Telecommun Syst
DOI 10.1007/s11235-010-9335-x

Multi-exit discriminator game for BGP routing coordination

Stefano Secci · Jean-Louis Rougier · Achille Pattavina · Fioravante Patrone · Guido Maier

Abstract Inter-Autonomous System (AS) links represent nowadays the real bottleneck of the Internet. Internet carriers may coordinate to efficiently balance the load, but the current practice is often based on an uncoordinated selfish routing. Firstly, we assess this issue by characterizing BGP route deviations across top-tier interconnections we could detect using recent Internet routing history data. Then, in order to improve the current practice, we present a novel game-theoretical framework to efficiently coordinate the routing on inter-AS links while modeling the non-cooperative carrier behavior. It relies on a coordinated use of the Multi-Exit Discriminator (MED) attribute of BGP, hence it is nicknamed ClubMED (Coordinated MED). We define the routing policy that shall be implemented upon Nash equilibria and Pareto-efficient profiles. We emulated the interconnection between the Internet2 and the Geant2 networks, comparing our proposition to the current BGP practice. The results show that the route stability can significantly be reinforced, the global routing cost can be significantly reduced, and the inter-AS link congestion can be avoided.

Keywords MED · BGP · Game theory · Deviation · Congestion

1 Introduction

The Internet backbone is composed of a few Autonomous Systems (ASs). To simplify, one may say it is composed of a few top-tier inter-continental carrier providers that provide transit connectivity to those regional providers the most part of customers and stub ASs are connected to. The Border Gateway Protocol (BGP) v.4 is the current inter-AS IP routing protocol. It includes criteria that allow implementing different interconnection settlements (namely based on transit, peering or sibling agreements). As a matter of fact, the inefficient way in which these criteria are currently used across top-tier interconnections overstresses inter-AS routing and overloads the inter-AS links. This is mainly due to the unpredictability of the aggregate IP flows and to the fact that some top-tier interconnection settlements (typically relying on peering and sibling agreements) releases an AS from following the neighbor's routing preferences [23]. This yields to selfish routing while, instead, coordination schemes may improve the bilateral routing efficiency.

In previous work about inter-carrier connection-oriented services [5], it sorted out that a form of *cooperation* among

A preliminary version of this paper has been published in the 2009 Next Generation Internet Networks (NGI 2009) conference proceedings (best paper award) [20].

S. Secci (✉)
LIP6, Université Pierre et Marie Curie – Paris VI, 4 place Jussieu, 75005 Paris, France
e-mail: stefano.secci@telecom-paristech.fr

J.-L. Rougier
Institut Télécom, Télécom ParisTech, LTCI CNRS, 23 av. d'Italie, 75013 Paris, France
e-mail: rougier@telecom-paristech.fr

A. Pattavina · G. Maier
Dip. Elettronica e Informazione, Politecnico di Milano, via Ponzio 34/5, 20133 Milano, Italy

A. Pattavina
e-mail: pattavina@elet.polimi.it

G. Maier
e-mail: maier@elet.polimi.it

F. Patrone
DIPTEM, Università di Genova, P. le Kennedy—Pad D, 16129 Genova, Italy
e-mail: patrone@diptem.unige.it

Published online: 07 July 2010

carriers is needed to overtake privacy, billing and monitoring issues. In this paper we argue how, instead, for connectionless IP services—for which such issues are not present—cooperation is not necessary in that *coordination* is enough. In particular, we concentrate on the coordination inter-AS routing issue to reduce congestions, routing cost and route deviations.

In Sect. 2 we link recent ideas in the area that motivated this work. Section 3 addresses the lack of coordination of current inter-AS routing by reporting BGP route deviations detected across top-tier interconnections. Willing to rely on the MED BGP attribute as the natural medium to convey coordination data, in Sect. 4 we define the ClubMED (Coordinated MED) framework, in which efficient strategy profiles can be detected in a non-cooperative game modeling. We define an effective routing policy relying on the concepts of Nash equilibria and Pareto-efficiency. We explain how, within the ClubMED framework, a form of load balancing can be implemented on selected strategy profiles for a subset of the destination networks whose traffic routing can be coordinated. We consistently integrate IGP weight optimization operations and inter-AS link congestion controls, which increases the number of possible Nash equilibria and, thus, the importance of a coordination scheme to select the most efficient ones. Section 5 reports the results from realistic simulations and comments the significant gains the ClubMED framework can offer with respect to the current practice. Finally, Sect. 6 summarizes the paper.

2 Rationales

2.1 BGP, route deviation and congestion

It is worth briefly reminding how the inter-AS route selection is performed via BGP. When multiple AS paths to a destination network prefix are available, a cascade of criteria is employed to compare them. The first is the "local preference" through which local policies, mainly guided by economic issues, can be applied: e.g., a peering link (i.e., free transit) is preferred to a transit link (transit fees). Marking routes with local preferences, an AS can thus implement peering and transit settlements. The subsequent BGP criteria incorporate purely operational network issues: smaller AS hop count, smaller MED, closer egress point (also called "hot-potato"), more recent route. If not enough, the AS path learned by the router with the smaller IP is selected (rule also called "tie-breaking"). Considering these criteria, BGP selects the best AS path which is the single one advertised to the neighbors (if not filtered by local policies).

Operationally speaking, carriers desire that AS paths pointing to them have been selected using the highest possible priority rule to obtain good performance, e.g., on the end-to-end delay for the connections along that AS path. The smaller AS hop count is a rude yet simple rule that avoids routing inefficiencies [7]. For a given AS path, if several border routes are available, the MED can be used by the downstream AS to suggest an egress router. However, it is rarely used: it is only for very specific cases or when requested by a client (see Sect. 2.3). In the absence of MED settings, IGP weights are compared and the closer egress point is selected (hot-potato).

The interaction between hot-potato routing and intra-AS routing represents a major issue. To react to non-transient network events, a carrier may re-optimize the IGP weights, inducing changes in the egress router selection so that congestions might appear where not expected. Reference [22] reformulates the egress routing problem and proposes to replace hot-potato with a more expressive and efficient rule. Reference [1] presents a comprehensive yet hard IGP Weight Optimization (IGP-WO) method aware of BGP hot-potato routing deviations, opportunely bounding them (they report that in real cases 70% of traffic could be affected). Reference [2] presents a similar proposition relying on graph expansion tricks. However, while effective, a problem seems to persist with the latter propositions: each time the BGP routes change, the BGP-aware IGP-WO is to be triggered. The scalability would be thus a practical issue: the occurrence of IGP-WOs, normally triggered only for intra-AS issues, would drastically increase given the frequency of BGP deviations. The reduction of the coupling between inter-AS and intra-AS routing is thus really an open issue [23].

2.2 From selfish to coordinated inter-carrier traffic engineering

With a more far-sighted standpoint, in [13] it is proven that, if part of the profits due to inter-carrier services were shared, the Internet carriers would behave less selfishly, yielding better global routing with lower routing cost than under the current practice. Using the Shapley value concept from *cooperative* games, they argue that profits and costs may be fairly imputed considering the importance of each AS in the interconnected "coalition" composed of ASs routing "common" inter-AS flows [14]. In this way, they prove that ASs have incentive to better route yielding to a common inter-domain routing cost lower compared with BGP routing.

More pragmatically, the authors in [21] show how much the hot-potato routing is far from being the ideal desirable solution. They compare it to a cooperative routing resulting from the maximization of a common utility for the two network configurations, i.e., the *bargaining problem* of the common utility, the (Nash) product of ASs' utilities, each one estimated somehow from the current intra-AS routing status (somehow withstanding possibly also a congestion

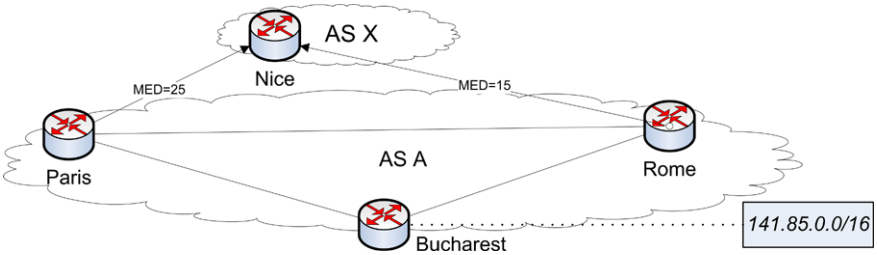


Fig. 1 Multi-exit discriminator signaling

risk, ignored in [13]). Then, the maximization of the common utility is solved by decomposition. However, to cope with multiple AS cooperative scenarios, their method should be at least redesigned given that the Nash product maximization solution—which can formally be extended[1] to the case of n/player games—does not take into account the role of subcoalitions [17]. They also show how their method outperforms a generic best-reply "Nash equilibrium" method from non-cooperative games, however not detailed, thus preventing the possibility of making a comparison.

Modeling the inter-AS BGP routing as a cooperative bargaining problem of a common utility may not be necessary. Besides appearing not expressive enough (hardly acceptable for operation engineers), the utility maximization result seems relying on an excessive abstraction of the real network status at the risk of losing the real routing optimality. In other words, the way in which the utility is computed may not be enough representative of the real operational status of the network. We propose, instead, a non-cooperative approach since it allows more straightforward solutions to implement w.r.t. the current practice (both technically and economically). In the next sections we explain how using MED signaling it is possible to implement strategies that are, in game theory parlance, *non-cooperative but coordinated*, i.e., that solve the inter-peers routing problem without binding agreements between peers.

2.3 The multi-exit discriminator (MED)

The MED is an integer metric that an AS can attach to route advertisements toward a potential upstream AS, to suggest an entry point when many exist. In this way the upstream AS has the choice of the entry point toward the advertised network. In Fig. 1, the upstream AS X selects a route for the network 141.85.0.0/16. It has two route alternatives through AS A: by the Paris router or the Rome router. MEDs are attached to the routes announced by AS A's Paris and Rome routers. If accepting MEDs, the AS X router will then select the route with smaller MED, hence the route passing by the Rome router. The default MED value is equal to the IGP cost of the corresponding intra-AS path.

[1]Extensions to cope with the role of subcoalitions have been devised by Harsanyi and Shapley [19], but they are not easily manageable.

Nowadays, the MED is often disabled. Even if a downstream AS uses it to suggest preferred entry points, the neighbor can discard its announcements. The MED can be used on transit or peering links. On transit links, subject to provider/customer agreements, the provider normally follows "MED-icated" routes suggesting the preferred entry points because the customers pay for. This is often not the case for peering agreement, and this is the main reason why the MED is usually not employed on peering links [15].

3 Detection of BGP route deviations

In the following, we aim to assess the importance of route deviations in the Internet backbone as an index of the dangerous lack of coordination in the Internet backbone.

We focus on those deviations that could be due to IGP/BGP routing interaction, and that happen across top-tier AS interconnections. This choice is guided by two main reasons. First, since top-ranked ASs dispose of higher path diversity, across top-tier interconnections the risk of deviations due to IGP path cost minimization (hot potato or least MED BGP rules) is higher. Second, top-tier borders are likely to rely on (or to have been or to become) peering settlements (i.e., two ASs agree in free-transit between their customers' networks only); BGP routes across peering links risk to be instable because generally one AS is not binded to follow the preferences of the peer, which produces a lack of routing coordination (e.g., no MED signaling, or rare capacity upgrade), hence the risk of sudden congestions and IGP reconfigurations is higher.

3.1 Analysis methodology

Many techniques have been defined in the literature to active measure the Internet topology. In [4] there is a thorough state of the art on measurement techniques (up to 2007). The detection of route deviations needs a history of routing maps. An Internet routing map is a collection of paths from a set of monitors to a set of destination hosts. A history of routing maps can be stored sampling sequentially source-destination routes during an observation period.

In [10], the authors present a measurement framework that allows building a history of Internet routing maps.

The framework stores traceroute-like samples, toward some thousands of destination hosts, collected from a few dozens of PlanetLab monitors [18]. Recorded 2008 data is now publicly available to the research community in [12]. For each PlanetLab monitor and sampling instant (round), a *tracetree* is stored as a file; a tracetree is a compact route tree from a source to many destinations that avoids some anomalies and useless ICMP signaling (replication of common paths among several destinations). We employed this data to build a history of Internet routing maps. As already mentioned, we focus on the detection of deviations across those top-tier AS interconnection (ideally peering interconnections) that are likely to suffer the most from these events. In order to select top-tier interconnections, we monitored all possible frontiers between the top-50 carrier providers in the Caida ranking [3]; this can be done grouping the ASs belonging to the same provider. The total number of monitored frontiers is around 7300 (excluding sibling frontiers).

We isolated the Radar data obtained from some monitors of different Planetlab sites. Each monitor has a random destination set of a few thousands of online IP hosts. We then extracted for each destination host the router-level and AS-level (or AS path) routes from the corresponding sources at each round. Then, we kept those crossing top-tier frontiers. If a route crosses more than one of such frontiers, we associate the route with the more ranked frontier.

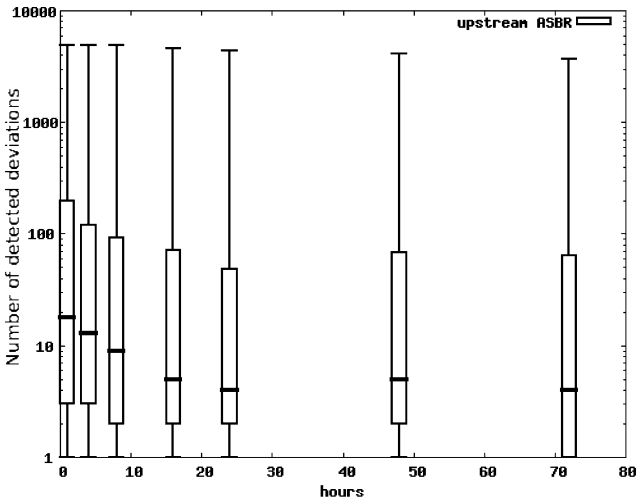
3.2 Intra-AS path deviations

In the following, we focus in those situations in which a BGP route deviation did cause an intra-AS path change, within a stable AS path. We only consider deviations occurring while the AS path remains stable during a chosen observation interval; in other words, for each stable AS path we study if there are intra-AS IP/router-level route deviations. We considered many AS path lifetimes: 1, 4, 8, 16, 24, 48 and 72 hours.

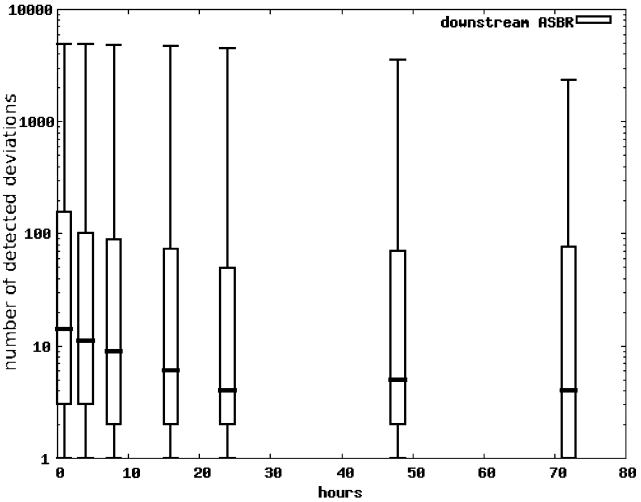
To detect intra-AS path deviations, for each IP-level route sample within a stable AS path, we isolate those crossing a top-50 frontier and deviating within one of the two ASs (again, if there are more than one top-50 frontier, we keep the more ranked one; if a deviating AS does not belong to a top-50 frontier, the deviation is not counted). Two kind of intra-AS deviations can be experienced:

- *internal deviation*: change of an intra-AS path with unchanged AS Border Router (ASBR). It is worth noting that such deviations can be caused by both BGP route deviations (different AS-level route, but same egress ASBR) and IGP route deviations or load balancing (same BGP route and egress ASBR).
- *ASBR deviation*: change of an intra-AS path with at least one different ASBR. When the ingress ASBR changes, the deviation is due to a change of the routing policy of the upstreaming AS; when the egress ASBR does, it is due to local policies. Besides to the hot potato and least MED rules, such deviations may also be due to the usage of BGP Multipath.

  If the deviating ASBR is the ingress one, we count the deviation as 'upstream ASBR', otherwise as 'downstream ASBR'.



(a) upstream ASBR



(b) downstream ASBR

Fig. 2 Boxplot statistics of the number of detected ASBR deviations

Because of the excessive risk of bias due to intra-AS IGP load balancing, we do not report internal deviation results (not enough pertinent). We focus instead on ASBR deviations that are less biased by load balancing.

Figure 2 reports the boxplot statistics (minimum, first quartile, median, third quartile, maximum) for the number of ASBR deviations and for both 'upstream' and 'downstream' types, as function of the observation period, and in logarithmic scale. We can observe that:

Fig. 3 AS path deviations from a PlanetLab monitor

– 'upstream' deviations are more numerous than 'downstream' ones;
– the number of deviations is less than 20 for 50% of the deviating routes (i.e., the median is always minor than 25);
– for AS path lifetimes of at least 24 h, 75% of the deviating routes faces less than 100 deviations (i.e., the third quartile is always minor than 100);
– the number of deviations decreases for AS paths stable for longer lifetimes. This highlights that a lot of AS paths with short lifetimes are likely to deviate in the long-run.

Behind the fact that upstream deviations are slightly more numerous than downstream ones, one reason worth discussing may reside in the usage of the MED attribute of BGP across the top-tier interconnection. Across a given monitored frontier, if the MED signaling is enabled, MED-icated routes would be sent by the downstream AS to the upstreaming one to suggest an entry point for the upstream flow. It is thus possible that a higher instability in the upstreaming AS is a symptom of a frequent MED reconfiguration by the downstreaming AS. However, it is worth mentioning that a route change in the upstream AS would often imply an ASBR route change in the downstream AS too.

Furthermore, some deviating destination may once suffer from an internal or ASBR deviation, and once an AS path deviation. This sort of deviation is likely to be related to the hot potato rule of BGP that compare all the routes (in fact, their IGP path cost) with no respect to their (downstreaming) AS neighbors. Such a deviation is not likely to be related to the least MED rule (MED = IGP transit path cost of the neighbor) since this last considers, instead, only the routes for the same (downstreaming) AS neighbor.

3.3 AS path deviations

In order to better characterize these phenomena, we then monitored the deviations inducing an AS path change. Only the deviations involving equal-length AS paths are considered. In this way, we can better target those deviations likely due to IGP path cost variations rather than those due to the least AS path length rule of BGP (being the least AS path length priority to the least MED and hot potato rules). Focusing on such a 1-hop AS-level deviation we can also target those situations in which an upstreaming AS discriminates between two downstreaming AS both containing the destination host in their destination cone.

With a rapid glance, for a dozen of different PlanetLab monitor traces considered for our analysis, from 3% to 10% of the AS paths deviate, and from 1% to 3% oscillates. The whole observation period changes with the monitor and ranges from 1000 to roughly 3000 rounds; rounds are delayed of roughly 10 minutes, a tracetree can take up to 5 minutes to be stored, thus the observation period is very approximately of 10–30 days.

Figure 3 gives a graphical representation of the AS path deviations detected from a sample PlanetLab monitor. The vertical axis is the time in the unit of round. At each round, a route toward each destination is recorded. In the horizontal axis we have different destination hosts identifiers (sequentially assigned). For each host, we have a vertical line composed of a sequence of points; a change of colour corresponds to an AS path deviation. Each colored point represents the crossing of one of the top-50 frontiers, while a white point represents a crossing of unmonitored frontiers. At a glance, we can observe that the deviations can vary from quite random to more regular ones, and that they affect a small yet non negligible part of the destination hosts.

Looking deeper into this data, Fig. 4 represents the duration distribution of the deviations, with the average length for each decile. We also report the duration distribution across different duration intervals. We focus now only on AS path deviations within the top-50 frontier area only, i.e.,

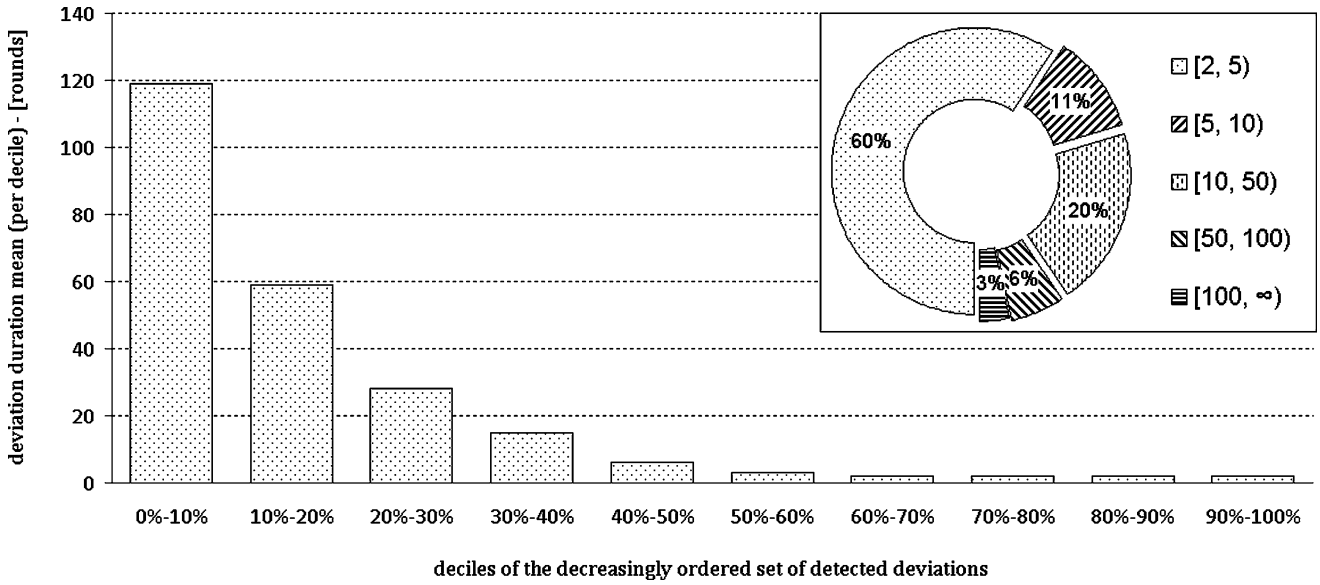


Fig. 4 Deviation duration decile distribution

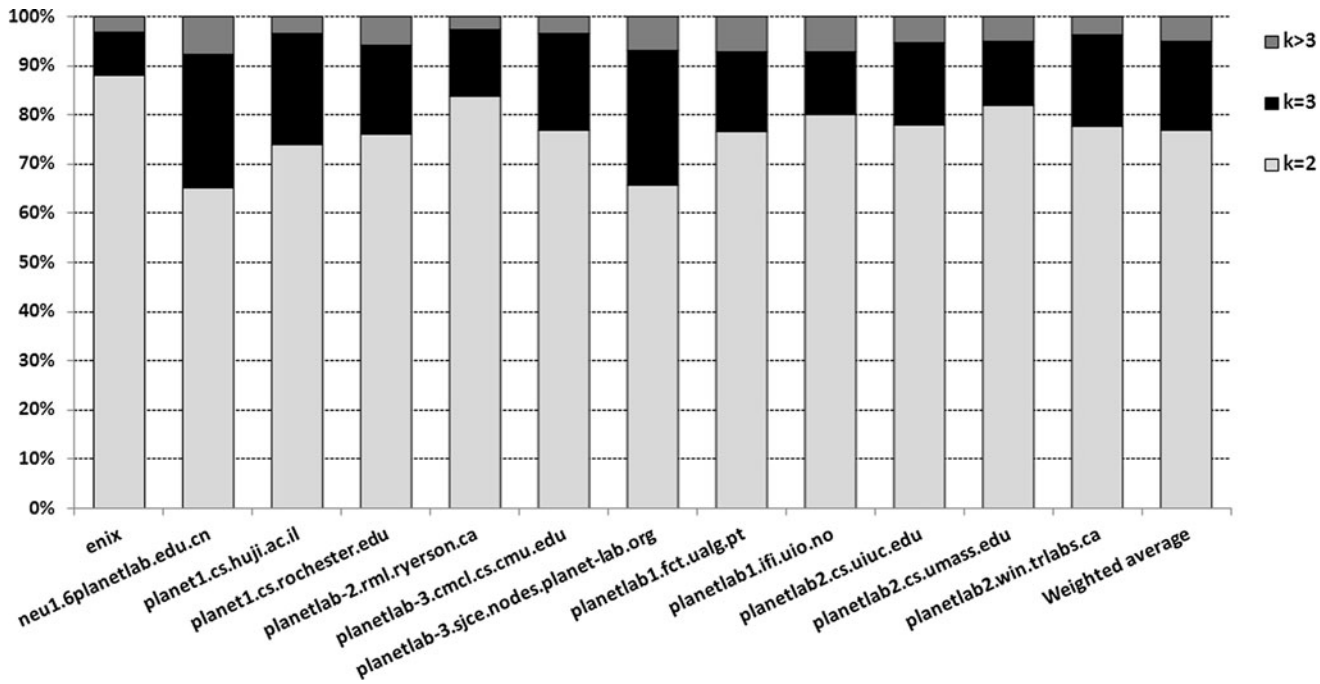


Fig. 5 Distribution of the number of deviations per Planetlab source

only when an AS path crossing a top-50 frontier deviates toward another top-50 frontier. We can assess that:

- for the 10% longer deviations, the mean duration is 120 rounds (roughly 24 h); they represent, however, less than 3% of all the deviations;
- 59% of the deviations lasts less than 5 rounds (roughly 1 h);
- 1/3 of the deviations lasts between 1 h and 10 h, or 1/3 more than 2 h.

Therefore, the large majority of the AS path deviations manifests with a daily occurrence, which is probably linked to IGP path cost variations. Those deviations with longer duration are probably due to topology changes and are more likely to happen only once during the observation period, hence their long duration. Isolating them for some monitors in Fig. 5, we indicate the percentage and the number of destinations whose route deviates k times, with $k = 2$, $k = 3$ and $k > 3$ (the last column is the weighted percentage av-

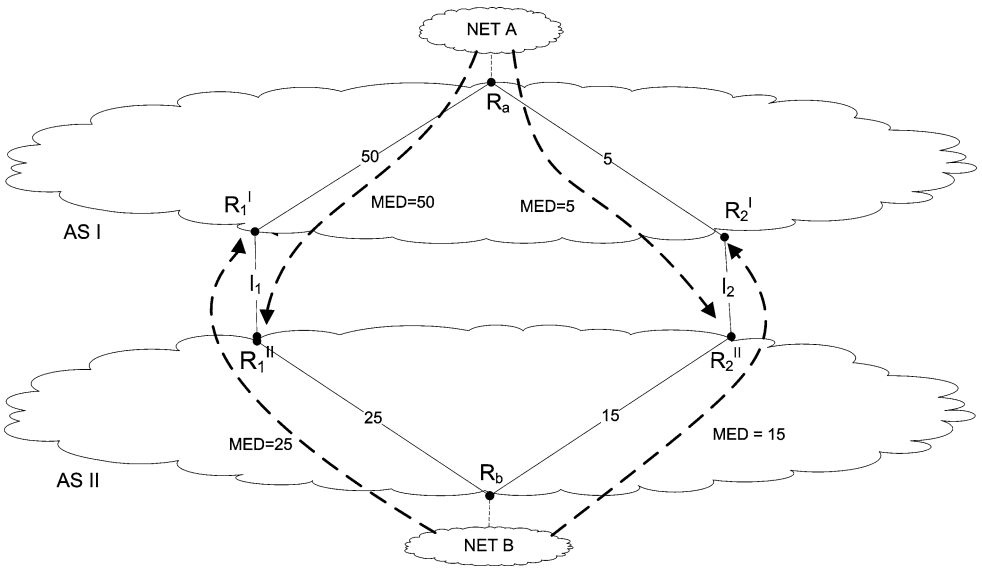


Fig. 6 Peering MED interaction example

erage). We can observe that there is a non negligible part of the destinations whose AS path deviates quite frequently, roughly 23% more than two times, and roughly 5% more than 3 times.

Table 1 A dummy game

I \ II	l_1	l_2
l_1	(**50**, **25**)	(**5**, **25**)
l_2	(**50**, **15**)	(**5**, **15**)

4 The ClubMED framework

The results showed in the previous section are the symptom of a lack of coordination in the Internet backbone. Such an occurrence of route deviations depends on a frequent reconfiguration of IGP routing costs at both sides in the routing interaction among neighbor AS carriers. Behind these events, there is probably a bad or absent congestion control for inter-AS links.

With the aim to improve the routing coordination across top-tier interconnections, in the following we model the MED signaling between neighbor ASs as a non-cooperative game wherein two ASs can implicitly coordinate their routing strategies. We nickname it the ClubMED (Coordinated MED) framework. For the sake of clarity, we first start with a simple but unrealistic model with 2 inter-AS links and bidirectional routing costs. Then, we generalize it to the complete realistic generic form, integrating IGP-WO operations and inter-AS link congestion controls.

4.1 MED-based coordination

In Fig. 6, AS I and AS II are two neighbor ASs. NET A and NET B are two destination networks whose flows are supposed to be equivalent (e.g., w.r.t. the bandwidth), so that their path cost can be fairly compared and their routing coordinated. Each AS would desire to minimize its routing cost for the incoming flow. The routing costs are indicated in Fig. 6. AS I and AS II announce NET A and NET B with the MED attribute set to the routing cost by the corresponding egress router. The routing interaction can be described with the strategic form in Table 1. The cost of each player is the MED of the route it announced, then selected by the neighbor. Each AS has the choice if routing the outgoing flow on link 1 (l_1) or on link 2 (l_2).

In non-cooperative games, a Nash equilibrium is to be selected by rational players because it yields stability to the strategy profile, the players not being motivated in deviating from it [17]. In Table 1 every profile is a Nash Equilibrium. We have a dummy game: whatever the other player's strategy is, there is no gain in changing its strategy. This somehow shows that a simple MED usage is dummy for such a case. We should enrich the dummy game considering the egress cost of the flow in the opposite direction, thus summing the routing costs of both the flows in opposite directions for each AS. However, in this way we would assume that both the NET A $\leftrightarrow$ NET B flows pass through the interconnection AS I-AS II, which would not be realistic (BGP policies can induce asymmetric routing). Moreover, traffic flows to care of are typically between content and "eyeball" providers (with a lot of clients) [14], which would not make the A $\leftrightarrow$ B flows equivalent. Instead of single prefix network, we should consider *destination cones* (i.e., groups of network prefixes). The cone prefixes shall belong to direct customers or stub ASs, whose entry point in a neighbor network is likely to be unique (even if they are multi-homed, they should have chosen backbone-disjoint providers, referring to disjoint core carriers).

Table 2 A ClubMED game

I \ II	l_1	l_2
l_1	(100, 50)	(55, 40)
l_2	(55, 40)	(**10**, **30**)

Table 3 2-link ClubMED game, sum of two games with potential

I \ II	l_1	l_2
l_1	(c_1^I, c_1^{II})	(c_1^I, c_2^{II})
l_2	(c_2^I, c_1^{II})	(c_2^I, c_2^{II})

$+$

	l_1	l_2
l_1	(c_1^I, c_1^{II})	(c_2^I, c_1^{II})
l_2	(c_1^I, c_2^{II})	(c_2^I, c_2^{II})

$$\begin{pmatrix} 0 & c_1^{II} - c_2^{II} \\ c_1^I - c_2^I & c_1^{II} - c_2^{II} + c_1^I - c_2^I \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$$

Therefore, in the complete strategic form in Table 2, each AS sums the costs due to the two community A $\leftrightarrow$ community B flows. (l_2, l_2) is the unique Nash equilibrium. Hence, rational ASs would implicitly coordinate as suggested by (l_2, l_2), which in this case corresponds to accept the suggestion to routing the flow toward the neighbor's preferred egress router.

Let c_i^I and c_i^{II} be the IGP costs between R_a and R_b (resp.) and l_i, $i \in E$. For the generic case of two inter-AS links, the cost vector for the strategy profile (l_i, l_j), $i, j \in \{1, 2\}$, is thus $(c_i^I + c_j^I, c_i^{II} + c_j^{II})$. The resulting ClubMED game (Table 3) can be described as $G = G_s + G_d$, sum of two games. $G_s = (X, Y, f_s, g_s)$, a selfish game, purely endogenous, where X and Y are the set of strategies and $f_s, g_s : X \times Y \to \mathbf{N}$ the cost functions, for AS I and AS II (resp.). In particular, $f_s(x, y) = \phi_s(x)$, where $\phi_s : X \to \mathbf{N}$, and $g_s(x, y) = \psi_s(y)$, where $\psi_s : Y \to \mathbf{N}$. $G_d = (X, Y, f_d, g_d)$, a dummy game, of pure externality, where $f_d, g_d : X \times Y \to \mathbf{N}$ are the cost functions for AS I and AS II (resp.). In particular, $f_d(x, y) = \phi_d(y)$, where $\phi_d : Y \to \mathbf{N}$, and $g_d(x, y) = \psi_d(x)$, where $\psi_d : X \to \mathbf{N}$. G_s is a cardinal potential game [16], i.e., the incentive to change players' strategy can be expressed in one global function, a potential function (P_s), and the difference in individual costs by an individual strategy move has the same value as the potential difference. G_d can be seen as a potential game too, with null potential (P_d). G has thus a potential $P = P_s + P_d = P_s$. In the bottom of Table 3 we report P_d and P_s. To explicate P_s (thus P) we use a form in which we set to 0 the minimum of ϕ_s and ψ_s, i.e., $P_s(x_0, y_0) = 0$ where: $\phi_s(x_0) \leq \phi_s(x) \ \ \forall x \in X$, and $\psi_s(y_0) \leq \psi_s(y) \ \ \forall y \in Y$. In potential games, the potential function minimum corresponds to a Nash equilibrium, but the inverse is not necessarily true. The next theorem proves that the inverse is also true for G.

Theorem 4.1 *A ClubMED Nash equilibrium corresponds to the strategy profile with minimum potential.*

Proof If (x^*, y^*) is an equilibrium, $P(x^*, y^*) \leq P(x, y^*)$, $\forall x \in X$. But: $P(x^*, y^*) = \phi_s(x^*) - \phi_s(x_0)$ and $P(x, y^*) = \phi_s(x) - \phi_s(x_0)$, $\forall x \in X$. Thus $P(x^*, y^*) \leq P(x, y^*)$, $\forall x \in X$, is equivalent to $\phi_s(x^*) - \phi_s(x_0) \leq \phi_s(x) - \phi_s(x_0)$, $\forall x \in X$, that is $\phi_s(x^*) \leq \phi_s(x)$, $\forall x \in X$. Hence x^* is a minimum for ϕ_s. Idem for y^*. So $P(x^*, y^*) = 0$, that is a minimum of P. □

Given that $P = P_s$, G_s fully guides the G Nash equilibrium.

Corollary 4.2 *G always possesses a Nash equilibrium.*

Indeed, authors in [16] prove that finite potential games always possess a (pure-strategy) equilibrium. The opportunity of using the minimization of the potential function represents a key advantage of the ClubMED solution. It decreases the complexity of the Nash equilibrium computation, which may be very high for large instances (especially for the generalized framework presented in the following). Therefore, for this base ClubMED modeling, if the equilibrium is unique it corresponds to hot-potato routing because G_s considers egress costs only, which somehow validates the current practice (however, we will explain how this differs in the generalized framework). When there are multiple equilibria, G_d can help in avoiding tie-breaking routing by the selection of an efficient equilibrium in the Pareto-sense (as detailed below).

Definition 4.3 A strategy profile s is *Pareto-superior* to another profile s' if a player's cost can be decreased from s to s' without increasing the other players' costs.

Remark: And s' is *Pareto-inferior* to s.

Definition 4.4 A strategy profile is *Pareto-efficient* if it is not Pareto-inferior to any strategy profile.

Remark: Pareto-efficient profiles form the *Pareto-frontier*.

In Table 2, (l_2, l_2) is Pareto-inferior to (l_1, l_2) because $2c_2^{II} < c_1^{II} + c_2^{II}$, and (l_1, l_2) forms a singleton Pareto-frontier.

It is worth noting that the MEDs of AS I and AS II are never compared, never summed together, hence they can be calculated over different integer scales. What is important to sort strategy profiles is the ordering of individual AS costs.

4.2 ClubMED game and BGP routing

Let $MED_I^{II}(l_i)$ be the MED advertised by AS II to AS I received via eBGP on the link l_i, and vice-versa for $MED_{II}^{I}(l_i)$.

Table 4 ClubMED strategic form with inverted AS I weights for Fig. 6

I \ II	l_1	l_2
l_1	(10,50)	(**55**, **40**)
l_2	(55,40)	(100,30)

Definition 4.5 The *MED ordering vector* $\overline{L}_I$ of the neighbor AS I is a vector of link indexes monotonically ordered with respect to $MED_I^{II}(l_i)$.

Remark: The same for $\overline{L}_{II}$ w.r.t. $MED_{II}^{I}(l_i)$.

Definition 4.6 Two *MED-aligned neighbors*, AS I and AS II, are such that $L_{II} = L_I$.

Remark: And *MED-disaligned* if not MED-aligned.

So, to study a critical example with MED-disaligned ASs, we can swap the R_a-R_1^I and R_b-R_2^I IGP weights for the previous example (Table 2). The resulting strategic form is in Table 4—with inverted AS I cost in (l_1, l_1) and (l_2, l_2), now 10 and 100 respectively. Swapping the R_a-R_1^I and R_b-R_2^I IGP costs (in Fig. 6) we have MED-disaligned ASs, and it is easy to verify that the Nash equilibrium is (l_1, l_2) with costs (55, 40). It is worth noting that: (l_2, l_1) has costs equal to (l_1, l_2), but it is not an equilibrium—because AS I and AS II would prefer l_1 to l_2 and l_2 to l_1, respectively, fixed the other AS strategy. The ClubMED game still behaves as hot potato routing, but in this case the MEDs of AS I are not respected by AS II.

Corollary 4.7 *The ClubMED Nash equilibrium for MED-aligned ASs with two inter-AS links is alike applying hot potato routing at both ASs.*

Proof MED-aligned neighbors can be such that:

$$c_2^I < c_1^I \ \wedge \ c_2^{II} < c_1^{II} \tag{1}$$

In (l_1, l_1) both ASs route against hot potato routing; it is an equilibrium if:

$$c_1^I + c_2^I \geq 2c_1^I \ \wedge \ c_1^{II} + c_2^{II} \geq 2c_1^{II} \tag{2}$$

not true given (1). In (l_2, l_1) and (l_1, l_2) a single AS routes against hot potato routing; they are equilibria if (resp.):

$$c_1^I + c_2^I \leq 2c_1^I \ \wedge \ c_1^{II} + c_2^{II} \leq 2c_2^{II} \tag{3}$$

$$c_1^I + c_2^I \leq 2c_2^I \ \wedge \ c_1^{II} + c_2^{II} \leq 2c_1^{II} \tag{4}$$

both not true given (1). In (l_2, l_2) both the ASs route alike hot potato routing and the other not; it is an equilibrium if:

$$c_1^I + c_2^I \geq 2c_2^I \ \wedge \ c_1^{II} + c_2^{II} \geq 2c_2^{II} \tag{5}$$

true given (1). □

Definition 4.8 Two neighbors, AS I and AS II, are *strongly MED-disaligned* if not MED-aligned and if $\forall i \neq j$:

$$MED_I^{II}(l_i) \neq MED_I^{II}(l_j) \ \wedge \ MED_{II}^{I}(l_i) \neq MED_{II}^{I}(l_j) \tag{6}$$

Corollary 4.9 *The ClubMED Nash equilibrium for strongly MED-disaligned ASs with two inter-AS links is alike applying hot potato routing at both ASs.*

Proof Strongly MED-disaligned neighbors can be such that $c_2^I > c_1^I \wedge c_2^{II} < c_1^{II}$. So, (5) and (3) and (2) are not satisfied but (4) is, which corresponds to (l_1, l_2), i.e., to route alike and against hot potato at both sides. Similarly, if $c_2^I < c_1^I \wedge c_2^{II} > c_1^{II}$, (5) and (4) and (2) are not satisfied, but (3) is, which corresponds to (l_2, l_1), i.e., to route alike hot potato. □

Differing from the MED-aligned case, in this case the MED is respected at one side only.

Definition 4.10 Two neighbors, AS I and AS II, are *weakly MED-disaligned* if not MED-aligned and if $\exists i \neq j$:

$$MED_I^{II}(l_i) = MED_I^{II}(l_j) \ \vee \ MED_{II}^{I}(l_i) = MED_{II}^{I}(l_j) \tag{7}$$

Corollary 4.11 *The ClubMED Nash equilibrium for weakly MED-disaligned neighboring ASs with two inter-AS links allows avoiding tie breaking routing.*

Proof Weakly MED-disaligned neighbors can be such that:

$$c_2^I = c_1^I \ \wedge \ c_2^{II} < c_1^{II} \tag{8}$$

Normally, the tie breaking would rule at AS I. Updating (2)–(5) with (7), ClubMED has two Nash equilibria, (l_1, l_2) and (l_2, l_2). AS II routes always alike hot potato routing while AS I, following a Nash equilibrium, routes bypassing thus the tie breaking rule given that the hot potato rule does not apply. □

4.3 Generalization to directed metrics and multiple links

So far, we assumed that the cost metric announced through the MED stands for the two directions, the incoming and the outgoing ones. Normally, it corresponds only to the incoming one (IGPs can manage directed costs). The modeling of directed costs intuitively does not change the ClubMED equilibrium and the Pareto-efficiency.[2] It decouples the (egress) costs used to form G_s from which the potential function is build from the (ingress) ones used to form G_d.

[2] We assume that composite MED attributes can be easily coded to transport both the ingress and the egress costs (and other costs mentioned hereafter).

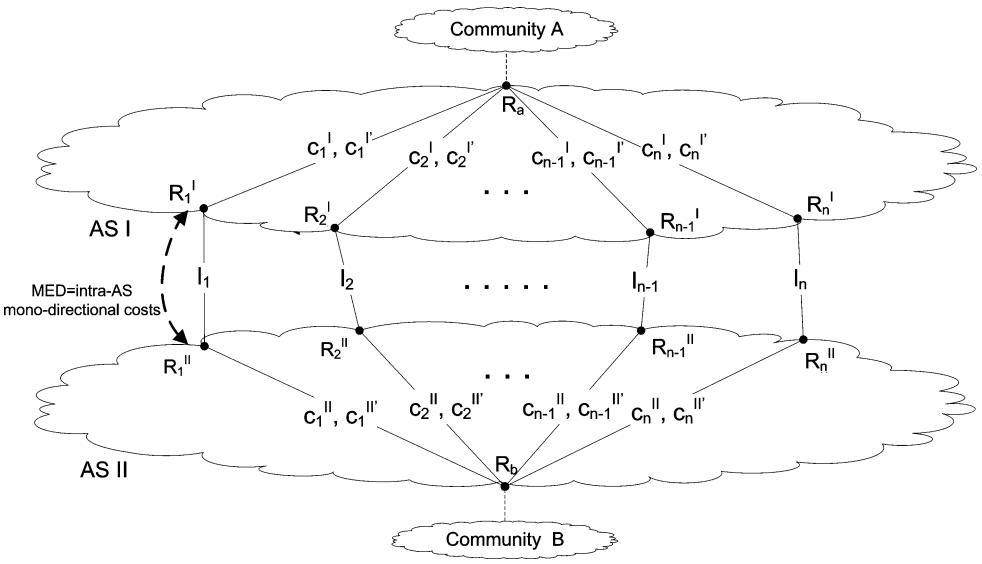


Fig. 7 ClubMED interaction example with multiple inter-AS links

Further generalizing, multiple inter-AS links are to be considered, as depicted in Fig. 7. Let $|E| = n$ be the link number, l_i and l_j, $i, j \in E$, the links chosen by AS I and AS II (resp.). Let c_i^I and $c_i^{I'}$ be the ingress and egress costs at link i for AS I, and idem c_i^{II} and $c_i^{II'}$ for AS II. The costs corresponding to the strategy profile (l_i, l_j) are $(c_i^I + c_j^{I'}, c_i^{II'} + c_j^{II})$.

The Nash equilibrium conditions in (l_i, l_j) are for AS I and AS II (resp.):

$$c_i^I + c_j^{I'} \leq c_k^I + c_j^{I'}, \quad \forall k \in E$$
$$c_i^{II'} + c_j^{II} \leq c_i^{II'} + c_k^{II}, \quad \forall k \in E \tag{9}$$

Corollary 4.12 *The ClubMED Nash equilibrium for not weakly MED-disaligned ASs with many inter-AS links is alike applying hot potato routing at both ASs.*

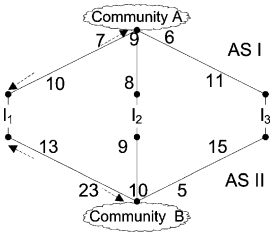
Proof For not weakly MED-disaligned neighbors, (9) have strict inequalities, thus $c_i^I < c_k^I$, $\forall k \in E$, and $c_j^{II} < c_k^{II}$, $\forall k \in E$, which correspond to hot potato routing. □

Corollary 4.13 *The ClubMED Nash equilibrium for not weakly MED-disaligned ASs with many inter-AS links allows avoiding tie breaking routing.*

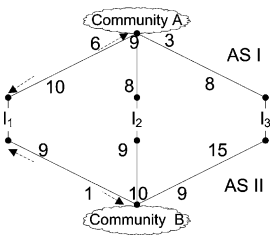
Proof For weakly MED-disaligned neighbors, (9) have at least one equality, i.e., $\exists k \in E | c_i^I = c_k^I \vee c_j^{II} = c_k^{II}$, which clearly avoids tie breaking routing. □

With multiple links, the occurrence of multiple equilibria increases. This happens under the necessary conditions:

$$\exists i, k \in E \mid i \neq k \wedge c_i^{I'} = c_k^{I'} \tag{10}$$
$$\exists i, k \in E \mid i \neq k \wedge c_i^{II'} = c_k^{II'} \tag{11}$$



I\II	l_1	l_2	l_3
l_1	$(17,36)^6$	$(19,32)^2$	$(16,38)^8$
l_2	$(15,23)^4$	$\mathbf{(17,19)^0}$	$(14,25)^6$
l_3	$(18,18)^7$	$(20,14)^3$	$(17,20)^9$



I\II	l_1	l_2	l_3
l_1	$(16,10)^2$	$(19,10)^2$	$(13,18)^8$
l_2	$\mathbf{(14,19)^0}$	$\mathbf{(17,19)^0}$	$(11,25)^6$
l_3	$\boldsymbol{(14,18)^0}$	$\mathbf{(17,18)^0}$	$(11,24)^6$

Fig. 8 3-link examples

for AS I and AS II (resp.). Given a link l_k, thus the ingress cost c_k^I, the strategy profiles (l_k, l_i) or (l_i, l_k)—many $i \in E$ may satisfy (10) or (11)—are eligible equilibria if both (10) and (11) are satisfied. Figure 8 depicts two examples with 3 inter-AS links and their strategic forms. The exponent indicates the corresponding potential value. Egress costs are given close to the egress points, while ingress costs are close to the destination communities. For the upper case, (10) and (11) are not satisfied: there is a single equilibrium (l_2, l_2). For the lower case, (10) and (11) hold: there are four equilibria. Which should be selected? The neighbors could easily coordinate in playing the Pareto-superior one, (l_3, l_1).

In ClubMED, a strategy profile (l_r, l_s) is Pareto-superior to a strategy profile (l_i, l_j) if:

$$\left(c_r^I + c_s^{I'} < c_i^I + c_j^{I'} \wedge c_r^{II'} + c_s^{II} \leq c_i^{II'} + c_j^{II}\right)$$
$$\vee \left(c_r^I + c_s^{I'} \leq c_i^I + c_j^{I'} \wedge c_r^{II'} + c_s^{II} < c_i^{II'} + c_j^{II}\right) \quad (12)$$

For the lower case, note that the Pareto-superior equilibrium is not Pareto-efficient, it is Pareto-inferior to (l_1, l_3) that is the single element of the Pareto-frontier—(l_1, l_3) is not an equilibrium because AS I will always prefer l_2 or l_3 to l_1 ($11 < 13$). This is due to the external effect of G_d. Indeed, it is possible that, after an iterated reduction of strategies, G assumes the form of a Prisoner-dilemma game, in which equilibria are Pareto-inferior to non-equilibrium strategy profiles.

4.4 Nash equilibrium multi-path (NEMP) coordination policy

Within the described framework two AS neighbors would rationally route according to a ClubMED equilibrium profile because it grants a rational stability to the routing decision.

By analyzing the exchanged composite MED values the neighbors can build the ClubMED strategic game form, and would rationally play its Nash equilibrium strategy profiles because they grant a rational stability to the routing decision, which would follow the hot potato routing at both sides when a single equilibrium exists, and avoid tie breaking routing when many exist yielding to a (more) rational routing decision. These proofs may seem not so useful at a first sight because ClubMED would improve the normal BGP solution only when hot potato rules can not be applied. However, with many links, mono-directional metrics and the needed practical extensions defined hereafter (see Sects. 4.5 and 4.6), the uniqueness of Nash equilibria is likely to be a rare event, hence the interest in defining at this point absolute ClubMED-based possible coordination strategies mixing the Nash equilibrium, the Pareto-efficiency and the minimum potential criteria. Let us further clarify these game-theoretical concepts:

- the ClubMED Nash equilibrium solution corresponds to the strategy profile(s) with minimum potential, which, however, may not be Pareto-efficient (please note that this is not true in general for potential games).
- With many equilibria, the Pareto-efficiency allows shrinking the Nash set to a unique element or to many equivalent ones, which may not belong to the G Pareto-frontier.
- A Pareto-efficient strategy profile that is not a Nash equilibrium, even if rationally advantageous it should not be played in a fully non-cooperative framework because the other neighbor would have a better move.

In a fully non-cooperative framework, the ClubMED implicit policy to which to coordinate is: *to play the equilibria in the Nash set*. Hence, it is feasible to natively implement a *Nash Equilibrium Multi-Path (NEMP)* inter-AS routing policy. No coordination signaling message is needed: NEMP can be applied at only one side, under the assumption that a rational agent, as a carrier shall be, would route accordingly to a Nash equilibrium. In the bottom of Fig. 8, e.g., AS I may balance the load on l_2 and l_3, being aware that AS II may or may not balance its load on l_1 and l_2. Especially top-tier carriers, interconnected at numerous Points of Presence worldwide, would benefit from NEMP avoiding so sudden bottlenecks at inter-carrier links. In fact, the NEMP policy implies multipath routing on inter-AS link when more than one equilibrium is selected. However, as mentioned above, the set of equilibria can be restricted to the Pareto-superior ones; but many Pareto-superior equilibria can exist, so the NEMP policy is to be used on the Pareto-superior profiles of the Nash set. Please note that there may not exist Pareto-superior equilibria: in this case, NEMP is performed on all the equilibria.

Finally, it is worth remarking that, from a computational standpoint, the NEMP policy is very efficient in that it simply requires the minimization of the potential value and a trivial Pareto-restriction of the Nash set, even if this contains "simple" equilibria of the one-shot game.

4.5 Dealing with incomplete cost information

Nowadays, IGP weights are frequently optimized and automatically updated rather than being manually configured. In this sense, we should assume that the ClubMED costs are subject to changes when the ingress/egress flow directions changes. The costs in Fig. 8, e.g., may be computed for the starting profile (l_1, l_1). A change of the AS II-head flow via l_2 or l_3 may cause a decrease of the ingress cost by l_1, because of available bandwidth decrease on the corresponding links, and/or an increase of the ingress cost at l_2 or l_3 for the opposite reason. It may also happen that the ingress/egress cost from an AS to another changes when an egress/ingress flow direction changes because, for large topologies, flows with inverse inter-domain directions may use core links in the same direction.

As currently formulated, with IGP-WO operations the ClubMED would converge to a stable configuration after some repetitions. The ClubMED decision shall be kept stable as long as needed to avoid too many route oscillations while assuring a good solution. In order to reach this purpose, we integrate IGP-WO operations as follows. Let $\delta_{i,j}^{k,I}$ and $\delta_{i,j}^{k,I'}$ be the variations of the egress and the ingress cost (resp.), at AS I for link k, passing from the current strategy profile to the profile (l_i, l_j). Similarly, $\delta_{i,j}^{k,II}$ and $\delta_{i,j}^{k,II'}$ for AS II. Once pre-computed, the δ may be conveyed within the MED attribute to refine the strategic form. How-

ever, while announcing loose intra-AS costs is not so critic because the intra-domain topology is fully abstracted, exchanging the δ may represent an excessive insight in a carrier's operations (that might allow an AS to partially infer the neighbor's network status). Hence, the δ should be abstracted. Using the current cost and its δ variations, each neighbor can just announce an error to give an interval of equivalence for the computation of the equilibrium. Let ϵ^I and ϵ^{II} be these egress cost errors for AS I and AS II (resp.). Being aware that IGP weights may significantly increase, we opt for an optimistic min-max computation:

$$\epsilon^I = \min_{k\in E}\left\{\max_{i,j\in E}\left\{\delta_{i,j}^{k,I}\right\}/c_k^I\right\} \tag{13}$$

Similarly for ϵ^{II}, and the ingress cost errors of AS I and AS II, i.e., $\epsilon^{I'}$ and $\epsilon^{II'}$ (resp.). The ϵ cost errors represent a good trade-off between network information hiding and coordination requirement: not announcing per-link errors avoid revealing the δ variations; announcing directed errors (ingress and egress) reflects the fact that upstream and downstream availability is likely to be unbalanced because of the asymmetric bottlenecks around inter-AS links.

The important effect of the errors is a larger number of equilibria. Indeed, they arise a *potential threshold* under which a profile becomes an equilibrium. That is, first the minimum potential strategy ($P(x^*, y^*)$) is found, then the other profiles that have a potential within the minimum plus the threshold (T_P) are considered as equilibria too. More precisely, in the worst case, each potential difference ΔP from strategy i to j can be increased by the amount (for AS I) $\epsilon^I(c_i + c_j)$. Generically, in the worst case the ΔP from (x_1, y_1) to (x_2, y_2) can be increased by $a_I(x_1, x_2) + a_{II}(y_1, y_2)$, where:

$$\begin{aligned} a_I(x_1, x_2) &= \epsilon^I[\phi_s(x_1) + \phi_s(x_2)] \\ a_{II}(y_1, y_2) &= \epsilon^{II}[\psi_s(y_1) + \psi_s(y_2)] \end{aligned} \tag{14}$$

It is reasonable to opt for the following optimistic threshold:

$$T_P = \min_{x_1,x_2\in X}\{a(x_1, x_2)\} + \min_{y_1,y_2\in Y}\{a(y_1, y_2)\} \tag{15}$$

All strategy profiles (x, y) such that $P(x, y) \le P(x^*, y^*) + T_P$ will be considered as equilibria. More straightforwardly, the Pareto-superiority condition (12) can be easily extended considering $\epsilon^{I'}$ and $\epsilon^{II'}$. For the upper case in Fig. 8, e.g., for simplicity let all the ϵ cost errors be equal to 12%. Besides the existing equilibrium (l_2, l_2), one new equilibria is added: (l_1, l_2). Besides anticipating possible future routing deviations, the potential threshold may also allow escaping selfish solutions mainly guided by G_s: Pareto-superior profiles may be introduced in the Nash set and then selected.

4.6 Dealing with multiple flows and inter-AS link congestion

In order to take broader decisions, it would result more useful to consider many pairs of inter-cone flows in a same ClubMED game. In this way the equivalence condition can be extended to all the pairs together, not necessarily related to a same couple of ClubMED routers. For 2 pairs and 2 links, the set of strategies X^2 or Y^2 becomes $\{l_1l_1, l_1l_2, l_2l_1, l_2l_2\}$. For m pairs and n links, the multi-pair game is the repeated permutation of m single-pair n-link games: $|X^m| = |Y^m| = n^m$. $G = (X^m; Y^m; f_s, f_d, g_s, g_d : X^m \times Y^m \to \mathbf{N})$.

In a multi-pair ClubMED framework, carriers shall control the congestion on inter-AS links. The more egress flows are routed on a inter-AS link, the more congested the link, and the higher the routing cost. We aim at weighting thus the inter-carrier links when congestion may arise due to the inter-AS flow routing. This may be done by modeling the inter-AS link in IGP-WO operations (e.g. [2]), but the requirement of separating intra-domain from inter-domain routing would not be met [23]. We add an endogenous congestion game $G_c = (X^m; Y^m; f_c, g_c : X^m \times Y^m \to \mathbf{N})$ to G, where $f_c(x^m, y^m) = \phi_c(x^m)$ and $g_c(x^m, y^m) = \psi_c(y^m)$. Let H be the set of inter-AS flow pairs, ρ_h the outgoing bitrate of the pair $h \in H$, and C_i the egress available capacity of l_i. G_c should not count when $\sum_{h\in H}\rho_h \ll \min_{i\in E}\{C_i\}$, otherwise it should do affecting the G equilibrium selection. The G_c costs are to be monotonically increasing with the number of flows routed on a same link. An effective and commonly agreed congestion cost convex function is $1/(C-\rho)$, where $(C-\rho)$ is the idle capacity [9]. We shall use (idem for $\psi_c(y^m)$):

$$\phi_c(x^m) = \sum_{i\in E|l_i\in x^m}\left\lceil K_i\frac{1}{C_i - \sum_{h\in H}\rho_h^i}\right\rceil \tag{16}$$

If $C_i < \sum_{h\in H}\rho_h^i$, $K_i = \infty$. Otherwise, K_i are constants to be scaled to make the cost comparable to IGP weights, e.g., such that it is 1 when the idle capacity is maximum, i.e., $K_i = C_i$. ρ_h^i is the fraction of the pair h flow that is routed on l_i.

5 Experimentation results

We created a virtual interconnection scenario among the Geant2 and the Internet2 ASs, depicted in Fig. 9, emulating their existing interconnection with three cross-Atlantic links. We considered six pairs of inter-cone flows among the routers depicted with crossed circles. The TOTEM toolbox [11] was used to run a IGP-WO heuristic, with a maximum link weight of 50 for both ASs. We used 360 successive traffic samples, over-sampling the datasets from [24] for

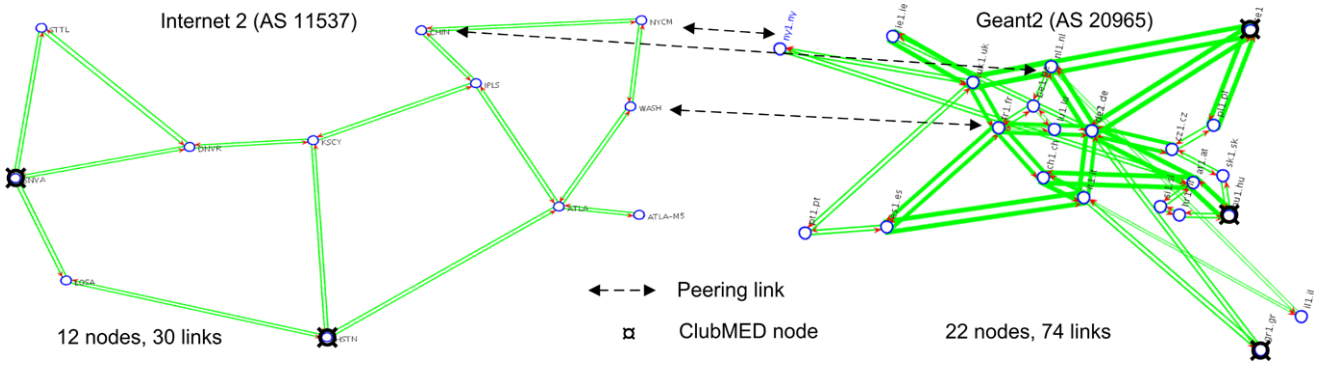


Fig. 9 Internet2–Geant2 interconnection scenario with 3 inter-AS links

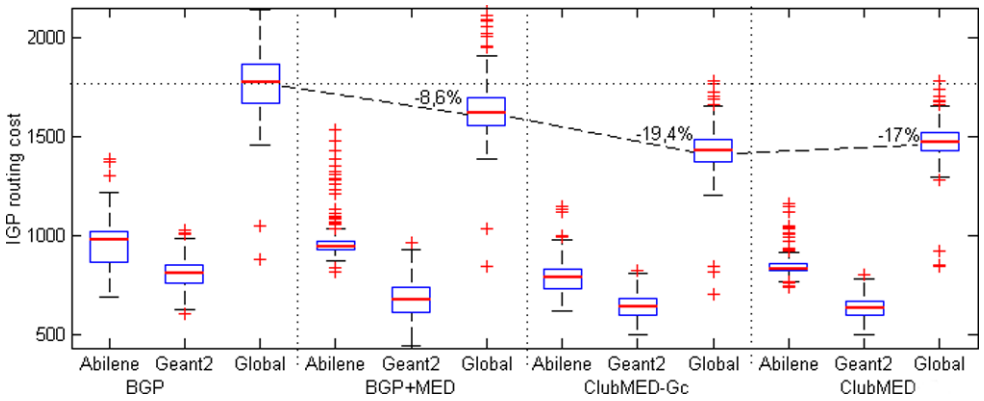


Fig. 10 Global routing cost Boxplot statistics

Geant2 and from [8] for Internet2 on a 8 h basis (to cover all the day times). The original intra-AS link capacities have been considered. The inter-cone routing generates additional traffic for the traffic matrices. We used a random inter-cone traffic matrix such that flows are equivalent with 200 Mb/s per direction, which corresponds to 2/3 of the total available interconnection capacity. To evaluate the effectiveness of the congestion game we considered inter-AS links with 100 Mb/s available per direction.

We compare ClubMED to the BGP solution, without and with ('BGP+MED' in the figures) MED signaling at both sides. Figure 10 reports the routing costs statistics in BoxPlot format (minimum; box with lower quartile, median, upper quartile; maximum; outliers). For each method, we display the Internet2, the Geant2 and the global routing cost. For the first two figures only, we considered two ClubMED solutions, without and with the congestion game G_c. Figure 11 reports the maximum link load in Boxplot statistics as seen by each neighbor, with the four above-mentioned methods. Figure 12 reports the number of ClubMED Nash equilibria and those Pareto-superior in a log-scale for all the rounds. When no Pareto-superior equilibria were found, NEMP was applied to all the Nash equilibria. Figure 13 reports the number of routing changes with respect to the previous round (with an upper bound equal to the total number of flows), together with the Boxplot statistics.

All in all, we can synthetically assess that:

- the median routing cost is reduced of roughly 17% (simple uncoordinated MED signaling already improves it by 8%, but ClubMED further improves it);
- the addition of the congestion game G_c slightly augments it, but allows nullifying the congestion on inter-AS links (that appear over-congested with a median between 130% and 200% with BGP, and a few congested with ClubMED without G_c);
- comparing BGP with BGP+MED, the latter seems improving the performance of Geant2 and Internet2 in terms of routing cost and maximum inter-AS link load respectively, and conversely; this is probably due to the higher global path cost for Internet2 and to the higher number of intra-AS connection requests for Geant2;
- the Pareto-superiority condition permits to pick a few efficient Nash equilibria over broad sets, whose dimension varies significantly in time (this reveals a high sensitivity to the routing costs);
- the routing stability is significantly improved thanks to the consideration of the loose cost errors in G and thus to the arise of the potential minimum threshold: we pass from a *median* of 4 routing changes per round on 12 possible ones with BGP, to a median of 0 with ClubMED; ClubMED can significantly increase routing stability;

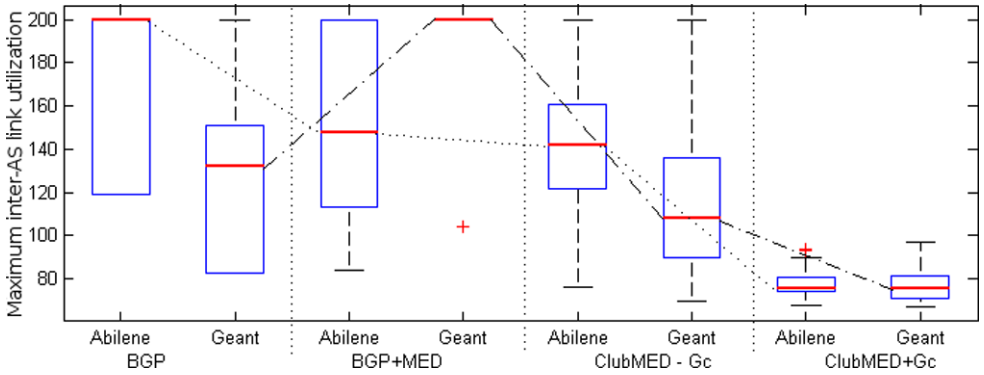


Fig. 11 Boxplot statistics of the maximum link load (%)

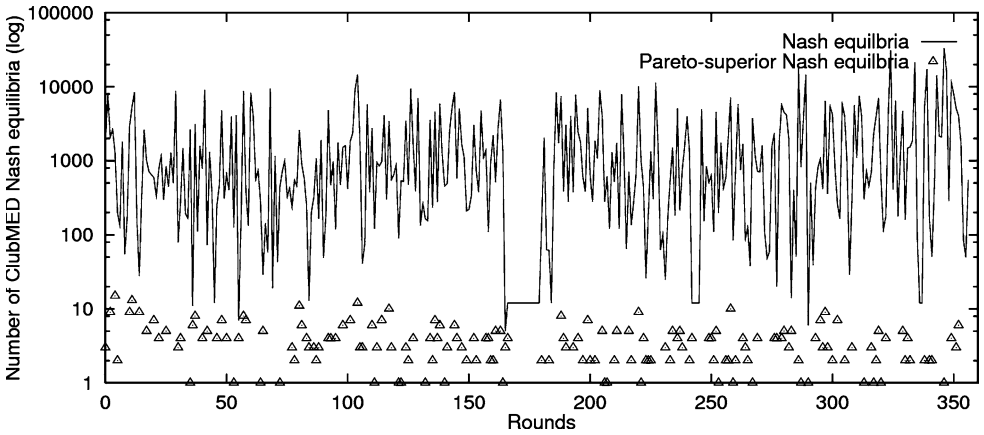


Fig. 12 Dynamics of the number of found Nash equilibria

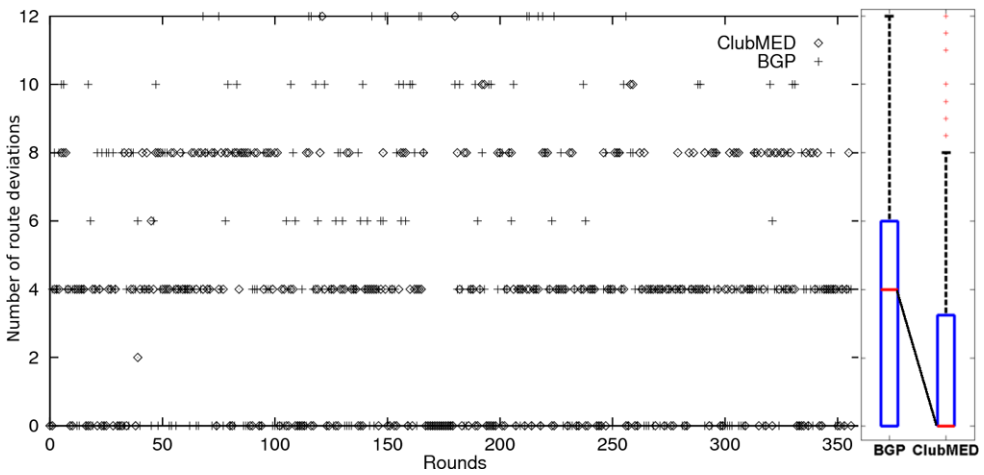


Fig. 13 Dynamics and Boxplot statistics of the routing deviations

– a better behavior in terms of routing stability seems corresponding to larger Nash sets (cf. Fig. 12 and Fig. 13).

6 Summary

We proceeded with a deductive analysis of the coordinated inter-AS routing interaction via the MED attribute of BGP. Firstly, by an in-depth analysis of an Internet routing map history, we characterized AS path route deviations across top-tier carrier interconnections. The analysis highlights a frequent occurrence of both intra-AS path and AS path deviations. Behind these events there is an uncoordinated coupling between IGP and BGP routing, which appears to be critical across top-tier interconnections. In order to improve current BGP routing, we modeled the bilateral interac-

tion among Autonomous Systems (ASs) as a routing game, named ClubMED; it is a composition of a potential game guiding the Nash equilibrium selection and of a dummy game affecting the Pareto-efficiency. The ClubMED game routes equivalent aggregates of inter-carrier flows over multiple links, includes cost errors due to IGP-WO operations and is able to control the congestion of inter-AS links. The frequent occurrence of multiple ClubMED Nash equilibria arises a need for a coordination policy. We presented the NEMP routing policy that shall be implemented on the Nash equilibria and Pareto-superior profiles. Finally, we validated the ClubMED framework emulating the interconnection between the European and North American research networks using real datasets.

The results show that the global routing cost can be reduced of roughly 17%, that the inter-AS link congestion can be avoided with the addition of an endogenous congestion game, and that the inter-AS routing can be stabilized. Besides this promising performance, in the ClubMED framework the carriers' selfish and non-cooperative behavior is respected as an imperative requirement. The ClubMED framework emerges as a pragmatic and effective solution between the current uncoordinated practice and other ideal yet unwise cooperative solutions. It may be implemented to overlay a special interconnection policy for critical inter-carrier flows above co-existing settlements concerning flows whose routing can not be coordinated. It may allow a still finer policy routing by the artificial addition of endogenous costs. It may freeze the wild bargaining that nowadays characterizes top-tier routing settlements [6], yielding to long-term and effective inter-carrier agreements for the future Internet.

Further work is needed to study coordination policies other than the presented NEMP one. The idea is to define suboptimal yet effective policies for the repeated ClubMED game. Moreover, we are working for the definition of an extended interconnection framework relying on a similar game-theoretical modeling, wherein the borders with multiple ASs are modeled as a classical bilateral interconnection border.

Acknowledgements The authors thank Matteo Marinoni and Eric Elena for their support in the code implementation for the detection of BGP routing deviations. This work has been funded by the INCAS (INter Carrier Alliance Strategies) S.JRA of the EU IST Euro-NF Network of Excellence, the ICF (Networks of the Future Lab) I-GATE (Internet- GAme-Theoretical Analysis of Traffic Engineering methods) project of the Institut Télécom, France, and the European FP7 Integrated Project ETICS (Economics and Technologies for Inter-Carrier Services).

References

1. Agarwal, S., Nucci, A., & Bhattacharyya, S. (2004). Controlling hot potatoes in intradomain traffic engineering. SPRINT RR04-ATL-070677.
2. Balon, S., & Leduc, G. (2008). Combined Intra and inter-domain traffic engineering using hot-potato aware link weights optimization. arXiv:0803.2824.
3. CAIDA ranking (website). http://as-rank.caida.org.
4. Donnet, B., & Friedman, T. (2007). Internet topology discovery: a survey. *IEEE Communications Surveys and Tutorials*, *9*(4), 2–15.
5. Douville, R., Le Roux, J.-L., Rougier, J.-L., & Secci, S. (2008). A service plane over the PCE architecture for automatic multi-domain connection-oriented services. *IEEE Comm. Magazine*, *46*(6), 94–102.
6. Faratin, P., et al. (2007). Complexity of Internet interconnections: technology, incentives and implications for policy. In *Proc. of TPRC 2007*.
7. Huffaker, B., et al. (2002). Distance metrics in the Internet. In *Proc. of IEEE international telecommunications symposium (ITS) 2002*.
8. By courtesy of Y. Zhang. Internet2/Abilene topology and traffic dataset. http://www.cs.utexas.edu/~yzhang/research/AbileneTM.
9. Larroca, F., & Rougier, J.-L. (2009). Routing games for traffic engineering. In *Proc. of IEEE ICC 2009*.
10. Latapy, M., Magnien, C., & Ouédraogo, F. (2008). A radar for the Internet. In *Proc. of ADN 2008*.
11. Lepropre, J., Balon, S., & Leduc, G. (2006). Totem: a toolbox for traffic engineering methods. In *Proc. of INFOCOM 2006*.
12. LIP6 complex networks website, radar traces. http://data.complexnetworks.fr/Radar.
13. Ma, R. T. B., et al. (2007). Internet economics: the use of Shapley value for ISP settlement. In *Proc. of CoNEXT 2007*.
14. Ma, R., et al. (2008). Interconnecting eyeballs to content: a Shapley value perspective on ISP peering and settlement. In *Proc. of SIGCOMM 2008*.
15. McPherson, D., & Gill, V. BGP MED considerations. RFC 4451.
16. Monderer, D., & Shapley, L. S. (1996). Potential games. *Games and Economic Behavior*, *14*(1), 124–143.
17. Myerson, R. B. (1991). *Game theory: analysis of conflict*. Harvard: Harvard University Press.
18. PlanetLab website. http://www.planet-lab.org.
19. Roth, A. E. (1988). *The Shapley value, essays in honor of Lloyd S. Shapley*. Cambridge: Cambridge University Press.
20. Secci, S., Rougier, J.-L., Pattavina, A., Patrone, F., & Maier, G. (2009). ClubMED: coordinated multi-exit discriminator strategies for peering carriers. In *Proc. of 2009 5th Euro-NGI/IEEE conference on next generation Internet networks (NGI 2009)*, Aveiro, Portugal, 1–3 July 2009.
21. Shrimali, G., et al. (2007). Cooperative inter-domain traffic engineering using Nash bargaining and decomposition. In *Proc. of INFOCOM 2007*.
22. Teixeira, R., et al. (2005). TIE breaking: tunable interdomain egress selection. In *Proc. of CoNEXT 2005*.
23. Teixeira, R., et al. (2008). Impact of hot-potato routing changes in IP networks. *IEEE/ACM Transactions on Networking*, *16*(6), 1295–1307.
24. Uhlig, S., et al. (2006). Providing public intradomain traffic matrices to the research community. *Computer Communication Review*, *36*(1), 83–86.

Stefano Secci is Associate Professor at the University Pierre et Marie Curie – Paris VI (LIP6 lab), France, since October 2010. He received a dual Ph.D. in computer science and networks from Politecnico di Milano, Italy, and from Telecom ParisTech, France, in 2009, and a M.Sc. in communications engineering from Politecnico di Milano, in 2005. He worked as postdoctoral research associate at NTNU, Norway, and George Mason University, USA. Before the Ph.D., he worked as network engineer at Fastweb Italia. His works space from optical network planning and switching to IP routing optimization and traffic engineering. He is author of 20+ articles in major international journals and conferences, and he is the recipient of the NGI 2009 Best Paper Award. His current interests are about future Internet routing, mobility and policy.

Jean-Louis Rougier received his engineering diploma in 1996 and his Ph.D. in 1999 from Télécom ParisTech (formerly called École Nationale Supérieure des Télécommunications). He joined the computer science and networks department of Télécom ParisTech in 2000 as an associate professor. He has been working ever since on routing and traffic engineering for networks in different contexts (Optical, Wireless Mesh networks, IP, Post-IP). He has been involved in several national and European projects and industrial collaborations. He is a technical advisor for several companies. His main research interest is currently on the evolution of architectures for the future Internet.

Achille Pattavina received the Dr.Eng. degree in electronic engineering from University La Sapienza of Rome (Italy) in 1977. He was with the same university until 1991 when he moved to the Politecnico di Milano, Milan (Italy), where he is now a Full Professor. He has been the author of more than 100 papers in the area of communications networks published in leading international journals and conference proceedings. He has been guest or coguest editor of special issues on switching architectures in IEEE and non-IEEE journals. He has been engaged in many research activities, including European Union-funded projects. Dr. Pattavina has authored two books, Switching Theory, Architectures and Performance in Broadband ATM Networks (Wiley, 1998) and Communication Networks (McGraw-Hill, 1st edn., 2002, 2nd edn., 2007, in Italian). He has been an Editor for Switching Architecture Performance of the IEEE Transactions on Communications since 1994 and Editor-in-Chief of the European Transactions on Telecommunications since 2001. He is a Senior Member of the IEEE Communications Society. His current research interests are in the areas of optical switching and networking, traffic modeling, and multilayer network design.

Fioravante Patrone received a degree (M.Sc.) in mathematics in 1974. Formerly full professor of Mathematical Analysis, is presently professor of Game Theory, Faculty of Engineering, University of Genoa (Italy). His recent research interests focus on applications of game theory to diverse fields: telecommunication, molecular biology, health care, natural resources. Author of more than 50 papers, he has been the promoter of the "Game Practice" meetings that started in 1998 and has served as Director of the Interuniversity Centre for Game Theory and Applications for many years. Member of the editorial board of international journals, has recently published (in Italian) an introductory book for Game Theory.

Guido Maier received his Laurea degree in electronic engineering and his Ph.D. degree in telecommunications in 1995 and 2000, respectively, both from the Politecnico di Milano, Italy. Through February 2006 he was a researcher and head of the Optical Networking Laboratory at CoreCom In March 2006 he joined Politecnico di Milano as an assistant professor. His main interests are optical network optimization, multidomain ASON/GMPLS, and photonic switching systems. He is the author of more than 70 papers in international journals and conferences in the area of optical networks. He has been or is involved in several research projects, including BONE, MUPBED, and NOBEL2.