Some Remarks to Recent Papers on Traffic Analysis

or: the case for public wiki-like platforms for commenting published papers

Fabio Ricciato
Forschungszentrum Telekommunikation Wien
Donau City Straße 1, Vienna, Austria, EU
ricciato@ftw.at

ABSTRACT

In this informal contribution we raise a couple of remarks and requests for clarification about some recent papers in the field of traffic analysis. These cases are illustrative of the kind of issues and open points that are encountered when reading, applying and working with published papers. The readers and followers of each published paper - especially of the best ones - form naturally a small community of interest. In most cases the remarks to the paper are of interest for them all. Based on these considerations we raise the following proposal to the research community: let each conference and/or journal editor maintain an open public wiki-like commenting platform for publishing comments and rebuttals after the paper publication.

1. INTRODUCTION

Research and science are collective enterprises. A researcher, say Bob, starts from the achievements of another researcher, say Anne, and makes one step forward (or at least tries to). In turn another researcher, say Carle, will follow building upon Bob's achievements. In theory, publishing papers into conferences and journals is one of the primary *means* of research, not the primary *goal* [1].

In our research, like Bob, we seek to apply methods and techniques that were previously published as a pre-condition to improve them further or simply to derive a real-world application out of them. In some cases we step into points deserving further clarification by the authors, or simply come up with remarks that might be useful to other readers as well. What to do then ? Following a bottom-up approach we first expose a few illustrative examples about recent published works. From there we derive more general lessons to be learned and make a concrete proposal: let each conference and/or journal editor maintain an open public wiki-like commenting platform for publishing comments and rebuttals after the paper publication.

2. A PROBLEM WITH LD PLOTS

Since the seminal work by Abry and Veitch [15] wavelet analysis has become an important tool for traffic analysis. There is now a long sequel of papers using Log-Diagram Plots (LDP for short) to summarize statistical traffic properties over many time-scales. The LDP reduces the whole time-series, typically a single realization of a complex traffic process, to a vector of 10-20 values. Some caution is due when interpreting LDP. It is hazardous to claim that two traffic traces have the same statistical properties based solely on the similarity of their respective LDP. For instance it has been shown that some artificial signals with particular forms of non-stationarity might well emulate the "typical" LDP behavior found in real traffic traces (see [18] and [4]). Nevertheless the fact that certain LDP patterns are recurrent in traffic traces captured on different networks suggests the presence of some invariant features in data traffic.

A simple test. In order to check the LDP signature of our traces from an operational 3G mobile network [1] we adopted the MATLAB scripts available from [5]. These scripts have been used by several different authors in a huge number of previous works, therefore they are supposedly well tested. Before peforming the analysis on the real traces we did a bit of preliminary training, playing with some simple synthetic signals in order to acquire confidence with the tool and "educate the eye" to look at time-series through the glasses of LDP.

During such preliminary stage we run the scripts over the following synthetic signal. We consider a flow arrival process where flow arrivals are Poisson with intensity λ and flow durations are Pareto distributed with parameters k (shape factor) and α (tail index). The signal of interest, denoted by $X(k)$, is the number of parallel flows sampled at equally-spaced instants t_k. According to the theory [11, p. 510] such process is *monoscaling* with Hurst parameter

$$H = (3 - \alpha)/2 \tag{1}$$

where α is the shape parameter (tail index) of the ON duration distribution. Accordingly the LDP should yield a straight line with slope H over all time-scales. In order to verify that we performed some simple MATLAB simulations. We considered two different process with different parameters for the flow duration: X_1 with $\alpha_1 = 1.25$, $k_1 = 5.34$ and X_2 with $\alpha_2 = 1.6$, $k_2 = 10$. Note that in both cases the mean flow duration is the same, i.e. $k\alpha/(\alpha - 1) = 26.67$. The total number of flows is N =1000 and the arrival times are uniformly distributed in a 1 hour periodequ, that is equivalent to a Poisson arrival process with intensity $\lambda = 0.28$ arrivals/sec. The sampling interval is $b = 0.01$ sec.

We applied the `LDestimate` function provided with the wavelet package in [5] to both processes. The output LDP are shown in Fig. 1. In agreement with the theory the re-

[1] In practice Bob's career will be evaluated to some extent based on his paper count. This might introduce a certain distortion and cause some confusion between the "means" and "goal" role of Bob's papers from his individual perspective. However this does not necessarily hamper the general role of papers as means for scientific dissemination.

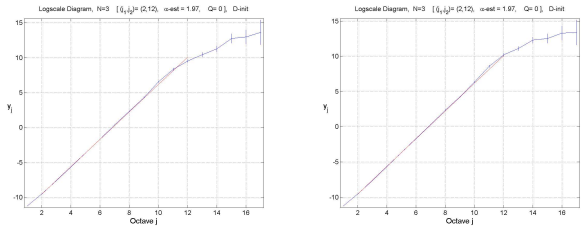

Figure 1: LDP for X_1 and X_2 (M/G/∞ with pareto distributed ON periods).

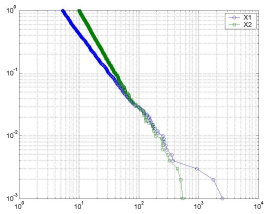

Figure 2: Empirical CCDF of flow-duration for for X_1 and X_2.

sulting curves in the LDP are straight lines across all timescales (the carving at the rightmost end is the border effect due to finite measurement period). But a careful look at their *slopes* (i.e. the Hurst parameters) reveals a mismatching with the theory: following equation 1 we would expect $H_1 = (3 - 1.25)/2 = 0.875$ and $H_2 = (3 - 1.6)/2 = 0.7$ respectively for X_1 and X_2, while instead the script report the same value (around 1.48) in both cases.

Comments. How to explain the mismatching between the theory and the results in Fig. 1 ? Possible hypothesis include:

1. We made some trivial error in our MATLAB simulation code and/or in using the `LDestimate` function. The code is pretty short and simple, it is publicly accessible[2] so that anybody can check for bugs and errors;

2. There is some problem with the MATLAB random number generator. In Fig. 2 we plot the empirical complementary CDF of the flow-duration in both simulations, they are consistent with the assigned distributions.

3. We mis-interpreted the theoretical results presented in [11, p. 510].

4. The `LDestimate` from [5] delivers a wrong output and needs to be revised: in order to check that it would be good firs step to produce a separate LDP computation for the same signal with a different code.

5. There is some other subtle explanation hidden in the simulation setup.

At the moment of writing we do not have an answer. Any input for solving the puzzle is wellcome.

[2]Available online from http://userver.ftw.at/~ricciato/open-comments/simpleLDcheck.m

3. UNWANTED TRAFFIC IN THE FLOW ARRIVAL PROCESS

One of the main interesting open points in the field of TCP/IP traffic characterization is the relationship between packets, flows and (perhaps) sessions. A flow is defined as a set of consecutive packets to the same 5-tuple (source / destination IP addresses and ports, protocol).

In a set of conceptually neighboring papers Veitch et al. ([12] [13] [14] and more recently [9]) adopt an interesting approach to investigate the relationship between certain traffic features at the flow and packet levels. Their approach is based on a set of trace manipulations called "semi-experiments": in each semi-experiment the arrival process of the flows and/or packets-within-flows are modified, eventually disrupting the internal correlation structure, if any. The LDP signatures are taken as a powerful summary for the statistical properties of the entire process. Two types of process are considered: packet arrival X and flow arrival Y.

When we tried to apply the same approach we immediately hitted against a practical problem, namely the presence of so called unwanted traffic (see [6] and references therein). In our traces we found a relatively small population of mobile terminals infected by scanning worms. During their scanning activity the worm agents produce high rates of TCP SYN to random IP addresses. Following strictly the classical flow definition based on the 5-tuple we should consider each of such packets as a new flow arrival. Similar consideration apply to other sources of unwanted traffic, e.g. TCP/UDP port scanners. We will use the term "pseudo-flow" to refer to these "non productive" connection attempts, the vast majority of which are not succesful.

In our traces we found that pseudo-flows can account for a large fraction of the total number of flows: in a recent dataset no less than 30% of all TCP SYN could be referred to unwanted traffic (mainly well-known scanning worms), but the share can raise up to 90% in certain periods (e.g. in the Dec'04 dataset evaluated in [8]). An important difference between the "unwanted" and the "legitimate" traffic components is that the former is never as statistically stable as the latter. In fact most of the unwanted traffic is typically produced by a relatively small sub-population of terminals (e.g. infected laptops) that is subject to relatively large fluctuations.

It is well-known that unwanted traffic is a constant component of the global Internet traffic since several years [16] [3]. It is possible that such traffic was present also in the datasets used by the group of papers referenced above, but it is not clear whether the authors recognized it and how they handled it.

These considerations raise a number of questions:

1. What is the fraction of pseudo-flows in the datasets analyzed by the cited papers ?

2. What is the impact of unwanted traffic on the LDP of the processes X and Y ?

3. Should future work discriminate legitimate flows from pseudo-flows and analyze them separately ? If yes, it would be convenient that the research community adopt a single a common definition for pseudo-flows (e.g. uncompleted connection attempts ?).

4. Should perhaps all scanning probes generated by a sin-

gle source during an active scanning period be considered as a single continuous flow ?

We believe that such questions are of interest for anybody involved in similar research activities.

4. CAPACITY ESTIMATION BEYOND THE BUFFERING ASSUMPTION

There is track of previous works and tools (e.g. MultiQ [17], Nettimer [10]) aimed at inferring the link capacity from passive measurements. A set of techniques have been developed that are based on the analysis of the dispersion of the packet arrival time (see [2] for an overview). The implicit assumption common to all such techniques is that the bottleneck links are *buffering* excess packets.

In some real networks however the bottleneck links might be *discarding* packets. This is typically the case for those network sections that use IP-over-ATM and/or IP-over-Frame Relay Virtual Circuits (VC), with L2 rate-limiters with token-bucket parameters configured on each VC. Furthermore it is common for ISPs to restrict the traffic rate of their edge customers by means of L3 rate-limiters set on the access element. Some Service Level Agreements foresee multiple rate thresholds (typically two: Guaranteed Rate and Excess Rate) and simply discard excess traffic. Finally, the bottleneck element can be a node instead of a link, e.g. an overloaded network element discarding the packets exceeding its processing capacity.

In summary there are two main classes of bottleneck links: buffering and discarding. The relative share of the two link types in the real Internet is unknown since to date no attempt has been made to measure it. It might be expected that buffering links are the majority nowdays, and that discarding links are more frequent in the access networks than in transit backbones.

The estimation of bottleneck capacities based exclusively on packet-dispersion techniques offers only a partial view as it consider exclusively the buffering links. These considerations raise the following open points that might be considered for future research:

1. How frequent are buffering and discarding links in the real Internet ? Is it possible to classify them based on external measurements (passive or active) ?
2. How to infer the bandwidth of discarding links ?
3. Considering a mixed environment where buffering and discarding links coexist, what is the impact of the latter onto the capacity estimations obtained by packet-dispersion techniques ?

We came across these questions during an earlier work where we analyzed the actual behaviour of a dicarding bottleneck link found in the live network [7].

5. CONCLUSIVE REMARKS

In the previous sections we have proposed to the attention of the research community a few open points found about published papers. As a general remark, it would be nice if such kinds of comments, requests for clarifications, research ideas etc. could be posted and replied on some public platform. The underlying idea is that any piece of scientific contribution (typically a paper) does not "die" but instead starts its lifetime right after the publication, i.e. attracting readers, followers and (why not?) criticisms. Consider the set of readers and followers of a generic published paper: they form naturally a community. It would be a convenient opportunity for the members of such community to refer to a common "collaboration tool" for exchanging comments, requesting clarifications, proposing collaborations, etc. The most simple, direct and efficient way to achieve that would be to let each conference and/or journal publisher maintain a wiki-like electronic platform where future readers can publicly post their comments about each published paper, and authors and/or other readers can post their replies and rebuttals.

Currently, with no such platforms in place, a reader can not do more than sending an email to the authors. Even in the best case that the latter find the time and the will to react (and this is not always the case) the value of the interaction is limited to that specific readers (and perhaps the authors) while in many cases its content would be highly valuable for many other readers and followers. Also, a wiki-like platform is even less intrusive than a private email to the authors, and in some cases can be replied even by other readers. Another advantage of a wiki-like public platform over private email exchange addresses the implicit cost-benefit balance of each query. Assume an author is queried with a "request for clarification" about one of his/her previous papers. Replying to the query has always a *cost* in terms of time and efforts to be processed (at least writing a small reply, or performing some small extra analysis, etc.). The satisfaction of a single reader might not be considered a sufficient *benefit* to compensate for such efforts. In this case the point-to-point email exchange might not proceed and queries are left unanswered [3]. On the other hand, with a pulic wiki-like platform the reply to each query about a paper would be visible to all future readers so as to increase the *benefit* of the reply, ultimately increasing the probability of receiving an answer.

In summary, public open wiki-like commenting facilities for published papers would be a convenient way to complement published material and increase the quality of scientific interactions. Conferences and journal pulishers are the best candidates to host such platforms on a regular basis.

6. REFERENCES

[1] DARWIN project home page http://userver.ftw.at/~ricciato/darwin.

[2] C. Dovrolis, P. Ramanathan, D. Moore. Packet dispersion techniques and capacity estimation. *IEEE/ACM Trans. on Net.*, 12(6), Dec 2004.

[3] C.C. Zou, W. Gong, D. Towsley. Code Red Worm Propagation Modeling and Analysis. *Proc. of CCS'02, Washington, USA.*, Nov 2002.

[4] D. Veitch, N. Hohn, P. Abry. Multifractality in TCP/IP traffic: the case against. *Journal of Computer Networks*, 48(3), June 2005.

[5] D. Veitch, P. Abry. Matlab code for the wavelet based analysis of scaling processes. *Available from ¡http:// www.cubinlab.ee.mu.oz.au/~darryl/secondorder_code.html¿.*

[6] F. Ricciato. Unwanted traffic in 3G networks. *Computer Communication Review*, 36(2), April 2006.

[3] This is especially the case for important well-known authors that are often very busy.

[7] F. Ricciato, F. Vacirca, P. Svoboda. Diagnosis of Capacity Bottlenecks via Passive Monitoring in 3G Networks: an Empirical Analysis. *Technical Report FTW-TR-2006-008 (submitted)*, May 2006. Available online from [1].

[8] F. Ricciato, P. Svoboda, E. Hasenleithner, W. Fleischer. On the Impact of Unwanted Traffic onto a 3G Network. *Technical Report FTW-TR-2006-006 (submitted)*, February 2006.

[9] J. Ridoux, A. Nucci, D. Veitch. The impact of the flow arrival process in Internet traffic. *IEEE INFOCOM'06, Barcelona*, April 2006.

[10] K. Lai, M. Baker. Nettimer: A Tool for Measuring Bottleneck Link Bandwidth. *Usenix*, 2001.

[11] K. Park, W. Willinger. *Self-similar Network Traffic and Performance Evaluation*. Wiley, 2000.

[12] N. Hohn, D. Veitch, P. Abry. Does Fractal Scaling at the IP Level depend on TCP Flow Arrival Processes? *Internet Measurements Workshop, Marseille, France*, November 2002.

[13] N. Hohn, D. Veitch, P. Abry. Cluster processes, a natural language for network traffic. *IEEE Transactions on Signal Processing, special issue on Signal Processing in Networking*, 51(8):2229–2244, 2003.

[14] N. Hohn, D. Veitch, P. Abry. The impact of the flow arrival process in Internet traffic. *Proc. of IEEE Int. Conf. on Acoust. Speech and Sig. Proc.*, 2003.

[15] P. Abry, D. Veitch. Wavelet analysis of long range dependent traffic. *IEEE Trans. on Information Theory*, 1(44):2–15, 1998.

[16] R. Pang et al. Characteristics of Internet Background Radiation. *Proc. of IMC'04, Taormina, Italy*, October 2004.

[17] S. Katti et al. MultiQ: automated detection of multiple bottleneck capacities along a path. *IMC'04, Taormina, Italy*, October 2004.

[18] S. Stoev, M. S. Taqqu, C. Park, J. S. Marron. On the wavelet spectrum diagnostic for Hurst parameter estimation in the analysis of Internet traffic. *Computer Networks*, 48(3), June 2005.