a single destination prefix. Each speaker applies the BGP selection process to pick its best path to d, after applying import policies to the routes that have been exported by its neighbors. BGP is an incremental protocol, where each speaker remembers the routes advertised by neighbors until they are withdrawn, and selects a best path from this set. In a stable state, a BGP speaker remembers precisely those routes that have been chosen by its neighbors. Hence, for studying convergence properties, it is sufficient to define the state of the BGP system in terms of the route chosen by each BGP speaker. That is, we assume that each speaker remembers only its own best route, selected from the set of routes exported by its neighbors. As such, we define the system state as a vector $s = (s_1, s_2, \ldots, s_n)$, where s_i denotes the route chosen by speaker $i = 1, 2, \ldots, n$.

Changes in the system state occur when one or more BGP speakers apply the route selection process. Formally, *activating* a speaker applies the export policies of the BGP speakers in neighboring ASs, the speaker's import policies, and the BGP path-selection process [20]. In particular, if the BGP speaker resides in AS_d, the route to d is a route (denoted as r_0) that contains a null AS path. Otherwise, the selection of s_i can be affected by the route chosen by any BGP speaker j that has a BGP session with a speaker $k \in a(i)$. This includes the BGP peers of speaker i, as well as the BGP peers of the other speakers in the same AS, since i could learn about these routes via iBGP sessions. The choices available to speaker i depend on the route s_j, the export policies of j, and the import policies of k:

$$Choices(i, s) = \begin{cases} r_0, & \text{if } a(i) = AS_d \\ \bigcup_{l=(k, j)\in E \wedge k \in a(i)} import(l, k) \\ [export(l, j)(s_j)], & \text{otherwise.} \end{cases}$$

Then, i selects a route $BestRoute(i, s) = Select(Choices(i, s))$. Note that the model assumes that each external neighbor's route is immediately available and that these routes are propagated via iBGP sessions. This simplifying assumption does not affect the BGP convergence properties, as the neighbors' updates would eventually become available (e.g., after finite propagation delay).

Since each BGP speaker operates independently, we cannot assume that every BGP speaker is activated at the same time. Instead, as in [20], we consider a subset $A \subseteq V$ of speakers that are activated at a given time. The remaining BGP speakers do not apply the path-selection process and, hence, do not change their best route. Therefore, the next state $s' = (s'_1, s'_2, \ldots, s'_n)$ has $s'_i = BestRoute(i, s)$ for $i \in A$, and $s'_i = s_i$ for $i \notin A$. We let $s \xrightarrow{A} s'$ denote the transition from state s to s' given the activation set A. The definition of the state of a BGP system, and the notion of an activation set, allows us to precisely define the notion of stability. Formally, a state s is *stable* if and only if $s \xrightarrow{A} s$ for any activation set A. That is, when the system is in state s, no AS would change to a different route.

To study convergence, we define an *activation sequence* as a (possibly infinite) sequence of activations. Let σ denote the activation sequence and $\sigma(j) \subseteq V$ denote the jth activation in σ. In studying convergence, we need to consider sequences that activate each AS several times. In particular, a *fair* activation sequence σ is an infinite sequence that has infinitely many elements j such that $i \in \sigma(j)$, for each BGP speaker $i \in V$. A BGP system *converges* for a particular activation sequence and initial state if it arrives at a stable state after the activation sequence. Formally, for an activation sequence σ and an initial state s^0, a BGP system converges if there is a finite j such that $s^0 \xrightarrow{\sigma(1)} s^1 \xrightarrow{\sigma(2)} \cdots \xrightarrow{\sigma(j)} s^j$ and s^j is a stable state.

Thus far, we have defined the notion of a stable state. However, some BGP systems have a stable state without necessarily converging. For example, Fig. 2 shows an example where three ASs are connected pairwise and AS 0 originates destination

[1] In practice, a BGP-speaking router may be configured to ignore MEDs, compare MEDs across advertisements with the same next-hop AS, or compare MEDs across all advertisements. If MED comparisons are limited to advertisements with the same next-hop AS, the comparison between advertisements is no longer associative. Route r may appear better than route s, which may appear better than route t, which may, in turn, appear better than route r. Recent work has shown that this can cause a BGP system to oscillate [23]. No general solution has been discovered for this problem.

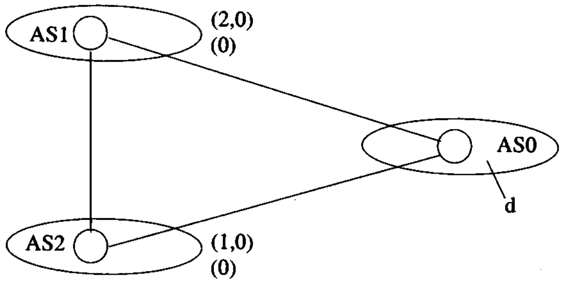


Fig. 2. BGP system has a stable state but might not converge.

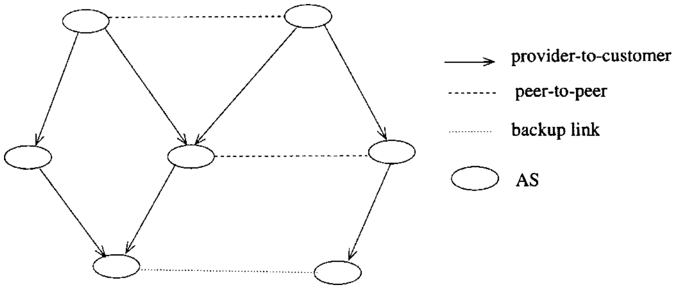


Fig. 3. Hierarchical AS interconnection.

prefix d [4]; with each AS, we list the set of possible routes in order of preference. Both AS 1 and AS 2 prefer the path through the neighbor over the direct route to reach d. The system has a stable state. For example, AS 2 could use the direct route (0) and AS 1 could use the route $(2, 0)$. However, the system could also oscillate between two unstable states. In the first state, both ASs have selected the direct route (0). Then, if activated simultaneously, both ASs switch to their indirect routes [e.g., AS 2 switches to $(1, 0)$]. Then, if activated again, both ASs return to their direct routes, and the process repeats. Whether or not the system eventually reaches a stable state depends on the exact timing of the reception and processing of the route updates. Hence, we define a stronger notion of a *safe* BGP system [20]. A BGP system is *safe* if it has a stable state and converges under any fair activation sequence and any initial state. Furthermore, we define an *inherently safe* BGP system as a BGP system that is safe and remains safe after removing any nodes and/or edges.

IV. Hierarchical AS Graph

Our policy configuration guidelines capitalize on the fact that ASs are interconnected in a hierarchical fashion. In this section, we describe the relationships between ASs and the resulting hierarchical structure.

A. Customers, Providers, and Peers

AS relationships arise from contracts that define the pricing model and the exchange of traffic. In a *customer–provider* relationship, the customer pays its provider for access to the rest of the Internet. The provider may, in turn, be a customer of another AS. In a *peer-to-peer* relationship, the two peers find it mutually advantageous to exchange traffic between their respective customers; typically, peers exchange a roughly even amount of traffic free of charge [9]. Each eBGP session defines a relationship between the two ASs it connects. Although there might be multiple eBGP sessions between two ASs, the relationship between the two ASs should be uniquely defined. An AS a may have multiple customers, providers, and peers. We define $customer(a)$, $peer(a)$, and $provider(a)$ as the set of customers, peers, and providers of a, respectively. We let $first(r.as\_path)$ denote the next-hop AS in $r.as\_path$. A route r is classified as a customer route of a if $first(r.as\_path) \in customer(a)$, a peer route if $first(r.as\_path) \in peer(a)$, or a provider route if $first(r.as\_path) \in provider(a)$. Two ASs may also have a bilateral *backup* agreement, as discussed in more detail in Section V-B.

The customer–provider and peer-to-peer agreements translate into several rules governing BGP export policies [8], [9]:

- **Exporting to a provider:** In exchanging routing information with a provider, an AS can export its routes and the routes of its customers, but can not export routes learned from other providers or peers. That is, an AS does *not* provide transit services for its provider.
- **Exporting to a customer:** In exchanging routing information with a customer, an AS can export its routes, as well as routes learned from its providers and peers. That is, an AS *does* provide transit services for its customers.
- **Exporting to a peer:** In exchanging routing information with a peer, an AS can export its routes and the routes of its customers, but can not export the routes learned from other providers or peers. That is, an AS does *not* provide transit services for its peers.

Drawing on our abstract model, consider a BGP speaker u and with a link l connecting to an AS $v \in provider(a(u)) \cup peer(a(u))$. For each r, if $first(r.as\_path) \in provider(a(u)) \cup peer(a(u))$, then $ex\_export(l, u)[\{r\}] = \{\}$. The Appendix presents a sample router configuration file that realizes these export policies.

B. Hierarchy

We assume that there is a hierarchical customer–provider relationship among ASs. The hierarchical structure arises because an AS typically selects a provider with a network of larger size and scope than its own. An AS u serving a metropolitan area is likely to have a regional provider v, and a regional AS v is likely to have a national provider w; it is very unlikely that a nationwide AS w would be a customer of a metropolitan-area AS u. That is, if $u \in customer(v)$ and $v \in customer(w)$, then $w \notin customer(u)$. AS v is a *direct* provider of u, whereas AS w is an *indirect* provider of u. Any direct or indirect provider of u cannot be a customer of u. To simplify the discussion, we define two directed graphs formed by the customer–provider relationships. In the *provider-to-customer graph*, the provider–customer edges are directed from provider to customer. The resulting subgraph formed by only provider–customer relationships should be a *directed acyclic graph* (DAG), as shown in the example in Fig. 3. In the *customer-to-provider graph*, the provider–customer edges are directed from customer to provider.

A route registry can be used to verify the hierarchical relationships. The registry could be populated in several ways. First, each AS a could supply its set $provider(a)$, updating the registry upon adding or deleting a provider. This approach requires the cooperation of the various autonomous systems in the Internet. Second, a registry could *infer* the AS relationships based on the BGP routing tables available throughout the Internet [10], [11], although this process could be vulnerable to incomplete information and incorrect inferences. Either way, the registry can check for a cycle whenever any AS changes its set of providers. This could happen when an AS adds or removes a provider, or when an AS changes its relationship with one of its neighbors; for example, a pair of ASs may transition from a customer–provider relationship to a peer-to-peer arrangement. The algorithm for checking whether there is a cycle in a directed graph takes $O(|N|+|E|)$ time [24], where $|E|$ is the number of edges and $|N|$ is the number of nodes of the directed graph. As of the spring of 2001, the AS graph has an estimated size of at least 11 500 nodes and 30 000 edges [25]. BGP permits at most $2^{16} = 65\,536$ AS numbers and the number of AS interconnections tends to grow linearly in the number of ASs [26]. Therefore, it is possible to run the cycle-detection algorithm whenever an AS updates its list of providers to ensure the conformity to the hierarchical relationships at all times.

If the provider-to-customer or customer-to-provider graph has a cycle, the registry can efficiently identify the sequence of ASs involved. If more detailed information is available about the routing policies of these ASs, the registry could check for possible convergence problems. Although checking for convergence is an NP-complete problem [4], the check would be applied on the subgraph, which would involve a much smaller number of vertices and edges than the initial AS graph. Alternatively, the registry could instruct the ASs in the cycle to coordinate amongst themselves to avoid policies that would cause convergence problems, or to force the use of a restrictive policy (such as shortest AS path) that would guarantee convergence.

V. BGP Policy Guidelines

This section presents policy guidelines that ensure that the BGP system is safe. To simplify the discussion, we initially consider only customer–provider and peer-to-peer relationships. We then extend the guidelines to include a simple form of backup relationships. Since the route selection process for each destination prefix is independent of other prefixes, it is sufficient to consider only one destination prefix d in describing and analyzing the guidelines.

A. BGP Systems With No Backup Links

In this section, we present the policy configuration guidelines for BGP systems that have only customer–provider and peer-to-peer relationships. We first consider the guideline for the case that any AS pair can have a peer-to-peer agreement. Then, we expand the set of local policies by imposing realistic restrictions on which AS pairs can have peer-to-peer relationships.

1) Unconstrained Peer-to-Peer Agreements: Our guideline requires an AS to prefer a route via a customer over a route via a provider or peer. Formally, we have Guideline A for the explicit import policy of each BGP speaker in AS a:

Guideline A

if $((first(r_1.as\_path) \in customer(a))$ and
$(first(r_2.as\_path) \in peer(a) \cup provider(a)))$
then $r_1.loc\_pref > r_2.loc\_pref$

Note that Guideline A does *not* restrict the preference among customer routes or among provider or peer routes, which leaves ISPs with significant flexibility in selecting local policies.In addition, ISPs have a financial incentive to follow the guideline since an ISP does not have to pay its customer to carry traffic. Guideline A allows a large number of possible configurations, much larger than policies based only on AS-path length. To implement the guidelines, an AS could allocate a range of local-pref values for each type of route (e.g., 86–100 for customer routes and 75–85 for peer and provider routes). The Appendix illustrates how to configure BGP sessions to obey Guideline A.

Guideline A ensures that the BGP system is safe. The proof draws on how the local-pref assignment affects how each BGP speaker picks its best route.

Theorem 5.1: For a BGP system that has only customer–provider and peer-to-peer relationships, if all ASs follow guideline A, then the BGP system is inherently safe.

We prove the theorem by two lemmas. The first lemma claims that the BGP system has a stable state. The second lemma claims that the BGP system converges to the stable state for any initial state and any fair activation sequence. Finally, we prove that the BGP system is safe after removing any nodes and/or edges.

Lemma 5.1: The BGP system has a stable state.

Proof: We prove the lemma by constructing an activation sequence σ^* that leads to a stable state for any initial state. Let d denote the destination prefix and AS_d denote the AS that originates prefix d. Since the activation order among the BGP speakers within an AS does not affect the best route selection of the BGP speakers, we activate all BGP speakers of an AS simultaneously. For simplicity of explanation, we use the activation of an AS to represent the activation of all BGP speakers in the AS. We activate ASs in two phases. In the first phase, a AS selects a customer route if one is available, following Guideline A. This is accomplished by activating the ASs in an order that conforms to the partial order in the customer-to-provider DAG. In the second phase, the ASs that do not have a customer route after Phase 1 get provider or peer routes. This is accomplished by activating ASs in an order that conforms to the partial order in the provider-to-customer DAG. Formally, we have a two-phase activation sequence σ^* as follows.

Phase 1: Activate ASs in a linear order that conforms to the partial order in the customer-to-provider DAG.

Phase 2: Activate ASs in a linear order that conforms to the partial order in the provider-to-customer DAG.

For the simplicity of the discussion, we partition the ASs into two classes; the first class consists of AS_d and ASs that select a customer route in Phase 1. The second class consists of the remaining ASs. We call ASs in the first class *Phase-1*

ASs and ASs in the second class *Phase-2 ASs*. Similarly, we call BGP speakers in a Phase-1 AS *Phase-1 BGP speakers* and BGP speakers in a Phase-2 AS *Phase-2 BGP speakers*. The activation sequence results a stable state independent of the initial state. We prove that each Phase-1 BGP speaker reaches a stable state after its activation in Phase 1 and each Phase-2 BGP speaker reaches a stable state after its activation in Phase 2. In other words, we prove the following two claims.

Claim 1: A Phase-1 BGP speaker reaches a stable state after its activation in Phase 1.

Proof: We prove by induction on the order that Phase-1 BGP speakers are activated in Phase 1. Clearly, among Phase-1 BGP speakers, BGP speakers in AS_d are the first to be activated. BGP speakers in AS_d reach a stable state as soon as AS_d is activated. Let Phase-1 BGP speaker i belong to AS_n. Suppose all Phase-1 BGP speakers that belong to an AS preceding AS_n in Phase 1 reach a stable state after their activation. BGP speaker i selects the best route amongst its customer routes. All of the customers precede AS_n in the activation sequence for Phase 1. Hence, each customer has either reached a stable state (earlier in Phase 1) or does not get a customer route in Phase 1. Any customer that does not get a customer route in Phase 1 does not export its route to BGP speaker i according to export policy rule. Hence, those customers' routing decisions do not affect BGP speaker i. Therefore, BGP speaker i reaches a stable state after its activation in Phase 1.

Claim 2: A Phase-2 BGP speaker reaches a stable state after its activation in Phase 2.

Proof: Following a similar approach, we prove by induction on the order that Phase-2 BGP speakers are activated in Phase 2. Let AS_0 be the first Phase-2 AS that is activated in Phase 2. Clearly, AS_0 does not have any Phase-2 provider. Since AS_0's BGP speakers are not Phase-1 BGP speakers, these BGP speakers can only get routes from AS_0's peers and providers. AS_0's peers either a) are stable after Phase 1 (if there is a customer route) or b) do not export their routes AS_0 (if the best route is a provider or peer route). The peers that fall in case a) are stable before AS_0 are activated. The peers that fall in case b) do not affect AS_0's BGP speakers' route. Since AS_0 does not have any Phase-2 provider, its providers are stable after Phase 1. Therefore, AS_0's BGP speakers are stable after their activation in Phase 2.

Let Phase-2 BGP speaker i belong to AS_n. Suppose all BGP speakers that belong to an AS preceding AS_n in Phase 2 reach a stable state after their activation in Phase 2. Since no customer route was learned in Phase 1, BGP speaker i must select a route from one of its providers or peers. Each provider has already reached a stable state (either in Phase 1, or earlier in the activation sequence of Phase 2). Each peer is either a Phase-1 AS or a Phase-2 AS. If a peer is a Phase-1 AS, the peer's route is available to BGP speaker i when it is activated in Phase 2. If a peer is a Phase-2 AS, then this peer selects a route from one of its providers or one of its other peers. The peer would not announce such a route to BGP speaker i and, hence, the routing decision would not affect BGP speaker i. Therefore, BGP speaker i reaches a stable state after its activation in Phase 2. ■

Lemma 5.2: The BGP system converges to the stable state for any initial state and any fair activation sequence.

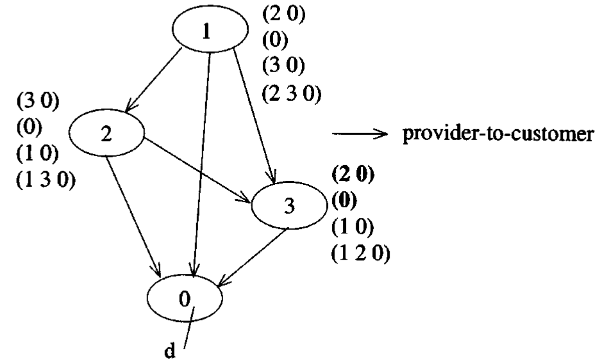


Fig. 4. BGP system that violates Guideline A.

Proof: Given any fair activation sequence σ, we prove by induction on the ASs in the order given by Phase-1 ASs followed by Phase-2 ASs where both Phase-1 and Phase-2 ASs are in the order of activation sequence σ^*. It is clear that each BGP speaker in AS_d reaches a stable state after a single activation. Suppose that all BGP speakers in the ASs that precede AS_n are stable after activation $\sigma(t)$. Let $\sigma(t')$ be the first activation set such that all BGP speakers in AS_n have been activated at least once between $\sigma(t)$ and $\sigma(t')$. Note that we can find t' since any fair activation sequence activates a BGP speaker infinitely many times. Using the same argument as above, we can prove that all BGP speakers in AS_n reach a stable state after $\sigma(t')$. Therefore, the system converges to the stable state after a finite number of activations in the fair activation sequence.

Finally, removing any nodes and/or edges from the BGP system do not affect the above two lemmas. Therefore, the BGP system is inherently safe. ■

Fig. 4 presents an example of a set of policies that violates Guideline A. The directed edges in the graph indicate the provider-to-customer relationships, and the routes of each AS are listed in the order of preference. AS 3 violates the guideline by preferring a provider route (via AS 2) over a customer route (via AS 0). This BGP system is not safe. Each AS initially selects route (0) and then decides to change to a route through its counterclockwise neighbor. This process can continue indefinitely. As another example, consider the BGP system given in Fig. 2. AS 1 and AS 2 are peers and both are providers of AS 0. Both AS 1 and AS 2 prefer the peer route over the customer route, which violates Guideline A. The resulting BGP system is not safe.

2) Constrained Peer-to-Peer Relationships: Guideline A assumes that any pair of ASs could have a peer-to-peer agreement. In this section, we make some realistic assumptions about peering agreements so as to relax the guideline. In particular, we allow peer routes to have the *same* local-pref as customer routes, to give ISPs greater flexibility in balancing network load. Typically, a peer-to-peer relationship is between two ASs with networks of similar size or scope that exchange a comparable volume of traffic. An AS is unlikely to have a peer-to-peer relationship with one of its (direct or indirect) providers. More generally, we group ASs that peer with each other directly or indirectly and call each group an *AS cluster*. An AS that does not peer with any other AS has its own cluster. In a *clustered AS graph*, each node is an AS cluster consisting of one or more

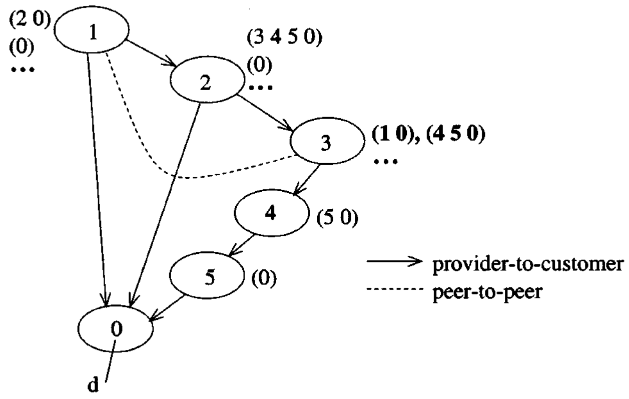


Fig. 5. BGP system that obeys Guideline B but violates Assumption P.

ASs. There is a directed edge from cluster t_1 to cluster t_2 if there is an AS a_1 that belongs to t_1 and an AS a_2 that belongs to t_2 where a_1 is a provider of a_2. A cluster has a self cycle (where t_1 and t_2 are the same) if any of the AS pairs in the cluster have a provider–customer relationship. Guideline A does not make any assumptions about the structure of the cluster graph. In formulating Guideline B, we assume that the cluster graph has a hierarchical structure. Formally, we assume that peer-to-peer relationships satisfy the following condition.

Assumption P: The clustered AS graph is a directed acyclic graph. That is, there is no cycle or self cycle in the clustered graph.

A routing registry can check for violations of Assumption P and notify the ASs involved, or force the system to abide by Guideline A.

Assumption P allows us to relax Guideline A to allow a peer route to have the same local-pref as a customer route. Formally, we have Guideline B for the explicit import policy of each BGP speaker in AS a:

Guideline B

if $((first(r_1.as\_path) \in customer(a))$ and
$(first(r_2.as\_path) \in peer(a))$
then $r_1.loc\_pref \geq r_2.loc\_pref$
if $((first(r_1.as\_path) \in customer(a))$ and
$(first(r_2.as\_path) \in provider(a))$
then $r_1.loc\_pref > r_2.loc\_pref$

Assumption P is essential for the stability of BGP system. For example, the BGP system in Fig. 5 violates Assumption P since there is a cycle between the cluster formed by AS 1 and AS 3 and the cluster formed by AS 2. Applying Guideline B, AS 3 assigns equal preference to the route (1, 0) through its peer and the route (4, 5, 0) through its customer; AS 3 ultimately favors the route $(1, 0)$ with the shorter AS path. However, AS 1 prefers the customer route $(2, 0)$ over its direct route (0), and AS 2 prefers the route $(3, 4, 5, 0)$ through its customer over the route $(1, 0)$ through its provider. Assume that initially none of the ASs have a route to d. After AS 5 and AS 4 have been activated, assume that ASs 1, 2, and 3 are always activated together. The first activation leads ASs $1, 2,$ and 3 to select routes (0), (0), and $(4, 5, 0)$, respectively. On the next activation, they switch to $(2, 0), (3, 4, 5, 0)$, and $(1, 0)$, and the process repeat indefinitely. This system would be safe if it followed Guideline *A* by requiring AS 3 to favor the customer route $(4, 5, 0)$ over the peer route $(1, 0)$.

Theorem 5.2: For a BGP system that has only customer–provider and peer-to-peer relationships and conforms to Assumption P, if all ASs follow guideline B, then the BGP system is inherently safe.

Proof: We prove the theorem by demonstrating that the BGP system has a stable state and converges to the stable state for any initial state and any fair activation sequence. Since the second part is similar to Theorem 5.1, we concentrate on proving that the BGP system has a stable state.

Similar to Lemma 5.1, we construct a two-phase activation sequence that leads to a stable state. We activate all ASs in a linear order that conforms to the partial order in the customer-to-provider DAG in Phase 1. We impose additional constraints on the order of AS activations in Phase 1 based on the peer-to-peer relationships and the AS-path length. Therefore, BGP speakers get their customer and peer routes in Phase 1. The BGP speakers that do not get a route in Phase 1 then select a route from a peer or a provider. Therefore, in Phase 2, ASs are activated in an order that conforms to the partial order given in the provider-to-customer DAG. Formally, we have a two-phase activation sequence σ^* as follows.

Phase 1: Activate ASs in a linear order that conforms to the partial order given by the clustered AS graph. In other words, if there is a directed edge from Cluster 1 to Cluster 2, ASs in Cluster 1 are activated after all ASs in Cluster 2. Among ASs in the same cluster, activate ASs in the following order. For each AS, select the best route among its customer routes according to the BGP route selection process *Select*. We call the selected route the *candidate route*. Activate the ASs according to the length of their candidate routes. An AS with a shorter candidate route is activated before an AS with a longer candidate route, breaking ties arbitrarily.

Phase 2: Activate ASs in a linear order that conforms to the partial order in the provider-to-customer DAG.

Note that we have the same Phase 2 as in Theorem 5.1. Our proof of the stability of ASs after Phase 2 follows the same argument. Therefore, we concentrate on Phase 1. We impose additional order on ASs so that an AS is activated only if all of its peers are stable or the routes of its unstable peers would not affect the routing decision. The order conforms to the length of the candidate route. Since a peer route never has a *larger* local-pref than a customer route, an AS never selects a peer route over a customer route with a *shorter* AS path. Hence, this additional restriction on activation order ensures that a Phase-1 AS is stable after its activation, following a similar argument as in Lemma 5.1.

Finally, removing any nodes and/or edges from the BGP system do not affect the above arguments. Therefore, the BGP system is inherently safe. ■

B. BGP Systems With Backup Links

Customer–provider and peer-to-peer are the two most common relationships between two ASs. However, an AS may also have a backup relationship with a neighboring AS. Having

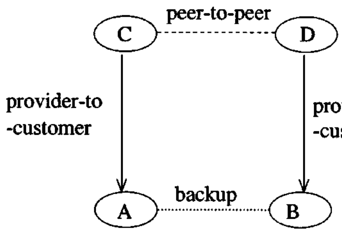


Fig. 6. Backup link between ASs A and B.

a backup relationship with a neighbor is important when an AS has limited connectivity to the rest of the Internet. For example, ASs A and B could establish a bilateral backup agreement for providing the connection to the Internet in the case that one AS' link to providers fails. AS C is a provider of AS A and AS D is a provider of AS B. ASs C and D have a peer-to-peer agreement, as shown in Fig. 6. Typically, A reaches others via C and B reaches others via D. If the link between A and C (or B and D) fails, the backup link between A and B is used for A (or B) to connect to and from the rest of the Internet. To provide a backup service to A, AS B must be willing to export A's routes to D and D's routes to A. Initially, we assume that an AS pair cannot have both a backup relationship and a customer–provider or peer-to-peer arrangement; we relax this assumption in Section VI-B.

Backup links are not meant to be used unless a failure occurs. Hence, routes involving backup links should have a lower local-pref than other routes. Note that a route through a backup link is a route that contains one or more backup links—the backup link does *not* have to be first hop. For example, the path (C, D, B, A) is a backup route because it traverses the backup link (B, A). We define a single local preference value of $backup\_pref$ for all backup routes. Formally, we have Guideline C for each BGP speaker:

Guideline C

```
if (($r_1$ does not contain a backup link)
    follow Guideline A or B to assign $r_1.loc\_pref$
    (where $r_1.loc\_pref > backup\_pref$)
if ($r_1$ contains a backup link)
    $r_1.loc\_pref = backup\_pref$
```

Note that, unlike Guideline A or B, enforcing Guideline C requires cooperation between ASs. An AS can not tell which routes involve backup links between other AS pairs. Hence, the BGP advertisements must identify these routes. This is typically achieved using the community attribute (c_set). Providers and customers agree on a community number that indicates which routes includes a backup link [27]. When the customer sends the provider a backup route, it assigns the community number to the route so that the provider can assign an appropriate loc_pref. See [21] for an example of the configuration specified using Routing Policy Specification Language (RPSL). Now, we prove that Guideline C ensures that the BGP system is inherently safe.

Theorem 5.3: If all ASs follow guideline C in setting up their policies, then the BGP system is inherently safe.

Proof: We prove the theorem for the case that all nonbackup routes follow Guideline A. A similar argument follows for the case that all nonbackup routes follow Guideline B. Let AS_d denote the AS that originates the destination prefix d. We construct an activation sequence that leads the BGP system to a stable state. We then prove that the system always converges to the stable state. The activation sequence first propagates routes using customer–provider and peer-to-peer links, and then propagates routes using backup links. There are three phases. The first two phases are the same as in Theorem 5.1. The last phase activates the remaining ASs. These ASs only have paths with one or more backup links. This phase activates the ASs in order of the length of the backup paths. Formally, we construct an activation sequence σ^* that leads to a stable state. The activation sequence activates the BGP speakers in each AS simultaneously.

Phase 1: Activate ASs in a linear order that conforms to the partial order given in the customer-to-provider DAG.

Phase 2: Activate ASs in a linear order that conforms to the partial order given in provider-to-customer DAG.

Phase 3: Activate the ASs that did not get a route in the first two phases in order of the length of their shortest backup path (shorter paths first).

Using the same argument as in Theorem 5.1, the first two phases ensure a stable state for all ASs that have a route to d without using a backup link. In the third phase, all remaining ASs reach a stable state using a backup path. All of these can be proven by induction, as in Theorem 5.1. Note that in Theorem 5.1, the activation sequence gives a linear order of ASs. Using the same argument, we can prove that the BGP system converges to the stable state for any fair activation sequence.

Finally, removing any nodes and/or edges from the BGP system do not affect the above arguments. Therefore, the BGP system is inherently safe. ■

Guideline C assigns a single local preference value to all routes with one or more backup links. Requiring each AS to select shortest-path backup routes may be overly restrictive in practice. For example, an AS might prefer a backup path through a customer over a backup path through a peer or provider. Alternatively, an AS might prefer a path with one backup link over a shorter path with two or more backup links. Guideline C can be generalized to support these more flexible backup policies, as discussed in more detail in [28].

VI. Practical Implications

In this section, we discuss the applicability of our guidelines to diverse and changing network topologies and routing policies. Then, we demonstrate how our methodology can be applied to more complex relationships between ASs, and describe how an AS pair can transition to a new relationship without disrupting system stability.

A. Robustness of the Guidelines

The network topology and routing policies are very dynamic in today's rapidly growing Internet. Router and link failures, and the deployment of additional network equipment, result in frequent changes to the underlying topology. ISPs often fine-tune their policy configurations to adapt to fluctuations in traffic demands and changes in their internal topology and connections to neighboring ASs. In addition, ASs periodically change their relationships by adding or removing customers, peers, or providers. Our guidelines ensure the stability of the BGP system even in this dynamic environment. Although these changes may trigger the exchange of new routing information, and may ultimately result in new routing decisions, the system remains safe. The inherent safety property ensures that the BGP system remains safe after the deletion of nodes and edges. The system remains safe after the addition of edges and nodes, as long as the new graph adheres to the policy guidelines. Alternate approaches [20] that establish convergence properties by performing a check on the topology and policy configurations would have to reconfirm these properties, with no guarantee that the new BGP system would be safe.

Similar to earlier work on BGP convergence properties [20], [4], [5], our guidelines focus on the application of local-pref to prefer some routes over alternatives with a shorter AS path. Since our work aims to prove *positive* results about the stability of the resulting BGP system, it is important to consider the impact of other BGP attributes and the possibility of an AS having multiple BGP speakers. The model in Section III, and the proofs of the theorems in Section V, allow each AS to have one or more BGP speakers. Speakers within the same AS do not necessarily choose the same route. The ultimate routing decision may also depend on AS path length (including paths with AS prepending), multiple exit discriminators, and cost information from the intradomain routing protocol. BGP speakers consider these attributes *after* applying local-pref to the routes learned from neighboring ASs. As such, these additional attributes only impact selection of routes *within* a preference class. For example, AS path length may determine *which* customer route is chosen but would not cause a BGP speaker to pick a provider route over a customer route.

B. Complex AS Relationships

As presented in Section IV, the hierarchical relationships apply at the level of AS pairs. That is, the discussion implicitly assumes that an AS pair has a customer–provider or peer-to-peer relationship for *all* destination prefixes. Since the path selection process proceeds independently for each prefix, this restriction is not actually necessary. In fact, allowing an AS pair to have their relationship depend on the destination prefix is important for expressing more complex policies. For example, two ASs may have both a peer-to-peer and a backup relationship, where each AS provides backup connectivity to the rest of the Internet in the event of a failure. This arrangement does *not* violate our guidelines, since the relationship is still uniquely defined for each destination prefix. The ASs have a peer-to -peer relationship for any prefixes belonging to either AS, and a backup relationship for all other prefixes. The ASs would need to use different ranges of local-pref values based on whether the routes were learned from customers or from providers and other peers.

Similarly, an AS may act as an intermediary between two ASs that would like to establish a peer-to-peer relationship. For example, consider two ASs u and w that would like to have a peer-to-peer arrangement. Suppose that u and w do not have dedicated connections to each other, but that they each have a peer-to-peer relationship with AS v. Normally, an AS would not advertise routes learned from one peer to another peer. But, AS v can agree to export routes learned from u to w (and routes learned from w to u). That is, routes r with $first(r.as\_path) = u$ would be exported to w, and routes with $first(r.as\_path) = w$ would be exported to u. AS v would not export these routes to any of its other peers or providers. This arrangement obeys our guidelines. AS v acts as a provider for u for routes to and from w (and as a provider for w for routes to and from u), and as a peer for all other routes. Hence, guideline A ensures the stability of the resulting BGP system. We believe that a similar approach can be used to analyze other potential relationships between ASs.

C. Changing AS Relationships

Over time, an AS may change the nature of its relationships with its neighbors. For example, a customer may grow large enough to renegotiate its relationship with a provider, and the AS pair may transition to a peer-to-peer relationship. As part of evolving to a new relationship, the two ASs may need to change their import and export policies. Ideally, these changes would occur simultaneously. However, in practice, each AS configures its routers independently of the other. As a result, the BGP system may go through a transition period where one AS has changed its configuration and the other has not. Since these changes occur on a human time scale, it is important to carefully study the influence of the transition period on system stability. Our methodology can be used to identify potential convergence problems, and to determine which AS should change its configuration first. We focus the discussion on a BGP system that obeys guideline A. Similar arguments apply under the other guidelines.

For example, consider a customer u and a provider v that transition to a peer-to-peer relationship. Each AS may change its configuration while remaining consistent with guideline A. AS u does not need to change its export policies since v remains in $provider(u) \cup peer(u)$. Similarly, guideline A does not require u to change its import policies. AS u may in fact modify its local-pref value for routes learned from v, but differences in local-pref within a preference class do not affect system stability. AS u does not need to coordinate with v in making these changes. In contrast, AS v needs to change its import and export policies. AS v stops exporting routes learned from its providers and peers. In addition, the import policy must apply a smaller local-pref to treat u as a peer, rather than a customer. This removes an edge in the provider-to-customer graph. Since removing an edge cannot introduce a cycle, the resulting graph is still a DAG.

Next, we consider a change in the opposite direction, from a peer-to-peer to a customer–provider relationship, where u is the customer and v is the provider. We assume that the final customer–provider configuration does not violate the hierarchy in the AS graph; that is, the final customer-to-provider and provider-to-customer graphs are DAGs. As in the previous example, AS u does not need to change its import and export policies. Hence, u does not need to coordinate with v. AS v changes its export policies to advertise routes learned from other providers and its peers. In addition, v changes its import policies to apply a higher local preference to routes learned from u. Since the changes are isolated to AS v, the BGP system remains safe. Stability problems may arise if multiple ASs transition from peer-to-peer to customer–provider relationships, if the resulting AS graph does not retain its hierarchical structure.

A routing registry could be consulted as each provider changes its configuration, and can flag proposed changes that would violate the hierarchical structure.

The transition is more complicated when a customer–provider relationship changes to a provider–customer relationship. This situation is extremely unlikely to happen in practice, and could be handled by performing two separate transitions from customer–provider to peer-to-peer, and from peer-to-peer to provider–customer. But, for the sake of completeness, we show how the AS pair can directly transition from customer–provider to provider–customer. Initially, u is the customer and v is the provider. Again, we assume that the final configuration does not violate our assumptions of a hierarchical relationship between ASs. We also assume that at most one AS pair changes its relationship at a time. Applying our methodology, we can show that the provider v should change its configuration first. For example, suppose that u changes its configuration first. Then, during the transition period, u sees v as a provider and v sees u as a provider. This introduces two problems. First, there is a cycle in the provider-to-customer graph. Second, both ASs export all routes to each other. The resulting BGP system may not be safe. For example, the two ASs are vulnerable to the scenario in Fig. 2.

Instead, suppose that u changes its configuration first. This removes an edge from the customer-to-provider graph and adds an edge to the provider-to-customer graph. Although the resulting provider-to-customer graph has a cycle, we can show that the BGP system is still safe during this transition period. The provider-to-customer graph has exactly one cycle—the cycle between u and v, since each AS considers the other as a provider. Consider a particular destination prefix d. We consider two cases depending on whether or not one (or both) of the ASs has a customer route to d. Without loss of generality, assume that AS u has a customer route to d. Then, applying guideline A, u would prefer this route over any route via v. Hence, the decision made by v has no influence on u, and the system is safe. In the second case, assume that neither AS has a customer route to d. Then, both u and v must select from routes learned from providers and peer routes. Neither u nor v would export such a route to each other, since a customer does not tell a provider about routes learned from peers or from other providers. Hence, the decision made by each AS does not affect the other, and the BGP system is safe. As such, our methodology demonstrates that the provider v should change its configuration first.

VII. Conclusions

In this paper, we present a detailed model of BGP, along with a set of guidelines for ASs to apply in configuring their BGP import policies. These guidelines capitalize on the commercial relationships between ASs, and provably guarantee route convergence for all possible initial states without requiring global coordination. As part of ongoing work, we are investigating how ASs can verify conformity with our proposed guidelines. Since router configuration files are typically managed by humans, the stability properties can be compromised by human errors. We propose to use the route registry that contains the hierarchical interconnection structure of ASs to check for consistency. For example, export policies should ensure that no AS path has a provider-to-customer link followed by either a customer-to-provider or peer-to-peer link. Using a routing registry, each ISP can verify the validity of a route announcement. The verification can be done statically by periodically checking routing updates or routing table entries; upon identifying an invalid route, the offending AS can be notified. In addition, an AS' router configuration files can be checked to ensure that local-pref values are consistent with the desired relationship with the neighboring AS (and the associated export policies). The iBGP configuration can be checked to ensure that techniques for reducing protocol traffic do not affect the routing decisions.

```
router bgp 7018
    neighbor 10.1.2.118 remote-as 65001
    neighbor 10.1.2.118 route-map INPEER in
    neighbor 10.1.2.118 route-map OUTPEER out
    neighbor 10.126.236.94 remote-as 65002
    neighbor 10.126.236.94 route-map INCUST in
    neighbor 10.126.236.94 route-map OUTCUST out
!
route-map INPEER permit 100
    set local-preference 80
    set community 0:1000
!
route-map OUTPEER permit 100
    match community 20
!
route-map INCUST permit 100
    set local-preference 90
    set community 0:2000
!
route-map OUTCUST permit 100
    match community 10 20
!
ip community-list 10 permit 0:1000
ip community-list 20 permit 0:2000
```

Fig. 7. Sample configuration of BGP sessions to a peer and a customer.

Appendix
Router Configuration

Network operators effect BGP policies by configuring the routers that communicate with neighboring autonomous systems. The Cisco Internet Operating System (IOS) serves as a de facto standard for router configuration. In Fig. 7, we present a small fragment of a configuration file to illustrate how to implement Guideline A outlined in Section V-A-1. The example omits the statements needed to configure the various interfaces on the router, and to associate the BGP sessions with these interfaces. A broader discussion of router configuration, and a more complete configuration example, are presented in [29]. The statements in Fig. 7 associate the router with AS 7018 and define BGP sessions with two neighboring autonomous systems. Each session has three neighbor statements that identify the IP address of the other end of the BGP session. For example, the router has a session to a peer at IP address 10.1.2.118 and a session to a customer at IP address 10.126.236.94. A BGP session with a provider would have the same configuration format as a session with a peer.

Each session is associated with route-maps for the import and export policies. The export policies for OUTPEER and OUTCUST reflect the commercial relationships outlined in Section IV-A. To classify the route advertisements received from neighbors, the import policies INPEER and INCUST tag routes with communities of 0 : 100 0 and 0 : 200 0, respectively. Consequently, any route learned from a peer would have a community of 0 : 100 0, whereas any route learned from a customer would have a community of 0 : 200 0. The export policies match on these community values to determine which routes to advertise to the neighbor. The OUTPEER route-map only exports customer routes, whereas the OUTCUST route-map exports both peer and customer routes. The match statement in each route-map identifies a list of community values defined in a separate community-list statement. For example, community-list 10 consists of the community 0 : 100 0.

The import policies INPEER and INCUST reflect Guideline A outlined in Section V-A-1. The INPEER route-map assigns a local-preference value of 80 to all routes learned from the peer, and the INCUST route-map assigns a local-preference value of 90 to all routes learned from the customer. Different BGP sessions may have different local-preference assignments, as long as all peer and provider route-maps assign a smaller value than all of the customer route-maps. Building on top of this basic example, the network operator could install more sophisticated route-maps that filter other routes or fine-tune the assignment of local preference values. For example, the operator could assign different local preference values to different routes on the same BGP session, as long as the configuration remains consistent with Guideline A.

Acknowledgment

The authors would like to thank A. Feldmann and T. Griffin for their comments on earlier drafts of this paper. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

[1] C. Huitema, *Routing in the Internet*. Englewood Cliffs, NJ: Prentice-Hall, 1999.

[2] Y. Rekhter and T. Li, "A Border Gateway Protocol 4 (BGP-4)," IETF, Request for Comments 1771, Mar. 1995.

[3] J. W. Stewart, *BGP4: Inter-Domain Routing in the Internet*. Reading, MA: Addison-Wesley, 1999.

[4] T. G. Griffin and G. Wilfong, "An analysis of BGP convergence properties," in *Proc. ACM SIGCOMM*, Sept. 1999, pp. 277–288.

[5] K. Varadhan, R. Govindan, and D. Estrin, "Persistent route oscillations in inter-domain routing," Univ. of Southern California Information Sciences Inst., Los Angeles, Tech. Rep. 96-631, Feb. 1996.

[6] D. Meyer, J. Schmitz, C. Orange, M. Prior, and C. Alaettinoglu, "Using RPSL in practice," IETF, Request for Comments 2650, Aug. 1999.

[7] L. Gao and J. Rexford, "Stable Internet routing without global coordination," in *Proc. ACM SIGMETRICS*, June 2000, pp. 307–317.

[8] C. Alaettinoglu, "Scalable router configuration for the Internet," in *Proc. IEEE IC3N*, Oct. 1996.

[9] G. Huston, "Interconnection, peering, and settlements," in *Proc. INET*, June 1999.

[10] L. Gao, "On inferring autonomous system relationships in the Internet," *IEEE/ACM Trans. Networking*, vol. 9, pp. 733–745, Dec. 2001.

[11] L. Subramanian, S. Agarwal, J. Rexford, and R. H. Katz, "Characterizing the Internet hierarchy from multiple vantage points," presented at the IEEE Infocom, June 2002.

[12] S. Halabi and D. McPherson, *Internet Routing Architectures*, second ed. Indianapolis, IN: Cisco Press, 2001.

[13] R. Govindan and A. Reddy, "An analysis of Internet inter-domain topology and route stability," in *Proc. IEEE INFOCOM*, vol. 2, Apr. 1997, pp. 850–857.

[14] C. Labovitz, A. Ahuja, and F. Jahanian, "Experimental study of Internet stability and wide-area network failures," in *Proc. Fault-Tolerant Computing Symp.*, June 1999, pp. 278–285.

[15] C. Labovitz, R. Malan, and F. Jahanian, "Internet routing stability," *IEEE/ACM Trans. Networking*, vol. 6, pp. 515–528, Oct. 1998.

[16] ——, "Origins of pathological Internet routing instability," in *Proc. IEEE INFOCOM*, vol. 1, Mar. 1999, pp. 218–226.

[17] C. Labovitz, A. Ahuja, A. Abose, and F. Jahanian, "Delayed Internet routing convergence," *IEEE/ACM Trans. Networking*, vol. 9, pp. 293–306, June 2001.

[18] R. Govindan, C. Alaettinoglu, G. Eddy, D. Kessens, S. Kumar, and W. Lee, "An architecture for stable, analyzable Internet routing," *IEEE Network Mag.*, vol. 13, pp. 29–35, Jan./Feb. 1999.

[19] T. Griffin and G. Wilfong, "A safe path vector protocol," in *Proc. IEEE INFOCOM*, vol. 2, Mar. 2000, pp. 490–499.

[20] T. G. Griffin, F. B. Shepherd, and G. Wilfong, "Policy disputes in path-vector protocols," in *Proc. Int. Conf. Network Protocols*, Nov. 1999, pp. 21–30.

[21] C. Alaettinoglu, C. Villamizar, E. Gerich, D. Kessens, D. Meyer, T. Bates, D. Karrenberg, and M. Terpstra, "Routing policy specification language (RPSL)," IETF, Request for Comments 2622, June 1999.

[22] R. Govindan, C. Alaettinoglu, K. Varadhan, and D. Estrin, "Route servers for inter-domain routing," *Computer Networks ISDN Syst.*, vol. 30, pp. 1157–1174, 1998.

[23] D. McPherson, V. Gill, D. Walton, and A. Retana. (2001, Mar.) BGP persistent route oscillation condition. IETF, Work in progress, Internet Draft. [Online]. Available: http://www.ietf.org/internet-drafts/draft-ietf-idr-route-oscillation-00.txt

[24] T. H. Cormen, C. E. Leiserson, and R. L. Rivest, *Introduction to Algorithms*. New York: McGraw-Hill, 1990.

[25] H. Chang, R. Govindan, S. Jamin, S. Shenker, and W. Willinger, "On inferring AS-level connectivity from BGP routing tables," , 2001, submitted for publication.

[26] M. Faloutsos, P. Faloutsos, and C. Faloutsos, "On power-law relationships of the Internet topology," in *Proc. ACM SIGCOMM*, Aug. 1999, pp. 251–261.

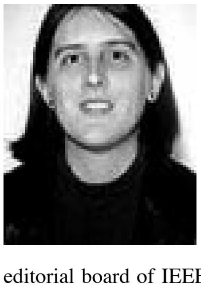
[27] E. Chen and T. Bates, "An application of the BGP Community attribute in multi-home routing," IETF, Request for Comments 1998, Aug. 1996.

[28] L. Gao, T. G. Griffin, and J. Rexford, "Inherently safe backup routing with BGP," in *Proc. IEEE INFOCOM*, vol. 1, Apr. 2001, pp. 547–556.

[29] A. Feldmann and J. Rexford, "IP network configuration for intradomain traffic engineering," *IEEE Network Mag.*, vol. 15, pp. 46–57, Sept./Oct. 2001.

Lixin Gao (M'96) received the Ph.D. degree in computer science from the University of Massachusetts, Amherst (UMass), in 1996.

She is currently an Associate Professor of Electrical and Computer Engineering at UMass. Her research interests include multimedia networking and Internet routing. She was a visiting researcher at AT&T Research Labs and DIMACS between May 1999 and January 2000. She is the founding director of the Multimedia Networking and Internet Research Laboratory at UMass. She is a principal investigator of a number of the National Science Foundation grants, including the NSF CAREER award.



Jennifer Rexford (S'89–M'96–SM'01) received the B.S.E. degree in electrical engineering from Princeton University, Princeton, NJ, in 1991, and the M.S.E. and Ph.D. degrees in computer science and electrical engineering from the University of Michigan, Ann Arbor, in 1993 and 1996, respectively.

She is currently a Member of the Internet and Networking Systems Center at AT&T Labs—Research, Florham Park, NJ. Her research interests include routing protocols, Internet traffic characterization, and multimedia streaming. She is a member of the editorial board of IEEE/ACM Transactions on Networking, and is coauthor of the book *Web Protocols and Practice* (Reading, MA: Addison-Wesley, 2001) with B. Krishnamurthy.