Inter-domain Traffic Engineering

Manuel Seixas Pinto
Instituto Superior Tecnico
Lisbon, Portugal
email: manuel.s.pinto@tecnico.ulisboa.pt

Abstract—Traffic Engineering (TE) techniques for inbound traffic control can seriously threaten the scalability of the Internet’s global routing system. This work provides an in-depth study to TE practices with focus on the scalability problem by doing an experimental assessment, impact evaluation and subsequently present a mitigating solution.

We first provide the tools need for the study by surveying the publicly available data sources and create the basic algorithms. Then, we portray the evolutionary and current state of the Internet with a prefix characterization. Given its importance for future studies, a web platform was created in order for these and other interesting statistics to be publicly available.

The assessment study uses real BGP data in order to observe the prevalence of TE on the Internet for both IPv4 and IPv6. We try to guess on its trending and briefly compare both IP versions. TE impact is evaluated for global reachability, routing table sizes and the changes to the route type and path length. Finally, we provide solutions to lighten the damaging consequences of TE by adapting and applying the Distributed Route Aggregation (DRAGON) strategy to the Internets inferred topology with very promising results.

Index Terms—Traffic Engineering, Prefix deaggregation, BGP, Scalability.

I. INTRODUCTION

Traffic engineering (TE) refers to the actions taken in order to control inbound IP traffic according to technical and economical necessities. BGP offers ASes two techniques to control inbound traffic. One is AS-PATH prepending (PP) where, by injecting redundancy of nodes in the route’s AS-PATH parameter, it is able to make paths longer and by doing so decrease the order of preference of those routes. The other technique is prefix deaggregation (PD). This method is characterized by an AS dividing its assigned prefixes in longer, more specific, prefixes and advertising them to different subsets of providers. Due to the longest match prefix rule, ASes that selective announce more specific prefixes to different neighbors can control and steer incoming traffic. This combination of prefix deaggregation with selective advertisements brings numerous advantages to the deaggregator AS. For example, the traffic bill depends on the peak usage, by balancing the traffic load between providers the AS can benefit from significant reductions in the monthly bill [1].

Figure 1 shows examples of the two traffic engineering techniques. In fig.1 (a), AS3 deaggregates its assigned prefix $p$ in the more specific $p_0$ and $p_1$ which it selectively announces to different providers. With this, the AS is able to balance the load of incoming traffic. With PP, 1 (b), AS3 is able to make the route through AS2 preferable by increasing the path length in the announced route for AS1. AS4 learns both routes from customer links so it will elect the one with the shorter path length and traffic will flow through AS2.

The Internet routing system should be scalable in order to accommodate an healthy growth. With prefix deaggregation, an AS generates more specific prefixes that are propagated to the whole network, therefore are part of every ASes forwarding table. This poses as a serious threat to the scalability as more aggressive deaggregation starts filling routing tables and slowly degrades the Internet’s global performance.

Scalability is not only related to technology constraints. As different individual entities share the same business ecosystem, the practices within this network should be respectful of each one’s intentions or else its impossible to mutually cohabit.

Route aggregation refers to the practice of substituting a set of routes pertaining longer prefixes with a aggregating route for a less specific prefix. This is achieved by filtering, but it should be done in a careful way so that the filtered state is as consistent as possible with the standard state.

This work provides a study to the TE practices focusing on its damages to the scalability. First, we survey the available data sources for the study and compute statistics for the prefix type. After, we start evaluating TE practices by doing an experimental assessment, impact evaluation and subsequently present a mitigating solution.

A. Related Work

The research community has devoted significant efforts to the study of the global impact of address fragmentation. The negative impacts of an increasing number of announced prefixes have been known for long [2]. Studies show that the growth of BGP routing tables pose a serious threat for
scalability of the global system with its decreasing packet forwarding speed and demand for more memory space.

Despite the negative side to it, prefix deaggregation can bring serious rewards. Recent works study the economical incentives for prefix deaggregation. Lulu et al. [1] shows that prefix deaggregation can have the advantageous side effect of reducing traffic expenses. Deaggregated prefixes allow for a better control of inbound traffic lowering fluctuations that can cause peak traffic usage taxed at higher rates. Lutu et al. [3] also showed in a followup work that prefix deaggregation by a customer can negatively impact the business of the associated provider. Bangera et al. [4] analyses incentives of transit providers for deaggregating customer’s prefixes in order to attract more traffic.

In order to reduce the number of routes due to prefix deaggregation one can look at route aggregation strategies. The state of the art in terms of route-aggregation is DRAGON [5] (Distributed Route Aggregation on the Global Network). DRAGON is able to achieve a set of goals that distances it from previous approaches namely that it works with BGP, its incrementally deployable and provides algorithms for reaching route consistent states, all of this while being able to reduce the number of prefixes in each AS by up to 80%.

II. SNAPSHOT OF THE INTERNET

The aim of this chapter is to provide the necessary tools to guide us on the TE study. First, we will explore the characteristics of public data sources available, and how these can be used to serve our purposes. The next step is to compute some statistics to better understand the current state of the Internet, in particular its prefix type distribution. Finally, it is presented a web platform that automatically generates some of these measurements and it is publicly available so it can be used for consultation in future studies.

A. Data Sources

1) Route Monitors: BGP data can be collected with special purposed routers known as monitors/collectors. Monitors are routers that, by establishing a peering session with an AS, can extract the exchanged routes and BGP updates. This routing data is captured over time, usually multiple times a day, then compiled and made publicly available on the Internet. Monitors are not restricted to the collection of data from directly connected neighbors, instead they use multi-hop BGP sessions to peer with geographically dispersed ASes as though they were external neighbors.

The most popular sources for route data are the RouteViews [6] (RV) and RIPE-RIS [7] projects.

2) Looking glass servers: Another source for adjacency information comes from looking glass servers (LG). LG servers allow the remote execution of non-privileged routing commands like pinging a specific address. In general, these servers do not allow full BGP table dumps as some information is intentionally kept secret.

3) Internet Routing Registeries: A more reliable source for adjacency are the Internet Routing Registeries (IRR). IRRs are public databases where AS administrators manually register adjacency and policy information. The data is inserted using the Routing Policy Specification Language (RPSL) [11]. The problem with this databases is that they are used in a voluntary basis which may lead to inaccuracies or even intentional misinformation. From these, one of the more trustworthy source is the RIPE database.

4) Internet’s Inferred topology: One of the more demanded applications for the routing data is the inference of the Internet’s AS relationships. Research community has devoted significant efforts in the search for algorithms that better portray the Internet’s topology, the ones with better results.
are the datasets provided by UCLA [12] and CAIDA [13]. CAIDA’s algorithm was recently (2013) improved in a way that it outperforms all the existing datasets, including UCLA’s [14]. Both use BGP data collected from RouteViews and RIPE-RIS.

5) Prefix to AS mapping: In order to study the prefix distribution and better characterize how these are being generated one can associate the advertised prefix to the origin AS. These mappings can be easily extracted from the monitor’s BGP data by combining the PREFIX entry with the last AS (origin) on the ASPATH. CAIDA [15] provides this prefix to AS mapping using the RV monitors as source.

B. Statistics

1) Prefix type: Let us introduce classes of prefixes to help us with the characterization.

- Lonely: A prefix is not itself covered by a less specific nor covers more specific prefixes. (123.45.0.0/16)
- Top: A prefix that is not itself covered by a less specific but covers at least one more specific prefixs. (123.44.0.0/16)
- Deaggregated: A prefix that is covered by a less specific prefix, and this less specific is originated by the same AS as the deaggregated prefix. (123.44.0.0/24)
- Deaggregated: A prefix that is covered by a less specific, and this less specific is not originated by the same AS as the delegated prefix. (123.44.1.0/24)

Using this procedure we got the following distribution of prefixes by the different types.

![Prefix type distribution](image)

Lonely class take the majority of the announced prefixes with an even bigger portion on IPv6. Scalability threats are strongly related with deaggregation, the situation is more severe for IPv4 with 37% of deaggregated prefixes. At this point it is hard to compare these two protocols and jump to conclusions as IPv6 is in an early phase as IPv4 still announces 24 times more prefixes than its upgraded version.

C. InterSnap web platform

We provide a web platform to show some statistics from this work through time as multiple snapshots of the Internet. The project goals include automatic updates, appealing interface and room for future integrations. Taking all of this into account the InterSnap [16] platform was created.

![InterSnap web architecture](image)

The InterSnap web server uses open source software hosted on a free cloud platform. As support, it uses an SQL database with pre-processed data that is automatically updated every year by directly fetching the files from the CAIDA’s directories and applying the statistics algorithms. It currently provides statistics on the Address Space, Prefix type to both IP versions and Network information for IPv4.

III. EXPERIMENTAL ASSESSMENT OF TE

A. Assessment methodology

It is not possible to have a full view of each AS advertisement policies in order to observe how TE is being used. RIR registries are not mandatory and some special adjustments are key to the ASes business strategy so they are kept secret. However, every AS has to speak BGP so, it is possible to get an idea on how routes were advertised by doing an in-depth study on real BGP-route data collected from route monitors even if that means working with gigabytes of raw information. From the elected RIB entries we will analyze some of its route parameters, the PREFIX, FROM, and ASPATH.

1) Selective advertisements: When deaggregation is used without TE intention, all the prefixes are advertised to all providers in order for those links to be equally preferred. When this occurs, in every monitor, the elected route for all the deaggregated as well as the aggregated prefixes must display the same last-hop neighbor. The only case when this is not true is if some nodes apply different policies based on the prefixes and not only on the route attribute, which we believe to be an unlikely scenario.

In the case where deaggregation is combined with selective advertisements, routes for the advertised prefixes may follow different paths. The consequence is that in every monitor, the elected route for the deaggregated and selectively advertised prefix must display the same last-hop neighbor.
2) *Path Prepending*: Prepending can be easily identified by repetitions of the same AS in the AS-PATH attribute. In this study, prepending is only to be considered when performed by the originating AS or by the second *last-hop* neighbor, case where an AS uses the BGP communities for the provider to prepend its routes.

A combination of both techniques can also be used to have a finer control of inbound traffic, but when prepending is used, it is not possible to distinguish selective from unselective advertisements. The problem arises from the fact that when prepending is used the preference of routes are altered in a way that in some circumstances it may be confused with selective advertisement practices. We cannot differentiate the techniques used but we can conclude that the routes were deaggregated for traffic engineering purposes.

B. Procedure

Table III-B summarizes the conditions used to reason on the TE practices for the whole prefix family, the aggregated and its deaggregated children. On the other hand, prepending is extracted per-prefix.

<table>
<thead>
<tr>
<th>Selective</th>
<th>Prepending</th>
<th>Each Monitor</th>
<th>Across Monitors</th>
</tr>
</thead>
<tbody>
<tr>
<td>NO</td>
<td>NO</td>
<td>all</td>
<td>Prop</td>
</tr>
<tr>
<td>YES</td>
<td>NO</td>
<td>&gt;=1</td>
<td>no</td>
</tr>
<tr>
<td>-</td>
<td>YES</td>
<td>yes</td>
<td></td>
</tr>
</tbody>
</table>

**TABLE I**  

**TE EVALUATION PROCEDURE**

C. Statistics and Analysis

1) *Path Prepending*: Path prepping takes on a significant portion in today’s TE practices as we were able to observe prepending in nearly half of the announced prefixes (41.38%) for IPv4 and a significantly less, 19.96%, for IPv6. In order to better understand where this percentages come from and observe its evolution in time let us take a look at figure 7.

We see that most of the prepended prefixes come from the lonely class in both IP versions. Lonely prefixes do not share address space with any other prefixes, in order to control traffic for this space, one can only prepend. From the deaggregated prefixes, in 2015 for IPv4, 41.7% of these (15.25% of total) show prepending. This can lead us to believe that deaggregation and prepending are significantly used in combination.

To what the evolution is concerned, both IP versions show an increase in the amount of total prepending where it mainly grows in the deaggregated and lonely class.

2) *Prefix Deaggregation*: How much of prefix deaggregation is used with TE intention? For each aggregated prefix with some other deaggregated child belonging to the same AS, table III-C2 shows how these family of prefixes can be attributed to TE.

For IPv4, there is a strong (79.59%) correlation between deaggregation and TE, suggesting that TE poses as a strong incentive for deaggregation. This relationship is is not so evident in IPv6 but still accounts for 63.7%.

We observed the evolution of deaggregated prefixes as displayed on the charts of figure 8.

Both IP versions show increasing deaggregated routes. Deaggregation, widely used in IPv4, seems to be spreading its popularity to IPv6, but the alarming rate at which it is growing poses as a serious threat and shows the lack of global awareness of the individual Internet’s entities.

![Fig. 7. Prefix type with prepending distribution](image)

![Fig. 8. Evolution of prefix deaggregation](image)
IV. IMPACT OF TE ON GLOBAL ROUTING

In this impact analysis, it is easier to to shift the notation that represents the Internet as a modeled graph where ASes are represented by nodes and the edges by the links connecting ASes, either customer-to-provider or peer-to-peer.

A. Impact analysis

1) Distortion to the route type: The route type could be customer, peer or provider based on the relationship with the AS from where the route was learned from. Different types have different orders of preference as an AS favors sending traffic to a customer where it has associated revenues than to pay a provider to reach a specific destination. These relationships also shape how the learned routes are exported following the GR model. For a given prefix, Path Prepending uses injections in the route ASPATH attribute to adjust the attractiveness of a provider link regarding inbound traffic. In relation to the no-TE situation, this change is felt in the route election process after the route type is evaluated when the modified path length is used to help reason on the most preferred from routes of the same type. In conclusion PP does not add distortion to the route type.

On the other hand, prefix deaggregation announces different, more specific prefixes to disjoint subsets of providers. One consequence is that routes flow through distinct paths and may arrive at nodes from links which have different associated route types. If this change is very significant, it can seriously impact the global business relationships. A practice that is beneficial to one AS can globally have a negative impact and therefore create a serious threat to scalability. To demonstrate the consequences of this action, let us take a look at the sample network depicted in figure 9.

<table>
<thead>
<tr>
<th>Without PD</th>
<th>With PD</th>
</tr>
</thead>
<tbody>
<tr>
<td>2RIB</td>
<td>2RIB</td>
</tr>
<tr>
<td>u1</td>
<td>u2</td>
</tr>
<tr>
<td>p</td>
<td>p0</td>
</tr>
<tr>
<td>u3</td>
<td>u4</td>
</tr>
<tr>
<td>p</td>
<td>p1</td>
</tr>
<tr>
<td>u5</td>
<td>u6</td>
</tr>
</tbody>
</table>

Fig. 9. Change in route type when using prefix deaggregation

In this simple network, node u4 deaggregates prefix p into p0 and p1 which it selectively advertises to different providers u2 and u3 in order to balance incoming traffic. For reliability reasons we will study in further sections, prefix p is also advertised to both providers. The situation with selective advertisements is depicted on the right side of the figure where straight arrows also show the flow of traffic pertaining prefix p1. In contrast, the left side shows the no-TE situation with only p being advertised.

Let us take a closer look at the differences. The network is policy connected, each node learns a route for every advertised prefix. In the TE situation, due to the longest match prefix rule, packets destined to p’s address space are routed to the forwarding neighbors of the p0/p1-route. Focusing on the expedition of packets for p1, node u2 has a provider route through u1 even though there is a usable direct customer link from u2 to u4. Node u2 now has to pay for traffic that would otherwise flow directly to its customer in the no-TE situation. Also, u2 even has a preferred customer p-route that it cannot use unless p1 is filtered. In order to satisfy u4’s intention to balance traffic, u2 suffered negative impact by the change in the route type from customer to provider regarding p1’s address space.

Considering that with PD we are restricting the paths of a prefix, each AS will have fewer route options in the RIB entry from which to elect the preferred route so PD can only worsen the type.

2) Distortion to the path length: Since the Internet is a vastly dense network, a data-packet can travel to any destination in the world having to transverse only through a small number of exchange points. Many services have strict Quality of Service, QoS, requirements that can be affected by the enlargement of paths.

For example, Voice over IP, VoIP communications are very sensitive to delay. The transit delay for the communication is a combination of several factors, but the biggest slice comes from the delay introduced by routers and switches in the path [17]. Each intermediate stop has an associated switching delay for table look up and also buffering delay because packets must wait in line before being processed.

The purpose of this part is to study the impact of TE on the change of the route’s path length. So far, TE was always seen as profitable practices an AS engages in that can have negative impact to the global system, here we are examining a situation that hurts the origin AS the most. To what QoS is concerned, because many factors contribute to the quality of a connection and have different significances depending on the nature of the service, a longer path doesn’t always mean that it is a worse path. The aim of this analysis is not to go any further and test specific QoS requirements, but to give a general understanding that they might be negatively affected.

In the route selection process, learned routes of the same route type use the ASPATH attribute to choose the most preferred route. Path prepending injects ASes in order to make the path longer and therefore less appealing. Note that the path is not longer in the sense that it travels through more nodes before arriving to the destination, the process blindly evaluates the path length and not the number of hops that take up the real length. To avoid any confusion we will refer path length as the number of ASes on the ASPATH attribute and use real length to the distinct number of ASes in the path attribute. For example, the ASPATH composed of AS1-AS1-AS2-AS3 is of path length 4, but the real length is only 3. In this way, prepending may change the preference of routes and a node can elect routes with larger real length thinking that it is electing a shorter path. Figure 10 portrays this situation.

Node u6 prepends 4 times its ASN in the path attribute before advertising to provider u4 in order to discourage traffic on this link. Node u1 receives two customer routes for the same prefix and is therefore left with the evaluation of the path length. The route from u4 is of path length 5 while the
one from $u_2$ is 4 thus $u_1$ will choose $u_2$’s learned route even though it is of longer real length.

There are no general guidelines for applying PP as it largely depends on the topology and the operator’s goals for a particular situation. That is why PP is usually executed in a succession of trials in an optimization process. Given the particularities of each case, it is not possible to conceive a simulation test that represents a realistic situation in order to measure the impact and extract relevant conclusions. However, we can still conclude that PP can only worsen the path length for a given prefix. When no PP is used, the elected route is the one with minimal real length. On the other hand, with PP, if the route changes, we know that the election process was tricked into choosing a route of equal or worse real length.

In the case of **prefix deaggregation**, we have already observed that the selective advertisements can significantly change the elected route at a node. In extreme cases, it can even change the route type, so it comes as no surprise that it also has an impact on the path length. Comparing to the no-TE situation, if the route type is the same, the path is of equal or worse length. The explanation for this is because in the no-TE situation, all routes of the same type were already evaluated and the one with shorter length was chosen. On the other hand, if the route type changes, the path length could be longer, equal or even shorter. Since in this case the route arrives via a different incoming link, it travels through a distinct path and as such it is not possible to find a relationship between situations and predict the outcome of the path length. Although the path could be shorter, because the route type is worse this is usually seen as an unfavorable situation. Figure 11 portrays an example of this two cases.

3) **Global Reachability:** One of the major concerns when applying TE is the fear that it may prevent routes from reaching portions of the internet. Before any traffic control, the network operator is interested in assuring that global reachability is satisfied so it is of most importance to understand in what circumstances TE can prevent this from happening.

For a given prefix, **Path Prepending** uses injections in the route ASPATH attribute to adjust the attractiveness of an incoming link. The different prepended routes are advertised to all providers. In relation to the no-TE situation, this change is only felt in the route election process after LOCALPREF attribute is evaluated. This means that the changed routes do not completely erase connections that may prevent global reachability. A link that is very unappealing is still a valid link. As such, PP will not change connectivity compared to the no-TE situation.

**Prefix deaggregation** is usually deployed by injecting different, more specific prefix routes to different provider links. One direct implication is that routes flow through distinct paths that may reach different sets of nodes in a network preventing global reachability. When the network is policy connected, this is not an issue as every route announced to whichever node reaches the remaining nodes. Even though we not advertising the prefixes to all providers, they will eventually receive the remaining routes.

To what global connectivity is concerned, PD does not diminish connectivity as long as a more specific covering the deaggregated prefixes is also advertised to all providers. This is also true for the filtering situations explained before, busyness dynamics may change but connectivity is assured.

4) **FIB & RIB - Table size:** All of the aforementioned aspects rely on the graph dynamics but nothing is more threatening to the scalability than exceeding the physical constrains of the technology. Forwarding tables do not have infinite capacity and take some time to process each entry that can be proportional to the number of entries. So, the undisciplined increase in number of prefixes and routes can seriously lead to a degradation in performance, routing instability, or impact availability of the global routing system.

FIB table size recently reached an important milestone. Some of the old deployed routers had a memory limit of 512K prefixes, a boundary that was exceeded in May of 2014 creating partial outages on the Internet. Although some workarounds are possible to extend capacity these are not responsible practices in the long run. This section is focused on understanding how TE can impact FIB & RIB table size.

Modified routes for **path prepending** pertain to a single prefix, which means they do not introduce a new entry in the forwarding table nor increase the amount of advertised routes. Although different-in-length routes are announced, each AS only advertises one elected route. On the other hand, **prefix deaggregation** injects more specific prefixes to the network meaning that it takes space from the FIB as well as the RIB. The severity of the impact depends on the deaggregation factor.

From this we see that PD is much more damaging to the scalability of the Internet. The fact that PD gives a better control over inbound traffic than PP, implicates that these
techniques cannot be substituted to achieve global scalability gains.

B. Statistics

1) Change in route type: Prefix deaggregation can impact the route type of ASes from the advertisement of a prefix to a subset of providers instead of using all links. When a prefix is selectively advertised, some of the nodes might experience a degradation on the type of route. Here we wish to reason on the change of routes that may negatively impact the global system.

Candidate ASes to preform PD are those with more than one provider. For each PD-AS as destination, we compute the elected route on all ASes using every provider link, this corresponds to the best route estimation and also the no-TE situation we want to compare. To simulate prefix deaggregation, let us ignore all the provider links except for one. We then compute the elected route and compare the differences. The process is repeated for each provider link and the differences are finally averaged.

Our simulation found that more than half (68.5%) of the ASes that deaggregated prefixes for TE purposes experienced global changes in the route type. Although this percentage is high, the changes on the other hand are not so severe. We already know that the estimation can only worsen the route type. Although this percentage is particularly deficient on estimating peering links. However, it is not possible to guess how it influences the results. The diminishing of nodes with peer routes through the hidden providers could be balanced by new peer links connecting these nodes to the selected provider lowering the overall impact. This could be seen as an incentive to establish more peering links, explaining the increasing popularity of IXPs.

2) Changes in path length: In order to evaluate the changes in path length, we used a procedure similar to the one for the route type. The changes in length are computed for each announced provider and compared to the advertisement on all links. Since the sign of the length difference is related to very different situations, averaging the results for each AS could produce wrongful conclusions so, as opposed to the route type simulation, each change is computed in relation to a provider link. This also gives us the results for the worst cases.

The path length can be extracted from the ASPATH attribute in the RIB announcements explored in chapter III. Results from these two approaches are depicted on figure 14. The path length explanation could be due to the great number of peering relationships on the internet, these can become unavailable once the route is selectively advertised. This precise case may slightly differ from reality since inferred topologies are particularly deficient on estimating peering links. However, it is not possible to guess how it influences the results. The diminishing of nodes with peer routes through the hidden providers could be balanced by new peer links connecting these nodes to the selected provider lowering the overall impact. This could be seen as an incentive to establish more peering links, explaining the increasing popularity of IXPs.

More than half of the routes (61.6%) are the same in length. The enlargement of paths which has negative length associated impact take on 30.1% of the provider links as opposed to the less 8.3% shorter paths with negative route type impact. The change is small and is mostly contained in the interval of [-2,+2] although in extreme cases it can reach the 12 hop difference.

Fig. 12. Cumulative distribution of deaggregating ASes with global impact for R2P route changes.

Fig. 13. Cumulative distribution of deaggregating ASes with global impact for R2P route changes.

Fig. 14. Change in path length from the use of TE.

The worst case scenario, where nodes change the type from customer to provider only affects 2 ASes (0.004%) for more than half (56.23%) of the deaggregators. The change from customer to a peer route has similar results with slightly higher impact.

From the chart analysis, one observation is that the impact is felt more intensively on the changes from peer to provider. Still, half of the deaggregators only impact 634 (1.27%) ASes but some may reach up to 5116 (10.21%) ASes. One explanation could be due to the great number of peering relationships on the internet, these can become unavailable once the route is selectively advertised. This precise case may slightly differ from reality since inferred topologies are particularly deficient on estimating peering links. However, it is not possible to guess how it influences the results. The diminishing of nodes with peer routes through the hidden providers could be balanced by new peer links connecting these nodes to the selected provider lowering the overall impact. This could be seen as an incentive to establish more peering links, explaining the increasing popularity of IXPs.

The path length can be extracted from the ASPATH attribute in the RIB announcements explored in chapter III. Results from these two approaches are depicted on figure 14.

Fig. 12. Cumulative distribution of deaggregating ASes with global impact for C2P and C2R route changes.

Fig. 13. Cumulative distribution of deaggregating ASes with global impact for R2P route changes.

Fig. 14. Change in path length from the use of TE.

More than half of the routes (61.6%) are the same in length. The enlargement of paths which has negative length associated impact take on 30.1% of the provider links as opposed to the less 8.3% shorter paths with negative route type impact. The change is small and is mostly contained in the interval of [-2,+2] although in extreme cases it can reach the 12 hop difference.

Comparing our simulated findings with the ones extracted from the RIB announcements we see that the results are very
close. Our simulation portrays a situation where every AS announced a different prefix to each of its providers, a very pessimistic scene which can explain the lower impact on the observed results.

V. Solution for Scalable TE

A scalable approach to this problem should be focused on global gains and preferably reach a solution with minimum changes from today’s deployed technologies in order to facilitate adoption. Most academic findings for FIB [18], [19] and RIB [20] reduction rely on local software optimizations to the table entries that only contribute to improve individual ASes and not the global system. Many clean slate approaches aim to redesign the Internet’s architecture and are therefore difficult to deploy without major changes which may not be feasible in the near future.

Inter-domain route aggregation strategies are of special interest as filtering gains propagate through the Internet. For this approach we will consider DRAGON as it is TE compatible. With this in mind, we will explore DRAGON with TE in a way that it is able to mitigate the negative impact of prefix deaggregation and therefore pose as a scalable solution. We will also evaluate the side-effects that may arise from the use of this approach, hoping that they seem insignificant given the global benefits that pave way for a healthy Internet’s future.

A. Scalable TE using DRAGON

Dragon proposes a route-aggregation solution for inter-domain routing. Its mechanisms rely on two basic rules:

- **Filtering code:** CR: If a node is not the origin of \( p \) and if attribute for elected \( q \)-route equals or is less preferred than the attribute of the elected \( p \)-route → filter \( q \)
- **Rule of announcement:** RA: The origin of \( p \) announces \( p \) with a route whose attribute is equal or less preferred than the attribute of the elected \( q \)-route.

Applying this algorithm guarantees an optimal route-consistent state for GR export policies. However, when using TE in the form of PD, DRAGON does not directly work as the \( q \) and \( p \) routes share the same attribute and can therefore be filtered. In order to understand the implications let us look at figure 15 that portrays a sample network and the standard stable state for \( p0 \).

![Fig. 15. Sample network to perform DRAGON](image)

In this Figure, node \( u_7 \) deaggregates prefix \( p \) into \( p0 \) and \( p1 \). Node \( u_7 \)'s providers, \( u_4 \) and \( u_5 \), elect a customer \( p \)-route. Executing DRAGON, CR filters the deaggregated prefixes because they are of a less or equally preferred route compared to the \( p \)-route. In this way, \( p0/p1 \) are not propagated to the network and \( u_7 \) is unable to engineer its incoming traffic despite the global filtering gains.

DRAGON can still be used with prefix deaggregation by an adjustment to reach an equivalent model. Another approach is to delegate the use of filtering to upstream providers.

**Provider's settlement (PS)**

If a customer deaggregating node, \( u_d \), wants to perform PD for a prefix \( p \) containing in space \( n \) deaggregated prefixes, \( q_n \), each provider of \( u_n \) should obey the following rules in order to reach an equivalent DRAGON model to respect \( u_d \)'s TE intention.

1) **Announce \( p \)** according to RA, i.e. with a route whose attribute is equally or less preferred than the elected \( q_n \) routes (\( p0/p1 \)-route in the example).

2) **Prefer \( p \)**-routes learned from other than \( u_d \) (\( u_7 \)).

![Fig. 16. Execution of DRAGON. (a) with Aggregation node, (b) without Aggregation node.](image)

To see how this works let us look at the following figure 16-(a). Node \( u_4 \), provider of \( u_7 \), elects both a customer \( p0 \)-route as well as a provider \( p1 \)-route. In order to obey rule 1 of the PS, \( p \) is announced as a provider route thus only to \( u_4 \)'s customers. Node \( u_6 \) learns a provider \( p/p0/p1 \)-route, executes CR and filters \( p0/p1 \). Globally this has little to no effect but note that node \( u_1 \) learns a customer \( p0/p1 \)-route. For that reason, \( u_1 \) could originate aggregation prefix \( p \) as a customer route, therefore announcing \( p \) to all neighbors. Node \( u_5 \) follows rule 2 of PS and prefers \( u_1 \)'s \( p \)-route. Now, there are more filtering opportunities and the final stable state is reached (fig. 16 right).

Comparing figure 16-(a)-right with figure 15 we see that not only is \( u_7 \)'s TE intention respected but also that the final state is route-consistent. DRAGON guarantees that route-consistency is met (for GR policies), so by applying a model that falls into DRAGON’s conditions it also ensures consistency to the final state.

An aggregation node, AN, is a node that elects a customer route for the selectively advertised deaggregated prefixes. Given the dense Internet graph, many nodes can fall into this category, for example a provider of an AN will also be an AN as customer routes are exported upstream. For greater global gains, we are interested in aggregating the closest as possible to the origin AS so that the aggregated route reaches more nodes that do not need to keep table entries for the deaggregated prefixes in order for those to be kept only within
a small vicinity of the origin. Note that closest is in relation to the GR policies and not necessary the shortest path. What this means is that the closest AN is the one where it is impossible to reach another AN following only customer routes from this node to the origin.

In order to discover if a node is an AN, one can simulate selective advertisements by computing the elected routes taking into consideration only one provider of the origin AS for each of the providers. If a node has a customer route for every of these computed routes then it is an AN (function isAN()). A provider of an AN is also an AN, as customer routes are exported to the providers so each AN will have a customer AN unless it is the first to aggregated routes from different customers in which case it is one of the shortest aggregation node. So, the algorithm for computing the shortest AN starts at an arbitrary candidate AN and explores its customers. If at least one of the customers is an AN, it will further evaluate this customer node and repeat the process. Has we keep on doing this, we are moving towards the destination following customer links until the shortest AN is found. For our study we are only interested in finding one shortest AN but multiple might exist.

Without aggregation node

There are cases where there is no upstream provider capable of aggregating routes, situation represented in fig. 16-(b) where link \( \{u_1,u_3\} \) was removed, incapacitating \( u_1 \) from generating aggregation prefix \( p \). In situation (a), node \( u_1 \) could announce \( p \) as a customer route in accordance to RA. Here, \( p \) could also be advertised but only as a peer or provider route meaning that it is exported only to the customers. If every node prefers a \( p \) route learned from a neighbor instead of being the one to generate \( p \), tier-1 nodes receiving customer routes for the deaggregated prefixes will be the ones to generate \( p \) but only to its customers. This creates filtering opportunities while maintaining route-consistency.

1) Robustness to link failures: Multi-homing could be used for load balancing but its main purpose is to increase reliability by ensuring that the node stays connected to the Internet even if there is a failure in a provider link. Let us see what happens to the previous network when the two-way link \( \{u_4,u_7\} \) fails.

Looking at the left of figure 17, prefix \( p0 \) stops being advertised, thus node \( u_1 \) cannot originate aggregation prefix \( p \). Node \( u_5 \) now has only a customer \( p1 \)-route so it can now announce \( p \) with customer route attribute exporting \( p \) to all neighbors. In this way, the stable state is reached and packets destined to the \( p0 \) space arrive at \( u_7 \) through \( u_5 \), maintaining global reachability.

This equivalent model offers great advantages to the global system as it creates many filtering opportunities while maintaining route consistency. The major downside comes from the fact that its not a completely distributed algorithm, as there needs to be a settlement with the providers in order to behave with special conditions and also some coordination from the all the nodes to advertise the aggregation prefix.

### B. Statistics and analysis

1) Equivalent model: In order to evaluate the filtering gains from applying DRAGON the \( FIB_{TE−gain} \) was defined as the relative difference on the FIB entries for a prefix and its deaggregated child.

\[
FIB_{TE−gain} = \frac{FIB_{initial} − FIB_{after}}{FIB_{initial}}
\]

For example, a multi-homed AS with 2 providers deaggregates its assigned prefix \( p \) into \( p0 \) and \( p1 \) which it selectively advertises to each provider. Without filtering, each AS will have a FIB entry for all the advertised prefixes \( p/p0/p1 \). When DRAGON is deployed, an AS that is able to filter \( p0 \) but not \( p1 \) will have FIB entries only for \( p/p1 \) resulting in a \( FIB_{TE−gain} \) of \( (3-2)/3 \) (33%).

Figure 18 computes the gains for the common situation of a multi-homed AS with 2 providers where a prefix is broke in two and these are selectively advertised to the different providers. Table V-B1 shows the average results for this case.

<table>
<thead>
<tr>
<th>%gain</th>
<th>0%</th>
<th>33%</th>
<th>67%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average %ASes</td>
<td>50%</td>
<td>3.09%</td>
<td>96.41%</td>
</tr>
</tbody>
</table>

Fig. 18. \( FIB_{TE−gain} \) for 2 providers. Deaggregating ASes are organized from the less to the better gains

| Average percentage of affected ASes for different gains in a situation of 2 TE providers |

Observing the previous results we see tremendous global gains. On average, almost all of the ASes (96%) only need
to keep a FIB entry for the aggregated prefix while maintaining route-consistency throughout the Internet. Only a small percentage (0.5%) have no gain on average.

This results are a huge improvement from today’s situation and it will hopefully be seen with a globally aware mindset to be a strong incentive for deployment. The gains are not only felt on the FIB entries but also RIB and reduced route updates. Small changes to the routing system that build big global gains, therefore creating room for a scalable Internet system.

VI. CONCLUSIONS AND FUTURE WORK

We showed great correlation between prefix deaggregation and traffic engineering. Also we observed an upward trend towards deaggregation throughout the years for both IP versions. The dramatic rate in which deaggregation is growing for IPv6 shows that if the lack of global awareness stays the same, IPv6 could face an even greater problem.

The two TE practices in study proved to have very different significances to the global impact. Path prepending is almost harmless apart from possibly increasing the number of transversed hops. Prefix deaggregation on the other hand is the core of the problem. For one, it can negatively impact the business relationships of distant ASes although the extent in which it does was found to be not so severe. The distortion to the route type mostly impacts changes from a peer to a provider route and when this happens, the worst case only affects 10.21% of the total ASes. We even showed that if deaggregation is not performed in a careful way, by also advertising as backup the aggregated prefix to all providers it could decrease reachability in extreme cases.

The most damaging consequence of PD is the bloating of forwarding tables. With this in mind, we showed that it is possible to achieve traffic engineering without overloading the routing system. By adapting the DRAGON aggregation strategy, one of the proposed solutions was not only capable of filtering most of the deaggregated prefixes but did so without compromising the global business relationships by maintaining a route consistent state. On average, 96.41% of the ASes were able to fully filter the deaggregated prefixes for the most common case.

We hope that this study at least served as a wake up call to the alarming consequences of undisciplined TE to the scalability. It is becoming increasingly difficult to find reasons to dismiss route aggregation strategies. Its global advantages far surpass the needed changes with deployment and should serve as sufficient incentive for adoption.

In the future, the web platform could be further improved to incorporate new measurements in order to serve a more complete range of studies. Also, we saw an increasing number of projects that aim to provide new kinds of Internet monitoring data. This data sources could be explored in future studies to maybe overcome the limitations of todays publicly available data sources.

As IPv6 keeps on maturing, soon academic studies will shift the attention from IPv4, as consequence we believe there will be IPv6 inferred topologies in a near future. Although both IP version share many features it is impossible to know if they will evolve in the same way so the results from this study that are valid for IPv4 may need a reassessment for IPv6. Although it may seem that IPv6 is far from its end, the sooner we start addressing scalability issues, the less likely they will become a serious problem.

REFERENCES