Scalable BGP prefix selection for scalable Interdomain TE

Wenqin Shao, Luigi Iannone
and Jean-Louis Rougier
Telecom ParisTech
23, Avenue d'Italie
Paris, France
{wenqin.shao, luigi.iannone, rougier}@telecom-paristech.fr

François Devienne
and Mateusz Viste
BORDER 6
101, Avenue du Général Leclerc
Paris, France
{francois.devienne, mateusz.viste}@border6.com

***Abstract*—Inter-domain Traffic Engineering (TE) in a multi-connected network faces a scalability challenge as the the size of global routing table (RIB) continue to grow at high pace. However, some routes/prefixes are responsible for much more traffic than others, which is known as the highly uneven internet traffic distribution. Therefore, a natural consequence is to perform TE only for those prefixes that matter most.**

However, traffic volume associated to a prefix can vary greatly over time. And, to our knowledge, we have little knowledge on the dynamism of traffic across BGP (Broad Gateway Protocol) prefixes. Moreover, predicting which prefixes will be the most significant is a complex issue, since sophisticated methods for predicting volume for each single prefix won't scale in this context.

In this paper, we show the relationships among prefix volume importance, stability and predictability based on recent real traffic traces from 9 networks of diverse profiles. With these observations, we propose three simple and resource-efficient metrics to select the forseeable prefixes of important volume. The proposed metrics yielded both satisfying volume coverage and pretty low prefix churn.Finally, once the prefixes are selected, the best route must be selected. Our measurements show that the performance in terms of RTT could differ a lot among different transit providers, which calls for fine-grained dynamic route selection mechanism. The route selection algorithm simulated in the work outperformed the best available transit by 20% on certain networks.

I. Introduction

Bringing enhanced connection reliability, reduced cost and improved transmission performance, multihoming has been long time practiced by large networks, such as enterprise networks, ISPs, Content Delivery Networks (CDN), etc. Smaller networks like residence networks and SOHO (Small Office Home Office) networks, though not allocated an ASN (Autonomous System Number), begin as well to use multiple WAN connections with the fall of Internet access price and growth in set-top box technology. In both cases, the local network is exposed with several next-hops in reaching a certain distant destination. Thus, one has to decide for each destination which next-hop to use so that transmission performance optimized and cost saved.

One major challenge in making Traffic Engineering (TE) decisions described above is to cope with the rapidly growing Internet routing table (RIB) [1]. To alleviate the burden caused by this RIB growth, forwarding table (FIB) caching, i.e. selectively install a smaller part of RIB on FIB, has been studied [2]–[7]. Same as the forwarding plane, TE is exposed as well to this continuous growth. Moreover, TE is associated with more complicated and resource consuming operations, such as performance monitoring, billing, route calculation, etc. Therefore, TE in the default-free routing zone faces as much scalability challenges vis-à-vis RIB size, which lacks the amount of attention that it deserves.

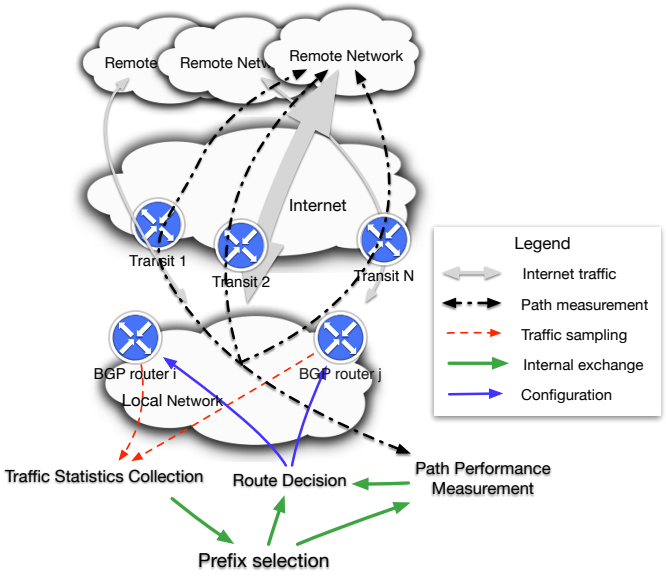


Fig. 1: Scalable TE with prefix selection, where path performance measurement and route decision are only conducted for selected remote prefixes with important traffic volume.

In this work, we propose improving the scalability of fine-grained inter-domain TE in the context of BGP, by concentrating on the most important BGP prefixes. This

is achieved by predicting which prefixes will correspond to the most important traffic volumes in the near future.

Fig. 1 gives a global picture of the different tasks involved in Traffic Engineering optimization of a multi-homed site. Traffic volume for each prefix is obtained from measurements at the border routers. The prefix selection is then performed based on these past measures. Finally, TE operations are only conducted for these significant prefixes, for the sake of scalability: first some paths characteristics (for each available path towards a given prefix) are obtained; then the best path is selected and enforced by means of configuration of the border routers.

The justification for this idea is based on the well known fact that the distribution of Internet traffic is highly skewed on multiple traffic aggregation levels and over multiple time spans [6], [8]–[10]. Yet, there are still challenges in fulfilling the goal. First to come is the scalability issue. A natural approach to predict prefixes with important traffic volume is to anticipate the traffic volume variation of each individual prefix, and then select top ranking ones according to the prediction values. Well established Time Series Forecasting (TSF) models and artificial neural networks (ANN) have been used previously in traffic prediction [10]–[12].

They all targeted on highly aggregated inter-PoP (Point of Presence) traffic for off-line tasks, such as dimensioning and long term MPLS-TE. These models are not only computationally heavy but also require pre-treatment and tuning to fit well to each individual trace, which makes them less applicable in the context of inter-domain TE, where more than 10^5 prefixes exist. Therefore, less complex prediction methods that work sufficiently well are needed. Second, the dynamism of traffic variation makes prediction difficult. It has been shown by Walleriche et al. [13] that the bandwidth ranking of a 5-tuple flow can change drastically from one moment to another. This indicates that the ensemble of flows that bring important traffic volume could be very different at different times. When it comes to traffic on BGP prefixes, no study, to our best knowledge, has so far given an in-depth investigation concerning its volume evolution in time.

The contribution of our work in addressing the challenges of predictive prefix selection lies in:

- We studied working traffic traces from networks of diverse profiles: ISP, content provider (CP) and hosting provider (HP). These networks locates in different countries of the world: USA, UK, Germany, Poland, Spain and France.
- A through investigation on the Internet traffic dynamism across BGP prefixes was conducted on these traffic traces which demonstrate different dynamism characters.
- Three scalable prediction metrics were designed based on the findings concerning traffic dynamism. These metrics are evaluated and compared to methods previously used in predictive FIB caching. The results show that methods proposed in our work outperformed previous models in both performance and scalability aspects.
- The performance of transit providers of each network is evaluated by probing selected prefix, which shows first there could be big difference among transit providers, and second, as a consequence, the potential gain of inter-domain TE is considerable.

The rest of the article is organized as follows: in § II, an in-depth study on the dynamism of traffic over BGP prefixes is given, based on which, three selection metrics are proposed and evaluated in § III; § IV evaluates the performance of different transit providers perceived by prefixes predictively selected; § V simulates the performance gain of a dynamic route selection algorithm based on RTT (Round-Trip Time) measurements toward selected prefixes; § VI concludes our results while comparing it to previous works; finally, we highlight the future directions in § VII.

II. Characters of Internet traffic over BGP prefixes

A. Data set

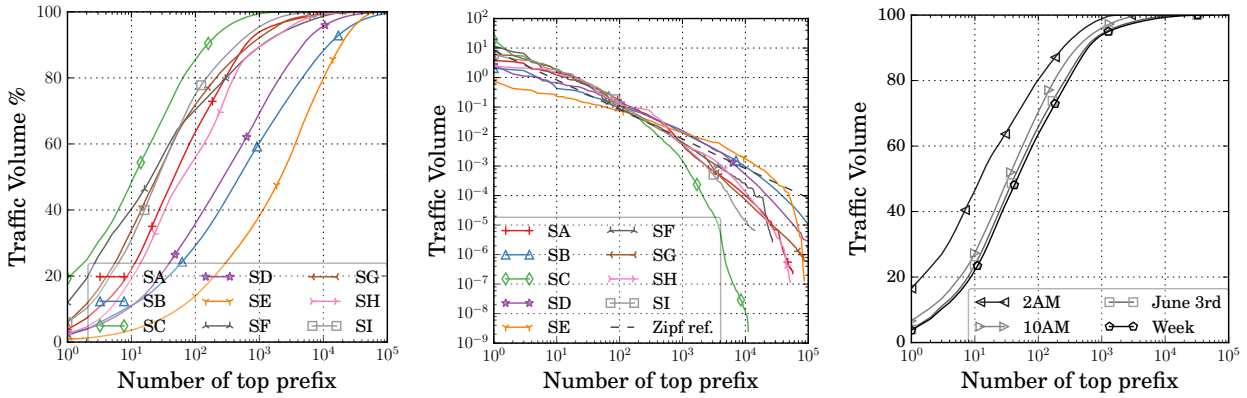
Name	Type	Hour volume (GB)	Sampling rate
SA	CP	133.49	1:1024
SB	ISP	528.43	1:512
SC	HP	6.63	?
SD	HP	1128.96	1:4096
SE	CP	1871.25	1:8192
SF	CP	5.09	1:512
SG	ISP	0.18	1:1
SH	CP	29.94	1:128
SI	HP	6.21	1:1024

TABLE I: Network profiles. Average hour volume over the week from June 1st, 2015 is given alongside with traffic sampling rate.

We base our study on working traffic traces collected from 9 networks of very different profiles listed in Table I. Traffic trace from each network covers a time span of two entire weeks, from May 25th, 2015 to June 8th, 2015, except for SB and SD, the traces of which only cover the second week starting from June 1st, 2015. These traces were sampled from real traffic, same practice as in previous works on FIB caching [4], [5]. It has been shown that bias introduced by sampling is negligible in this kind of use cases.

We obtained the out-bound traffic volume count over real BGP prefixes by mapping the sampled traffic to the longest prefix in BGP FIB of the corresponding network. Therefore, BGP prefix volume in our work actually indicates the traffic volume associated to a certain BGP route in use. For example, we have following routes/prefixes and volume counts:

- `10.0.0.0/24`, 10GB,
- `10.2.0.0/16`, 5GB,



(a) Week volume share, CDF. (b) Week volume share, log-log. (c) CDF of different time scales, SA.

Fig. 2: Traffic distribution among BGP prefixes.

- `10.0.0.0/12`, 3GB,
- `0.0.0.0/0`, 0.5GB,

where the prefixes are overlapped, which is likely the case in real BGP FIB. IPs in `10.0.0.0/24` are responsible for 10GB of traffic. 5GB traffic is sent to IPs in `10.2.0.0/16` but outside the `10.0.0.0/24`. The same is with `0.0.0.0/0`, which doesn't mean the entire Internet traffic but rather the traffic forwarded by default route. Bigger the volume associated to a prefix (route entry), more important it is, and hence worthy of attention to traffic engineer the next-hop for this prefix (route entry).

Further, traffic volume count for each prefixes is accumulated over 1 hour period. In comparison, previous studies concerning FIB caching [5], [6] group volume count over much shorter time bins, from 1 second to 10 minutes, to capture instant changes of traffic variation. In our case, we argue that 1 hour is an appropriate update interval for performance monitoring system to keep long term evolution records and to avoid unnecessarily frequent route changes which might lead to packet reordering and even route convergence issues in inter-domain routing with BGP.

B. Traffic distribution over BGP prefixes

Basing on the Zipf property of Internet traffic, we are able to control a majority of traffic by leveraging a small number of BGP prefixes. We examine in this section that our traffic traces demonstrate this property of concentration at different time scales.

Fig. 2a plots the cumulative week volume fraction over the week starting from June 1st. Except for SB, SD and SE, top 1000 prefixes cover more than 80% of week total traffic. Fig. 2b shows the traffic volume fraction associated to each BGP prefix over the same week. Prefixes are decreasingly sorted along the X-axis according to their cumulative volume fraction over the week. We observe that the week volume share of BGP prefixes on these networks can be approximately described by a reference Zipf's distribution with $N = 10^5, s = 1$ (dashed line)[1]. Fig. 2c compares the traffic distribution over different time periods for SA. The week line corresponds to the same week in two previous sub-figures. On top of that, traffic distributions over the entire day of June 3rd, that over one hour starting from 2AM (non-peak hour) and 10AM (peak hour) of the same day are also given. All these graphs confirm, as expected, that Internet traffic are highly concentrated on a few prefixes over multiple different periods of time, even after years' of rapid RIB increase. Furthermore, Fig. 2c demonstrates as well that the level of traffic concentration over BGP prefixes is not stable but rather varies within a day, which leads to the study in the following section on traffic dynamism.

C. Dynamism of traffic over BGP prefixes

In this section, we investigate the dynamism of traffic over BGP prefixes from two perspectives: volume variation and presence in *core* of each single prefix. These two aspects are first studied at week scale and then at hour scale. The terms used are defined and discussed here below.

As mentioned in § II-A, we collected traffic traces during two weeks and group volume count over 1 hour time bins. We use the later part of the traces, i.e. the week from June 1st, in the study on traffic dynamism. For each prefix P ever active during the week, it has a time series, $v(P)$, of 168 in length that stores its traffic volume each hour. We define the Coefficient of Variation (c_v) of prefix P's hour volume series over the week as:

$$c_v(P) = \frac{\delta(v(P))}{\mu(v(P))}, \tag{1}$$

where δ calculates the standard deviation of the hour volume series; μ gives the mean hour volume over the week. c_v is a measure of hour volume variation in relation to its

[1]Zipf's law defines that the k^{th} most popular element among total N elements has an occurrence share of $f(k, s, N) = \frac{1/k^s}{\sum_{n=1}^{N} 1/n^s}$.

mean hour value over the week[2]. Bigger hour volume c_v a prefix has, in a larger range its hour volumes oscillate around its mean hour volume over the week, and consequently more difficult to anticipate [14].

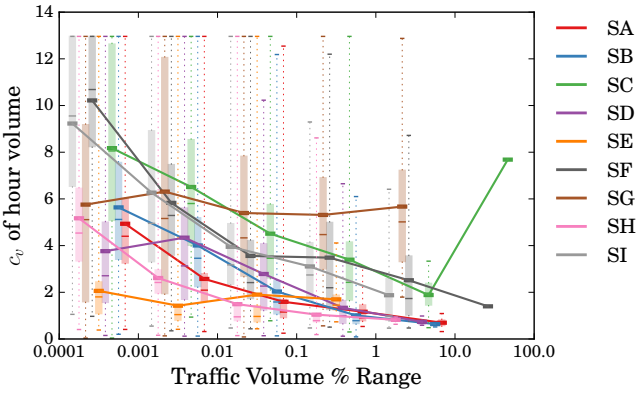


Fig. 3: Relation between c_v of hour volume and week volume fraction of BGP prefixes. Prefixes are grouped by their week volume percentage. Six bins are plotted in the graph from $[10^{-4}, 10^{-3})$ to $[10, 100)$. For each bin, its range, 25^{th}, 75^{th} percentile, median and mean values are given. Mean of each bin is marked with thick rod and chained with each other with solid line.

Fig. 3 plots the c_v of hour volume series of a prefix (Y-axis) against its accumulated volume fraction over the week (X-axis). We observe, for all the networks except SG and SE, that hour volume c_v of prefixes that contribute a bigger amount of traffic over the week tend to be constraint in a narrower range around a smaller mean value than those prefixes of less week volume. This indicates that prefixes with important volume over the week, in most cases, have relatively less variant hour volumes. Due this smooth variation, we can capture a large part of these prefixes using its mean hour volume, and consequently cover a big part of overall traffic.

Next, we describe traffic dynamism from another perspective: *core*. We define *core* as a prefix set containing least prefixes that represent 95% of total traffic each hour, i.e. top ranking prefixes by their hour volume. Table II lists the average number of prefixes included in the *core* and its percentage with regard to the average number of active prefixes each hour.

If a prefix is inside the *core* at a certain hour, it can be regarded important for bringing a big amount of traffic. We formulate this *core* presence for prefix P as:

$$cp(P)_i = \begin{cases} 1 & \text{when } P \in core_i, \\ 0 & \text{when } P \notin core_i, \end{cases} \tag{2}$$

where $core_i$ is the *core* prefix set at hour i. $cp(P)$ is a time series of 168 in length, with which we can further

[2]The maximum c_v for a hour volume series of 168 in length is $\sqrt{167}$, corresponding to the case where the prefix in question is active during only one single hour throughout the entire week.

Name	Avg. prefix #	Avg. prefix %	Max prefix #
SA	629	17.87	1051
SB	5264	9.45	13934
SC	73	4.59	177
SD	2481	17.35	3757
SE	15501	53.61	20900
SF	377	30.73	772
SG	570	7.76	1766
SH	965	19.42	1731
SI	175	21.00	415

TABLE II: Average number of prefixes within *core* each hour, its percentage in relation to average active prefixes each hour and maximum *core* size over the week from June 1st, 2015.

describe the frequency or likelihood of prefix being present in *core* over the week, i.e. *core* presence intensity, denoted at $I_{cp}(P, 168)$:

$$I_{cp}(P, 168) = \frac{\sum_{i=1}^{168} cp(P)_i}{168}. \tag{3}$$

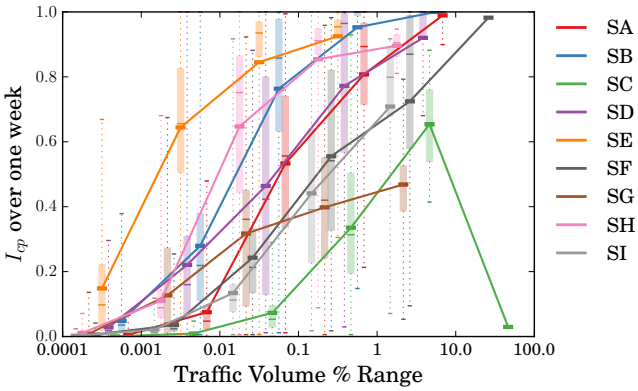
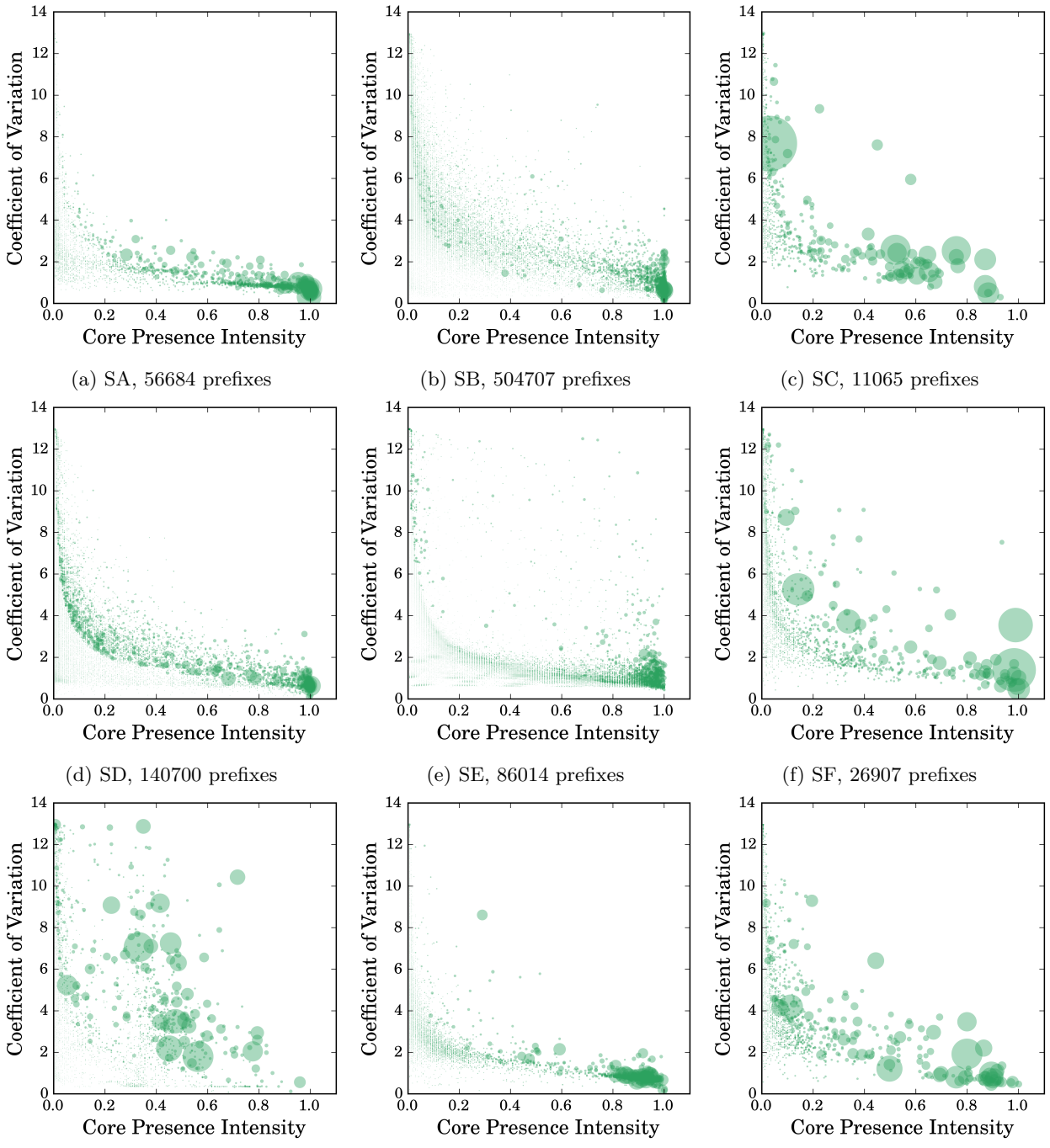


Fig. 4: Relation between I_{cp} over the week and week volume fraction of BGP prefixes. Prefixes are grouped by their week volume percentage. Six bins are plotted in the graph from $[10^{-4}, 10^{-3})$ to $[10, 100)$. For each bin, its range, 25^{th}, 75^{th} percentile, median and mean values are given. Mean of each bin is marked with thick rod and chained with each other with solid line.

Fig. 4 plots the I_{cp} over one week of each prefix (Y-axis) against its week volume share (X-axis). For all the networks, we can conclude that prefixes with bigger week volume are less likely to have a low I_{cp} over the week, thus frequently appear in *core*. This implies that by focusing on prefixes intensely appear in the *core* prefix set, we will be able to capture a large part of these prefixes associated with important traffic volume over the week, and thus cover a large part of overall traffic.

Further, we study the correlation between c_v of hour volume throughout the week and I_{cp} over the same period by plotting them together in Fig. 5, where each circle



(a) SA, 56684 prefixes (b) SB, 504707 prefixes (c) SC, 11065 prefixes

(d) SD, 140700 prefixes (e) SE, 86014 prefixes (f) SF, 26907 prefixes

(g) SG, 104425 prefixes (h) SH, 52061 prefixes (i) SI, 14189 prefixes

Fig. 5: Relation between I_{cp} over one week and c_v of hour volume over the week from June 1st, 2015. Each circle stands for a prefix. Number of active prefixes plotted is each sub-graph throughout the week is also given. Circle size is proportional to the week volume fraction of the prefix the circle represents and is of the same scale for all networks.

represents a prefix and its radius is proportional to the prefix's week volume share. Most visibly big circles concentrate in the lower right corner of each sub-graph, which corresponds to the findings made previously. However, exceptions there are, especially on SC, SF, SG and SI. We notice circles of big sizes having low I_{cp} and relatively high c_v. These prefixes bring bursty traffic of significant volume share within a short duration, which makes predictive prefix selection difficult. Finally, it's not a surprise to see that the c_v of prefixes with big week volume is, to a certain extent, inversely correlated to their I_{cp}. That is, less variation in hour volume, more often a prefix stays in *core*, if the prefix brings sufficiently large amount of traffic over the week.

The above analysis at week scale concludes that most prefixes associated with a big volume over a long term tend to be stable in hour volume variation and present often in *core*, though a few exceptions exist. Next, we explore these two measures at hour resolution by visualizing them in Fig. 6, where each column corresponds to top prefixes ranked by their hour volume with big prefixes at the top and small ones at the bottom. For each prefixes we use grey scale to represent its c_v or I_{cp} value of the week. Red line in each sub-figure, indicates the number of prefixes composing *core* each hour. Only graphs for SA are shown for the patterns demonstrated are most informative and less noisy due to less bursty traffic.

With regard to c_v, area above the red line, i.e. *core*, has observably lighter tone in color than the lower part. This implies that each hour, prefixes in the *core* have more stable hour volume over a long term, i.e. closer to its mean value, than those outsiders with little volume significance. Plus, it demonstrates as well a time-of-day pattern. During late night and early morning, prefixes in the *core* have deeper color, thus more volatile, than those in the day time, which indicates that important prefixes composition are not quite the same over these two periods. However, the above two phenomenon are sometimes difficult to be perceived on other networks with more bursty traffic. Still, the tone for *core* prefix is not visibly lighter nor deeper on these networks. It means that the prefixes bringing important volume each hour is not significantly more stable in volume than the rest. Since the volume associated to these prefix are big, we should all the same be able to pick them out using their mean value. The lack of clear diurnal pattern is possibly related to their business types and client populations they serve, which is out of the scope of this work.

When it comes to I_{cp}, it is true for all networks that top ranked prefixes each hour are more likely to frequently appear in the *core* prefix set, which leads to deeper color at the top. Time-of-day pattern can also be clearly seen. It confirms again that prefixes active over night are different from those during day time. Each hour, *core* prefixes enclosed by red line have in general more intense presence in the core than the outsiders. However, still we can witness light spots each hour in the upper part of the graph, which corresponds to bursty traffic. Further, the color difference across the red line is not obvious. This indicates I_{cp} alone is not informative enough to tell whether a prefix brings a big amount of traffic at a certain hour, though for long term it demonstrate strong correlation with volume importance as seen in Fig. 4.

D. Quantitative index of traffic burstiness

In this section, we quantify the overall traffic burstiness of a network. Throughout the above analyses, we had the impression that some networks, most obviously SC, SF, SG and SI, carry more bursty traffic than the others. It is mainly because, in Fig. 5, we observe big prefixes (circles) with low I_{cp} on these networks. In order to be more exact on this notion of traffic burstiness than measurements done with eyes, we define the the burstiness index β of prefix P at a certain hour i of the week under investigation as:

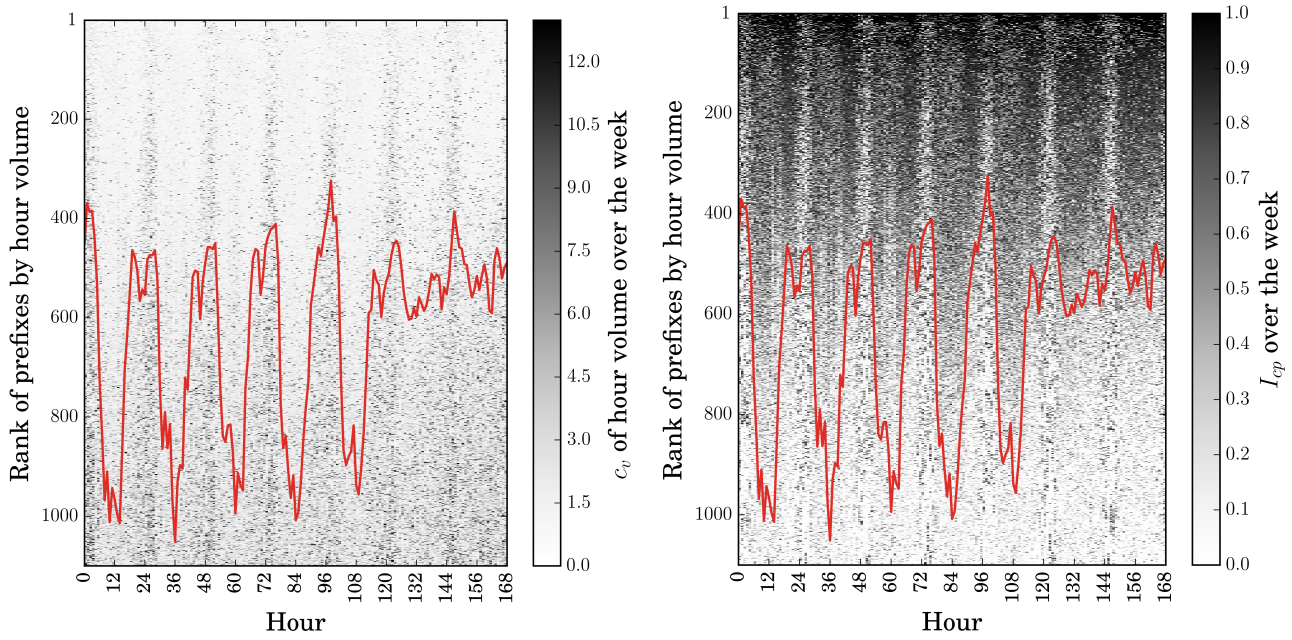
$$\beta(P)_i = \begin{cases} -\log(I_{cp}(P)) \times vp(P)_i & \text{if } I_{cp}(P) > 0 \\ 0 & \text{if } I_{cp}(P) = 0 \end{cases} \quad (4)$$

where $vp(P)_i$ is the hour volume percentage of prefix P at hour i. The logarithmic term [3] derived from I_{cp} aims at amplifying the volume contribution from prefixes with rare *core* presence, thus bursty, and attenuating the influence of prefixes being intensively in the *core*, thus easy to catch. If the I_{cp} of a prefix equals to zero, it means that the prefix never appeared in the *core* over the week, hence of few importance. Bigger the β is, larger the impact brought by the bursty traffic of a prefix at that moment, as it is hard to anticipate and possibly with big volume share. In order to estimate the overall impact brought by bursty traffic at hour i, we sum up the $\beta(P)_i$ for each P inside the *core* of that hour, denoted as BI_i, more formally:

$$BI_i = \sum_{P \in core_i} \beta(P)_i. \quad (5)$$

For all the networks, we estimate their traffic burstiness with the mean and maximum value of BI series over the week in Table III. The maximum β of all time is also given. Mean BI over the week measures the overall impact brought by bursty traffic, while maximum BI describes the degree of burstiness in worst cases. In correspondence to the observation made from Fig. 5, there are big prefixes with fairly low I_{cp} on SC, SF, SG and SI, whence the much bigger maximum β value. What is less evident in Fig. 5 is that SD suffers actually a lot from bursty traffic, even more than SC in general. For the rest networks, i.e. SA, SB, SE and SH, their mean BI over the week is around 30 or lower. Their corresponding sub-graphs in Fig. 5 manifest as well much less big circles on the left side where I_{cp} is low. More specifically SB suffers more from bursty traffic than SA,

[3]Logarithm function is used in the cost function of logistic regression to scale (magnify and compress) the estimation error in $(0, 1)$. We find it a good fit for scaling I_{cp} as well.



(a) c_v (b) I_{cp}

Fig. 6: c_v and I_{cp} over one week for top ranked prefixes at each hour on SA. Prefixes are ranked by their hour volume along the column (big prefixes at the top). Their c_v or I_{cp} over the week are represented by the grey scale. Red line indicates the number of prefixes in *core* prefix set each hour.

Network	**Mean** BI	**Max** BI	**Max** β
SA	14.61	37.79	7.44
SB	31.09	46.85	4.08
SC	40.57	145.07	145.05
SD	42.14	69.34	18.10
SE	20.91	44.30	20.17
SF	44.10	98.69	78.77
SG	51.21	125.41	102.05
SH	15.91	35.06	16.29
SI	38.59	85.47	56.56

TABLE III: Traffic burstiness.

therefore bigger value in both maximum and mean BI over the week, which is as well not easy to tell directly from Fig. 5. Still, the maximum β on SA is larger than that of SB, which is due to the fact that traffic on SB is more evenly distributed among active prefixes (see Fig. 2), and the faction of traffic associated to each prefix is generally smaller.

E. Indications for predictive prefix selection

Over period of one week, we have demonstrated the relationship among week volume share of a prefix, its c_v of hour volume series and I_{cp} of the week. We found that bigger volume a prefix brings over a long term, more likely it has little variation at hour scale and frequently present in *core* set. This suggests if we use mean hour volume, or *core* presence intensity as an indirect measure for prefix volume importance, we will be able to capture a big part of overall traffic.

More details are revealed, when we look at hour scale. Fig. 6 confirms that the c_v of top ranking prefixes each hour is not obviously worse than those outside the *core* prefix set. This implies using mean hour volume, we will be able to capture a big part of them at hour scale. As for I_{cp}, we argued that it might be not informative enough to indicate the hour volume importance, for bursty prefixes bringing large amount of traffic can have fairly low *core* presence intensity.

Nonetheless, I_{cp} conveys important information concerning traffic burstiness, which is crucial to prefix selection. Intuitively, if a network have more bursty traffic in general, it will be hard to achieve a satisfying overall traffic volume coverage by selected prefixes. If a network suffers at certain moments big amounts of bursty traffic, or there are several prefixes being extremely bursty, the minimum traffic fraction covered by predictively selected prefixes is endangered.

We make use of the above discoveries in the following section to predict prefixes of important volume in a scalable manner and verify if these conclusions hold true.

III. Predictive prefix selection

Based on the findings in previous sections, we propose using following metrics as indirection measures of the volume importance for the next hour $i+1$ (supposing that we are now at hour i). The calculation of each metrics is based on the hour volume records of last L hours.

Mean Volume

$$MV(P,L)_{i+1} = \frac{\sum_{j=i-L+1}^{i} v(P)_j}{L}. \quad (6)$$

With this metric, we predict that at hour $i+1$, the volume importance of prefix P is indicated by it's mean hour volume over last L hours. It is based on the observation from Fig. 3 and Fig. 6, that top ranking prefixes over the week tend to have smaller hourly volume variation around their mean volume and prefixes top ranked each hour have no worse volume stability compared to the less significant ones.

Core presence Intensity

$$I_{cp}(P,L)_{i+1} = \frac{\sum_{j=i-L+1}^{i} cp(P)_j}{L}. \quad (7)$$

Metric I_{cp} predicts prefixes basing on the *core* presence intensity over last L hours. It derives from the observation made from Fig. 4, that high ranked prefixes by their week volume are more likely to have intense *core* presence.

Core Volume

$$CV(P,L)_{i+1} = \frac{\sum_{j=i-L+1}^{i} cp(P)_j \times v(P)_j}{L}. \quad (8)$$

CV is a marriage between MV and I_{cp}. CV has the potential to be more resource economic compared to MV, as it is only calculated for prefixes ever appeared in *core* over last L hours, while MV is for all active prefixes. Depending on networks, *core* size is in average only 5% to 50% of all active prefixes.

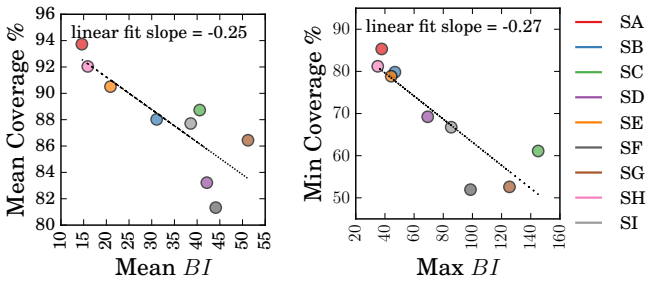
In previous work by Zhange et al. [5] on FIB caching, a grey differential model $GM(1,1)$ is employed to predictvely select BGP prefixes with big packet counts at 5 minute interval using historical records of last 50 minutes. We implemented the $GM(1,1)$ as well in our work to compare here above proposed metrics to it. A brief introduction to this model and how we apply it in our context is given below. Mathematical details of this model can be find in the work by Deng [15].

$GM(1,1)$ predicts the cumulative hour volume v^1 instead of hour volume directly:

$$v^1(P,L)_i = \sum_{j=i-L+1}^{i} v(P)_j, \quad (9)$$

where $v^1(P,L)_i$ is the cumulative hour volume of prefix P over last L hours at hour i. The purpose is to derive $v(P)_{i+1}$, i.e. the volume in the following hour, from estimations of cumulative volumes of hour $i+1$ and i:

$$\hat{v}(P)_{i+1} = \hat{v}^1(P,L)_{i+1} - \hat{v}^1(P,L)_i, \quad (10)$$



(a) Average level (b) Worst case

Fig. 9: The relationship between burstiness index BI and traffic volume coverage of selected prefix using CV metric.

where $\hat{v}$ and $\hat{v}^1$ are all estimation values given by the model. $GM(1,1)$ predicts that the cumulative hour volume at hour $i+1$ equals:

$$\hat{v}^1(P,L)_{i+1} = (v(P)_{i-L+1} - \frac{b}{a})e^{-aL} + \frac{b}{a}, \quad (11)$$

where a and b are parameters that can be estimated with least square method (symbol with hat are all estimations).

$$\hat{\mathbf{a}} = \begin{bmatrix} \hat{a} \\ \hat{b} \end{bmatrix} = (\mathbf{B^T B})^{-1}\mathbf{B^T Y}, \quad (12)$$

where

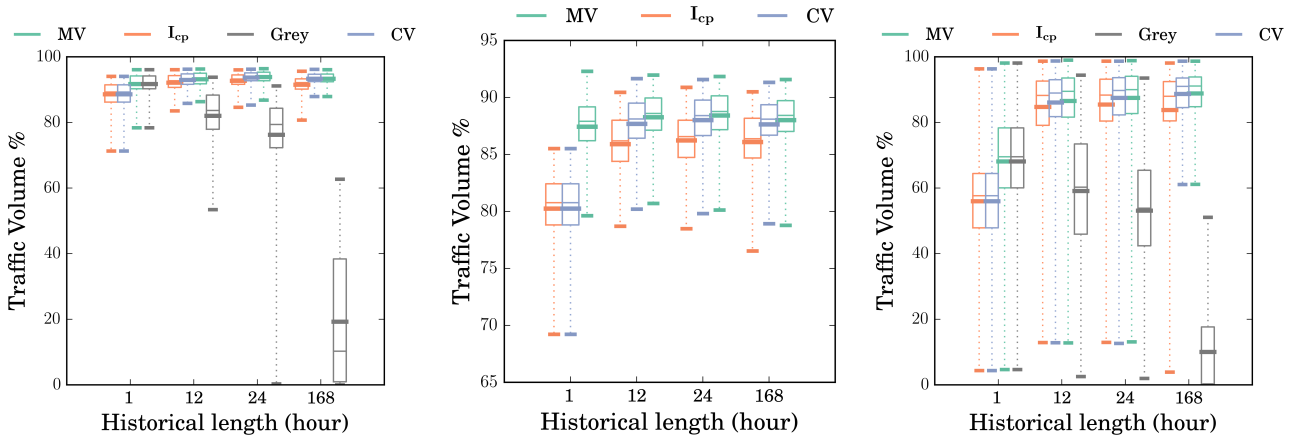
$$\mathbf{B} = \begin{bmatrix} -0.5(v^1(P,L)_{i-L+2} + v^1(P,L)_{i-L+1}) & 1 \\ -0.5(v^1(P,L)_{i-L+3} + v^1(P,L)_{i-L+2}) & 1 \\ \cdots & \cdots \\ -0.5(v^1(P,L)_i + v^1(P,L)_{i-1}) & 1 \end{bmatrix}, \quad (13)$$

$$\mathbf{Y} = \begin{bmatrix} v(P)_{i-L+2} \\ v(P)_{i-L+3} \\ \cdots \\ v(P)_i \end{bmatrix}. \quad (14)$$

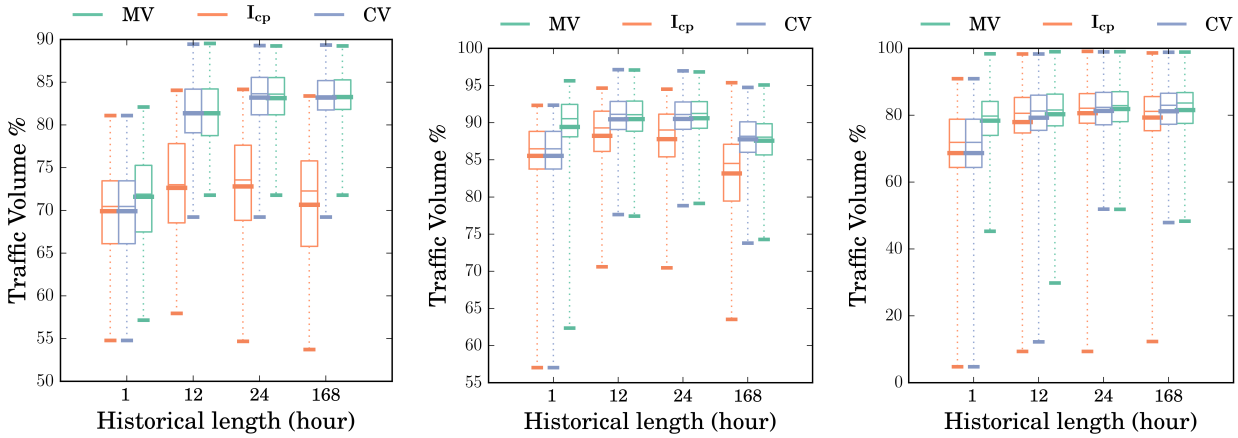
We can see that for each single prefix, two parameters are to be estimated, a and b, at each hour based on a new volume series that slides over time. Computationally, estimation with $GM(1,1)$ is much heavier than the three metrics proposed above, the calculation of which is identically to all prefixes without parameters to be customize.

We evaluate the performance of these four methods from two aspects: the hour volume **coverage** by selected prefixes set and the prefix **churn** of selected prefix set at each hour. More prefixes we select for the next hour, higher the volume coverage will be. However, a selection set of arbitrary big size is meaningless as it invalidates the purpose of making TE operation scalable and may as well conflict with engineering constraints in deployment. In this work, we set the selection set size to the maximum *core* size over the week, so as to be reasonable in relation to the smallest number of prefixes needed to cover 95% of total traffic each hour.

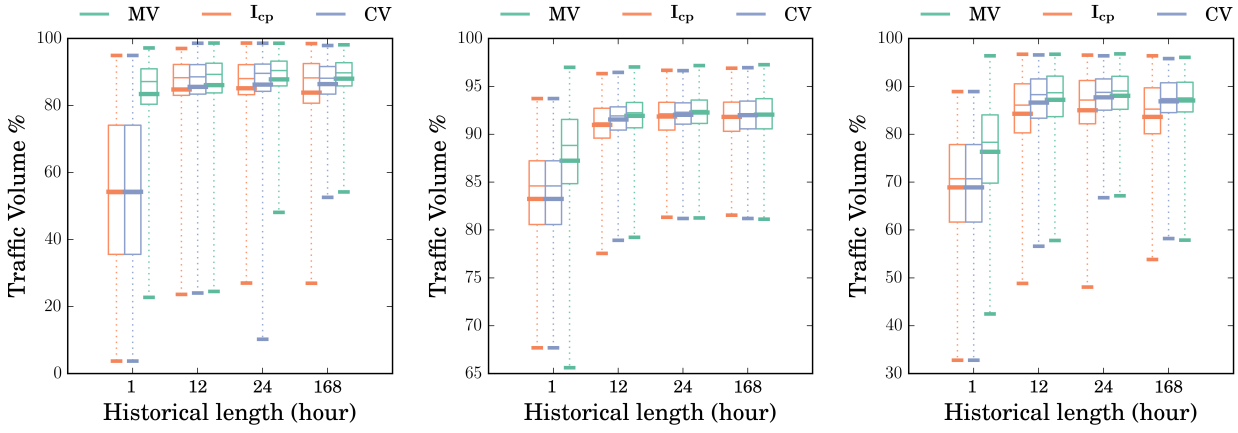
Fig. 7 illustrates the hour volume coverage by the four methods in form of boxplot, where the whiskers represent



(a) SA (b) SB (c) SC

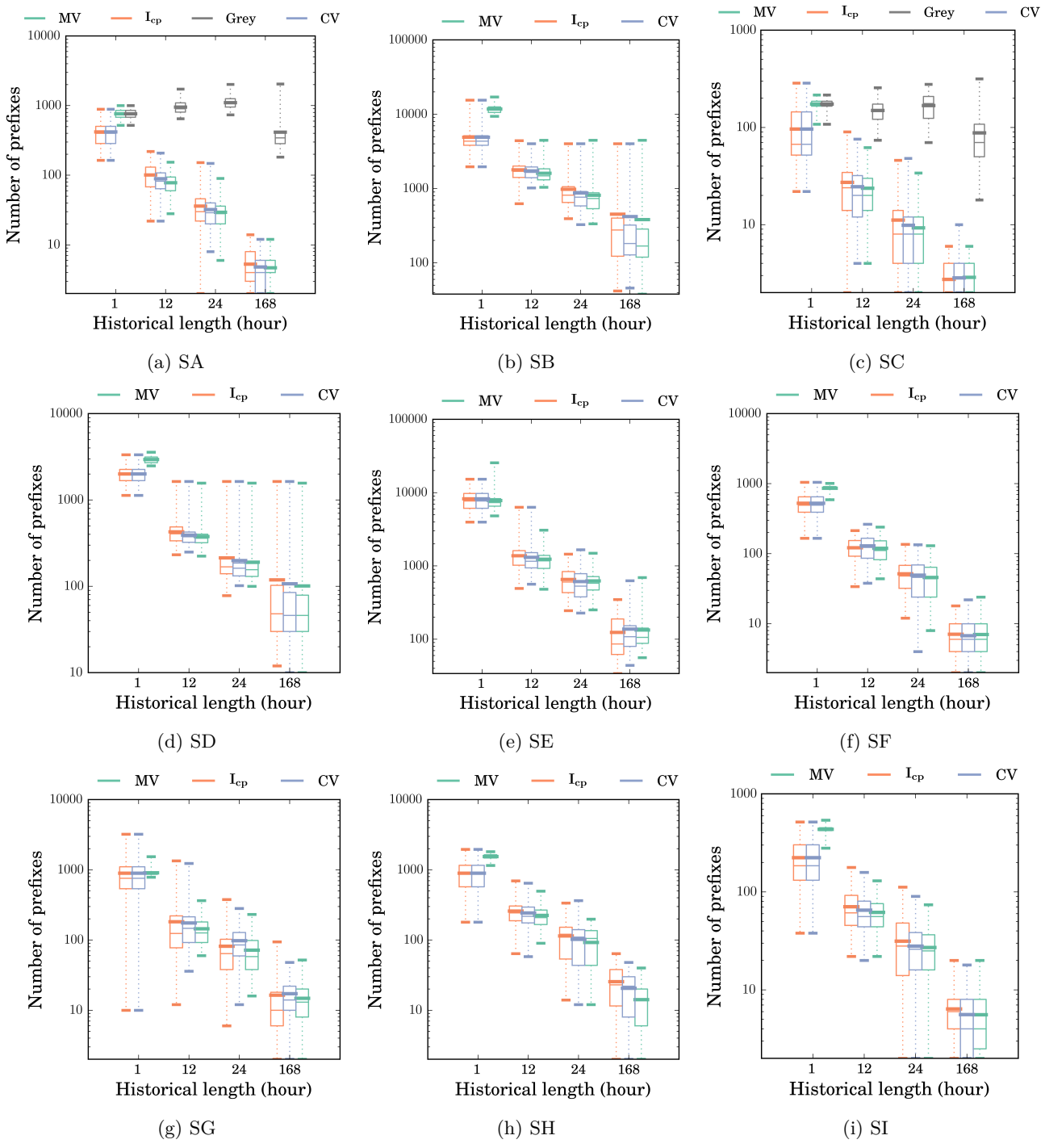


(d) SD (e) SE (f) SF



(g) SG (h) SH (i) SI

Fig. 7: Hour volume fraction covered by prefixes predictively selected using historical records of different lengths. The selection set size of each network is set to the maximum *core* size over the week starting from June 1st, 2015, see in Table II.



(a) SA (b) SB (c) SC

(d) SD (e) SE (f) SF

(g) SG (h) SH (i) SI

Fig. 8: Hour churn of the prefix set predictively selected using historical records of different lengths. The selection set size of each network is set to the maximum *core* size over the week starting from June 1st, 2015, see in Table II.

Network	L	Mean Cvg.	Min Cvg.	Mean Churn	Max Churn
SA	24	93.73	85.32	32.23	148
SB	24	88.02	79.82	873.11	4026
SC	168	88.73	61.11	2.84	10
SD	168	83.22	69.23	107.06	1642
SE	24	90.51	78.85	613.72	1664
SF	24	81.33	51.94	49.52	134
SG	168	86.44	52.60	17.24	48
SH	24	92.04	81.21	102.60	364
SI	24	87.71	66.75	28.01	90

TABLE IV: Hour volume coverage in percentage and hour prefix churn of selection set using core volume metric CV with record length L yielding the highest mean coverage.

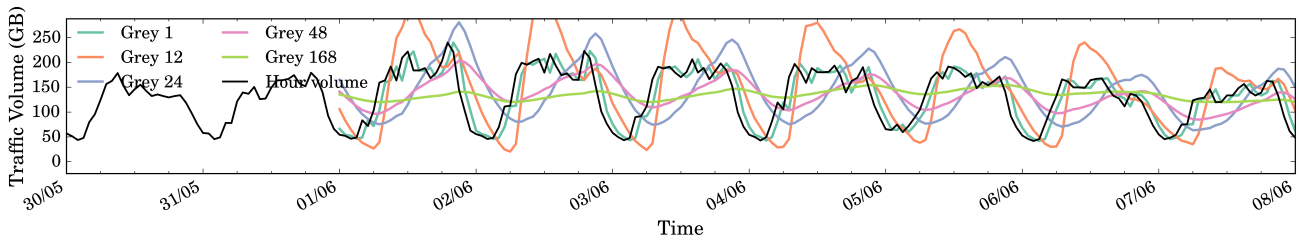
maximum and minimum value over the week; upper and lower edge of the box indicate the 25^{th} and 75^{th} percentile values; thin bar at the middle of the box marks medium value; and thick rod locates the mean. Results with different historical records length L are as well shown.

Among proposed metrics, we find that the CV is very close to MV in terms of volume coverage, proving that it is a good approximation of the later while being more scalable. Among all the networks, we achieved the best mean volume coverage under CV metric ($> 92\%$) on SA and SH, see in Table IV. This attributes to the fact that these two networks suffers the lest from bursty traffic. More generally, Fig. 9 reveals that the the mean/minimum coverage achieved with CV metric is inversely proportional to the mean/maximum BI index. This implies that it is difficult to cover as much traffic for networks with more bursty traffic.

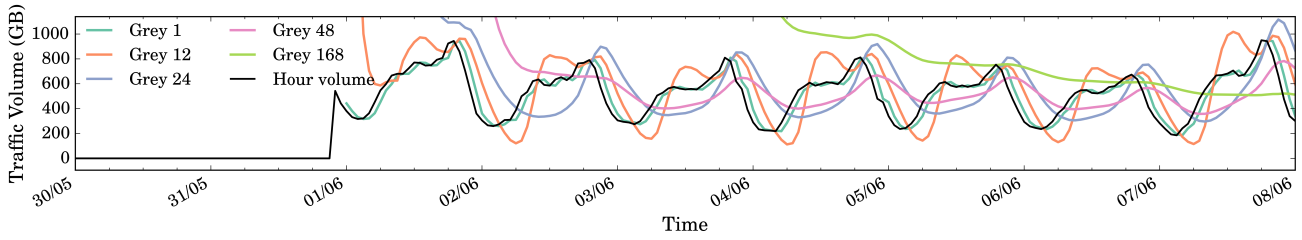
Basing solely on the last 1 hour records, all methods yield already a mean volume coverage $> 80\%$ on SA, SB, SE and SH, which implies a strong continuity on prefix volumes between two consecutive hours. On SC and SG, using records of last 168 hours, i.e. a week, offers much better minimum volume coverage than shorter records. This is due to the fact that at certain hour, SC and SG undergo a great amount of bursty traffic, which is shown in Table III and Fig. 5. By increasing the historical length, the selection metrics are able to have better visibility into the past and capture some of these bursty prefixes, and finally improve the minimum coverage. This gain in minimum volume coverage by using long historical records can actually be observed on all networks, between 1 hour and 12 hours, also between 12 hour and 24 hour. However from 24 hour to 168, this gain doesn't necessary happen on all sites, which is due to the fact that the total volume brought by some highly bursty prefix is diluted by the long time span using MV and CV metrics. In order to capture them, a larger selection set size is need, which inevitably includes more prefixes of few significance.

On the other hand, the hour volume coverage by grey model drops as we increase the historical records length and is in general much worse than the metrics proposed in this work. In order to understand the underlying reason, we used $GM(1,1)$ model described above to dynamically predict the total hour volume of all prefixes, which is normally much more regular and smoother than the volume series of individual prefixes. Results are shown in Fig. 10, Fig. 11 and Fig. 12. On SB and SD, we miss the hour volume data for the week starting from May 25th. $GM(1,1)$ model using records of last 168 suffer a lot from this 'abnormal' value and converge extremely slow to values in reasonable range. However this kind of 'abnormal' value can be quite normal for bursty prefixes that bring huge amount of traffic within in a short duration and then remain silent over days, which is common on SC, SF, SG judging from their burstiness index in Table III. This explains why grey model leads to fairly low volume coverage in Fig. 7c. Furthermore, in Fig. 10b, Fig. 10c and as well in others, we can see that grey model reacts to volume variations in a delayed manner. Longer the historical records length is, more obvious and longer grey model delays the variations of actual trace. This behavior could be fatal in the presence of bursty prefixes. We can observe considerably large pikes in both directions several hours after the volume burst in Fig. 10c, which might greatly disturb the prefix selection during those periods and consequently give rise to low volume coverage. Compared to grey model, metrics proposed in our work do not directly predict hour volumes of each prefix but rather are indirect measures of prefix volume importance. And analysis in § II-C showed that the overall coverage using these metrics are guaranteed by the traffic character itself.

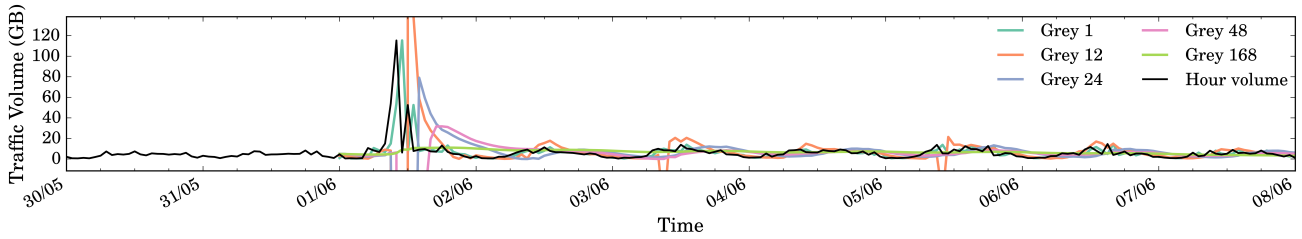
Fig. 8 gives the results concerning hour prefix churn of selection prefix set. The boxplot legend is the same as Fig. 7. The drop in prefix churn by increasing historical length is tremendous with the three proposed metrics, while the churn level of grey model is not much impacted by record length and remains at a high position. Sarrar et al. [6] argued that small prefix churn leads to low communication overhead in network architecture with decoupled forwarding plane and control plane, such as SDN (Software Defined Networking), thus more scalable. In that sense, using long historical records can be a wise choice in practical uses. On top of that, for networks with relatively few bursty traffic, e.g. SA, the mean volume coverage with last 168 hour records is extremely close to that with last 24 hour records, seen in Fig. 7a. Moreover,



(a) SA



(b) SB



(c) SC

Fig. 10: Predict total hour volume using $GM(1,1)$ with different historical record lengths for the week starting from June 1st, 2015. SA-SC.

for networks with highly bursty traffic, SC, using long records has the potential to obviously improve worst-case volume coverage.

In the purpose of lowering churn, Sarrar et al. [6] proposed selecting top prefixes over time bins of different lengths, ranging from 1 second to 10 minute. We find that in our case the difference in mean volume coverage using record lengths larger than 1 hour is marginal, thus little gain can be expected from this method. Meanwhile, the churn of selection set from mixed time bin is surly larger than that using the longest records length according to the tendency shown in Fig. 8.

IV. RTT probing and transit performance evaluation

In this section, we present the method used to evaluate the transit performance in terms of RTT (Round-Trip Time) perceived by selected prefixes.

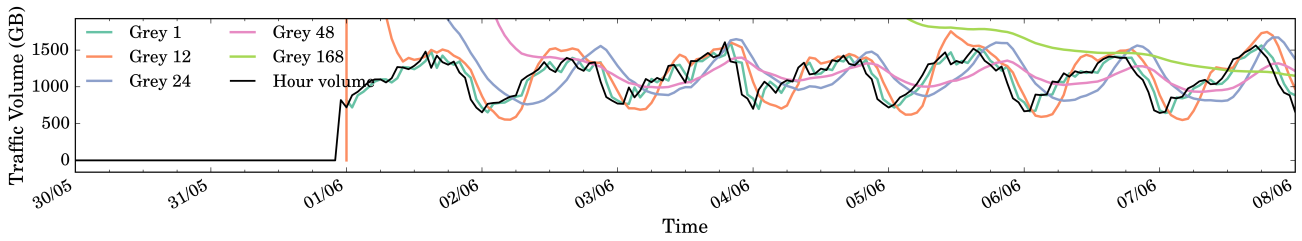
TCP SYN scanning was employed to measure RTT between selected prefixes and local network. The probe traffic toward a selected prefix was steered on all contractual transit providers, by means of explicit routing. For a pair of selected prefix and transit provider, a probe is scheduled at 240 second interval in average with 30% randomization left or right. The probe data covers the entire week starting from June 1st.

Our performance evaluation method follows the same principle as the one proposed by Akella et al. [16], which normalizes the RTT measure and aggregates results to different destinations by mean. More formally, the computation is done as such:

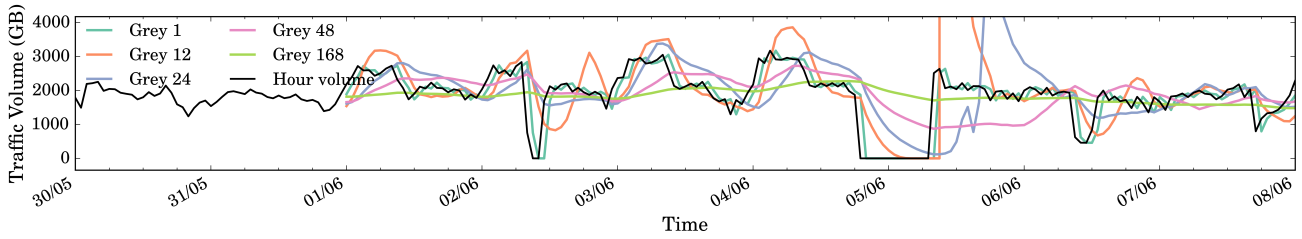
$$NP_{t_i}^{Tx} = \frac{\sum_{P \in SP} M_{t_i}^{Tx}(P) / \min_{Tj \in T} M_{t_i}^{Tj}(P)}{|SP|} \tag{15}$$

where $NP_{t_i}^{Tx}$ is the normalized RTT performance for transit provider Tx at probe tickle t_i. $M_{t_i}^{Tx}(P)$ denotes the RTT measured toward selected prefix P. This measure is fist normalized over the best measure among all available transit providers T at the same probe tickle. Then we average this normalized RTT over all prefixes inside the selected prefix set SP. If a transit provider offers the smallest RTT to all selected prefixes, it should have a normalized RTT performance equaling to 1.

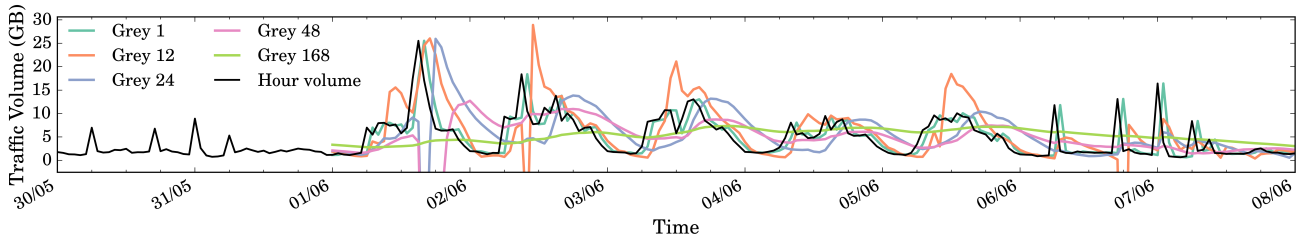
Fig. 13 gives the results of transit performance evaluation for SA, where the whiskers stand for 5^{th} and 95^{th} percentile values. Values below and beyond the whiskers are marked by + symbol and can be regarded as outliers.



(a) SD



(b) SE



(c) SF

Fig. 11: Predict total hour volume using $GM(1,1)$ with different historical record lengths for the week starting from June 1st, 2015. SD-SF.

Along the X-axis, available transit providers are aligned increasingly according to their mean normalized RTT performance (marked by square). $l1/r$ denotes a virtual transit provider simulated by our dynamic route selection algorithm, which is presented in the following section. The number of prefixes probed on each network and traffic volume represented are as well given in Fig. 7a.

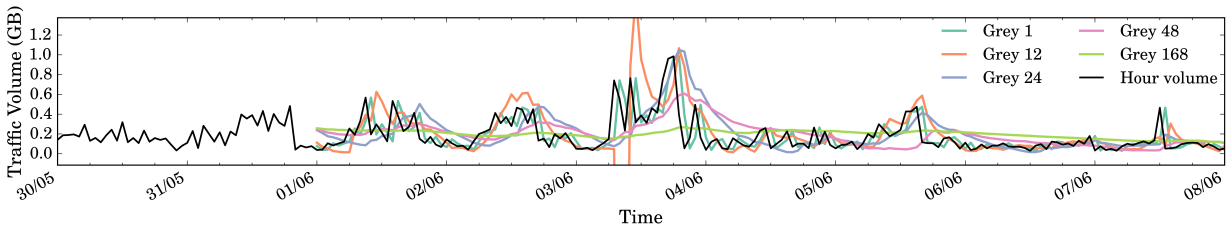
We observe that the performance differences among different transit providers is evident on some networks, e.g. SB, SC, SG, SH and SI. This confirms that mutlihoming can still provide significant performance improvement in modern days. However this gain in performance is not inherently given in the context of BGP. Even for transit provider that offers the best best mean NP, there are moments that its performance deviates far above 1, which means that traffic toward some selected prefixes are suffering from RTTs much larger than other available transit providers. This implies that a network may undergo temporal performance derogation if it uses transit providers in a static and indistinguishable manner for each individual prefixes. Therefore a dynamic route selection method that decides the best outgoing next-hop for each individual selected prefixes is needed, which leads to our next section.

V. Dynamic route selection

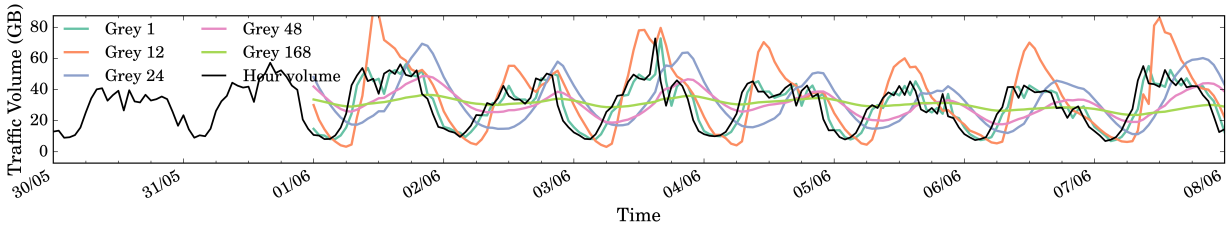
Algorithm 1 $l1/r$: selection based on last single measurement, if not available then pick randomly

for $t_i \in Probe\_tickle$ **do**
  for $P \in SP$ **do**
    if at least one transit has $M_{t_{i-1}}(P)$ **then**
      find Tb s.t. $M^{Tb}_{t_{i-1}}(P)$ being the smallest
      $M^{l1r}_{t_i}(P) \leftarrow M^{Tb}_{t_i}(P)$
    else
      $M^{l1r}_{t_i}(P) \leftarrow M^{Tr}_{t_i}(P)$, Tr is randomly chosen
    end if
  end for
end for

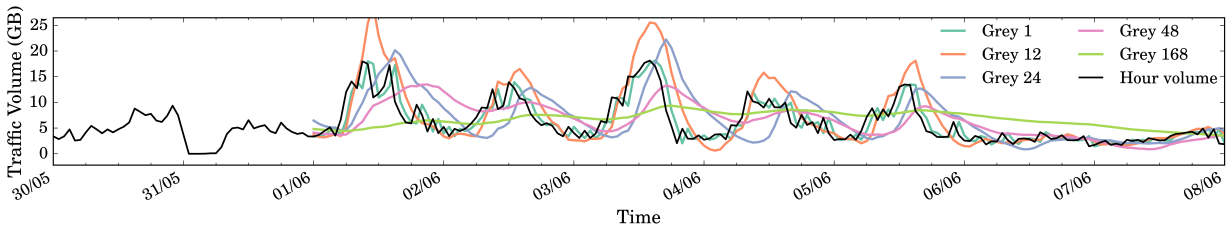
In this section we present a simple dynamic route selection algorithm, described in Algorithm 1. The method selects the transit provider that provides the smallest RTT in last round of probing for each selected prefix, which bases on the hypothesis that RTT of a path demonstrates temporal locality and thus is closely related to its most recent measurement [17]. The performance of the algorithm is shown in Figure 13. For all networks except SG, the simulated algorithm out-performs all physical transit



(a) SG



(b) SH



(c) SI

Fig. 12: Predict total hour volume using $GM(1,1)$ with different historical record lengths for the week starting from June 1st, 2015. SG-SI.

providers. On SB, the overall RTT drop for selected prefixes approaches 20% compared to the best physical transit provider. Still the NP of this virtual provider oscillates within a wide range, which calls for further investigation on the characters of RTT variation in time.

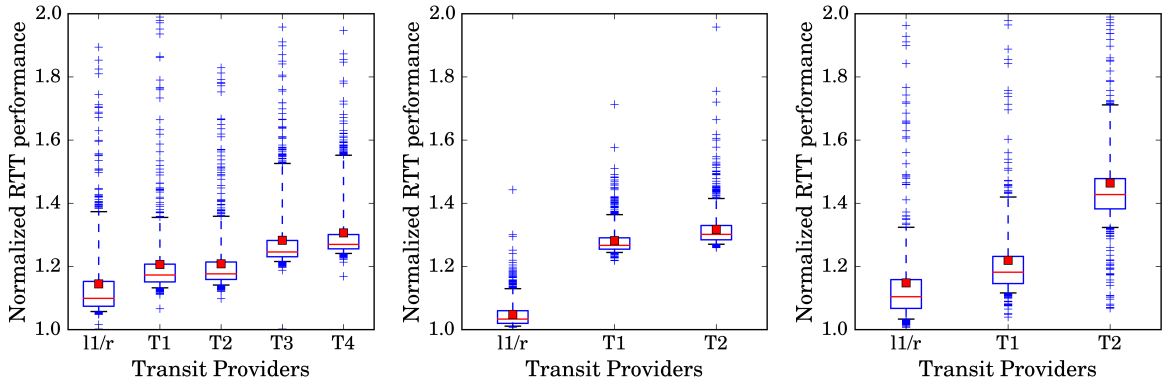
VI. Related work and Conclusions

Multiple previous works have shown that Internet traffic follows Zipf's law [8], [9], [13]. We revisited this conclusion on BGP prefixes and confirmed that it held true on multiple time scales with working traffic traces from networks of diverse profiles.

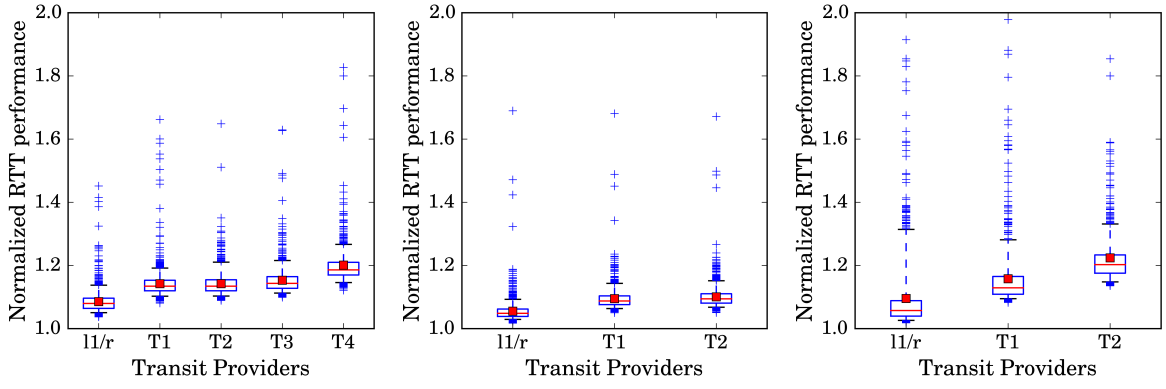
We proposed exploiting this feature to control a majority of traffic with a small amount of prefixes in Inter-domain TE with BGP. Some previous work in this domain [9], [17], [18] acknowledged the importance of preforming inter-domain TE only for destinations that matter. However, no general solution predictively selecting BGP prefixes with important traffic volume is proposed. Some other works leveraged the skewed distribution of Internet traffic to downsize FIB [2]–[7], where understanding of traffic dynamism on much shorter time scale, e.g. seconds and minutes, are required. In our work, we used traffic traces of coarser time resolution over longer periods in order to accommodate to the more complicated operations associated to TE.

In order to arrive at a viable solution for TE on BGP prefixes, we dedicated a large part of this article exploring traffic dynamism at hour interval spanning over a week. Zhang et al. [5] and Papagiannaki et al. [19], both teams studied the relationship between stability and popularity, and drew conflicting conclusions. We showed in our study that the hour volume variation of top ranking prefixes over the week is smaller in average than those tail prefixes. We also extended the findings made by Wallerich et al. [13] by demonstrating that traffic evolution can also be radical over week time span due to bursty traffic. Furthermore, we quantify the traffic burstinees of each network, which corresponds well to visual observations.

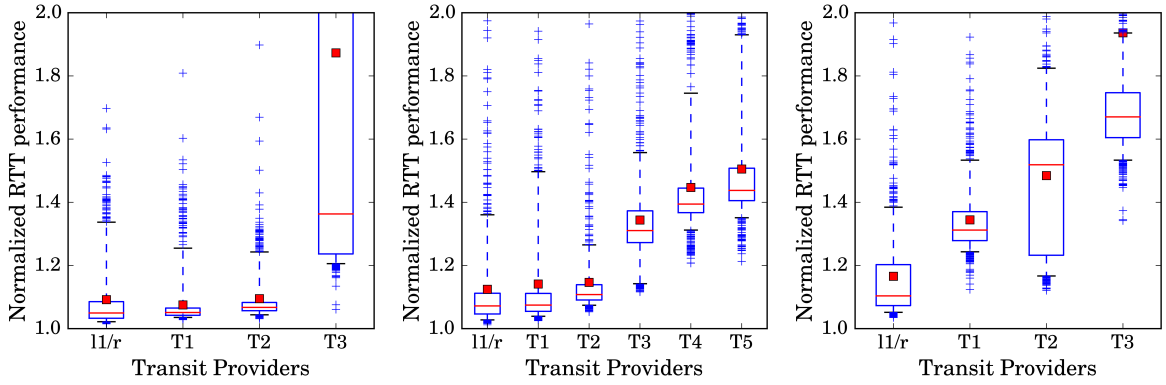
Based on the observations made in this work, we proposed predictive prefix methods demanding much less storage and computation resources than Time Series Forecasting (TSF) models, such as Holt-Winter, autoregressive moving average (ARMA) and artificial neural network, which are generally applied to highly aggregated Inter-PoP traffic [10]–[12]. We showed that the metrics proposed in our work is more robust facing sudden traffic volume changes and delivers better volume coverage than grey model used by Zhang et al. [5] under our specific context. The results showed that the methods proposed in our work



(a) SA, 866 prefixes, 85.69% traffic (b) SB, 935 prefixes, 48.49% traffic (c) SC, 762 prefixes, 44.23% traffic



(d) SD, 2841 prefixes, 49.18% traffic (e) SE, 2287 prefixes, 38.59% traffic (f) SF, 879 prefixes, 78.49% traffic



(g) SG, 924 prefixes, 63.08% traffic (h) SH, 73 prefixes, 2.83% traffic (i) SI, 642 prefixes, 67.51% traffic

Fig. 13: Normalized RTT performance with active probing. Average number of prefixes probed and average traffic volume fraction represented by these prefixes each hour are given.

offered both satisfying volume coverage and pretty low prefix churn when using long historical records, i.e. last 168 hours.

Akella et al. [16] quantified the performance gain from multi-homing using traces from a large CDN network more then ten years ago. We revisited this measurement in modern network environment and showed that there is still considerable performance differences among different providers. Further, Akella et al. [17] evaluated a dynamic route selection system on a testbed with only 100 destinations. IPs instead of prefixes are selected and probed for route decision making, which is a good fit for the network environment they worked in, i.e. CDN, but fails to be universal in inter-domain networking where routing and forwarding are prefix based. From the perspective of selected BGP prefixes with important volume share, we showed that overall transmission performance of the network can be improved by 20% compared to the best available transit provider with dynamic route selection based on real-time measurements.

VII. Future directions

The volume coverage by selected prefixes using metrics proposed in this work is to certain extent decided by the traffic burstiness. The volume coverage, both mean and minimum, on networks with quite a lot bursty traffic e.g. SC, SD and SF is obviously poorer than that of SA. Capturing these bursty prefixes is the key to arrive at a better selection method. A possible direction in doing so is to first group prefixes by their activity profiles. For each group, we design methods that is adapted to the its traffic dynamism characters. For example, when dealing with prefixes with regular volume patterns and small hour variation, the simple metrics proposed in the work function already well. For those bursty ones, we might need more sophisticate model that extracts more activity features, for example periodicity. The burstiness index β proposed in the work can be used in this kind of prefix classification. We also showed that prefixes that are important during late night are quite different from those over daytime, which calls for classification on one other axis. Predictively select prefixes that are of importance to TE operations makes the later scalable and effective, which is crucial in better engineering the performance of inter-domain transmission.

References

[1] BGP Routing Table Analysis Reports. [Online]. Available: http://bgp.potaroo.net

[2] L. Iannone and O. Bonaventure, "On the cost of caching locator/ID mappings," *CoNEXT '07*, p. 1, 2007.

[3] H. Ballani, P. Francis, T. Cao, and J. Wang, "Making routers last longer with ViAggre," *NSDI '09*, pp. 453–466, 2009.

[4] C. Kim, M. Caesar, A. Gerber, and J. Rexford, "Revisiting route caching: The world should be flat," *Lecture Notes in Computer Science*, vol. 5448, pp. 3–12, 2009.

[5] W. Zhang, J. Bi, J. Wu, and B. Zhang, "Catching popular prefixes at AS border routers with a prediction based method," *Computer Networks*, vol. 56, no. 4, pp. 1486–1502, Mar. 2012.

[6] N. Sarrar, S. Uhlig, A. Feldmann, R. Sherwood, and X. Huang, "Leveraging Zipf's law for traffic offloading," *ACM SIGCOMM CCR*, vol. 42, no. 1, p. 16, Jan. 2012.

[7] Y. Liu, V. Lehman, and L. Wang, "Efficient FIB caching using minimal non-overlapping prefixes," *Computer Networks*, vol. 83, pp. 85–99, 2015.

[8] W. Fang and L. Peterson, "Inter-AS traffic patterns and their implications," in *GLOBECOM '99*, vol. 3. IEEE, 1999, pp. 1859–1868.

[9] N. Feamster, J. Borkenhagen, and J. Rexford, "Guidelines for interdomain traffic engineering," *ACM SIGCOMM CCR*, 2003.

[10] K. Papagiannaki, N. Taft, Z.-L. Zhang, and C. Diot, "Long-term forecasting of internet backbone traffic." *IEEE transactions on neural networks*, vol. 16, no. 5, pp. 1110–24, Sep. 2005.

[11] P. Cortez, M. Rio, M. Rocha, and P. Sousa, "Internet Traffic Forecasting using Neural Networks," in *IJCNN '06*. IEEE, 2006, pp. 2635–2642.

[12] T. Otoshi, Y. Ohsita, M. Murata, Y. Takahashi, K. Ishibashi, and K. Shiomoto, "Traffic Prediction for Dynamic Traffic Engineering Considering Traffic Variation," *GLOBECOM '13*, pp. 1592–1598, Dec. 2013.

[13] J. J. Wallerich and A. Feldmann, "Capturing the variability of internet flows across time," *INFOCOM '06*, pp. 1–6, 2006.

[14] Q. He, C. Dovrolis, and M. Ammar, "On the predictability of large transfer TCP throughput," in *SIGCOMM '05*, vol. 35, no. 4. New York, New York, USA: ACM Press, Aug. 2005, p. 145.

[15] D. Julong, "Introduction to Grey System Theory," *The Journal of Grey System1*, vol. 1, pp. 1–24, 1989.

[16] A. Akella, B. Maggs, S. Seshan, A. Shaikh, and R. Sitaraman, "A measurement-based analysis of multihoming," in *SIGCOMM '03*. New York, New York, USA: ACM Press, 2003, p. 353.

[17] A. Akella, B. Maggs, S. Seshan, and A. Shaikh, "On the Performance Benefits of Multihoming Route Control," *IEEE/ACM Transactions on Networking*, vol. 16, no. 1, pp. 91–104, Feb. 2008.

[18] D. K. Goldenberg, L. Qiuy, H. Xie, Y. R. Yang, and Y. Zhang, "Optimizing cost and performance for multihoming," *ACM SIGCOMM CCR*, vol. 34, no. 4, p. 79, Oct. 2004.

[19] K. Papagiannaki, N. Taft, and C. Diot, "Impact of Flow Dynamics on Traffic Engineering Design Principles," in *INFOCOM '04*, vol. 4, 2004, pp. 2295–2306.