Traffic Matrix Reloaded: Impact of Routing Changes

Renata Teixeira
U. California–San Diego
teixeira@cs.ucsd.edu

Nick Duffield   Jennifer Rexford
AT&T Labs–Research
{duffield,jrex}@research.att.com

Matt Roughan
University of Adelaide
matthew.roughan@adelaide.edu.au

I. Introduction

The design and operation of IP networks depends on a good understanding of the offered traffic. Internet Service Providers (ISPs) usually represent the traffic as a matrix of load from each ingress point to each egress point over a particular time interval. Although well-provisioned networks are designed to tolerate some fluctuation in the traffic matrix, large variations break the assumptions used in most designs. In this paper, we investigate the *causes* of the traffic matrix variations. Identifying the reasons for these disruptions is an essential step toward predicting and planning for their occurrence, reacting to them more effectively, or avoiding them entirely.

The traffic matrix is the composition of the *traffic demands* and the *egress point selection*. We represent the traffic demands during a time interval t as a matrix V, where each element $V(i,p,t)$ represents the volume of traffic entering at ingress router i and headed toward a destination prefix p. Each ingress router selects the egress point for each destination prefix using the Border Gateway Protocol (BGP). We represent the BGP routing choice as a mapping ε from a prefix to an egress point, where $\varepsilon(i,p,t)$ represents the egress router chosen by ingress router i for sending traffic toward destination p. At time t each element of the traffic matrix $\mathcal{T}M$ is defined as:

$$\mathcal{T}M(i,e,t) = \sum_{p \in P : \varepsilon(i,p,t)=e} V(i,p,t). \qquad (1)$$

where P is the set of all destination prefixes.

Figure 1 presents a simple network with one ingress router i, two egress routers e and e', and two external destination prefixes p_1 and p_2. Given traffic demands $V(i,p_1,t)$ and $V(i,p_2,t)$ and a prefix-to-egress mapping $\varepsilon(i,p_1,t) = \varepsilon(i,p_2,t) = e$, the traffic matrix for this network is $\mathcal{T}M(i,e,t) = V(i,p_1,t) + V(i,p_2,t)$ and $\mathcal{T}M(i,e',t) = 0$.

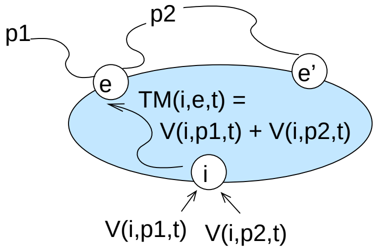

Fig. 1. Example of traffic matrix.

Fluctuations in the traffic demands and changes in the prefix-to-egress mapping cause the traffic matrix to vary. This paper considers the natural question: *what are the causes of large variations in the traffic matrix?*

Most previous work on measuring [1–4] and analyzing traffic matrices [5, 6] has assumed that the prefix-to-egress mapping ε is stable. However, relying on periodic snapshots of routing data runs the risk of associating some traffic measurements with the wrong elements in the traffic matrix, obscuring real variations in the traffic. In this paper, we study how *changes* in ε impact the traffic matrix. A previous analysis of five traces of 6–22 hours in duration on the Sprint network [7] shows that most BGP routing changes do not lead to large traffic shifts. However, given that large traffic variations are infrequent (yet significant) events, we believe that longer traces are necessary to draw meaningful conclusions.

In this paper, we study the impact of routing changes on the traffic matrix over a *seven-month* period in a tier-1 ISP network. Using Cisco's Sampled Netflow feature [8] and feeds of internal BGP (iBGP) messages, we compute the traffic demands V and the prefix-to-egress mapping ε for eight ingress routers, as discussed in Section II. We also monitor the intradomain routing protocol to identify the changes in ε caused by internal network events. In Section III, we demonstrate that, although most routing changes do not cause large traffic shifts, many of the large shifts *are* in fact triggered by routing changes. We also show that changes in ε caused by internal network events tend to have more impact on the traffic matrix than the external BGP events. Section IV concludes the paper with a discussion of our ongoing work.

II. Measuring Traffic Matrix Variation

Studying the variation of traffic matrix elements over time requires collecting fine-grained measurements of traffic and routing. We analyze data collected from a tier-1 ISP network for 173 days from March to September 2004. We collect data from eight aggregation routers that receive traffic from customers destined to peers and other customers. The eight routers are located in major Points of Presence (PoPs) that are spread throughout the United States.

We compute eight *rows* of the traffic matrix, considering all traffic from these eight ingress aggregation routers to all of the egress PoPs. This section describes how we compute the prefix-to-egress mapping $\varepsilon(i,p,t)$ from the BGP data and the traffic demands $V(i,p,t)$ from the Netflow data. Once we have computed ε and V, we use Equation 1 to compute the elements of the traffic matrix $\mathcal{T}M(i,e,t)$. The BGP monitor and the Netflow collection servers are NTP-synchronized, allowing us to use the timestamps to join the two datasets.

A. Prefix-to-Egress Mapping

A BGP monitor collects internal BGP update messages directly from each vantage point. Configured as a route-reflector client of each vantage point, the BGP monitor receives updates reporting any change in the best BGP route at each router for each destination prefix. The monitor records each BGP update with a timestamp at the one-second granularity.

We group BGP updates that happen less than 70 seconds apart to account for transient routing changes during the BGP convergence process (as in previous studies [9, 10]). That is, we focus on the changes from one stable route to another and not on the short-lived routes that exist during the transition.

Based on an initial BGP table dump and a sequence of BGP updates, we generate the prefix-to-egress mapping $\varepsilon(i, p, t)$ for any given time. The egress point corresponds to a *PoP* rather than a specific router. We associate each egress router with a PoP based on the router name and configuration data.

B. Traffic Demands

Every vantage point has the Cisco's Sampled Netflow feature [8] enabled on all links that connect to access routers and exports flow records to a collection server at the same location. The collection server samples the flow records using the technique presented in [11] in order to reduce processing overhead, and computes 10-minute aggregated traffic volumes for each destination prefix. We use these aggregated reports to extract $V(i, p, t)$ for each vantage point i and destination prefix p at every 10-minute interval. Consequently, a reference to a time t indicates the end of a 10-minute interval[1].

Because of sampling, the volumes $V(i, p, t)$ are random quantities that depend on the sampling outcomes. Through a renormalization applied to the bytes reported in sampled flow records, the quantities $V(i, p, t)$ are actually unbiased estimators of the volumes of the original traffic from which they were sampled, i.e., their average over all possible sampling outcomes is the original volume. The standard error associated with an aggregate of size V is bounded above by $\sqrt{k/V}$ for some constant k that depends on the sampling parameters [11]. For the parameters employed in the current case, $k < 21$MB. Note that the standard error bound decreases as the size of the aggregate increases. This property aligns well with our focus on the largest changes in traffic rates: these are the most reliably estimated. As an example, for a 10-minute aggregate of traffic at a rate of 10 MB per second, the standard error due to sampling is no more than 6%.

III. Causes of Large Traffic Variations

In this section, we explore the contributions of changes in the traffic demands V and prefix-to-egress mapping ε to the variations in the traffic matrix elements $\mathcal{TM}$. Our analysis shows that, although most changes in ε have a small effect on the traffic matrix, many of the large variations in the traffic matrix are caused by changes in ε. Also, we show that, while most changes in ε are caused by external routing events, the small number of internal routing events are more likely to cause larger shifts in traffic.

A. Definition of Traffic Variations

Figure 2 shows an example of how two traffic matrix elements (with the same ingress point i) change over the course of a day. The total traffic entering at the ingress point varies throughout the day, following a typical diurnal cycle. For the most part, the traffic $\mathcal{T}M(i, e_1, t)$ has the same pattern, keeping the proportion of traffic destined to e_1 relatively constant. For most of the day, no traffic travels from ingress i to egress point e_2. The most significant change in the two traffic matrix elements occurs near the end of the graph. The traffic leaving via egress point e_1 suddenly decreases and, at the same time, traffic leaving via egress point e_2 increases. This shift occurred because a routing change caused most of the traffic with egress point e_1 to shift to egress point e_2. The egress point e_2 also starts receiving traffic that had previously used other egress points (not shown in the graph), resulting in an increase for e_2 that exceeds the decrease for e_1. In the meantime, the total traffic entering the network at ingress i remained nearly constant.

The traffic experiences other relatively large downward spikes (labeled as load variation). These spikes may very well be associated with a routing change in another AS in the Internet that caused traffic to enter at a different PoP (this kind of traffic variation was called an “ingress-shift anomaly” in [6]). In this paper, we analyze traffic shifts caused by routing changes experienced by our network. Finding a signature of routing-induced traffic variations for one network is an important first step to infer other traffic variations that are caused by routing changes in other networks.

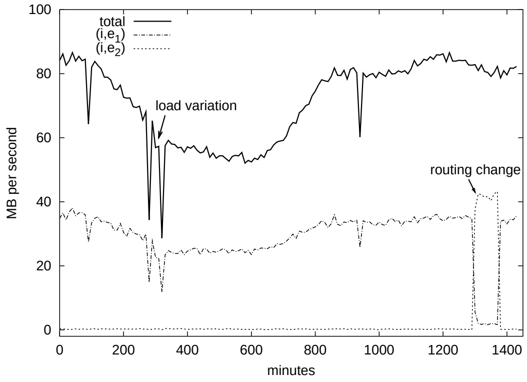

Fig. 2. Sample traffic volume from one ingress to two egresses.

To analyze these kinds of traffic fluctuations, we define the variation of a traffic matrix element at an interval t as:

$$\Delta \mathcal{T}M(i, e, t) = \mathcal{T}M(i, e, t) - \mathcal{T}M(i, e, t-1).$$

[1] If the mapping $\varepsilon(i, p, t)$ changes more than once in a 10-minute interval, then we cannot distinguish the volume of traffic affected by each of them individually. Therefore, we exclude those cases from our analysis by ignoring intervals with prefixes that have more than one stable routing changes in that bin; this excludes 0.05% of the (i, e, t) tuples from our study. We also exclude all traffic for the small number of flows that had no matching destination prefix in the BGP routing tables or update messages; we verified that these flows corresponded to an infinitesimal fraction of the traffic.

B. Changes in Traffic Demands vs. Egress Points

The variation of a traffic matrix element ($\Delta\mathcal{TM}$) is composed of the load variation (ΔL), which represents volume fluctuations on the traffic demands V, and the routing shifts (ΔR), which accounts for changes in the prefix-to-egress mapping ε:

$$\Delta\mathcal{TM}(i,e,t) = \Delta L(i,e,t) + \Delta R(i,e,t)$$

$\Delta L(i,e,t)$ represents the change in the volume of traffic for all destination prefixes that did *not* change their egress point from the previous time interval (i.e., $\varepsilon(i,p,t) = \varepsilon(i,p,t-1) = e$):

$$\Delta L(i,e,t) = \sum_{\substack{p \in P: \\ \varepsilon(i,p,t) = e \\ \varepsilon(i,p,t-1) = e}} V(i,p,t) - V(i,p,t-1)$$

Fluctuations in the traffic demands may occur for a variety of reasons, such as changes in user or application behavior, adaptations caused by end-to-end congestion control, or even routing changes in other domains.

The routing variation $\Delta R(i,e,t)$ considers the destination prefixes that shifted *to* egress point e during time interval t or shifted *from* e to another egress point in t:

$$\Delta R(i,e,t) = \sum_{\substack{p \in P: \\ \varepsilon(i,p,t) = e \\ \varepsilon(i,p,t-1) \neq e}} V(i,p,t) - \sum_{\substack{p \in P: \\ \varepsilon(i,p,t) \neq e \\ \varepsilon(i,p,t-1) = e}} V(i,p,t-1)$$

Note that if a routing change occurs within the time interval t, we associate *all* of the traffic associated with that prefix in that time interval with the new egress point.

Not all traffic matrix elements carry the same volume of traffic, and the volume of traffic from an ingress to an egress PoP varies over time. How do we judge if a change in the traffic is "large"? There is no absolute standard: one approach might be to judge the size of the change in traffic matrix element relative to the average traffic for that element. However, this is not useful here, because the traffic process itself is non-stationary. It has daily and weekly cycles, as well as level shifts resulting from routing changes. On the other hand, we should consider what type of process we observe, namely, a difference process. Over short time periods, we can approximate the traffic with a linear process $y_t = \alpha + \beta t + x_t$, where x_t is a zero-mean stochastic process, with variance σ^2. We observe the differences $\Delta y_t = y_t - y_{t-1}$, which will form a *stationary* process, with mean β and variance $2\sigma^2$. Thus we can approximate the difference process by a stationary process, and measure deviations from the mean, relative to the standard deviation of this process. We measure $2\sigma(i,e)^2$ on the traffic variation process $\Delta L(i,e,\cdot)$ (using the standard statistical estimator), and use this to normalize the traffic variations, i.e. we then observe $\Delta\tilde{L}(i,e,t) = \Delta L(i,e,t)/\sqrt{2}\sigma(i,e)$, and $\Delta\tilde{R}(i,e,t) = \Delta R(i,e,t)/\sqrt{2}\sigma(i,e)$.

If the variance of the process x_t was time dependent, it might make sense to use a moving average to estimate the process variance at each point in time, i.e. $\sigma(i,e,t)^2$, and use this to normalize the traffic variations. We tried such an approach, but it made little difference to the results, and so we use the simpler approach described above.

Figure 3 presents a scatter plot of $\Delta\tilde{\mathcal{TM}}(i,e,t)$ versus $\Delta\tilde{R}(i,e,t)$ for all the valid measurement intervals t. The high density of points close to zero shows that large traffic variations are not very frequent (99.88% of the traffic variations are in the $[-4,4]$ range). Points along the horizontal line with $\Delta\tilde{R}(i,e,t) = 0$ correspond to traffic variations that are not caused by routing changes, whereas points along the diagonal line correspond to variations caused almost exclusively by routing changes. Points in the middle are caused by a mixture of routing changes and load variation. Figure 3 shows that both load and routing are responsible for some big variations. Routing changes, however, are responsible for the *largest* traffic shifts. Indeed, one egress-point change made a traffic matrix element vary more than 70 times the standard deviation.

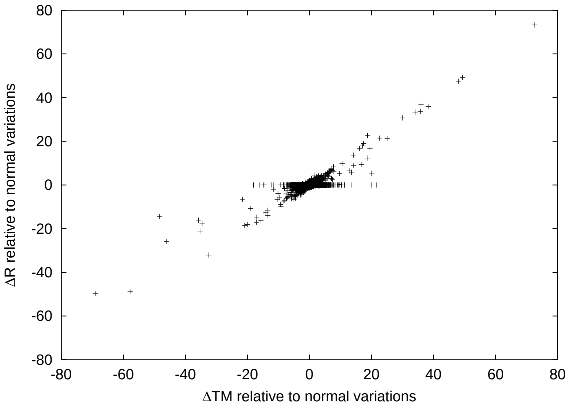

Fig. 3. Scatter plot of $\Delta\tilde{\mathcal{TM}}$ versus $\Delta\tilde{R}$ for all traffic matrix elements over the seven-month period.

C. Internal vs. External Routing Changes

The prefix-to-egress mapping ε may change because of either internal or external routing events. *External routing changes* represent any changes in the set of egress points that an AS uses to reach a destination prefix. For example, in Figure 1, the neighbor AS might withdraw the route for $p2$ from the router e, resulting in a change in ε. External routing changes may be caused by a variety of events, such as an internal routing change in another domain, a modification to the local BGP routing policy, or a failure at the edge of the network. In contrast, *internal routing changes* stem from changes in the routing inside the AS, due to equipment failures, planned maintenance, or traffic engineering. These events affect the prefix-to-egress mapping because the intradomain path costs play a role in the BGP decision process through the common practice of *hot-potato routing*.

When selecting a best BGP route, a router first considers BGP attributes such as local preference, AS path length, origin type, and multiple exit discriminator. If multiple "equally good" routes remain, the router selects the route with the "closest" egress point, based on the intradomain path costs. Since large ISPs typically peer with each other in multiple locations, the hot-potato tie-breaking step almost always drives the final routing decision for destinations learned from peers, although this is much less common for destinations advertised by customers. In the example in Figure 1, a link failure might make router i's intradomain path cost to e suddenly *larger* than the path to e'. This would change the prefix-to-egress mapping for $p2$, causing a shift in traffic from egress point e to e'. Using the methodology described in [9], we identified which changes in ε were caused by internal events.

Figure 4 shows the cumulative distribution functions of $\Delta\tilde{R}$ caused by hot-potato routing and by external BGP changes. For comparison, we also present the CDF of a normal distribution, which is drawn from randomly generated Gaussian data with standard deviation equal 1. Although the routing events are rare (only 0.66% of non-zero $\Delta\widetilde{\mathcal{T}M}$ are caused by eBGP changes and 0.1% by hot-potato changes), this result shows that there are significant cases where these events are big, to very big. In particular, approximately 5% of traffic shifts caused by hot-potato routing are at least one order of magnitude bigger than normal variations. A single internal change is more likely to affect a large number of destination prefixes [9], including the popular destinations receiving large amounts of traffic.

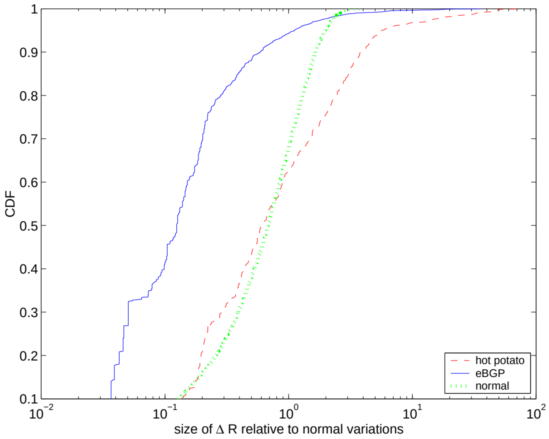

Fig. 4. Cumulative distribution function of $\Delta\tilde{R}$ caused by hot-potato routing and eBGP.

By analyzing the source of traffic variation for individual traffic matrix elements, we see that the likelihood of changes in the prefix-to-egress mappings can vary significantly from one ingress router to another. Some traffic matrix elements have no traffic variation caused by routing changes, whereas other have few very large egress shifts. This is because the likelihood of hot-potato routing changes varies significantly from one ingress point to the other [9], depending on the location in the network and the proximity to the various egress points. For our eight ingress points, the fraction of routing changes caused by internal events varies from 1% to 40%. As a result, the likelihood of large traffic shifts caused by hot-potato routing varies significantly from one traffic matrix element to another.

This observation makes the analysis of the impact of routing on the traffic matrix very dependent on where the data are collected. For example, the study in [10] showed that popular destination prefixes do *not* experience BGP routing changes for days or weeks at a time. In addition to studying RouteViews and RIPE BGP feeds, the analysis included iBGP data from two of the eight routers used in our current study. In our analysis, these two routers did not experience many hot-potato routing changes. Had the analysis in [10] analyzed a router that experiences several hot-potato routing changes a day, the conclusions might have been quite different. In fact, hot-potato routing changes can affect a large number of prefixes [9], both popular and not, so we might reasonably expect popular destinations to experience changes in their egress points. To verify this hypothesis, we plan to repeat the analysis of [10] using all eight vantage points.

IV. Conclusion

Our study shows that large traffic variations, while rare, do sometimes happen. Although most routing changes typically do not affect much traffic, routing is usually a major contributor to large traffic variations. This implies that network operators need to design the network to tolerate traffic variations that are much larger than standard traffic variations. In addition, research on traffic engineering and anomaly detection should take into account the impact of routing on the traffic matrix. Since both the traffic demands V and the prefix-to-egress mapping ε are necessary to compute an accurate traffic matrix, we believe it is more accurate to operate on V and ε directly, rather than simply on $\mathcal{TM}$.

Our ongoing work focuses on quantifying the inaccuracies introduced in studies of routing and traffic stability when changes in ε are ignored. We are also studying the duration of the traffic shifts. If traffic shifts are short-lived, then network operators should just over-provision to tolerate them. If they are long-lived, however, adapting the routing protocol configuration may be a better approach for alleviating congestion.

References

[1] J. Cao, D. Davis, S. V. Wiel, and B. Yu, "Time-varying network tomography," *J. American Statistical Association*, December 2000.

[2] A. Medina, N. Taft, K. Salamatian, S. Bhattacharyya, and C. Diot, "Traffic matrix estimation: Existing techniques and new directions," in *Proc. ACM SIGCOMM*, August 2002.

[3] Y. Zhang, M. Roughan, N. Duffield, and A. Greenberg, "Fast, accurate computation of large-scale IP traffic matrices from link loads," in *Proc. ACM SIGMETRICS*, June 2003.

[4] Y. Zhang, M. Roughan, C. Lund, and D. Donoho, "An information-theoretic approach to traffic matrix estimation," in *Proc. ACM SIGCOMM*, August 2003.

[5] A. Lakhina, K. Papagiannaki, M. Crovella, C. Diot, E. Kolaczyk, and N. Taft, "Structural analysis of network traffic flows," in *Proc. ACM SIGMETRICS*, June 2004.

[6] A. Lakhina, M. Crovella, and C. Diot, "Characterization of Network-Wide Anomalies in Traffic Flows," in *Proc. Internet Measurement Conference*, October 2004.

[7] S. Agarwal, C.-N. Chuah, S. Bhattacharyya, and C. Diot, "Impact of BGP dynamics on intra-domain traffic," in *Proc. ACM SIGMETRICS*, June 2004.

[8] Sampled Netflow. `http://www.cisco.com/univercd/cc/td/doc/product/software/ios120/120newf%t/120limit/120s/120s11/12s_sanf.htm`.

[9] R. Teixeira, A. Shaikh, T. Griffin, and J. Rexford, "Dynamics of hot-potato routing in IP networks," in *Proc. ACM SIGMETRICS*, June 2004.

[10] J. Rexford, J. Wang, Z. Xiao, and Y. Zhang, "BGP routing stability of popular destinations," in *Proc. Internet Measurement Workshop*, November 2002.

[11] N. Duffield, C. Lund, and M. Thorup, "Estimating flow distributions from sampled flow statistics," in *Proc. ACM SIGCOMM*, August 2003.