Inter-domain Traffic Engineering as Bi-level Network Flow Optimization

Praveen K. Muthuswamy, Koushik Kar, Aparna Gupta
Rensselaer Polytechnic Institute
Troy, NY 12180, USA
Email: {muthup,kark,guptaa}@rpi.edu

Murat Yuksel
University of Nevada, Reno
Reno, NV 89557, USA
Email: yuksem@cse.unr.edu

***Abstract*—In this work, we study the inter-domain traffic engineering problem with the perspective of attaining socially optimal flows across the Internet. In accordance with the two-level (intra- and inter-AS) hierarchy that naturally exists in the Internet architecture, we use bi-level network flow decomposition to derive inter-domain traffic engineering solutions that do not require ISPs to reveal their intra-domain topologies in their information exchange with each other. The distributed solutions that we obtain through convex optimization techniques require ASes to exchange destination-specific rate information across the inter-domain links (i.e., between adjacent gateway routers belonging to different ASes), and determining the intra-domain traffic rates between the edge-routers based on the state of congestion of the edge-nodes in the AS, and the cost of traffic routing through the AS which in turn depends on the AS's intra-domain routing/traffic engineering conditions. We investigate the optimality and convergence properties of these solutions through simulations on networks generated using realistic inter- and intra-domain topology models.**

I. Introduction

The existing routing architecture of the Internet follows a two-level model, which are commonly referred to as *intra-domain* and *inter-domain* routing. Routing at the intra-domain level often employs proactive approaches focusing on reliability and quality. Intra-domain traffic engineering has become common practice, with numerous methods such as MPLS [1], DiffServ [2], RSVP [3], and link weight optimization for Intra-Gateway Protocols (IGPs) such as OSPF and IS-IS [4]. Using these methods, researchers have proposed both online and offline algorithms that consider network throughput optimization, congestion avoidance, interference minimization, and increasing network reliability [5], [6], [7], [8], [9], [10]. In contrast, inter-domain routing/traffic engineering (TE) practices have not changed much. The current inter-domain routing and pricing strategies are typically policy driven, and do not in general optimize the overall use of Internet resources. Inter-domain routing follows the BGP standard [11] that only attempts to minimize the number of AS-hops on the path of a flow (in addition to following local policy considerations and certain heuristic rules). Inter-domain TE techniques are mostly constrained to outbound traffic load balancing; and other TE goals and desired networking practices are only expressed indirectly. Common expression methods include increasing LOCALPREF, ASPATH, or MED at peering points where less inbound traffic is desired [26], [13]. In current practice, traffic pricing is typically based on the total traffic offered by the customer ISP, irrespective of their destination (point-to-anywhere pricing) [14]. It can be argued that point-to-point or destination-based TE and pricing can result in better efficiency of network resource usage [15]. Moreover, recent measurements also show fast-changing patterns for inter-domain traffic [16]; inter-domain TE and pricing solutions need to be flexible enough to adapt to such dynamically changing traffic patterns. Another challenge in this context is the counter-productive effects of intra-domain and inter-domain TE policies if they are not designed with careful consideration of their interaction with each other [17], [18].

This paper focuses on the design of cooperative mechanisms between ISPs, or the constituent autonomous systems (ASes) or "domains" managed by them, towards attaining optimal TE objectives in the overall Internet. Even if cooperation between ISPs, or their conformity to a chosen standard, is assumed, several challenges exist in designing effective inter-domain TE mechanisms. Firstly, for scalability and compliance with existing inter-domain practices, the TE solutions must be implementable in a *fully distributed* manner, requiring message exchanges only between *neighboring* ASes (domains). Secondly, to satisfy ISPs' confidentiality/security requirements (which is key to the acceptability of a chosen solution/protocol by the ISPs, the major stakeholders), the chosen solution must require an ISP to only share minimal information about its topology, current conditions or traffic statistics of its (intra-domain) networks. It should also allow each ISP the full flexibility of implementing its own intra-domain TE solutions.

We design inter-domain TE solutions that satisfy the above requirement, and seek to attain socially (globally) optimal flows in the Internet. In accordance with the two-level (intra- and inter-AS) hierarchy that naturally exists in the Internet architecture, this question can naturally be posed as a bi-level traffic network flow optimization question. The lower level problem corresponds to the TE solution followed by each AS (which may differ across ASes, being based on the AS's intra-domain preferences and practices), and determines the path followed by the traffic inside the AS. The upper level problem corresponds to inter-domain TE, that determines how much traffic the AS accepts at its edge (gateway) routers from its neighboring ASes. This decision is also coupled with how the transit traffic accepted at an edge (ingress) node of an AS is split across the other edge (egress) nodes. In this paper, we focus on this upper level problem, which determines the

inter-domain TE solution, given the solutions to the lower level problem (i.e., the intra-domain routing/TE solution is pre-determined). We use separable convex optimization to develop distributed iterative solutions that attain optimal (or near-optimal) traffic rates in the sense that they can minimize some separable convex cost for the overall Internet (aggregated across all domains), given the intra-domain TE solutions of the individual ASes. The proposed solution requires each AS to declare (to its neighboring ASes) how much traffic rate it wants to send or accept *for each destination* across an extra-domain (inter-domain) link, and is accordingly called the *destination-rate vector* inter-domain TE approach. This TE solution can be realized as an extension to existing inter-domain message exchange protocols, and allows ASes to run their own intra-domain routing/TE methods, and keep their topology and intra-AS state information private. It is fully distributed in its nature, and converges to yield inter-domain traffic flow rates that minimize the traffic routing cost across the entire Internet, given the intra-domain routing/TE policies of the different ASes. In the following, we briefly survey related work on this topic in Section II. Sections III and IV formulates the inter-domain TE problem and described our destination-vector TE solution approach. Simulation based evaluation of the convergence of the proposed TE solution on realistic Internet topologies is provided in Section V.

II. Related Work

For inter-domain routing/TE, much literature has been devoted to instability issues of BGP, and the choice of stable routes for path-vector based routing protocols like BGP, e.g., [19], [20], [21], [22]. Research in inter-domain TE has gained momentum over the last decade, although the bulk of this work has been devoted to inter-domain TE solutions in the context of BGP, through intelligent use of some of its parameters/flexibilities. [23] and [24] discuss methods for doing inter-domain TE by careful control of BGP route advertisements, while [25] discusses how the BGP AS-Path attribute can be manipulated for that purpose. General guidelines for TE within the context of BGP are discussed in [26]. [27] discusses an extension to BGP to enable multipath inter-domain routing.

Limitations of inter-domain TE have attracted many researchers to studying cooperative TE mechanisms where neighboring ISPs work on feasible traffic management outcomes for the benefit of both parties [28], [32]. Promising benefits of cooperative TE have driven research community to develop signaling protocols to mediate such negotiation and distributed decision making [30], [31]. Among existing work, [32], [33], [18] are the most closely related to the decomposable network flow optimization based perspective that we adopt for cooperative inter-domain TE. These approaches, like the one we envision, do not require the individual ISPs to reveal their intra-domain topology or state directly, but only in the form of transit/forwarding costs. A cooperative optimization-based approach to inter-domain TE based on dual decomposition and Nash bargaining is described in [32]; however the analysis is done only for a set of two ISPs. A symbiotic optimization based inter-domain TE approach is discussed in [33], but no formal analysis is presented on the optimality of the approach. The authors in [18] study flow equilibrium efficiency with optimal routing at the intra-domain level and selfish (non-cooperative) routing across domains. [18] shows that Braess' paradox may exist in this case, in that optimal intra-domain routing need not necessarily improve overall network performance, and show that the worst case inefficiency is bounded by 25%. There is a growing body of literature in non-cooperative inter-domain routing and pricing questions [34], [35], [36], [37], [41], [42]. In contrast, we take a cooperative approach to inter-domain TE, in line with the current inter-domain practices.

III. System Model and Preliminaries

We model each domain or Autonomous System (AS) as a collection of routers connected via capacitated links into an arbitrary topology. A set of these routers are *edge-routers*, each of which can serve as an ingress or egress of traffic that the AS receives from, or sends to, its neighboring ASes. In other words, traffic can enter or leave the domain only through these edge (gateway) routers. Therefore, from the perspective of other ASes, or for the purpose of inter-domain TE, we can abstract each AS as a collection of the edge-routers of the AS, and "logical links" (directed) connecting the AS's edge-routers to each other, and to other edge-routers belonging to neighboring ASes. This abstraction, which is discussed in greater detailed in our recent paper [43], is illustrated in Figure 1. We call these logical links between edge-routers *edge-links*. The edge-links connecting two edge-routers of two different (neighboring) ASes are called *extra-domain* or *e*-edge-links; the edge-links connecting edge-routers of the same AS are termed *intra-domain* or *i*-edge-links. Although not necessary, an *e*-edge-link will often span just a single physical link between the two edge-routers (of different ASes) it connects. An *i*-edge-link may correspond to one of more multi-hop paths between the two edge-routers it connects, and the traffic on that edge-link is mapped to those physical paths based on the intra-domain TE policy that the AS follows.

Let the set of all ASes in the Internet be denoted by $\mathcal{N}$. Let $\mathcal{R}_j$ denote the set of edge routers of the AS $j \in \mathcal{N}$, and let $\mathcal{L}^i_j = \{\{k, k'\} : k \neq k' \in \mathcal{R}_j\}$ denote the set of *i*-edge-links of AS j. Let $\mathcal{R} = \cup_{j\in\mathcal{N}}\mathcal{R}_j$ and $\mathcal{L}^i = \cup_{j\in N}\mathcal{L}_j$ respectively denote the set of all edge-routers, and the set of all *i*-edge-links, in the Internet. Let $\mathcal{L}^e$ be the set of all *e*-edge-links in the Internet. Let $\mathcal{L}^e_j = \{(k, k') \in \mathcal{L}^e, k \text{ or } k' \in \mathcal{R}_j\}$ denote the set of *e*-edge-links that are incoming to, or outgoing from, the edge-routers in AS j. Let $\mathcal{L} = \mathcal{L}^i \cup \mathcal{L}^e$ denote the set of all edge-links in the Internet.

Let D denote the set of all destination networks for the overall Internet, each of which is connected to one of the edge-routers. Let e_d denote the edge router to which the destination network d is attached. Traffic is generated from source networks, each of which is also connected to one of the edge-routers. For any edge-router k, let λ^d_k denote the rate of traffic generated at the source network(s) connected to the edge-router, that is destined to destination network d.

For any edge-link (k, k'), let $g_{kk'}$ denote the aggregate traffic (across all destination networks) that is sent on the edge-link, i.e., from edge-router k to edge-router k'. Let

$\mathbf{g}_j^i = \{g_{kk'} : (kk') \in \mathcal{L}_j^i\}$ denotes the traffic rate vector on the i-edge-links of AS j. For feasibility, the edge-link traffic rate vector $\mathbf{g}_j^i$ must belong to a set $\mathcal{G}_j^i$, so that the traffic rate vector is routable through AS j's network, without violating the capacity constraints on any of the physical links. Clearly, the feasibility set $\mathcal{G}_j^i$ would depend on AS j's network topology and intra-domain routing/TE policy. We make the reasonable assumption that the set $\mathcal{G}_j^i$ is convex for any j. We also assume that carrying the traffic rate vector $\mathbf{g}_j^i$ incurs AS j a cost of $V_j(\mathbf{g}_j^i)$, which is convex in $\mathbf{g}_j^i$. This cost can be viewed as the "congestion cost" or "forwarding cost" associated with routing the traffic through the AS. As an example, let us assume that all flow on an i-edge-link follows a single, fixed intra-AS path (say the min-hop path between the corresponding two edge-routers in the AS, as typically done by intra-domain routing protocols like RIP and OSPF). Let $\Theta_{kk'}$ denote the set of physical links on the path over which the traffic flow of i-edge-link (k, k') is routed. Then the traffic load on any physical link (l, l') is $\nu_{ll'} = \sum_{(k,k')\in\mathcal{L}_j^i:(l,l')\in\Theta_{kk'}} g_{kk'}$. If we interpret total (average) delay of traffic in the AS as the measure of the congestion (forwarding) cost, then the cost could be approximated as (the M/M/1 delay formula), $\sum_{(l,l')\in L_j} 1/(c_{ll'} - \nu_{ll'})$, where L_j is the set of physical links in AS j's network, and $c_{ll'}$ is the link capacity of (l, l'). Note that this cost is convex in the traffic rate vector $\mathbf{g}_j^i$. In this case the constraint set $\mathcal{G}_j^i = \{\mathbf{g}_j^i = \{g_{kk'}, (kk') \in \mathcal{L}_j^i : \sum_{(k,k')\in\mathcal{L}_j^i:(l,l')\in\Theta_{kk'}} g_{kk'} \le c_{ll'}\}$. Note that this constraint set is automatically satisfied if we consider delay as the minimization objective; however, for other cost functions, the constraint $\mathbf{g}_j^i \in \mathcal{G}_j^i$ must be imposed separately.

Since e-edge-links correspond to capacitated paths (often just a single physical link), we associate a cost of $v(g_{kk'})$ with sending a flow of $g_{kk'}$ on e-edge-link (k, k'). We assume that e-edge-link (k, k') is associated with a capacity of $C_{kk'}$, which corresponds to the maximum or reserved capacity on the extra-domain link or path associated with e-edge-link (k, k'). Therefore, for any e-edge-link (k, k'), $g_{kk'} \le C_{kk'}$ for feasibility. For simplicity, in our model we ignore the capacities of the access links that connect the source and destination networks with their closest edge-routers.

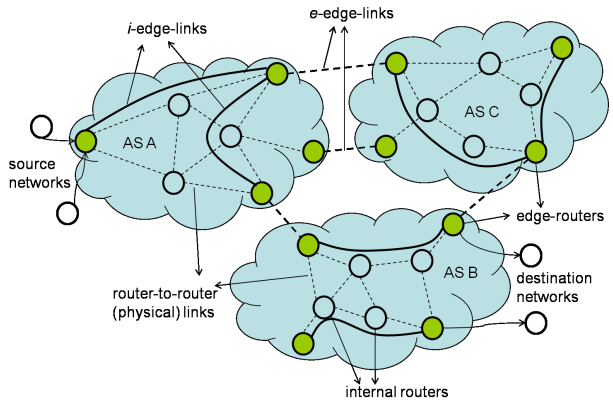

Fig. 1. For inter-domain TE, an AS is abstracted as set of edge(gateway)-routers and logical edge-links connecting them.

IV. Destination-Rate Vector Inter-domain TE

For each edge-link (k, k'), let $f_{kk'}^d$ denote the traffic rate on the edge-link corresponding to destination network d. Let $\mathbf{f}_j^i = \{f_{kk'}^d, (k, k') \in \mathcal{L}_j^i, d \in D\}$ denote the traffic rate vector on all i-edge-links of AS j. Also, let the vector $\mathbf{f}_j^e = \{f_{kk'}^d, (k, k') \in \mathcal{L}_j^e, d \in D\}$ denote the traffic rate vector on all (incoming plus outgoing) e-edge-links of AS j. Then the inter-domain policy of AS j corresponds to specifying $\mathbf{f}_j^e$, or in other words, determining the traffic rates (for each destination) that AS j can send from, or accept at, its edge-routers, and exchanging such information among neighboring edge-routers. We naturally call this model the Destination-Rate Vector (DRV) model. As we will see shortly, the inter-domain TE solution also depends on how the traffic accepted by the AS for forwarding is split across the intra-domain edge-links, and the cost of routing the traffic through the AS's network which in turn depends on the intra-domain TE policy.

A. TE Objective Formulation

In this framework, the global (inter-domain) TE objective would be to minimize the traffic congestion (forwarding) cost over all ASes, subject to the necessary flow conservation and physical link capacity constraints. In other words, the inter-domain TE problem in our destination-rate vector (DRV) model, **DRV-TE**, can be posed as:

$$\textbf{DRV-TE}: \quad \text{minimize} \sum_{j\in\mathcal{N}} V_j(\mathbf{g}_j^i) + \sum_{(k,k')\in\mathcal{L}^e} v(g_{kk'}), \quad (1)$$

$$\text{subject to:} \quad g_{kk'} = \sum_{d\in D} f_{kk'}^d, \ \forall (k,k') \in \mathcal{L}; \quad (2)$$

$$\mathbf{g}_j^i \in \mathcal{G}_j^i, \ \forall j \in \mathcal{N}; \quad (3)$$

$$g_{kk'} \le C_{kk'}, \ \forall (k,k') \in \mathcal{L}^e; \quad (4)$$

$$\sum_{k':(k',k)\in\mathcal{L}} f_{k'k}^d + \lambda_k \le \sum_{k':(k,k')\in\mathcal{L}} f_{kk'}^d, \forall k \in \mathcal{R} \setminus e_d, \ \forall d \in D; \quad (5)$$

$$f_{kk'}^d \ge 0, \ \forall (k,k') \in \mathcal{L}, \forall d \in D. \quad (6)$$

In the objective function (1), the first term $\sum_{j\in\mathcal{N}} V_j(\mathbf{g}_j^i)$ represents the total intra-domain congestion (forwarding) cost; while the second term, $\sum_{(k,k')\in\mathcal{L}^e} v(g_{kk'})$, represents the total extra-domain congestion (forwarding) cost. Constraint (3) along with (2) correspond to feasibility conditions on the i-edge-link traffic rates, resulting from physical link capacity constraints in AS j, as discussed in Section III. Inequalities (4) are capacity constraints on the e-edge-links. Constraints (5) correspond to the flow balance constraints at the edge-routers, and state that the total incoming traffic rate at an edge-router k must be no greater than the total outgoing rate at the edge-router. One could argue that this relation should be an equality; it is easy to see, however, that there is no loss of generality in relaxing it with an inequality constraint, as we have done here. In fact, since the cost functions $V_j(\cdot)$ and $v(\cdot)$ will in general be strictly increasing functions of the traffic rate vector, the relation will be satisfied as an equality at optimum; representing it as an inequality aids in developing the TE solution that we present next.

B. Distributed Solution Approach

Next we outline an approach that attains the above-described TE objective in an iterative, distributed manner. The approach is derived from considering a gradient descent approach (with the destination-specific traffic rates on the edge-links, $f^d_{kk'}$, as the variables) to the penalized objective function of **DRV-TE**, where the penalty function is computed from constraints (5). To derive the approach, first note that constraints (3) and (2) for the *i*-edge-links are equivalent to a set of feasibility constraints on the traffic rate vector $\mathbf{f}^i_j$, for each AS j; let these feasibility constraints be represented by $\mathbf{f}^i_j \in \mathcal{F}^i_j$. Since we assume that $\mathcal{G}^i_j$ is a convex set, it is easy to show that $\mathcal{F}^i_j$ is convex as well. Constraints (4) along with (2) for the *e*-edge-links, can be represented by the feasibility sets $\mathcal{F}^e_{kk'} = \{f^d_{kk'}, d \in D : \sum_{d\in D} f^d_{kk'} \leq C_{kk'}\}$, for all $(k,k') \in \mathcal{L}^e$. We denote $\mathbf{f}_{kk'} = \{f^d_{kk'}, d \in D\}$, and write the feasibility (capacity) constraints on the *e*-edge-links compactly as $\mathbf{f}_{kk'} \in \mathcal{F}^e_{kk'}$. Note that since $\mathbf{g}^i_j$ is completely specified by $\mathbf{f}^i_j$ (from (2)), we can write $V_j(\mathbf{g}^i_j)$ in the objective function as $\hat{V}_j(\mathbf{f}^i_j)$ (which is also convex in its argument), for each AS j. Similarly, we can write $v(g_{kk'})$ in the objective function as $\hat{v}(\mathbf{f}_{kk'})$, for each *e*-edge-link (k,k').

Let $\hat{\mathbf{f}}^d_k = \{f^d_{k'k}, (k',k) \text{ or } (k,k') \in \mathcal{L}\}$ be the set of incoming and outgoing edge-link traffic rates for destination network d at edge-router k. Then the constraints (5) corresponding to edge-router k and destination network d ($k \in \mathcal{R} \setminus e_d$) can be represented as $h_{k,d}(\hat{\mathbf{f}}^d_k) \leq 0, \forall$, where $h_{k,d}(\hat{\mathbf{f}}^d_k) = \sum_{k':(k',k)\in\mathcal{L}} f^d_{k'k} + \lambda_k - \sum_{k':(k,k')\in\mathcal{L}} f^d_{kk'}$, or the "overload" in destination network d's traffic at edge-router k. We associate an increasing penalty function P with this overload, $P(h_{k,d}(\hat{\mathbf{f}}^d_k))$. In general, the penalty should be "small" (or close to zero), when $\hat{\mathbf{f}}^d_k$ is such that $h_{k,d}(\hat{\mathbf{f}}^d_k) \leq 0$, and "high" when $h_{k,d}(\hat{\mathbf{f}}^d_k) > 0$. We then seek to solve a modified (approximate) version of the problem **DRV-TE**, parameterized by the chosen penalty function $P(\cdot)$, where the constraints (5) are removed by incorporating them in the objective through the penalty function. This problem, which we call **approx-DRV-TE**(P), can be posed as:

approx-DRV-TE(P): minimize

$$\sum_{j\in\mathcal{N}} \hat{V}_j(\mathbf{f}^i_j) + \sum_{(k,k')\in\mathcal{L}^e} \hat{v}(\mathbf{f}_{kk'}) + \sum_{d\in D}\sum_{k\in\mathcal{R}\setminus e_d} P(h_{k,d}(\hat{\mathbf{f}}^d_k)), \quad (7)$$

$$\text{subject to: } \mathbf{f}^i_j \in \mathcal{F}^i_j, \ \forall j \in \mathcal{N}; \quad (8)$$

$$\mathbf{f}_{kk'} \in \mathcal{F}^e_{kk'}, \ \forall (k,k') \in \mathcal{L}^e; \quad (9)$$

$$f^d_{kk'} \geq 0, \ \forall (k,k') \in \mathcal{L}, \forall d \in D. \quad (10)$$

Under certain reasonable assumptions, it is possible to design the penalty function P in such a way that the original problem **DRV-TE** is equivalent to minimizing the penalized objective [40](Chapter 4.7); in other words, the problems **DRV-TE** and **approx-DRV-TE**(P) are equivalent. In general, however, **approx-DRV-TE**(P) represents an approximation to **DRV-TE**. The inter-domain TE policy that we now propose corresponds to iteratively updating the traffic rate variables in the steepest descent direction of the objective function (7), subject to projection on the sets $\mathcal{F}^i_j \ \forall j \in \mathcal{N}$, and $\mathcal{F}^e_{kk'} \ \forall (k,k') \in \mathcal{L}^e$, and the positive orthant, to account for the constraints (8), (9) and (10), respectively.

Let us now compute the different components of the gradient of the objective function in (7). In the gradient of the total cost term $\sum_{j\in\mathcal{N}} \hat{V}_j(\mathbf{f}^i_j) + \sum_{(k,k')\in\mathcal{L}^e} \hat{v}(\mathbf{f}_{kk'})$, the component corresponding to $f^d_{kk'}$ for an *i*-edge-link belonging to AS j is simply $\nabla \hat{V}_j(\mathbf{f}^i_j)|_{f^d_{kk'}}$, i.e. the component corresponding to $f^d_{kk'}$ in the gradient of $\hat{V}_j(\mathbf{f}^i_j)$. Note that $\nabla \hat{V}_j(\mathbf{f}^i_j)|_{f^d_{kk'}} = \nabla V_j(\mathbf{g}^i_j)|_{g_{kk'}}$, or the component corresponding to $g_{kk'}$ in the gradient of $V_j(\mathbf{g}^i_j)$. Similarly, for an *e*-edge-link (k,k'), in the gradient of the total cost term, the component corresponding to $f^d_{kk'}$ is simply $\nabla \hat{v}(\mathbf{f}_{kk'})|_{f^d_{kk'}}$, which equals $v'(g_{kk'})$. Finally, $\Phi^d_{kk'}$, defined as the partial derivative of the penalty term $\sum_{d\in D}\sum_{k\in\mathcal{R}\setminus e_d} P(h_{k,d}(\hat{\mathbf{f}}^d_k))$ with respect to the traffic rate variable $f^d_{kk'}$ for any edge-link (intra- or extra-domain) (k,k'), is expressed as

$$\Phi^d_{kk'} = \begin{cases} P'(h_{k',d}(\hat{\mathbf{f}}^d_{k'})) - P'(h_{k,d}(\hat{\mathbf{f}}^d_k)) & \text{if } k,k' \neq e_d; \\ P'(h_{k',d}(\hat{\mathbf{f}}^d_{k'})) & \text{if } k = e_d; \\ -P'(h_{k,d}(\hat{\mathbf{f}}^d_k)) & \text{if } k' = e_d. \end{cases}$$

We next describe the gradient projection update step for the optimization problem **approx-DRV-TE**(P), which results in our inter-domain TE policy. First, based on the above discussion, we define $\Delta^d_{kk'}$, a generic component in the gradient of the objective function (7) as follows. If $(k,k') \in \mathcal{L}^i$, then for any $d \in D$,

$$\Delta^d_{kk'} = \nabla V_j(\mathbf{g}^i_j)|_{g_{kk'}} + \Phi^d_{kk'}. \quad (11)$$

Otherwise, i.e. if $(k,k') \in \mathcal{L}^e$, then for any $d \in D$,

$$\Delta^d_{kk'} = v'(g_{kk'}) + \Phi^d_{kk'}. \quad (12)$$

Now define $\mathbf{\Delta}^i_j$, the vector of the gradient component terms for all per-destination traffic rates on the *i*-edge-links of AS j, as $\mathbf{\Delta}^i_j = (\Delta^d_{kk'}, (k,k') \in \mathcal{L}^i_j, d \in D)$, where $\Delta^d_{kk'}$ is defined in (11). Also, for any *e*-edge-link (k,k'), define $\mathbf{\Delta}_{kk'}$, the vector of the gradient component terms for all per-destination traffic rates on the *e*-edge-link, as $\mathbf{\Delta}_{kk'} = (\Delta^d_{kk'}, (k,k'), d \in D)$, where $\Delta^d_{kk'}$ is defined in (12). The gradient projection method for vector $\mathbf{f}^i_j$ for any AS j is now written as

$$\mathbf{f}^i_j(t+1) = [\mathbf{f}^i_j(t) - \alpha_t \mathbf{\Delta}^i_j]_{\mathcal{F}^i_j \cap \Re^+}, \quad (13)$$

where α_t is the step-size at iteration t. Similarly, the gradient projection method for vector $\mathbf{f}_{kk'}$ for any *e*-edge-link (k,k') is now written as

$$\mathbf{f}_{kk'}(t+1) = [\mathbf{f}_{kk'}(t) - \alpha_t \mathbf{\Delta}_{kk'}]_{\mathcal{F}^e_{kk'} \cap \Re^+}. \quad (14)$$

Note that in (13) the projection $[\cdot]_{\mathcal{F}^i_j \cup \Re^+}$ is due to constraints (8) and (10). Similarly, in (14) the projection $[\cdot]_{\mathcal{F}^e_{kk'} \cup \Re^+}$ is due to constraints (9) and (10).

The above set of iterations can be shown to converge to the optimum of **approx-DRV-TE**(P) under some reasonable conditions on the step-sizes α_t and cost functions $V_j(\cdot)$ and $v(\cdot)$ and penalty function P. The reader can refer to [39] (Chapter 2) for convergence conditions of the gradient projection method. In particular, if the step-sizes $\alpha_t = \alpha \ \forall t$,

where the common step-size α is sufficiently small, and the objective function in (7) satisfies Lipschitz conditions [39], convergence to the optimum of **approx-DRV-TE**(P) can be guaranteed. In other cases, under certain reasonable conditions, approximate "convergence" to a neighborhood of the optimum whose size depends on the step-size α can be shown.

Relation (13) represents how the traffic rate variables are updated along the i-edge-links of AS j at iteration t. Note that $\mathbf{\Delta}_j^i$ as defined by (11) can be computed locally by AS j, as it requires $V_j(\cdot)$ and rate vector $\mathbf{g}_j^i$ that are known to AS j, and the penalties associated with congestion at the edge-routers of AS j, which we can assume each individual edge-router keeps track of. This implies that the step (13) can be computed by the different ASes individually based on information that is local to the AS. The updating of the traffic rates on an e-edge link (k, k'), as in (14), can also be done locally at edge-router k, possibly requiring information exchange only with edge-router k'. To see this, note that the computation of $\mathbf{\Delta}_{kk'}$ defined by (12) in this case, requires knowledge of the penalty terms at edge-routers k and k', and the aggregate traffic rate on e-edge-link (k, k'), $g_{kk'}$, which are all local variables.

Note that in the formulation of **DRV-TE** and **approx-DRV-TE**(P), we have placed no restriction on the inter-domain paths that are taken from source to destination. This necessitated keeping track of traffic rate variables for all destinations at each edge-link. In general, however, inter-domain routing protocol and policy may only allow certain paths to be chosen and not others. Paths that have a large number of AS-hops or have loops may be avoided; next-hop forwarding restrictions may also be placed due to customer-provider relationship between ISPs and other ISP peering and transit rules. Such restrictions can be easily incorporated in our framework by omitting certain edge-links (that cannot be on any forwarding path to the destination) from consideration while maintaining and updating the per-destination traffic rate variables. This would in general reduce the number of variables and likely to speed up the convergence process, but may result in increased congestion (forwarding) cost due to reduced choice of paths.

V. Simulation Results

We perform numerical simulation study to evaluate the convergence properties of the inter-domain TE solution as described in Section IV. The intra-domain cost functions Vi_j corresponds to the sum of delays across all physical links of AS j; the extra-domain cost function v represents the delay on single physical link for the corresponding extra-domain edge-link. Delay on a physical link is approximated by the M/M/1 delay formula, as in the example described towards the end of Section III. We choose a quadratic penalty function, with $P(h_{k,d}(\hat{\mathbf{f}}_k^d)) = b \times (h_{k,d}(\hat{\mathbf{f}}_k^d) + a)^2$, if $h_{k,d}(\hat{\mathbf{f}}_k^d) + a > 0$; and 0 otherwise. In our study, we choose $a = 1$ and $b = 1$. We use a constant step-size $\alpha = 0.25$. Due to the complexity involved in computing the iterative TE procedure in a single processor system and that of solving **DRV-TE** optimally (needed for benchmarking), we focus on a moderate number of ASes (upto 50) and a small number of destinations (upto 8). The intra-domain routing policy simply chooses to route along the minimum number of physical hops.

We generate realistic Internet topologies according to the Albert-Barabasi model (for both intra-AS topologies and inter-AS connectivity) using the popular BRITE topology generator [38]. The number of routers within each AS is kept the same for all ASes and equals 10. Each edge-router in the network has a source network attached to it for each destination. Initially, we consider two destination networks, but later study how the convergence time scales with increasing number of destinations. The traffic demand between each source-destination pair is assumed to be 10 units.

We run our inter-domain TE algorithm for 15000 iterations and obtain an estimate of the converged total cost as the average of the total cost during the last 1000 iterations. We also solve the **DRV-TE** optimization problem on each network subject to physical link capacity and flow conservation constraints and obtain the optimum cost. Table I shows the average and maximum ratio of the converged cost to the optimum cost for different number of ASes in the network (from 10 to 30). For a given network size, we generate 10 network instances for calculating the average and maximum ratios. From the table, we find that the maximum and average ratios are close to 1. This suggests that the converged cost (of the proposed inter-domain TE solution) is very close to the **DRV-TE** optimum.

Number of ASes	10	20	30
Average	1.0039	1.0057	1.0059
Maximum	1.0046	1.01	1.0093

TABLE I
RATIO OF CONVERGED TOTAL COST TO OPTIMUM TOTAL COST

In Figure 2, we illustrate the total cost convergence as attained by our inter-domain TE solution on a sample network topology with 20 ASes and 10 routers within each AS. As seen, the total cost starts from a low initial value and gradually settles down as the number of iterations increases. The horizontal line in Figure 2 represents the **DRV-TE** optimum.

Figure 3 shows the average convergence time when the number of ASes in the overall network is increased. As before, each AS consists of 10 routers. The total cost is said to have converged at time t_c, if the total cost is within 0.5% of the estimated converged cost (i.e. within 0.25% above and 0.25% below) during the window $[t_c, 15000]$. For the results shown in Figure 3, the average convergence time is obtained by averaging the results over 20 networks of given size. We also plot the error bars that show the range of values (max-to-min) for the convergence time. The convergence time seems to increase sub-linearly with the number of ASes in the network.

Finally, we study the average convergence time when the number of destination networks in the overall network is increased from 2 to 8. Note that increasing the number of destination networks by 1 doubles the number of traffic rate variables in the optimization problem. However, from the results shown in Figure 4, the average convergence time (along with the max-to-min range) seems to increase sub-linearly with the number of destination networks.

Acknowledgment: This work was supported by NSF through the NeTS-FIND program (awards CNS-0721609, CNS-0721600, CNS-0831830, and CNS-0831957).

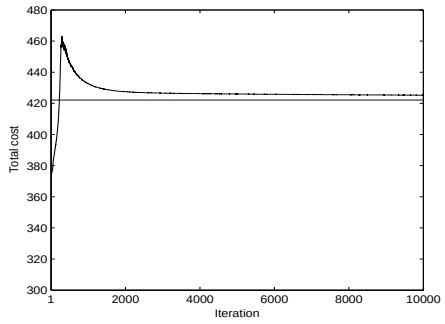

Fig. 2. Convergence of the inter-domain TE solution on a sample network topology.

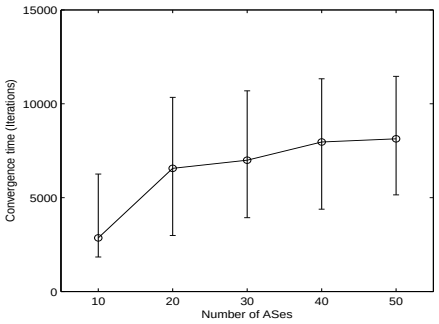

Fig. 3. Average convergence time and max-to-min range for different number of ASes (20 samples each).

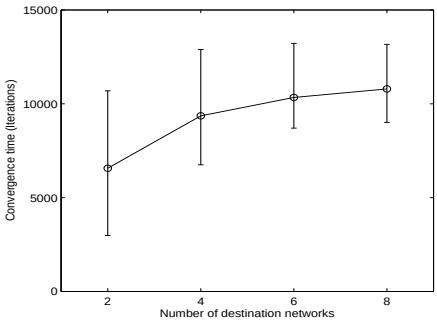

Fig. 4. Average convergence time and max-to-min range for different number of destination networks (20 samples each).

REFERENCES

[1] A. Viswanathan, E. Rosen, and R. Callon, "Multiprotocol label switching architecture (MPLS)," RFC 3031, www.ietf.org, January 2001.

[2] S. Blake, D. Black, M. Carlson, E.Davies, and Z.Wang, and W.Weiss, " An architecture for differentiated services," *IETF Internet RFC 2475*, Dec 1998.

[3] R. Braden and et al, "Resource Reservation Protocol (RSVP) - V1 functional Spec," *IETF Internet RFC 2205*, Sep 1997.

[4] B. Fortz and M. Thorup, "Internet traffic engineering by optimizing ospf weights," in *Proc. of IEEE INFOCOM*, 2000.

[5] K. Kar, M. Kodialam, T. V. Lakshman, "Minimum Interference Routing of Bandwidth Guaranteed Tunnels with MPLS Traffic Engineering Applications," *IEEE Journal on Selected Areas in Communications*, Vol. 18, No. 12, Dec 2000.

[6] A. Elwalid, C. Jin, S. Low, and I. Widjaja, "MATE: MPLS Adaptive Traffic Engineering," *Computer Networks*, Dec 2002, pp. 695-709.

[7] H. Wang *et al.*, "Cope: traffic engineering in dynamic networks," in *Proc. of SIGCOMM*, 2006.

[8] Tao Ye, H.T. Kaur, S.Kalyanaraman, and M.Yuksel, "Large-scale network parameter configuration using an on-line simulation framework," in *IEEE/ACM Trans. on Networking*, 16(4):777 –790, 2008.

[9] J.He, M.Bresler, M.Chiang, and J.Rexford, "Towards robust multi-layer traffic engineering: Optimization of congestion control and routing," in *IEEE Journal on Selected Areas in Communications*, 25(5):868–880, June 2007.

[10] J. He *et al.*, "Rethinking internet traffic management: from multiple decompositions to a practical protocol," in *Proc. of CoNEXT '07*, pages 17:1–17:12, New York, NY, USA, 2007. ACM.

[11] Y. Rekhter and T. Li, "A Border Gateway Protocol 4 (BGP-4)," RFC 4271, January 2006.

[12] N. Feamster, J. Borkenhagen, and J. Rexford, "Guidelines for interdomain traffic engineering," in *SIGCOMM Comp. Comm. Rev.*, 33:19–30, 2003.

[13] S. Uhlig and B. Quoitin, "Tweak-it: Bgp-based interdomain traffic engineering for transit ass," in *Next Generation Internet Networks*, pages 75 – 82, April 2005.

[14] J. Sommers, P. Barford, N. Duffield and A. Ron, "Accurate and Efficient SLA Compliance Monitoring," in *Proc. of SIGCOMM*, 2007.

[15] M. Yuksel, A. Gupta, K. Kar, and S. Kalyanaraman, "Contract-Switching for Managing Inter-Domain Dynamics," Book chapter, in *Next-Generation Internet Architectures and Protocols*, Eds. B. Ramamurthy *et al.* Cambridge University Press, pp.136-153, March 2011.

[16] C. Labovitz, S. Iekel-Johnson, D. McPherson, J.Oberheide and F. Jahanian, F., "Internet Inter-Domain Traffic," *Proc. of SIGCOMM*, New Delhi, India, August 2010.

[17] S. Balon and G. Leduc, "Combined intra- and inter-domain traffic engineering using hot-potato aware link weights optimization," in *Proc. of ACM SIGMETRICS*, pages 441–442, 2008.

[18] D. Acemoglu, R. Johari, and A.E. Ozdaglar, "Partially optimal routing," in *IEEE Journal on Selected Areas in Communications* 25 (6): 1148-1160 (2007).

[19] T. G. Griffin, F. B. Shepherd, and G. Wilfong, "The stable paths problem and interdomain routing," *IEEE/ACM Trans. on Networking*, 10(22):232-243, Apr. 2002.

[20] J. L. Sobrinho, "Network routing with path vector protocols: Theory and applications," in *Proc. of ACM SIGCOMM*, Karlsruhe, Germany, Aug. 2003.

[21] H. Wang *et al.*, "Stable egress route selection for interdomain traffic engineering: model and analysis," *Proc. of ICNP*, 2005.

[22] Y.R. Yang *et al.*, "On route selection for interdomain traffic engineering," *IEEE Network*, Vol. 19, No. 6, pp. 20-27 (2005).

[23] B. Quoitin *et al.*, "Interdomain Traffic Engineering with BGP," *IEEE Communications Magazine*, Vol. 41 (2003).

[24] B. Quoitin, S. Tandel, S. Uhlig and O. Bonaventure, "Interdomain traffic engineering with redistribution communities," *Computer Communications*, pp. 355-363 (2004).

[25] R. Gao, C. Dovrolis, and E. W. Zegura, "Interdomain Ingress Traffic Engineering Through Optimized AS-Path Prepending," *NETWORKING 2005*: 647-658.

[26] N. Feamster, J. Borkenhagen, and J. Rexford, "Guidelines for interdomain traffic engineering," in *Proc. of SIGCOMM Computer Communication Review*, vol. 33, no. 5, pp. 19-30 (2003).

[27] W. Xu and J. Rexford, "MIRO: multi-path interdomain routing," in *Proc. of SIGCOMM Computer Communications Review*, Vol. 36, No. 4, 2006.

[28] R.Mahajan, D.Wetherall, and T.Anderson, "Negotiation-based routing between neighboring ISPs," in *Proc. of USENIX NSDI*, 2005.

[29] G. Srimali, A. Akella, A. Mutapcic, "Cooperative Interdomain Traffic Engineering Using Nash Bargaining and Decomposition," in *IEEE/ACM Trans. on Networking*, Vol. 18, No. 2, pp. 341-352, 2010.

[30] N.Bitar, JP.Vasseur, R.Zhang, and JL.Le Roux, "A backward-recursive pce-based computation (brpc) procedure to compute shortest constrained inter-domain traffic engineering label switched paths," RFC 5441, www.ietf.org, April 2009.

[31] J.De Clercq and et al, "BGP-MPLS IP virtual private network (VPN) extension for ipv6 vpn," *IETF Internet RFC 4659*, September 2006.

[32] G. Srimali, A. Akella, and A. Mutapcic, "Cooperative Interdomain Traffic Engineering Using Nash Bargaining and Decomposition," in *IEEE/ACM Trans. on Networking*, Vol. 18, No. 2, pp. 341-352 (2010).

[33] M. Roughan and Y. Zhang, "GATEway: Symbiotic inter-domain traffic engineering," *Telecommunication Systems*, 47(1-2): 3-17, 2011.

[34] T. Roughgarden, "Routing Games," Chapter 18 in: N. Nisan *et al.* (eds.), *Algorithmic Game Theory,* Cambridge University Press.

[35] A. Ozdaglar and R. Srikant, "Incentives and Pricing in Communication Networks," Chapter 22 in: N. Nisan *et al.* (eds.), *Algorithmic Game Theory,* Cambridge University Press.

[36] C. Papadimitriou and G. Valiant, "A New Look at Selfish Routing," *Innovations in Computer Science*, 2010.

[37] S. C.M. Lee and J. C.S. Lui, "On the Interaction and Competition among Internet," *IEEE Journal on Selected Areas in Communications*, Vol. 26, No. 7, Sep 2008, pp. 1277-1283.

[38] A. Medina, A. Lakhina, I. Matta, and J. Byers, "BRITE: An Approach to Universal Topology Generation," *Proc. of MASCOTS*, 2001.

[39] D. P. Bertsekas, "Nonlinear Programming", Athena Scientfic, 1995.

[40] N.Z. Shor, "Minimization Methods for Non-differentiable Functions," Springer-Verlag, New York, 1985.

[41] E. Anshelevich, A. Hate, and K. Kar, "Strategic Pricing in Next-hop Routing with Elastic Demands," in *Proc. of SAGT*, Salerno - Amalfi Coast, Italy, Oct 2011.

[42] A. Hate, E. Anshelevich, and K. Kar, "Stable and Efficient Pricing for Inter-domain Traffic Forwarding," Poster paper, To appear in *Proc. Sigmetrics 2012*, London, UK, June 2012.

[43] P.K. Muthuswamy, K. Kar, A. Gupta, H.T.Karaoglu, M.Yuksel, "ISPs as Nodes or Sets of Links?," To appear in *Proc. of ICC 2012*, Ottawa, Canada, June 2012.