iBGP Deceptions: More Sessions, Fewer Routes

Stefano Vissicchio[†§] Luca Cittadini[†] Laurent Vanbever[‡] Olivier Bonaventure[‡]
[†] Roma Tre University [§] GARR Consortium [‡] Université catholique de Louvain
[†] {vissicch,ratm}@dia.uniroma3.it
[‡] {laurent.vanbever,olivier.bonaventure}@uclouvain.be

***Abstract*—Internal BGP (iBGP) is used to distribute interdomain routes within a single ISP. The interaction between iBGP and the underlying IGP can lead to routing and forwarding anomalies. For this reason, several research contributions aimed at defining sufficient conditions to guarantee anomaly-free configurations and providing design guidelines for network operators.**

In this paper, we show several anomalies caused by defective dissemination of routes in iBGP. We define the *dissemination correctness* property, which models the ability of routers to learn at least one route to each destination. By distinguishing between dissemination correctness and existing correctness properties, we show counterexamples that invalidate some results in the literature. Further, we prove that deciding whether an iBGP configuration is dissemination correct is computationally intractable. Even worse, determining whether the addition of a single iBGP session can adversely affect dissemination correctness of an iBGP configuration is also computationally intractable. Finally, we provide sufficient conditions that ensure dissemination correctness, and we leverage them to both formulate design guidelines and revisit prior results.

I. Introduction

The Border Gateway Protocol (BGP) has two different modes of operation: external BGP (eBGP), which is used among Internet Service Providers (ISPs), and internal BGP (iBGP), which is used within a single ISP to distribute externally learned routes. Routing information in iBGP is exchanged on transport connections called iBGP sessions. Vanilla iBGP did not allow an iBGP router to relay messages to other routers, hence a full mesh of iBGP sessions was needed to ensure correct route distribution. Two mechanisms were proposed to allow iBGP topologies to scale: route reflection and BGP confederations. In this paper, we focus on route reflection as it is the most widely adopted mechanism.

Route reflection trades scalability for correctness guarantees, as it is prone to routing and forwarding anomalies due to the interaction between iBGP and the underlying Interior Gateway Protocol (IGP) [1]. Routing anomalies consist in routing oscillations that prevent iBGP from settling to a stable state. Forwarding anomalies can result in forwarding loops.

In the last decade, the research community has devoted significant effort to design techniques preventing such anomalies. Griffin et al. [1] formalized the absence of routing and forwarding anomalies by introducing two fundamental properties of iBGP configurations, named signaling and forwarding correctness. *Signaling correctness* ensures that BGP will always converge to a stable routing state, while *forwarding correctness* guarantees the absence of packet deflections along the forwarding path. Later, several authors proposed solutions to guarantee signaling and forwarding correctness, either by enforcing special properties on the iBGP configuration (e.g., [2], [3], [4]) or by modifying the protocol itself (e.g., [5], [6]).

In this paper, we show that route propagation rules also play a fundamental role in ensuring correctness of iBGP configurations with route reflection. We give simple examples in which traffic blackholes can be created by the combined effect of iBGP route propagation rules and the iBGP route selection algorithm. Even worse, our examples show that distinct destination prefixes cannot always be analyzed separately. To model the absence of anomalies due to iBGP route propagation rules, we define a new correctness property, called *dissemination correctness*. We show how dissemination correctness fills the gap between signaling and forwarding correctness. Unfortunately, we find that checking dissemination correctness is computationally intractable even when adding a single iBGP session to an existing dissemination correct iBGP configuration. Therefore, we propose sufficient conditions that ensure dissemination correctness and use them to define iBGP design guidelines. In particular, we find that the absence of a special type of iBGP session (*spurious OVER* sessions, as defined in Section IV) guarantees no iBGP route propagation anomalies. Even if uncommon, spurious OVER sessions are sometimes deployed in real-world networks [7], [8]. Indeed, they can be used by network operators to fix forwarding issues and improve route diversity, as suggested in some recent research work (e.g., [9], [10]). Unfortunately, most previous work incorrectly assumes that signaling correctness implies dissemination correctness. Since the presence of spurious OVERs affects the generality of this assumption, we review the state of the art, especially discussing how it relates to dissemination correctness.

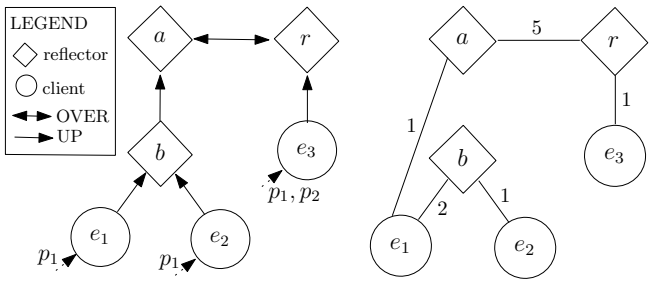
This paper is organized as follows. Section II summarizes iBGP route reflection and shows simple iBGP networks in which routes are not correctly distributed. Section III introduces the model and the notation we adopt in this paper. Section IV analyzes the impact of spurious OVER sessions on iBGP configuration correctness. Section V studies the computational complexity of checking dissemination correctness. Section VI proposes sufficient conditions for dissemination correctness. Section VII revises previous work. Finally, Section VIII concludes the paper.

II. Tweaking iBGP Route Reflection

The behavior of an iBGP router mainly consists of three phases: first, it collects routing information from neighboring

Step	Criterion
1	Prefer routes with higher `local-preference`
2	Prefer routes with lower `as-path` length
3	Prefer routes with lower `origin`
4	Among the routes received from the same AS neighbor, prefer those having lower `MED`
5	Prefer routes learned via eBGP
6	Prefer routes with lower IGP metric
7	Prefer routes having the lowest `egress-id`
8	Prefer routes with shorter `cluster-list`
9	Prefer the route coming from the neighbor with lower IP address

TABLE I
BGP DECISION PROCESS.



(a) iBGP topology (b) IGP topology

Fig. 1. A simple network that exhibits visibility issues.

iBGP routers; second, it selects the best route; third, it selectively propagates its best route to neighboring routers.

With route reflection, the iBGP neighbors of each router are split into three sets: *clients*, *peers* and *route-reflectors*. Each iBGP router propagates its best route according to the following rules: if the route is learned from a peer or from a route-reflector, then it is relayed only to clients, otherwise it is reflected to all iBGP neighbors. Organizing routers in a hierarchy of clients and route-reflectors allows iBGP to scale. A *cluster* consists of one or more route-reflectors and all their clients. Whenever not explicitly stated, we assume that every cluster has a single route-reflector. Each cluster is identified through a unique `cluster-id`. Messages carry a `cluster-list` attribute, which accounts for the iBGP path and is used to avoid control-plane loops.

In the following, we refer to a session between a client and a route reflector as an UP session if it is traversed from the client to the route-reflector, as a DOWN session otherwise. Also, we refer to a session between iBGP peers as an OVER session. We call the organization of iBGP sessions *iBGP topology*.

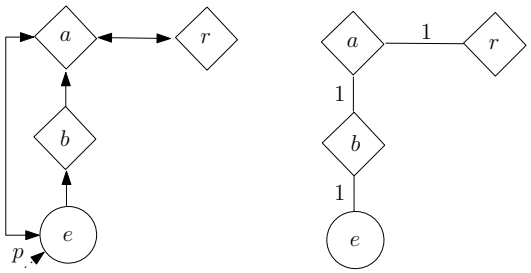
Best route selection is performed at each iBGP router according to the BGP decision process summarized in Table I. The BGP decision process consists of a set of rules: whenever there are ties for a rule, the next rule is applied. We refer the reader to [11], [12] for a detailed description of the BGP decision process. The evaluation of Steps 1-4 is the same at every iBGP router, since those steps consider global attributes, usually not modified in iBGP [1]. Throughout the paper, we only consider routes that are equally preferred according to the first four steps of the BGP decision process. We denote the routers that receive an eBGP route for a given prefix as *egress points* for that prefix. Each iBGP router has an identifier, i.e., a `router-id`. The `egress-id` of a route is the `router-id` of the egress point that announces that route.

As an example of iBGP network, consider the network depicted in Fig. 1(a). The graphical convention adopted in the figure will be used throughout the paper. Circles represent iBGP clients, while diamonds represent iBGP route-reflectors. UP and OVER sessions are depicted with single and double arrow links, respectively. The dashed arrows labeled p_1 entering routers e_1, e_2 and e_3 represent the fact that e_1 and e_2 are egress points for prefix p_1. Similarly, e_3 is an egress points for both p_1 and p_2. The underlying IGP graph is depicted in Fig. 1(b): lines represent IGP links and labels represent the IGP weight assigned to a link.

Consider prefix p_1. Due to step 5 of the BGP decision process, routers e_1, e_2 and e_3 will select their external route denoted as R_1, R_2 and R_3, respectively. Therefore, they will advertise their best route to all their iBGP neighbors, namely b (for e_1 and e_2) and r (for e_3). Router b will collect routes from its clients, select its best route, and propagate it to its neighbors. By step 6 of the BGP decision process, b will select route R_2 because e_2 is a closer egress point than e_1. Therefore, by iBGP propagation rules b will advertise R_2 to its route-reflector a. Each router will keep performing route collection, route selection and route dissemination until BGP converges and no further messages are propagated. After convergence, router r will select route R_3 and router a will select route R_2.

Observe that router a has no knowledge of route R_1, because it only receives route R_2 from b and route R_3 from r. In fact, route reflection introduces suboptimal route visibility and limits the amount of route diversity available at router a. Another side effect induced by route reflection is the packet deflection that happens when a sends traffic to prefix p_1. More precisely, a believes that the traffic will exit from egress point e_2 and forwards it to e_1 because it is the next hop to e_2. However, e_1 is itself an egress point for prefix p_1, so it will deflect traffic outside the ISP. The combination of multiple deflections can result in forwarding loops [1].

Whenever issues due to suboptimal route visibility arise, fixing them by adding additional iBGP sessions may look like an easy and tempting solution for a network operator. In our example, adding an iBGP session between routers a and e_1 will provide a with increased route diversity and will make it able to select its optimal egress point. The addition of OVER sessions to increase route diversity in iBGP has been already proposed in [9], [10], e.g., to support recently proposed techniques for reducing iBGP convergence time [13]. Indeed, quantitative studies have already shown that route reflection leads to very poor route diversity [14]. This, in turn, can cause high convergence time in case of failure or interdomain routing changes. Moreover, additional sessions can provide better route visibility to routers, thus making it easier for a network operator to fix its iBGP configuration in order to comply with state of the art guidelines [3].



(a) iBGP topology (b) IGP topology

Fig. 2. OVER-RIDE GADGET

Observe that, in general, additional iBGP sessions do not need to be OVER sessions, i.e., they could be UP sessions as well. However, network operators might prefer to deploy OVER sessions, in order to lower memory overhead and update churn, as only a subset of reflected routes is announced on OVER sessions.

Unfortunately, adding OVER sessions to an iBGP topology may have undesirable side effects. Consider the iBGP network in Fig. 2 (OVER-RIDE GADGET) which is a simplified version of the one in Fig. 1. An additional OVER session exists between routers a and e. Since e is the only egress point for prefix p, a will prefer the route that it learns on the OVER session because of step 8 of the BGP decision process. Then, since its best route was learned from a peer, a will not propagate it to r, so r will have no route to prefix p.

Now if r has a route for a less specific prefix that includes prefix p (e.g., a default route), it will use that route for traffic destined to p, possibly generating forwarding deflections and loops. Consequently, it is not safe to assume that prefixes are independent in iBGP. Otherwise, if r has no route for a less specific prefix than p, r will create a traffic blackhole. Observe that both kinds of anomalies are due to the iBGP topology alone. The IGP topology is irrelevant in this case because there is only one egress point for p. For this reason, the OVER-RIDE GADGET complies with the conditions of [3], yet it is subject to anomalies. Even worse, such anomalies could be triggered by external events, e.g., if an egress point fails.

III. A MODEL FOR IBGP CORRECTNESS

We now present the model we use in the rest of the paper. We model an IGP graph as an undirect weighted graph $I = (V, E)$, with a weight associated to each edge $(u, v) \in E$. We denote with $dist(u, v)$ the total weight of the shortest path from u to v. Moreover, we model an iBGP topology as a directed labeled multigraph $B = (V, E)$ where nodes in V represent routers and edges in E represent iBGP sessions. Each edge (u, v) is associated with a label which is either UP, DOWN, or OVER. We use $u \leftarrow v$, $u \rightarrow v$, and $u \leftrightarrow v$ to indicate that the label of edge (u, v) is DOWN, UP or OVER, respectively. Because of the way iBGP relationships are defined, $u \leftarrow v \Leftrightarrow v \rightarrow u$, and $u \leftrightarrow v \Leftrightarrow v \leftrightarrow u$.

Due to the iBGP route dissemination rules, not every path on B can be used to distribute a BGP route announcement. We define a *valid signaling path* as a path $(u \ \ldots \ v)$ on B that can be used to advertise routes from u to v (or vice versa). A *valid signaling path* consists of zero or more UP sessions, followed by zero or one OVER session, followed by zero or more DOWN sessions. This means that a valid signaling path matches regular expression UP*OVER?DOWN* [4]. The presence of a valid signaling path between u and v is a necessary condition for u to learn routes announced by v, even if we show in Section IV that it is not a sufficient condition. Throughout the paper, we assume that the iBGP graph B is connected, that is, $\forall u, v \in B$ there is a valid signaling path from u to v, otherwise obvious forwarding anomalies can arise (routes are not propagated network-wide). Whenever it is clear from the context, we use a signaling path to refer to the route advertised over that signaling path (e.g., we say that a router receives a path, or prefers a path over another).

Route reflection topologies are usually organized in a hierarchy where there are no cycles consisting of UP sessions only. Indeed, such cycles are a sign of bad topology design and can create routing anomalies [1]. In a hierarchy, each BGP router can be assigned to a layer. We denote the set of routers in the top layer of an iBGP topology B as T_B. A router belongs to the top layer T_B if it has no route-reflector.

It has been shown [1] that the suboptimal route visibility introduced by route reflection can cause both routing and forwarding anomalies. Routing anomalies can prevent BGP to settle to a stable state because of *routing oscillations*. Moreover, inconsistent routing decisions between the forwarding plane and the control plane can create forwarding deflections and loops. A BGP configuration is said to be *signaling correct* if it is free from routing anomalies, i.e., if BGP is guaranteed to always converge to a single predictable stable state. A signaling correct configuration is *forwarding correct* if it is always free from forwarding anomalies. Observe that there are no guarantees that all the routers have a route towards all the prefixes even in a signaling correct BGP configuration.

A. Known Sufficient Conditions for Correctness

The following set of sufficient conditions guarantees that an iBGP topology B is both signaling and forwarding correct [1].

1) B has no cycles consisting of UP sessions only;
2) any route-reflector prefers paths propagated by its clients over paths propagated by non-clients; and
3) all shortest paths must also be valid signaling paths.

Conditions 1 and 2 ensure that the iBGP configuration is signaling correct, while Condition 3 guarantees forwarding correctness. Although interesting from a theoretical perspective, such conditions can be too constraining to be applied in real networks. For example, Condition 3 practically forces the BGP topology to be congruent to the IGP one, in such a way that even a full-mesh of iBGP sessions is not compliant. We discuss the applicability of Condition 2 in Section VI.

In [3], [4] the concept of *fm-optimality* is introduced as a relaxed sufficient condition to ensure forwarding correctness in a signaling correct iBGP configuration. To understand fm-optimality, we need to define white routers and white paths [4].

Given an iBGP topology B, a router r and an egress point e, a router r' is said to be a *white router* for pair (r, e) if there is no egress point e' in B such that $dist(r, e') > dist(r, e)$ and $dist(r', e') \leq dist(r', e)$. A *white path* between a router r and an egress point e is defined as a valid signaling path between r and e that contains only white routers for pair (r, e). An iBGP topology is *fm-optimal* if for each router r and for each egress point e there exists at least one white path.

IV. Unveiling iBGP Deceptions

In this section, we introduce the concept of spurious OVER sessions. Also, we show how their side effects can invalidate simple assumptions that apparently hold in any iBGP topology, and have been used in previous research work.

Definition 1: Given an iBGP topology B, an OVER session $x \leftrightarrow y$ is *spurious* if one of the two routers is not in the top layer, i.e., if $x \notin T_B$ or $y \notin T_B$.

Spurious sessions are not frequent in today's ISP networks. Vendor guidelines also suggest to not deploy them [11]. Nevertheless, spurious sessions have been proposed to solve visibility issues [9], [10], and previous work showed that large ISPs sometimes use them [7], [8]. Moreover, spurious OVERs can be unintentionally introduced in iBGP reconfigurations. For example, current best practices to replace an iBGP full-mesh with route reflection [15] suggest to progressively introduce UP sessions before removing the full-mesh. Hence, OVER sessions initially in the full-mesh are likely to become spurious in intermediate configurations.

A. Route Dissemination Deceptions

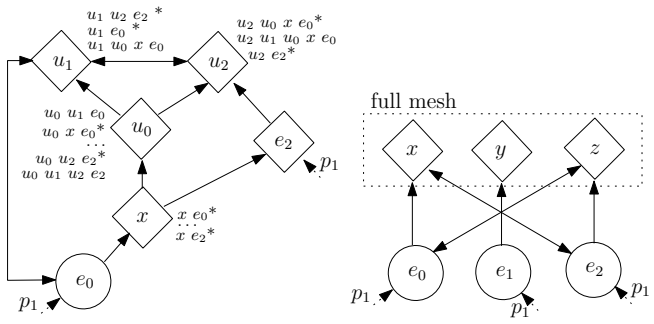
As discussed in Section II, the OVER-RIDE GADGET provides an example of how a spurious OVER improves egress point visibility at some routers, but potentially worsens visibility at other routers. In the gadget, the side effect of adding a spurious OVER is counter-intuitive because it induces a change in the route dissemination process at router r without affecting the egress point selected by r. This contradicts the intuition that a connected iBGP topology guarantees that every router eventually learns at least one route for any given prefix.

Unfortunately, some previous work is based on that intuition. In particular, [9], [10] assume that adding an OVER session can only improve route visibility, while [3], [4] assume that a route-reflector r can "hide" a route to a neighboring router v only if it has a closer alternative egress point.

More generally, spurious OVER sessions show that the concept of valid signaling path is not a good abstraction to study the actual ability of a router to learn a route to a given prefix. In order to better understand this property, we introduce the concept of *dissemination correctness*.

Definition 2: Let B be a signaling correct iBGP topology. Then, B is *dissemination correct* if all the routers in B are guaranteed to receive at least one route to prefix p in the stable state, for any non-empty set of egress points for p.

Observe that dissemination correctness does not depend on interdomain routing nor on the set of egress points currently learning routes for given prefixes. That is, it is a topological property. Dissemination correctness differs from both signaling and forwarding correctness. Indeed, a signaling correct topology is not guaranteed to be dissemination correct. Moreover, a dissemination correct topology is not guaranteed to be forwarding correct. The three properties actually complement each other: signaling correctness deals with routing anomalies that can prevent BGP from converging; dissemination correctness deals with issues in the route propagation process; forwarding correctness deals with forwarding anomalies caused by the interaction between iBGP and IGP.



(a) A spurious OVER can create routing oscillations.

(b) A spurious OVER can cause forwarding loops.

Fig. 3. Two cases in which adding a spurious OVER creates signaling and forwarding anomalies.

B. Signaling and Forwarding Correctness Deceptions

Beside affecting dissemination correctness, a single spurious OVER can even prevent an iBGP topology to be either signaling or forwarding correct, as shown in Fig. 3.

Consider Fig. 3(a). Every router is equipped with a list of valid signaling paths, sorted in decreasing order of preference. Observe that (u_1, e_0) is a spurious OVER session. We now show that iBGP cannot converge in this configuration. Assume by contradiction that a stable state exists, and consider the routing choice at router u_2. Since u_2 receives a route directly from e_2, it is not possible that u_2 does not select any route for prefix p_1. Hence, we have the following cases.

- u_2 steadily selects $(u_2\ e_2)$. In this case, u_1 will use its most preferred path $(u_1\ u_2\ e_2)$, preventing u_0 from selecting $(u_0\ u_1\ e_0)$. Thus, u_0 will select $(u_0\ x\ e_0)$, and eventually announce it to u_2. Because of path preferences, u_2 should switch to $(u_2\ u_0\ x\ e_0)$, yielding a contradiction.
- u_2 steadily selects $(u_2\ u_1\ u_0\ x\ e_0)$. This involves that u_1 steadily selects $(u_1\ u_0\ x\ e_0)$, leading to a contradiction, since path $(u_1\ e_0)$ is always available at u_1 and is more preferred than $(u_1\ u_0\ x\ e_0)$.
- u_2 steadily selects $(u_2\ u_0\ x\ e_0)$. This implies that u_0 steadily selects $(u_0\ x\ e_0)$, and u_1 is forced to select $(u_1\ e_0)$, since it does not receive path $(u_2\ e_2)$ from u_2. This leads to a contradiction, since u_0 will eventually learn and select path $(u_0\ u_1\ e_0)$, preventing u_2 from steadily selecting $(u_2\ u_0\ x\ e_0)$.

All the cases lead to a contradiction, hence a stable state does not exist in the topology in Fig. 3(a). Observe that the

path preferences highlighted in the figure can result from the standard BGP decision process (Table I) if the IGP topology is such that $dist(x, e_0) < dist(x, e_2)$, $dist(u_0, e_0) < dist(u_0, e_2)$, $dist(u_2, e_0) < dist(u_2, e_2)$, and $dist(u_1, e_0) = dist(u_1, e_2)$. In this case, x, u_0, and u_2 prefer paths based on the closest egress point, while u_1 prefers eBGP routes received from e_2 over those received from e_0 for `egress-id`. Ties are broken by shorter `cluster-list` and lower neighbor address criteria. Also notice that, in such a configuration, the iBGP topology in Fig. 3(a) is fm-optimal as defined in [3]. For each router, its white paths for egress points e_0 and e_2 are marked with an asterisk.

Forwarding correctness can also be affected by the presence of spurious OVER sessions. Consider the topology in Fig. 3(b), and assume that x steadily selects path $(x\ e_2)$, while z steadily selects path $(z\ e_0)$, because of the IGP distances. Since those paths are learned via an OVER session, x and z will not propagate their best route to y, hence y will be forced to select the route from e_1. If y is on x's shortest path to e_2 and x is on y's shortest path to e_1, then a loop arises for p_1.

V. Checking Dissemination Correctness is Hard

In this section, we study the computational complexity of deciding whether a given iBGP topology is dissemination correct. Unfortunately, we find that such a problem is computationally intractable. Even worse, we show that the problem of deciding if the addition of a single session can affect the dissemination correctness of an iBGP topology is also computationally intractable.

We formally define the problems we consider as follows.

Dissemination Correctness Problem (DCP): Given a signaling correct iBGP topology B and the underlying IGP topology I, decide if B is dissemination correct.

One More Session Problem (OMSP): Given a dissemination correct iBGP topology $B = (V, E)$, the underlying IGP topology I, and a spurious OVER session $o = (x, y), x, y \in V$, decide if $B' = (V, E \cup (x, y))$ is dissemination correct.

Observe that *DCP* is the iBGP equivalent of the REACHABILITY problem defined for eBGP in [16].

A. Dissemination Correctness is coNP-Hard

We now prove that *DCP* is coNP-hard [17]. Intuitively, computational complexity of *DCP* mainly depends on the fact that all the non-empty sets of egress points have to be checked in the general case. In the following, we show that the *3-SAT COMPLEMENT* problem [17] can be reduced to *DCP* in polynomial time. Consider an instance of *3-SAT COMPLEMENT* and let F be a logical formula in conjunctive normal form. Moreover, let $C_1, \ldots, C_n$ be the clauses in F, and let $X_1, \ldots, X_m$ be the boolean variables appearing in the clauses. Each clause C_i is the logical disjunction ("or") of exactly 3 literals L_{ij} with $j = 1, 2, 3$. A literal L_{ij} can be either a variable X_l or a negated variable $\bar{X}_l$. The *3-SAT COMPLEMENT* problem consists in deciding if F is unsatisfiable, that is, if no boolean assignment makes F true.

We now build the corresponding instance of *DCP* (see Fig. 4), following an intuition similar to that used in [1] for proving that signaling correctness is NP-hard. The skeleton of the iBGP topology $B = (V, E)$ consists of 4 nodes, e, s, r, and b connected as in the OVER-RIDE GADGET. In particular, $e \rightarrow s$, $s \rightarrow r$, and $e \leftrightarrow r$. Moreover, $r \leftrightarrow b$ since $b, r \in T_B$. For each variable X_i, we add two *literal nodes* x_i and $\bar{x}_i$ to V, representing the two literals associated to X_i. For each clause C_j, we add a clause node c_j and three nodes v_{j1}, v_{j2}, and v_{j3}. We add OVER sessions between c_j and v_{ji}, $i \in \{1, 2, 3\}$. Also, each $c_j \in T_B$, hence it is in the top layer full-mesh. We also add an UP session from e to c_j. Moreover, two UP sessions (x_k, v_{ji}) and $(\bar{x}_k, v_{ji})$ are added to E iff either X_k or $\bar{X}_k$ is the i^{th} literal appearing in clause C_j. Finally, we add an UP session between each v_{ji} and r.

Fig. 4(b) shows an example of the IGP topology resulting from a clause C_1 in which X_j (X_l, resp.) appears negated (unnegated, resp.). In particular, for any clause C_j, if variable X_i appears unnegated in the k^{th} literal of C_j, then we set $dist(c_j, x_i) < dist(c_j, e) < dist(c_j, \bar{x}_i)$, and $dist(v_{jk}, \bar{x}_i) < dist(v_{jk}, x_i)$. For any router $n \neq e, x_i, \bar{x}_i$, we set $dist(c_j, \bar{x}_i) < dist(c_j, n)$ and $dist(v_{jk}, x_i) < dist(v_{jk}, n)$. Otherwise, if variable X_i appears negated in the k^{th} literal of C_j, we set IGP metrics such that x_i is replaced with $\bar{x}_i$ and vice versa in the above inequalities. Finally, we set IGP metrics in such a way that r and s prefer routes announced by e over all other routes, and the shortest paths from r to e and to any x_i and $\bar{x}_i$ traverse s.

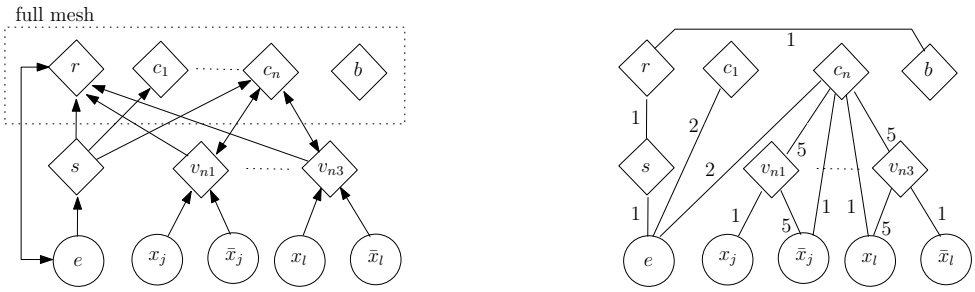
Intuitively, a boolean assignment M corresponds to a set S_M of egress points for a given prefix p. Router x_i ($\bar{x}_i$, resp.) belongs to S_M iff X_i is true (false, resp.) in M. Also, router e always belongs to S_M.

A *3-SAT COMPLEMENT* instance can be reduced to a *DCP* one in polynomial time, since each clause and each variable is mapped to a polynomial number of routers and links. We now show that the reduction is correct.

Lemma 1: B is signaling correct. Also, if e is not an egress point for a prefix p, all routers in B are guaranteed to receive a route to p; otherwise, b may not receive a route to p.

Proof: Consider prefix p and let $S \neq \oslash$ be the set of egress points for p. Abusing the notation a bit, we refer to routers x_i and $\bar{x}_i$ as to x-routers, and similarly we refer to v-routers and c-routers. We have two cases: $e \in S$ and $e \notin S$.

First, assume $e \notin S$. In this case, all x-routers in S steadily select an eBGP route to p because of step 5 of the BGP decision process. The v-routers that have at least one client in S steadily select the route propagated by one of their clients, because of the IGP metrics. Router r receives routes from all v-routers that have at least one client in S. Since $S \neq \oslash$, we conclude that r is able to select a route to p announced by a v-router. Router r's best route is then forwarded to all r's neighbors, because it was learned from a client. Observe that all the shortest paths from r to a router in S contain s, which implies that s will select the same route as r. For this reason, e will receive the same route from s and from r and will steadily select it. Every c-router learns a route from r



(a) iBGP topology.

(b) IGP topology.

Fig. 4. Reduction from *3-SAT COMPLEMENT* to *DCP*

and possibly additional routes from its v-peers. In any case, c-routers' best routes can only be propagated to s due to iBGP route reflection rules. This cannot affect the route selected by s. Router b and v-routers having no clients in S receive a single route, i.e., the one announced by r. For this reason, x-routers that are not in S receive at least one route. Those routers do not propagate their best route further because it is learned from a route-reflector. Hence, iBGP is guaranteed to converge to a state in which all routers have one route to p.

We now consider the case in which $e \in S$. Again, all routers in S steadily select their eBGP route and announce it to their respective route-reflectors. The v-routers that have at least one client in S steadily select the route propagated by one of their client, because of IGP metrics. Let R_e be the route learned by e. IGP metrics imply that routers r and s steadily select R_e. Every c-router learns at least R_e from router s, and possibly additional routes from their v-peers. Such additional routes (if any) can be propagated to s only, which prefers R_e due to IGP metrics. Again, v-routers having no clients in S receive a single route from r, i.e., R_e. Also, x-routers that are not in S receive at least one route, but they do not propagate their best route further because of the iBGP topology. Observe that router b may or may not learn a route to prefix p. Since b's decision cannot influence any other router, we conclude that iBGP is guaranteed to converge. ■

Theorem 1: *DCP* is coNP-hard.

Proof: Consider a logical formula F and construct the corresponding *DCP* instance $B = (V, E)$, as described above. By Lemma 1, B is signaling correct. Also, Lemma 1 states that B is dissemination correct if router e does not receive an external route. Thus, we focus on cases in which e receives an external route. We prove the statement in two parts.

If F is unsatisfiable then B is dissemination correct.

Assume by contradiction that B is not dissemination correct, i.e., there exists a set of egress points S_M for prefix p such that at least one router in B does not receive any route to p. By Lemma 1, such a router must be b. We now build a boolean assignment M that satisfies F, yielding a contradiction.

Since b does not receive any route, every c-router must select a route learned by one of their peers, with $k = 1, 2, 3$.

Let C_j be a clause and assume that X_i appears unnegated in the k^{th} literal of c_j. Then, router c_j selects a route propagated by v_{jk} only if v_{jk} selects the route originated by x_i, since $dist(c_j, x_i) < dist(c_j, e) < dist(c_j, \bar{x}_i)$. In turn, router v_{jk} selects the route originated by x_i only if x_i is an egress point for p and $\bar{x}_i$ is not, since $dist(v_{jk}, \bar{x}_i) < dist(v_{jk}, x_i)$. Symmetrical considerations hold if X_i appears negated in the k^{th} literal of C_j. In both cases, we are able to find a boolean assignment to variable X_i that makes clause C_j true.

Iterating the same argument on all the clauses, we can map S_M to a boolean assignment M which satisfies F.

If F is satisfiable then B is not dissemination correct.

Let M be a boolean assignment that satisfies F. We now show that B is not dissemination correct, since there exists a set S_M of egress points such that if a prefix p is learned at S_M then b receives no route to p.

By definition of M, all clauses are satisfied in M, hence for any clause C_j at least one literal must be true. Assume, without loss of generality, that the k^{th} literal of C_j is true. If the k^{th} literal of C_j is X_i, then we impose that router x_i receives an eBGP route R to prefix p, while router $\bar{x}_i$ does not receive any eBGP routes to p. Since $dist(v_{jk}, x_i) < dist(v_{jk}, e)$, router v_{jk} selects route R and propagates it to router c_j. Similarly, since $dist(c_j, x_i) < dist(c_j, e)$, c_j selects route R. Otherwise, if the k^{th} literal of C_j is $\bar{X}_i$, we can apply the same argument by replacing x_i with $\bar{x}_i$. In both cases, c_j selects a route propagated by an iBGP peer.

Since the above argument applies to all clauses, we have that every c_j selects a route learned from an iBGP peer. Router r also selects a route learned from an iBGP peer, because of the presence of OVER session (r, e) (see Section II). Hence, every router which is a neighbor of b selects a route learned from an iBGP peer, thus b receives no route for prefix p. ■

B. Distinguishing Harmless Sessions is coNP-Hard

A similar reduction to that described in Section V-A can also be used to show that *OMSP* is coNP-Hard.

Starting from a logical formula in conjunctive normal form, we build the *OMSP* instance as follows. B coincides with the BGP topology in Fig. 4(a) without OVER session (r, e), I is as depicted in Fig. 4(b), and $o = (r, e)$.

Using the same arguments as in the proof of Lemma 1, it can be shown that B is dissemination correct. However, deciding if $B' = (V, E')$, with $E' = E \cup \{o\}$, is dissemination correct is coNP-hard, because of Theorem 1. In other words, we cannot exploit the knowledge that an input iBGP network

is dissemination correct to efficiently check whether adding an arbitrary OVER session preserves dissemination correctness.

VI. Guaranteeing Dissemination Correctness

In this section, we propose sufficient conditions for dissemination correctness, and we discuss their applicability. Firstly, we prove that one of the sufficient conditions for signaling correctness given in [1] also ensures dissemination correctness. Unfortunately, we find that this condition is hard to enforce, especially when current design best practices [18] are followed. Hence, we present a simpler sufficient condition.

A. Sufficient Conditions for Dissemination Correctness

Either of the following conditions guarantees a signaling correct iBGP topology B to be dissemination correct.

1) *prefer-client*: all iBGP routers in B prefer routes propagated by clients (on a UP* path) to any other route.
2) *no-spurious-OVER*: B contains no spurious OVER.

In order to prove our results, we need the following lemma.

Lemma 2: Given a signaling correct iBGP topology B, if for any prefix p at least one router in the top layer T_B selects a route for p that was learned over an UP* path, then B is dissemination correct.

Proof: Consider any prefix p, and let $\bar{r} \in T_B$ be the router that selects a route $\bar{R}$ to p which was learned over an UP* valid signaling path $(e \dots \bar{r})$ (possibly $e = \bar{r}$). By iBGP route propagation rules, $\bar{r}$ propagates route $\bar{R}$ to all routers in T_B. Since B is signaling correct and all routers in T_B receive at least one route for p, all routers in T_B will eventually select a route. Independent of the neighbor from which the best route was learned, routers in T_B will propagate their best route to all their clients, which are then guaranteed to receive a route for p. These routers, in turn, will announce their own best route to their clients, and so on until routers in the bottom layer are reached. Then, we conclude that every router receives at least one route for prefix p, hence B is dissemination correct. ■

In the following theorems, we prove that *prefer-client* and *no-spurious-OVER* guarantee dissemination correctness.

Theorem 2: Given a signaling correct iBGP topology B, if B complies with the *prefer-client* condition, then B is dissemination correct.

Proof: We now prove that for any prefix p at least one router r in T_B selects a route to p over an UP* path. Then, the statement follows because of Lemma 2.

Let p be a prefix and e_p be an egress point for p receiving an eBGP route R. Because of step 5 of the BGP decision process, e_p selects R. If $e_p \in T_B$, then $r = e_p$. Otherwise, there must exist a router r_1 such that $r_1 \leftarrow e_p$, by definition of T_B. Because of iBGP dissemination rules, r_1 receives at least route R from e_p. Let R' (possibly $R' = R$) be the route that r_1 selects in the stable state. Since r_1 receives route R from a client, the *prefer-client* condition implies that route R' is also received from a client. Again, if $r_1 \in T_B$ then $r = r_1$. Otherwise, iBGP dissemination rules force r_1 to propagate route R' to all its route-reflectors. Let r_2 be one of the route-reflectors of r_1, that is, $r_2 \leftarrow r_1$. Observe that r_2 must exist since $r_1 \notin T_B$. Again, r_2 receives at least route R' from its client r_1, so we can apply the same argument to r_2. We can iterate the argument until we reach a router r in T_B that learns a route from one of its clients. Because of iBGP propagation rules, that route must be learned over an UP* path. ■

Theorem 3: Let B be a signaling correct iBGP topology with no spurious OVER. B is dissemination correct.

Proof: We now prove that for any prefix p at least one router in T_B selects a route to p over an UP* path.

Let e_p be a router that receives an eBGP route R towards p. Because of step 5 of the BGP decision process, e_p selects R. If $e_p \in T_B$, then the statement follows by Lemma 2. Otherwise, there must exist a router r_1 such that $r_1 \leftarrow e_p$. Because of iBGP dissemination rules, r_1 receives at least route R from e_p. Let R' (possibly, $R' = R$) be the route that r_1 selects in the stable state. We have the following cases.

- $r_1 \in T_B$ and r_1 learned R' from one of its clients. By the iBGP propagation rules, R' must be learned over an UP* path.
- $r_1 \in T_B$ and r_1 learned R' from a peer r_2. In this case, r_2 must have received R' over an UP* path, otherwise it would not have propagated it to r_1.
- $r_1 \notin T_B$ and r_1 learned R' from one of its clients. Then, r_1 forwards route R' to all its route-reflectors.
- $r_1 \notin T_B$ and r_1 learned R' from one of its route-reflectors.

Observe that the *no-spurious-OVER* condition implies that r_1 cannot learn R' from a peer if $r_1 \notin T_B$.

In the first two cases, the statement follows by Lemma 2. In the last two cases, there must exist a router r_2, with $r_2 \leftarrow r_1$, such that r_2 learns a route for prefix p. Hence, we can iterate the same argument on r_2. Since the number of layers in B is finite, we eventually find a router in T_B for which one of the first two cases applies, yielding the statement. ■

B. Applicability of the Sufficient Conditions

We now discuss how the sufficient conditions presented above can be enforced in real-world iBGP topologies.

In theory, the *prefer-client* condition can be enforced by carefully designing iBGP topologies. However, we find that this condition is too constraining for real-world topologies. In fact, in order to satisfy the *prefer-client* condition each router should rank the routes it receives according to the first hop in the iBGP signaling path, while the BGP decision process uses tie-breaking criteria that are based on the last hop in the signaling path (i.e., `egress-id`) or on the length of the path itself (i.e., `cluster-list`). In particular, a direct consequence of condition *prefer-client* is that, if a router r has a valid signaling path $P = (r\ s \dots e)$ with $r \leftarrow s$ (possibly $s = e$), then any other valid signaling path between r and e must either have a client of r as next-hop or be longer than P. Hence, satisfying the *prefer-client* condition requires a deep evaluation of all the decision steps in the iBGP decision process. For this reason, it becomes a really hard task when deploying redundant route-reflectors, even on very simple topologies. Consider, for example, the configuration in Fig. 5,

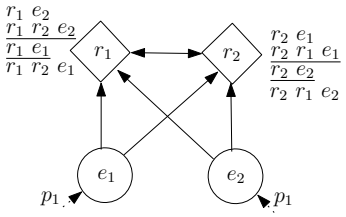


Fig. 5. Redundant topologies hardly satisfy the *prefer-client* condition.

which is the simplest redundant route reflection topology designed according to current best practices [11], [18]. Clients e_1 and e_2 are connected to both route-reflectors r_1 and r_2, and r_1 and r_2 belong to different clusters. Both e_1 and e_2 are egress points for prefix p_1. Even in such a simple scenario, the *prefer-client* condition does not hold, whatever the IGP topology is: underlined paths highlight violations of the *prefer-client* condition. In fact, consider router r_1, and assume that e_2 is its closest egress point according to IGP metrics. In this case, r_1 prefers all the routes received by e_2 to all the routes received by e_1, because of step 6 in the BGP decision process. Hence, r_1 prefers routes learned over path $(r_1\ r_2\ e_2)$ to those over path $(r_1\ e_1)$. This violates the *prefer-client* condition. A similar violation happen if $dist(r_1, e_1) < dist(r_1, e_2)$. This kind of violations of the *prefer-client* condition can be solved by a wiser design of route-reflector clusters. Indeed, if r_1 and r_2 belong to the same cluster, then r_1 always discards routes propagated by r_2 and vice versa [12].

Guideline A: In redundant iBGP configurations, redundant route-reflectors must belong to the same cluster in order to enforce the *prefer-client* condition.

Observe that current best practices for cluster design [18] do not comply with Guideline A.

The *no-spurious-OVER* condition is relatively easier to enforce, since it only imposes constraints on the iBGP topology and does not require to evaluate the whole BGP decision process at every router. However, there might be cases in which additional (spurious) sessions are desirable to locally fix forwarding issues or to improve route diversity, as discussed in Section IV. In such cases, UP sessions can be deployed instead of spurious OVERs, without adversely affecting dissemination correctness.

Guideline B: Whenever an additional session is needed to solve visibility issues, an UP session should be deployed, in order to enforce the *no-spurious-OVER* condition.

Observe that using UP sessions is not free from possibly undesired side effects, e.g., shortening the *cluster-list* of existing signaling paths, change the layering of the hierarchy, impact router memory, etc. Some of these side effects can be mitigated, e.g., by configuring route filters that allow route propagation in one direction only.

VII. Revisiting the State-of-the-Art

In this section, we discuss how dissemination correctness relates to previous work on iBGP correctness properties and topology design. We find that dissemination correctness was often overlooked, so extra conditions (see Section VI) are needed to keep the validity of the results. This section collects previous contributions to the best of our knowledge, hence there might be other results affected by incorrect assumptions on iBGP route dissemination.

Signaling and forwarding correctness have been introduced and analyzed by Griffin *et al.* in [1]. The authors show that checking either of the two properties is NP-hard and give sufficient conditions to enforce both of them. While the concept of dissemination correctness is not envisaged in [1], we find that the proposed sufficient conditions also guarantee dissemination correctness, since they encompass the *prefer-client* condition as formulated in Section VI. However, as discussed in Sections III and VI, these conditions are very constraining for real-world networks.

In [2], Rawat and Shayman give a set of sufficient conditions that guarantee signaling and forwarding correctness and also prevent MED-induced routing oscillations. In particular, one of the conditions in [2] imposes that, for any router, IGP distances to clients must be shorter than IGP distances to non-clients. While this conditions is intended to be a variant of the *prefer-client* condition, it is not enough to prevent dissemination anomalies caused by multiple valid signaling paths to the *same egress point*, as the OVER-RIDE GADGET demonstrates. Moreover, Fig. 3(b) shows an example which matches the conditions of [2] but is not forwarding correct.

In [5], Flavel and Roughan propose a modified BGP decision process that evaluates the length of the `cluster-list` before comparing IGP weights. Such a variant of iBGP is proved to always converge. However, no guarantee is given for dissemination correctness. Actually, the OVER-RIDE GADGET is a simple example where the modified iBGP protocol cannot provide all routers with a route for every prefix.

In [3], [4], Buob *et al.* introduce the concept of fm-optimality, which models the visibility issues that arise when two routers in a valid signaling path disagree on which egress point is the closest one. Fm-optimality is said to guarantee forwarding correctness. Unfortunately, the fm-optimality concept does not account for visibility issues caused by iBGP route propagation rules, e.g., in presence of spurious OVER sessions. In other words, even if all routers on the signaling path agree on which egress point is the closest one, dissemination correctness is not guaranteed. As an example, the OVER-RIDE GADGET is fm-optimal but not dissemination correct.

In [9], [10] Pelsser *et al* propose to add spurious OVER sessions to locally fix visibility issues. Our results show that such a local fix comes at the cost of potential visibility issues on remote routers. Section VI discusses alternatives to spurious OVER sessions that provides similar benefits with no impact on dissemination correctness.

A more general consequence of our work is that the presence of a valid signaling path P between a router r and an egress point e is not sufficient to ensure that r has visibility of routes announced by e (e.g., in the OVER-RIDE GADGET). In fact, depending on both the IGP and the iBGP topology, there might be some routers in P that do not propagate to

r the route announced by e. Observe that such a counter-intuitive behavior affects Lemma 3 of [19], where the presence of an UP*DOWN* path for each pair of routers is said to guarantee full visibility. On the contrary, since only best routes are propagated, the iBGP topology design technique proposed in [19] guarantees signaling and dissemination correctness, but cannot guarantee forwarding correctness. Also, conclusions drawn in [7] are similarly affected. Indeed, configuring a top layer full-mesh (as prescribed by Theorem 4.1 in [7]) guarantees a valid signaling path for each pair of iBGP routers, but does not imply dissemination correctness.

Despite the concept of dissemination correctness had not been formalized before, we find that some results in the literature guarantee it as a side effect.

Modifications to the iBGP protocol as proposed in [6] and fine tuning of attributes in iBGP messages as proposed in [20] can be leveraged to enforced the *prefer-client* condition. In both cases, however, the likelihood of incurring suboptimal routing increases, since client routes are preferred, no matter what are the IGP distances of the corresponding egress points.

Recently, BGP Add-Paths [21] has been proposed to allow routers to propagate multiple routes. It is important to note that the advertisement of multiple routes guarantees dissemination and forwarding correctness only if all the routes that are equally preferred according to the first four steps of the BGP decision process (so called *AS dominant routes*) are propagated network-wide. However, the higher number of routes handled in iBGP could cause router memory and update churn penalties [22]. Raszuk *et al.* [23] propose to add special route-reflectors in order to distribute multiple routes. Unfortunately, since this technique relies on additional route-reflectors, it does not guarantee the advertisement of all the AS dominant routes, and thus it is not sufficient for dissemination correctness. Packet encapsulation is suggested in both cases to solve forwarding anomalies when not every AS dominant route is propagated. Observe that both proposals are still in the development stage.

VIII. Conclusions

iBGP route reflection provides network operators with good scalability at the cost of possibly introducing routing and forwarding anomalies. In this paper, we show that iBGP route propagation anomalies are also possible, triggering unexpected side effects like traffic blackholes and forwarding loops. Moreover, the ability of iBGP to correctly distribute routing information within an ISP can be affected by the addition of even a single iBGP session. This is particularly relevant as prior contributions proposed to fine tune iBGP by adding extra sessions. Hence, we introduce the concept of dissemination correctness to model visibility issues caused by iBGP route propagation rules. We study the computational complexity of checking dissemination correctness and provide sufficient conditions to enforce it in real-world configurations.

We thoroughly review previous work and discuss how existing results relate to dissemination correctness, finding that some contributions need to be revisited. In our opinion, this study shows that iBGP semantics are actually more complex than what is commonly assumed, and provides new motivation to recent efforts (e.g., [21], [24], [25]) for decoupling route propagation from route selection in iBGP.

IX. Acknowledgements

We thank Steve Uhlig, Marc-Olivier Buob, Cristel Pelsser, and INFOCOM anonymous reviewers for constructive comments. Laurent Vanbever is supported by a FRIA scholarship. Stefano Vissicchio is partially supported by MIUR PRIN Project AlgoDEEP.

References

[1] T. Griffin and G. Wilfong, "On the correctness of IBGP configuration," in *Proc. SIGCOMM*, 2002.
[2] A. Rawat and M. Shayman, "Preventing persistent oscillations and loops in IBGP configuration with route reflection," *Comput. Netw.*, vol. 50, pp. 3642–3665, 2006.
[3] M. Buob, M. Meulle, and S. Uhlig, "Checking for optimal egress points in iBGP routing," in *Proc. DRCN*, 2007.
[4] M. Buob, S. Uhlig, and M. Meulle, "Designing optimal iBGP route-reflection topologies," in *Proc. Networking*, 2008.
[5] A. Flavel and M. Roughan, "Stable and flexible iBGP." in *Proc. SIGCOMM*, 2009.
[6] R. Musunuri and J. Cobb, "A complete solution for iBGP stability," in *Proc. ICC*, 2004.
[7] N. Feamster and H. Balakrishnan, "Detecting BGP configuration faults with static analysis," in *Proc. NSDI*, 2005.
[8] J. H. Park, P. chun Cheng, S. Amante, D. Kim, D. McPherson, and L. Zhang., "Quantifying i-BGP Convergence inside Large ISPs," UCLA, Tech. Rep., 2011.
[9] C. Pelsser, T. Takeda, E. Oki, and K. Shiomoto, "Improving route diversity through the design of iBGP topologies." in *Proc. ICC*, 2008.
[10] C. Pelsser, S. Uhlig, T. Takeda, B. Quoitin, and K. Shiomoto, "Providing scalable NH-diverse iBGP route redistribution to achieve sub-second switch-over time," *Comput. Netw.*, vol. 54, no. 14, pp. 2492–2505, 2010.
[11] R. Zhang and M. Bartell, *BGP Design and Implementation*. Cisco Press, 2003.
[12] T. Bates, E. Chen, and R. Chandra, "BGP Route Reflection: An Alternative to Full Mesh Internal BGP (IBGP)," RFC 4456, 2006.
[13] C. Filsfils, P. Mohapatra, J. Bettink, P. Dharwadkar, P. D. Vriendt, Y. Tsier, V. Van Den Schrieck, O. Bonaventure, and P. Francois, "BGP Prefix Independent Convergence (PIC)," Cisco, Tech. Rep., 2007.
[14] S. Uhlig and S. Tandel, "Quantifying the BGP routes diversity inside a tier-1 network," in *Proc. Networking*, 2006.
[15] G. Herrero and J. Van der Ven, *Network Mergers and Migrations: Junos Design and Implementation*. Wiley Publishing, 2010.
[16] T. Griffin and G. T. Wilfong, "An analysis of BGP convergence properties," in *Proc. SIGCOMM*, 1999.
[17] C. Papadimitriou, *Computational complexity*. Addison-Wesley, 1994.
[18] P. Smith, "BGP Techniques for Service Providers," NANOG 50, 2010.
[19] M. Vutukuru, P. Valiant, S. Kopparty, and H. Balakrishnan, "How to Construct a Correct and Scalable iBGP Configuration," in *Proc. INFOCOM*, 2006.
[20] L. Cittadini, G. Di Battista, and S. Vissicchio, "Doing don'ts: Modifying BGP attributes within an autonomous system," in *Proc. NOMS*, 2010.
[21] D. Walton, A. Retana, E. Chen, and J. Scudder, "Advertisement of multiple paths in BGP," Internet Draft, July 2011.
[22] V. Van den Schrieck, P. Francois, and O. Bonaventure, "BGP Add-Paths: The Scaling/Performance Tradeoffs," *Journ. on Select. Areas in Comm.*, vol. 28, no. 8, pp. 1299 – 1307, October 2010.
[23] R. Raszuk, R. Fernando, K. Patel, D. McPherson, and K. Kumaki, "Distribution of diverse BGP paths," Internet Draft, 2011.
[24] I. Oprescu, M. Meulle, S. Uhlig, C. Pelsser, O. Maennel, and P. Owezarski, "oBGP: an Overlay for a Scalable iBGP Control Plane," in *Proc. IFIP Networking*, 2011.
[25] R. Chen, A. Shaikh, J. Wang, and P. Francis, "Address-based Route Reflection," in *Proc. CoNEXT*, 2011.