Networks Modelling Using Sampled Statistics

Hamed Haddadi and Lionel Sacks

University College London

Abstract: The fast increase in deployment of Wide Area Networks (WAN) and global enterprise and research networks has increased the demand for network measurement and monitoring applications to generate realistic statistics of the status of networks, help in resource management, billing and provisioning and most importantly traffic characterisation. These enable the network operator's to have a near real-time view of the network, be aware of any attempts to break into the networks via detecting anomalies, and be able to keep the Quality of Service (QoS) service levels. The network traffic traces recently made available to public by projects such as DANTE [1]. In this paper the author presents a novel method of realistic network simulation in different scales in such a way that it resembles the collected statistics from a real network. This simulation method is then used to analyse the effects of sampling on detailed network statistics.

1 Introduction to Flow Characterisation

Flow characterisation and packet classification is an area which is closely related to any monitoring activity. Network operators, and exceedingly users of services, would like to know how their network or connection to provider is utilised, what applications are consuming the bandwidth and what ports they operate on. This is a major area of research and it leads to activities such as billing, network tomography and anomaly detection. A better understanding of the nature and origin of flow rates in the Internet is important for several reasons. First,to understand the extent to which application performance would be improved by increased transmission rates, we must first know what is limiting their transmission rate. Flows limited by network congestion are in need of drastically different attention than flows limited by host buffer sizes. Further, many router algorithms to control per-flow bandwidth algorithms have been proposed, and the performance and scalability of some of these algorithm depends on the nature of the flow rates seen at routers. Thus, knowing more about these rates may inform the design of such algorithms. Finally, knowledge about the rates and their causes may lead to better models of Internet traffic. Such models could be useful in generating simulation workloads and studying a variety of network problems and are studied in [2]. The authors claim their findings confirm what has been observed previously, the distribution of flow rates is skewed, but not as highly skewed as flow sizes and that flow rates strongly correlated with flow sizes.

2 Modelling Networks by Measurement Statistics

Internet and the networks connected to it are constantly subject to topological changes, bandwidth upgrades, new applications and new threat models. These trends have all contributed to the sudden increase in interest in network measurement and monitoring. It is only by thorough analysis of packets traversing the networks, their source, destination, protocols and distribution of sizes that network operators can enable provisioning of networks and facilities for future applications and growth in networks. On the other hand, it is becoming increasingly difficult to obtain traffic traces. Network operators are extremely cautious in giving access to researchers. This is even more the case in enterprise networks, where more complicated techniques such as MPLS and VPN provisions make it more complicated to analyse the source and destination of packets even when access to routers are gained. However it is possible to model the simulations based on the data available from research networks such as DANTE [1]. Figure 1 displays the destination IP address prefixes that have been recorded to be access form the point of presence in Cambridge, UK, on 24^{th} of November 2004.

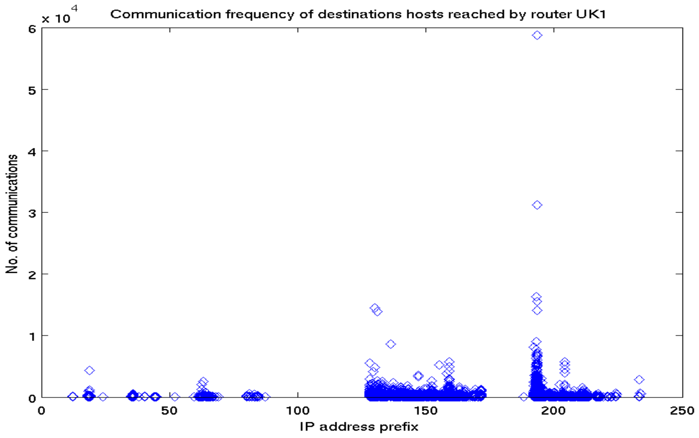

Figure 1: The number of destination IP addresses reached to by UK router

Albeit it has to be noted that these statistics are all based on NetFlow exported statistics and hence are biased against small flows, many of which may be missed. These will be discussed in more detail in the next section. Figure 2 include the source IP address prefixes of hosts on DANTE behind the UK router, trying to get access to the rest of the world.

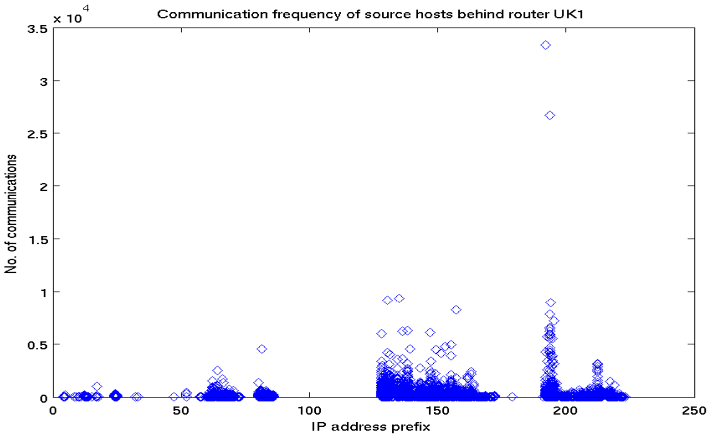

Figure 2: The number of source IP addresses originating traffic from UK router

3 Flows and Sampling

A flow of traffic is a set of packets with a common property, known as the flow key, observed within a period of time. Many routers construct and export summary statistics on packet flows that pass through them. Ideally, a flow record can be thought of as summarising a set of packets that arises in the network through some higher level transaction, for example, a remote terminal session or a web page download. Flow statistics are collected at routers and exported periodically by the router.

The main resource constraint for the formation of flow statistics is at the router flow cache. To perform lookup of packet keys and counter increment at line rate would require the flow statistics to be stored in fast memory. However, core routers will carry increasingly large number of concurrent flows, necessitating large amount of fast memory: this would be expensive. By sampling the packet stream in advance of the construction of flow statistics, the time window available for flow cache lookup is prolonged, enabling storage to be carried out in slower, less expensive, memory.

These collected statistics are the only resource available to the researchers looking at the core of the network which allows them to analyse the network performance. As an example, figure 3 displays the number of flows originated from the Austrian nodes of the DANTE network on 24^{th} of November 2004.

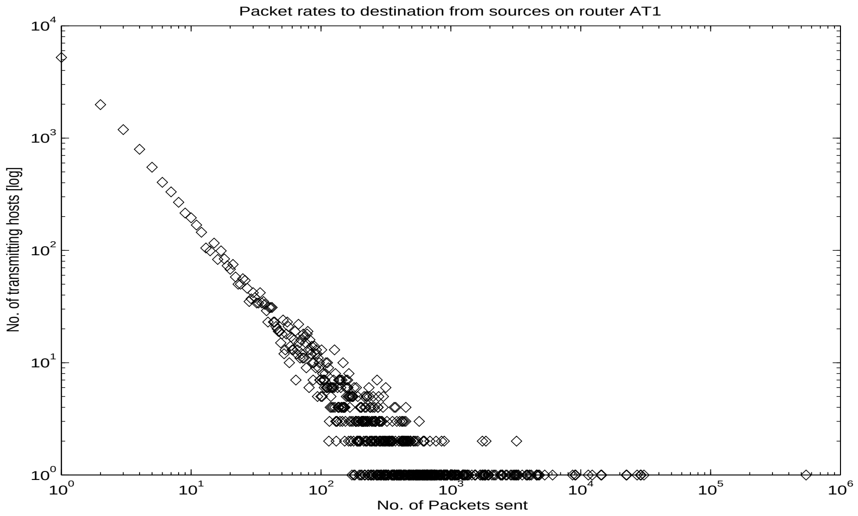

Figure 3: CDF of packets per flow, original file

Figure 4 shows the probability distribution function of the number of packets sent on the Slovakian point of presence of Dante, which carries the least traffic compared to the other locations. it can be seen that a major part of the traffic is comprised of smaller packets, and this would have been more the case if sampling had not been applied at the router, which inevitably misses many of the small flows.

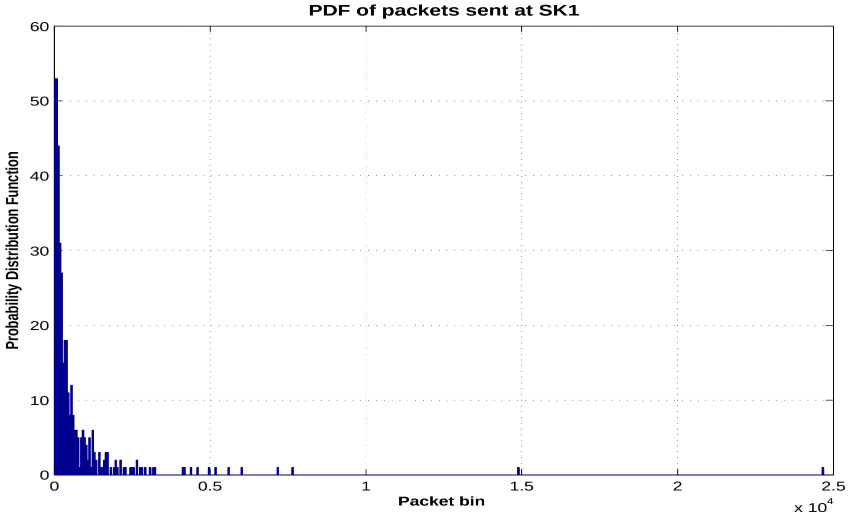

Figure 4: CDF of packets per flow, original file

Figure 5 displays the cumulative distribution function of the number of packets sent by the Slovakian router. This CDF is used to generate the simulation parameters for the network modelling.

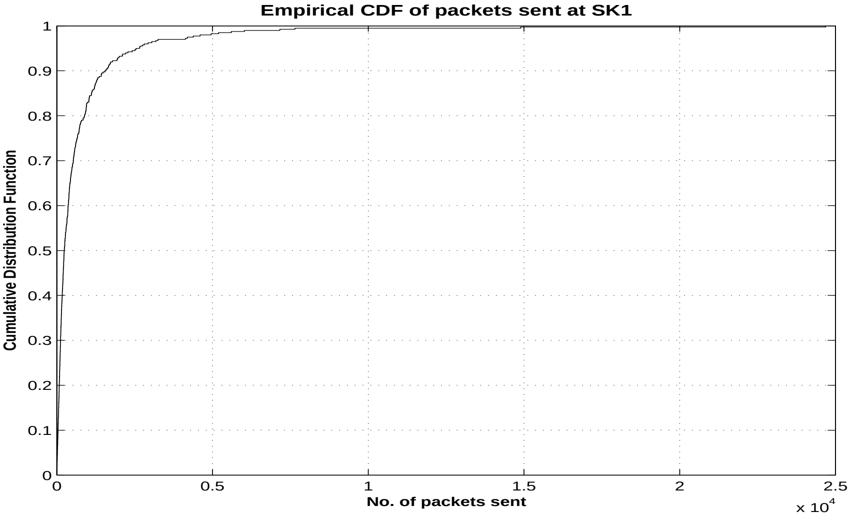

Figure 5: CDF of packets per flow, original file

4 Conclusions

To date there has been little work on recovering detailed properties of the original unsampled packet stream, such as the number and lengths of flows. It may be advantageous to adjust flow delineation criteria with sampling rate in order to match the flow definition to the underlying nature of the transactions that generate the traffic.

In an adaptive sampling scheme, if the sampling rate is 1 in N, based on the average number of packets per flow in a given interval τ, we can predict the traffic characteristics and set the sampling rate accordingly to capture more small flows, or to increase the time-out in order to get full-lengths of large flows. The simulation has been modified to allow end nodes capture their own packet trace and store them. The on-going in this context will enable sampling methods which will enable inferring of the original traffic profile from a given subset.

Acknowledgements

This work was done while author was visiting Intel Research Cambridge. The authors would like to acknowledge Gianluca Iannaccone of Intel Research Cambridge and Richard Mortier of Microsoft Research Cambridge for help and advice on obtaining the datasets and design of the simulation tool.

References

[1] DANTE (Delivery of Advanced Network Technology to Europe): `http://www.dante.net/`

[2] Y. Zhang, L. Breslau, V. Paxson, and S. Shenker. On the characteristics and origins of internet flow rates. In Proceedings of ACM Sigcomm, August 2002

[3] Will Leland, Murad Taqqu, Walter Willinger, and Daniel Wilson, ”On the Self-Similar Nature of Ethernet Traffic (Extended Version)”, IEEE/ACM Transactions on Networking, Vol. 2, No. 1, pp. 1-15, February 1994

[4] Cisco, Netflow services and applications, Available from `http://www.cisco.com`