Estimating Routing Symmetry on Single Links by Passive Flow Measurements

Wolfgang John
Chalmers University of Technology, Sweden
johnwolf@chalmers.se

Maurizio Dusi
Università degli Studi di Brescia, Italy
maurizio.dusi@ing.unibs.it

kc claffy
CAIDA, UCSD
San Diego, USA
kc@caida.org

ABSTRACT

The assumption of routing symmetry is often embedded into traffic analysis and classification tools. This paper uses passively captured network data to estimate the amount of traffic actually routed symmetrically on a specific link. We propose a Flow-Based Symmetry Estimator (FSE) – a set of metrics to assess symmetry in terms of flows, packets and bytes, which disregards inherently asymmetrical traffic such as UDP, ICMP and TCP background radiation. This normalized metric allows fair comparison of symmetry across different links. We evaluate our method on a large heterogeneous dataset, and confirm anecdotal reports that routing symmetry typically does not hold for non-edge Internet links, and decreases as one moves toward core backbone links, due to routing policy complexity. Our proposed metric for traffic asymmetry induced by routing policies will help the community improve traffic characterization techniques and formats, but also support quantitative formalization of routing policy effects on links in the wild.

Categories and Subject Descriptors

C.2.2 [**Computer-Communication Networks**]: Network Protocols; C.2.3 [**Computer-communication Networks**]: Network Operations—*Network monitoring*; C.4 [**Performance of Systems**]: Measurement techniques

General Terms

Measurement

Keywords

Internet, network, measurement, traffic, analysis, characterization, classification, flow, routing, symmetry, asymmetry

1. INTRODUCTION

In today's Internet, path stability is not guaranteed, i.e. many nodes along a path offer alternative routes to the same destination. If packet streams between two endpoints follow the same physical links[1] between intermediate nodes for both forward and reverse direction, they are *symmetrically routed*. Otherwise, routing between this pair is asymmetric. A common cause of routing asymmetry is "hot-potato routing", the business practice of configuring traffic crossing one's own network to exit as soon as possible. Another cause is link redundancy within networks or multipath routing. Since routing decisions occur independently for each flow[2], load-balancing may cause different flows destined for the same endpoint to follow different physical links, even if all the intermediate nodes are the same.

Literature on routing asymmetry has mainly considered an end-to-end perspective, inferred by active measurements of delay or path differences between endpoints [1, 2, 3, 4]. To our knowledge, using passive measurement to quantify routing asymmetry observed on a specific link has only received tangential reference [5]. We propose a technique that uses passive measurements [6] to quantify the amount of traffic routed (a)symmetrically on specific network links, in terms of flows, packets and bytes. Using passively captured network data, the Flow-Based Symmetry Estimator (FSE) method provides an effective way to exclude traffic that is canonically asymmetric, such as ICMP traffic or *nonproductive TCP background radiation* [7], allowing a fair comparison of routing symmetry across different links with substantially different traffic decomposition.

Knowledge of the fraction of symmetric flows on specific links is especially important to traffic analysis and characterization tasks, which are often performed on data collected on single measurement points. Researchers and developers often embed an assumption of traffic symmetry in tools and analyses [8, 9, 10], an assumption only safe for stub access links, otherwise quite harmful [11].

We wanted to provide the community with a technique and accompanying open source tool for measuring flow symmetry, as well as raise awareness about macroscopic symmetry characteristics by providing statistics from running such tools over a variety of data. We evaluated our technique on traffic traces from four varied locations (Tier-2 to Tier-1 backbone) in two countries (USA and Sweden) over a period of four years (from 2006 till 2009), to provide a baseline global data set on routing symmetry. Such data

[1]Optical links, generally composed of a pair of unidirectional fibers or wavelengths, are considered as one physical link.
[2]To our best knowledge, most routing is done on a flow- or IP-Pair level in order to minimize jitter and out-of-order packets within sessions.

IWCMC'10, June 28–July 2, 2010, Caen, France.

```
1:  given a time-interval of traffic trace:
2:    consider TCP data traffic (TCP packets w/ data)
3:    T_f (T_b) = set of tuples going forward (backward)
4:    T_f ∩ T_b = set of symmetric tuples T_S
5:    pkts(bytes) in T_S=set of symmetric pkts(bytes)
```

Figure 1: The FSE method. After collecting a unique list of unidirectional flows for each direction of a link, FSE classifies 5-tuples as symmetric if they appear on both lists. Packet (byte)-level symmetry is the fraction of packets (bytes) sent between tuples classified as symmetric, so that the degree of symmetry can be quantified in three dimensions: 5-tuple flows, packets, bytes.

sets will allow tracking of macroscopic Internet trends. Our main contributions are: (i) a simple method to assess and fairly compare routing symmetry on specific links; (ii) an open source tool for analyzing flow symmetry based on our method; and (iii) symmetry statistics for a large heterogeneous set of network traces.

Section 2 explains our choice and implementation of FSE to analyze flow symmetry. Section 3 and 4 describe the data and the results of applying FSE to the data, resp. Section 5 validates the method and Section 6 concludes the paper.

2. FLOW-BASED SYMMETRY ESTIMATOR

In this section we present the *Flow-based Symmetry Estimator (FSE)*, a simple method (depicted in Figure 1) and associated tool[3] to estimate the level of routing symmetry from passively measured flow data that takes unidirectional 5-tuple flow data as input. We could have computed symmetry based on IP pairs (2-tuples), but most traffic classification and engineering methods deal with flows [8, 9, 10], so we chose the flow granularity. We used CoralFlow (part of CoralReef [12]) to extract interval-based 5-tuples of source and destination IP, port numbers and protocol. Due to its simplicity, most traffic analysis tools [13] prefer this method to tracking TCP connection state, although we use TCP connection information extracted from packet level-data [5] to validate our technique in Section 5.

2.1 Removing inherently asymmetric traffic

Our first step is to remove from the traces any traffic that is inherently asymmetric, such as UDP and ICMP flows that do not always expect packet recipients to reply[4], and which would mislead symmetry comparisons if they appear in different magnitudes across networks. TCP background radiation, such as network scanning and probing, can also be a substantial fraction of total inherently asymmetric flows on some links, although it is usually a much lower proportion of bits [7, 15]. FSE discards ICMP, UDP, and TCP signaling packets with no data. As a heuristic for the TCP category, we keep only TCP packets without signaling flags (SYN/FIN/RST) but with the ACK bit set, thereby removing unreplied single-packet probes, scans, or attacks using SYN, FIN, or RST flags. We call the post-filtered data *TCP data traffic*, reflecting the dominant transport activity on the Internet [16, 17], at least so far.

[3]Available at http://www.cse.chalmers.se/∼johnwolf/FSE/

[4]While many application protocols communicate in bidirectional request/respond fashion over UDP (e.g. DNS), related work has shown that UDP flows on some links are dominated by single-packet flows with no observed response, such as P2P signaling and unsolicited traffic (scanning, DoS) [14].

2.2 Observation time interval

We use CoralFlow to create flow 5-tuples for a given observation interval. CoralFlow defines flows by timeout interval, i.e., two packets sharing the same tuple belong to the same flow if their timestamps are within a given time interval. CoralFlow splits traces into chunks according to the specified time interval and collates unique lists of 5-tuples for each direction. The results might be affected by border effects, i.e. long flows spanning many intervals, or short symmetric flows that seem asymmetric because packet exchange occurs at the edge of an interval. We will evaluate these effects by varying the time interval, described in Section 4.2.

3. DATASETS

Table 1 lists the packet-level datasets we considered. The data from GigaSUNET was collected on a backbone close to the edge of the Internet, on an OC192 link which was the primary link from the region of Gothenburg to the main Internet outside Sweden. The link mainly carried traffic from major universities and large student residential networks, but also from a regional access point exchanging SUNET traffic with local ISPs. TCP was responsible for 42% of flows, which corresponded to 93% of packets and 97% of bytes. UDP carried 55% of flows (6% of packets and 3% of bytes). Other transport protocols, such ICMP, GRE and ESP, represented minor traffic amounts.

In the current OptoSUNET, customers are redundantly connected to a central Internet access point. Besides some local exchange traffic, the traffic routed to the international commodity Internet is carried on two links (40Gb/s and 10Gb/s) between SUNET and NorduNet. The data used in this study was collected on the 10Gb/s link, which according to SNMP statistics carried 50% of all inbound but only 15% of the outbound traffic volume. Around 20% of flows on the link during the measurement interval were exchanged via TCP, corresponding to 82% of packets and 89% of bytes, while 79% of connections (16% of packets, 9% of bytes) were UDP flows.

The two core links are part of an OC192 Tier1 backbone operated by a commercial ISP in the U.S. The first link connects Chicago and Seattle, monitored at an Equinix datacenter in Chicago. The other one connects San Jose and Los Angeles, monitored at a datacenter in San Jose. On those links, TCP is responsible for about 50% of flows, which was 85% of packets and 93% of bytes on average. UDP carried about 45% of flows (13% of packets and 6% of bytes).

4. EXPERIMENTAL RESULTS

We apply FSE to the datasets of Table 1 and discuss the impact of traffic composition, observation interval and flow granularity on routing symmetry estimation.

4.1 Impact of inherently asymmetric traffic

To evaluate the impact of flows that are inherently asymmetric on traffic symmetry estimates, we first apply the method to all IP traffic, then on TCP traffic (i.e. disregarding UDP, ICMP and other traffic) and finally on the proposed category: TCP data traffic. The last category

excludes nonproductive, inherently asymmetric TCP background radiation. Table 2 provides the excluded TCP-signaling fractions, a reasonable estimate for the amount of (asymmetric) TCP background radiation on our links, consistent with other studies [7, 15].

Figure 2a provides box-plots[5] of flow-based symmetry estimates (FSEs) for 10-minute samples of traffic filtered in three ways. Due to space constraints we only show symmetry in terms of flows and bytes. As expected, the fractions of symmetric tuples increase when excluding inherently asymmetric traffic (e.g. from a median of 53% to 69% for GigaSUNET 2006-04 and from 2.7% to 5.5% for Eq-Chicago 2008-05). But the filtering operation only slightly affects symmetry in terms of bytes (e.g., from 8.7% to 9.0%) and packets (e.g. from 73% to 74%, not shown here), since packets carrying TCP signaling flags are a minor fraction of the total TCP packets and typically carry no data (see Table 2).

Figure 2a also suggests that the degree of routing symmetry radically decreases as we move toward the core of the Internet. On GigaSUNET, inside a Tier2 network close to the edge of the Internet, most traffic we observed was routed symmetrically (around 70%). The asymmetric traffic fraction here is caused by hot-potato routing due to local peering and the underlying ring architecture which does not guarantee shortest-path transport. One step closer to the core, on the OptoSUNET link connecting a Tier2 to a Tier1 network, only about 10% of the observed flows were symmetrical. On this link asymmetry can be explained by the load-balancing policy applied on the redundant route between SUNET and NorduNet (see Section 3) as well a regional exchange point introducing some hot-potato routing. On the two Tier1 ISP backbone links, hot-potato routing and other peering artifacts in aggregation induce high asymmetry: only 4-5% of tuples generate traffic routed symmetrically.

4.2 Impact of observation intervals

The observation interval used for the analysis impacts flow, and thus symmetry, assessments. Short intervals introduce border effects, such as causing short symmetric flows to seem asymmetric if packet exchange occurs at the edge of an interval. Large intervals increase the probability of incorrectly aggregating multiple sessions with identical 5-tuples into one flow within the interval.

To evaluate the impact of these effects, we split each traffic trace into feasible chunks[6] of 1, 5, and 10-minutes, and apply FSE to filtered TCP data traffic within each observation interval. Figure 2b shows box-plots of the FSEs, reflecting symmetrically routed traffic in terms of tuples and bytes for each time interval (we omit packets again). Observation intervals shorter than 10 minutes have little effect on routing symmetry estimates, which are stable (low interquartile-range) over the entire dataset samples (six 10-minute samples across one month for SUNET data, and within one continuous hour for Equinix data). Moreover, we observe that the symmetry estimate computed on TCP data traffic remains stable on each location over time (comparing FSEs of data samples separated by seven months for GigaSUNET, two months for the other locations), and this observation also holds for all IP traffic as well as for all TCP traffic (not shown here).

In recent work [18], Lee and Brownlee studied traces measured during 24 hours on the network boundary of the University of Auckland in 2006, and showed that around 98% flows last less than 10 minutes. In the rest of this paper we will consider 10-minute samples, which minimize border effects but represents a meaningful statistical data sample.

[5]Boxes represent median, lower and upper quartile, plus whiskers and outliers.

[6]intervals >10min on large backbone traces may exhaust memory (e.g. 10min of SanJose0807: 2.7GB for 23M flows).

Table 3: Mean FSEs computed by considering TCP data traffic exchanged between 2-tuples (IPpairs) and 5-tuples (TCP flows), and how this aggregation granularity affects FSEs. Higher symmetry values in the IPpairs follow from the fact that the method counts all traffic generated by two 5-tuples with the same source and destination IP as symmetric even if only one 5-tuple is actually observed as symmetric. In fact, the total number of packets (bytes) remains unchanged regardless of granularity. In terms of tuples, traffic granularity affects the degree of symmetry, depending on the fraction of flows that share the IP pairs.

TCP-data traffic 10-min samples		% of tuples		% of packets		% of bytes	
		flow	IPpair	flow	IPpair	flow	IPpair
GigaSUNET	06-04	69.4	79.4	73.6	73.7	73.9	73.9
	06-11	72.3	77.9	78.1	78.1	77.9	77.9
OptoSUNET	09-01	10.1	10.7	25.3	25.4	33.8	33.9
	09-02	10.9	11.7	24.5	24.6	34.7	34.8
Eq-Chicago	08-04	4.0	3.3	9.0	10.3	9.6	11.6
	08-05	5.5	5.2	9.9	11.8	9.0	11.7
Eq-SanJose	08-07	4.1	3.5	9.3	11.8	11.0	13.8
	08-08	3.6	4.2	10.7	14.0	12.7	16.3

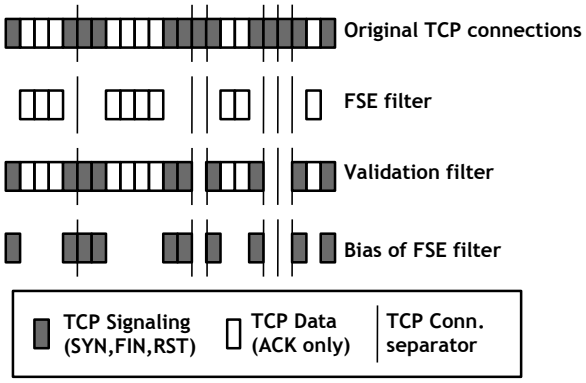


Figure 3: FSE removes purely signaling and scanning packets prior to flow creation. The validation method filters out TCP background radiation by retaining only connections with at least one non-signaling packet.

4.3 Impact of traffic granularity

In this subsection we compare routing symmetry between two levels of traffic granularity: IP pairs, more relevant to routing questions [2]; and flows, more relevant to traffic analysis and classification techniques [8, 9, 10].

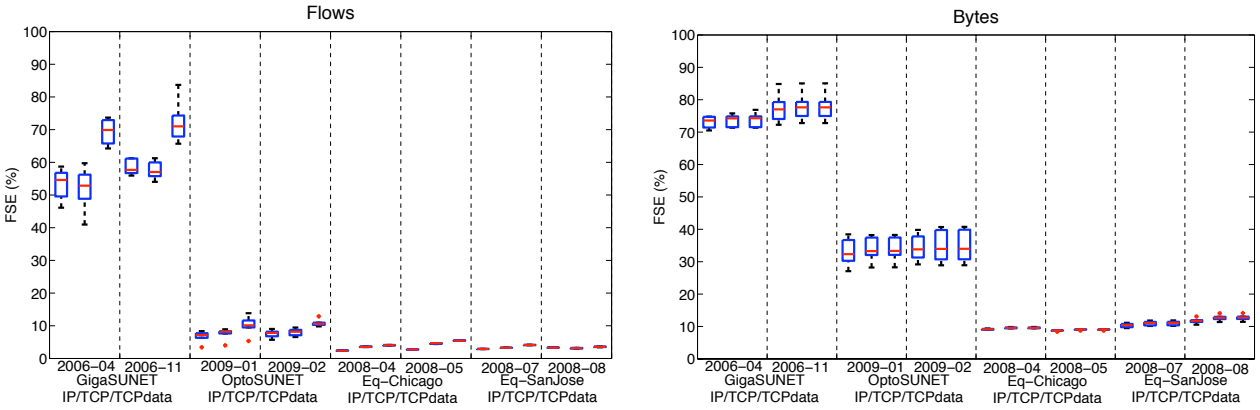
Table 3 lists the mean values of the FSE metric calculated for 10-minute observation intervals of TCP data traffic. In terms of packets and bytes, IP pairs (which have higher levels of aggregation) often exhibit higher symmetry, indicating that flows between the same IP pairs may follow different paths.

Table 1: **Dataset description: Two datasets are from OC192 links in Swedish networks - GigaSUNET, operative until 2007, and OptoSUNET's current connection to NorduNet. The latter two are from OC192 backbone links of a Tier1 ISP in the U.S.**

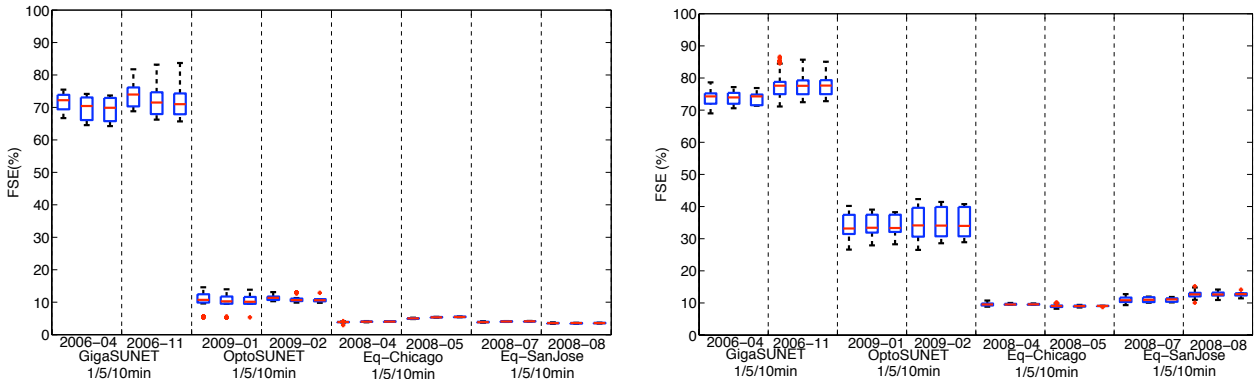
		Time interval	#flows	pkt/s	bit/s	Network loc.
GigaSUNET	2006-04	6x10min	8.9M	142Kp/s	790Mbit/s	Tier2 backbone
	2006-11	6x10min	15.6M	176Kp/s	1008Mbit/s	(Sweden)
OptoSUNET	2009-01	6x10min	57M	358Kp/s	1700Mbit/s	Tier2-Tier1 connection
	2009-02	6x10min	62M	442Kp/s	2000Mbit/s	(Sweden)
Eq-Chicago	2008-04	1x1hour	119M	717Kp/s	3970Mbit/s	Tier1 backbone
	2008-05	1x1hour	134M	936Kp/s	6100Mbit/s	(Illinois-Washington)
Eq-SanJose	2008-07	1x1hour	145M	680Kp/s	3000Mbit/s	Tier1 backbone
	2008-08	1x1hour	139M	664Kp/s	3040Mbit/s	(California)

Table 2: **TCP traffic carrying only signaling packets, as removed by the TCP data filter. The numbers are good estimates for the amount of nonproductive TCP background radiation on the links.**

Dataset		% flows	% packets	% bytes	Dataset		%flows	% packets	% bytes
GigaSUNET	2006-04	32.36	4.85	0.15	Eq-Chicago	2008-04	19.19	5.60	0.51
	2006-11	27.86	1.95	0.15		2008-05	23.62	4.31	0.34
OptoSUNET	2009-01	34.81	2.05	0.08	Eq-SanJose	2008-07	25.27	8.04	0.83
	2009-02	34.74	2.05	0.09		2008-08	19.41	7.75	0.78



(a) **Comparison of FSEs for IP, TCP and TCP-data flows (left) and bytes (right): statistics for 10-minute samples. FSEs increase when excluding inherently asymmetric traffic. FSEs in terms of bytes and packets (not shown here) are only slightly affected. Routing symmetry decreases towards the core of the Internet (from GigaSUNET close to the edge to Tier-1 backbone links Eq-Chicago and Eq-SanJose).**



(b) **Comparison of FSEs for TCP-data flows (left) and bytes (right): statistics for 1, 5 and 10-minute samples. Observation intervals have little effect on FSEs. FSEs are relatively stable both over short periods (low interquartile-ranges) and over long periods (between different months on the measured links).**

Figure 2: Box-plots of flow-based symmetry estimates (FSEs).

5. VALIDATION

This section validates our FSE metric against an approach using explicit TCP flags to distinguish bidirectional sessions, as described in John *et al.* [5]. The validation method considers TCP traffic in both directions, inspecting TCP-signaling flags (SYN, FIN, RST) to distinguish TCP flows. We define the percentage of symmetric 5-tuples as the fraction of connections with at least one packet for each direction. The amount of packets (bytes) carried by symmetric tuples yields packet- (byte-)symmetry. This flow definition classifies scanning behavior that re-uses 5-tuples as a series of 1-SYN-packet flows, while many common timeout-based flow definitions [12, 13] (often used as input for traffic classification tools [9, 10]), will label it as a single flow with multiple SYN packets. Figure 3 outlines the difference between FSE and the validation method applied to original TCP connections. The validation method filters out TCP background radiation by retaining only connections with at least one non-signaling packet. FSE filters out all signaling packets prior to flow creation. The filter discards scanning traffic, reducing the size of legitimate TCP sessions by its signaling packets and the respective header data.

Validation performed on a smaller validation dataset of one 10-minute interval from each dataset in Table 1 revealed that the interval-based flow definition as applied in FSE led to significant underestimation (between 14% and 31%) of the number of TCP connections. This underestimation derives from our aggregation of TCP connections into one flow if the exact five-tuple is re-used within the timeout interval. However, when considering (filtered) TCP data traffic, the underestimation is much slighter, 0.15%-0.45%. Table 4 shows the small impact of the FSE filter on symmetry assessments. These results indicate that legitimate TCP traffic (i.e. connections including SYN packets, data packets and RST/FIN termination), in contrast to TCP background radiation (often consisting of one signaling packet like SYN only), rarely reuses the same five-tuple for connections within 10 minutes, which demonstrates the utility of the proposed traffic filter. This fact further suggests that FSE is robust against varying flow definitions (i.e., timeout-based vs. signaling-based), at least for intervals less than 10 minutes.

In terms of packets and bytes, the validation shows that their absolute numbers are slightly higher than FSE estimated, since FSE aggressively discards signaling packets (see Figure 3). Table 5 shows this discrepancy during a ten-minute interval. On the complete validation dataset, FSE removed 1-7% of TCP packets, corresponding to 0.1-0.6% of bytes, before computing its symmetry estimates. However, this bias in absolute numbers has negligible effect on corresponding symmetry estimates, which shows the validity of the estimation.

Using the validation method to characterize background radiation in the datasets (quantified in Table 2), we can confirm that in our data background radiation is indeed mostly asymmetric: it is mainly composed of 1-pkt flows. Between 85% and 95% of the discarded connections are 1-SYN-pkt flows. Verification of the number of ICMP destination unreachable packets shows that no more than 15% of the 1-SYN-pkt TCP flows receive ICMP packets in response. If we did not remove these sources of strong bias from the symmetry estimate, even exclusive access links (100% symmetric) could be erroneously perceived as having substantial routing asymmetry.

Table 4: Flow level symmetry by FSE (F) vs. validation method (V) for all TCP [left] and TCP data [right] traffic. Flow symmetry differs greatly for all TCP traffic, but negligibly for TCP data traffic. Thus, TCP data traffic is robust against the different flow definitions (timeout vs flags).

10-min sample		TCP all		TCP data	
		F (%)	V (%)	F (%)	V (%)
GigaSUNET	2006-04	48.9	41.9	65.8	64.7
	2006-11	55.8	42.0	74.1	73.9
OptoSUNET	2009-01	8.0	6.8	9.7	9.7
	2009-02	9.5	7.8	12.9	12.8
Eq-Chicago	2008-04	3.5	3.0	3.9	3.9
	2008-05	4.7	4.0	5.6	5.5
Eq-SanJose	2008-07	3.3	3.0	4.2	4.2
	2008-08	3.2	3.0	3.7	3.7

Table 5: FSE (F) vs. validation method (V). A small bias is introduced by the FSE TCP-data filter when discarding signaling packets. However, symmetry estimates are hardly affected.

Eq-Chicago 2008-05

Packets

	Sym.	Tot.	Sym.%
V	47.2M	469.8M	10.05%
F	45.7M	455.5M	10.04%
Diff.	1.5M	14.3M	

Bytes

	Sym.	Tot.	Sym.%
V	39.4G	433.6G	9.09%
F	39.3G	432.5	9.09%
Diff.	0.1GB	1.1G	

To further validate our estimation method (FSE), we collected two 10 minute traffic samples on the 100Mb/s single access link which connects the edge router of the University of Brescia to the Internet [19]. Each sample includes about 60 thousand flows, carrying 3.5GB of data, with TCP responsible for about 43% of the flows and 98% of the data volume. Traffic that flows on this link is 100% symmetric, i.e. all outgoing and incoming packets follow this link, so this data can serve as ground truth to assess the effectiveness of the FSE mechanism. Estimating flow symmetry based on all IP traffic on the link resulted in an FSE of only 79%. Considering TCP traffic resulted in an FSE of 84%, which is closer to ground truth (100%) but still a significant underestimation. However, when assessing routing symmetry on our proposed category of TCP data traffic, FSE for flows resulted in >98%, and almost 100% of bytes and packets (>99.99%). These estimates are very close to ground truth and thus highlight the bias introduced if TCP background radiation is not discarded during flow symmetry estimation. The remaining underestimation of <2% of flows, which the FSE erroneously classified as asymmetric, can be attributed to border effects due to the observation interval. Specifically, connections established/terminated just before/after the interval, which happen to send only one data packet within the interval, appear as asymmetric flows. Since this link carries relatively little P2P traffic (around 10%) [19], thus also little P2P signaling traffic (1-pkt UDP flows) [14], we believe that the positive effect of the TCP data filter could be even stronger for other links with more inherently asymmetric traffic.

6. SUMMARY AND CONCLUSIONS

In order to shed light on the assumption of routing symmetry often embedded into traffic analysis and classification methods, we provided insight into symmetric routing on a flow granularity using observations from a variety of Internet links. We do so by proposing a simple flow-based symmetry estimation method, FSE, providing a normalized metric allowing to assess and compare routing symmetry of links on flow level by utilizing passive measurements. We provide an open source tool implementing the proposed method, and apply it to a large heterogeneous dataset, resulting in valuable reference data points on routing symmetry.

We designed FSE to leverage available tools providing traditional, timeout-based 5-tuple flows (e.g. CoralFlow). Since TCP is an inherently bidirectional protocol and still the dominant protocol carrying traffic on today's observable Internet, we established a TCP-based metric. We further filtered out the inherently asymmetric TCP traffic (i.e. TCP background radiation), leaving only TCP packets without signaling flags. This process allows for fair comparison of symmetry across links with substantially different traffic decomposition.

We did use TCP signaling flags to validate our simplified metric against ground truth measurements, allowing us to demonstrate that our flow-based symmetry estimate (FSE) is robust against multiple flow definitions. We quantified the small bias of the filter and confirmed that most of the filtered nonproductive flows are asymmetric, carrying one packet only. A validation of the FSE method on ground truth data showed that non-filtered data results in strongly biased symmetry estimates, where even exclusive access links (100% symmetric) could be erroneously perceived as having substantial routing asymmetry.

We also found that in the data we examined, spanning over four years, four measurement locations on two continents, 5-tuples carrying legitimate TCP data traffic are rarely reused within ten-minutes observation intervals. Shorter observation intervals do not significantly alter symmetry estimates. Aggregating traffic by IP pairs instead of flows often results in greater symmetry. Unsurprisingly, routing-based symmetry seems to be stable over hours and even months, and decreases as one moves from edge links to highly aggregated backbone, which also hinders examination of complete, bidirectional flows on a single link. This result implies that traffic analysis tools and methods (e.g. traffic classification) should assume little routing symmetry unless intended only for stub access links with no path diversity.

Acknowledgments

The authors would like to thank Emile Aben for valuable discussions and Luca Salgarelli and Tomas Olovsson for useful feedback and comments. This research was supported in part by SUNET, the Swedish University Network, as well as DHS PREDICT Contract NBCHCC040159, NSF CONMI CRI-0551542, and CAIDA members.

7. REFERENCES

[1] Z. M. Mao, L. Qiu, J. Wang, and Y. Zhang, "On AS-level Path Inference," in *ACM SIGMETRICS*, Banff, Alberta, Canada, 2005.

[2] Y. He, M. Faloutsos, and S. Krishnamurthy, "Quantifying Routing Asymmetry in the Internet at the AS Level," in *IEEE GLOBECOM*, 2004.

[3] S. Savage, A. Collins, E. Hoffman, J. Snell, and T. Anderson, "The End-to-end Effects of Internet Path Selection," *SIGCOMM Computer Communication Review*, vol. 29, no. 4, 1999.

[4] V. Paxson, "End-to-end Routing Behavior in the Internet," *IEEE/ACM Transactions on Networking*, vol. 5, no. 5, 1997.

[5] W. John and S. Tafvelin, "Differences between In- and Outbound Internet Backbone Traffic," in *TERENA Networking Conference*, Copenhagen, DK, 2007.

[6] W. John, S. Tafvelin, and T. Olovsson, "Passive Internet Measurement: Overview and Guidelines based on Experiences," *Computer Communications*, vol. 33, no. 5, pp. 533 – 550, 2010.

[7] R. Pang, V. Yegneswaran, P. Barford, V. Paxson, and L. Peterson, "Characteristics of Internet Background Radiation," in *ACM Internet Measurement Conference*, Taormina, Sicily, Italy, 2004.

[8] L. Bernaille, R. Teixeira, and K. Salamatian, "Early Application Identification," in *ADETTI/ISCTE CoNEXT*, Lisboa, Portugal, 2006.

[9] A. McGregor, M. Hall, P. Lorier, and J. Brunskill, "Flow Clustering Using Machine Learning Techniques," in *Passive and Active Measurement Conference*, Antibes Juan-les-Pins, France, 2004.

[10] S. Zander, T. Nguyen, and G. Armitage, "Automated Traffic Classification and Application Identification using Machine Learning," in *IEEE Conf. on Local Computer Networks (LCN)*, Sydney, Australia, 2005.

[11] M. Crotti, F. Gringoli, and L. Salgarelli, "Impact of Asymmetric Routing on Statistical Traffic Classification," in *IEEE GLOBECOM*, Honololou, Hawaii, USA, 2009.

[12] K. Keys, D. Moore, R. Koga, E. Lagache, M. Tesch, and k claffy, "The Architecture of CoralReef: An Internet Traffic Monitoring Software Suite," in *Passive and Active Measurement Workshop*, Amst.,NL, 2001.

[13] B. Claise, "Cisco Systems NetFlow Services Export Version 9," RFC 3954 (Informational), Internet Engineering Task Force, Oct. 2004. [Online]. Available: http://www.ietf.org/rfc/rfc3954.txt

[14] W. John, S. Tafvelin, and T. Olovsson, "Trends and Differences in Connection-behavior within Classes of Internet Backbone Traffic," in *Passive and Active Measurement Conference*, Ohio, USA, 2008.

[15] M. Allman, V. Paxson, and J. Terrell, "A Brief History of Scanning," in *ACM Internet Measurement Conference*, San Diego, USA, 2007.

[16] W. John and S. Tafvelin, "Analysis of Internet Backbone Traffic and Header Anomalies observed," in *ACM Internet Measurement Conference*, San Diego, CA, USA, 2007.

[17] "Real Time CoralReef Report Generator," http://www.caida.org/data/realtime (accessed 2010-01-14).

[18] D. Lee and N. Brownlee, "Passive Measurement of One-way and Two-way Flow Lifetimes," *SIGCOMM Computer Communication Review*, vol. 37, no. 3, 2007.

[19] A. Este, F. Gringoli, and L. Salgarelli, "Support Vector Machines for TCP Traffic Classification," *Computer Networks*, vol. 53, no. 14, 2009.