A Tricky Problem

Darryl Veitch

Principal Research Fellow

http://www.cubinlab.ee.mu.oz/∼darryl

Department of Electrical & Electronic Engineering
The University of Melbourne

Preamble

Abstract

I will discuss an example of a 'structure identifiability problem' where without prior knowledge, the 'right model' may be impossible to obtain.

This talk was originally delivered on the whiteboard, with no prepared slides. Here I give a reproduction of my presentation, hopefully retrospectively enhanced.

The idea for this talk came when I was reflecting on a point process model for packet arrivals which my colleagues and I developed recently. In that model the idea of *flows of packets* holds a central place. It struck me that without inside knowledge of the existence of flows, it may not have been possible to even detect their existence, even though their impact is very profound.

The Packet Arrival Process, Flows and Clusters

We want to understand and model packet arrivals, viewed as a point process. The first figure shows a sample path of packet passage times past some monitoring point (imagine time as the horizontal axis):

Through our understanding of networks, we know that *flows*, that is groups of packets that share some common end-to-end identity, exist. Flows can be defined for example as TCP connections, or more generally by using the common 5-tuple plus timeout. It we colour-code the above example by flow, the ‘apparent’ clustering we might have imagined takes on a more concrete meaning.

In fact the first picture was really only schematic in relation to the ‘parameter values’ found in networks. In reality the ‘packet-bodies’ of flows interleave each other, giving us a more complex picture:

Clearly, if the packets weren’t coloured here, we would have far more trouble in picking out the clusters due to flows. We might be led to define different clusters based purely on time locality.

Clusters Lost in Clustered Fog

In fact, in practice the situation is much more extreme than the flow ‘mixing’ shown in the previous picture. In backbone links for example, we have tens of thousands of flows all aggregated together. A huge amount of interleaving occurs, leading to a very different picture:

There is now only 1 flow (the light green one) which has more than one of its packets in the picture. Each packet now finds itself surrounded by strangers, its neighbors are not in the same flow. A neighborhood about a given packet large enough to capture its closest brothers would also include (with high probability) a large number of other packets, belonging to many different flows.

Now imagine that a cluster-hungry researcher monitors packet arrivals, but has never heard about flows and does not measure header information so he (or she) cannot discover them. What he will see is:

The researcher will be able to find clusters, and to construct black box cluster models describing some statistics of the overall packet arrival process. However, it seems intuitively clear that he is unlikely to find those corresponding to the underlying flow clustering, although it truly exists, given the noise of the clustering, real or ‘imagined’, arising from packets across flows.

Furthermore, even if the researcher *knew* about flows but was unable to measure them directly, he would be hard put to identify them without extraneous imformation.

But, does it matter? If the flow structure is so ‘faint’ as to be practically invisible, maybe its impact is as small and this faintness would suggest?

A Cluster Model Based on Flows

In our recent work, we were able to show that the packet arrival point process can in fact be well modelled by a known point process class, the *Barlett-Lewis Poisson Cluster Process*.

In the BL-PCP, seeds (the flow arrivals) fall as a homogeneous Poisson process, and associated to each seed is a cluster (the packets in the flow), which begins at the seed and whose points are distributed as a 'finite' renewal processes (i.e. after a finite number of points, they terminate).

In this picture flows are independent from each other, and the variance of the overall process is generated by the variability of the renewal process' inter-arrival distribution A, and the properties of the distribution of P, the number of packets per flow. These interact to give a change of variance with timescale which can be predicted analytically. Long-range dependence is included in a natural way through heavy tailed P.

The point here is that that this model, which is built fundamentally upon the existence of flows, explains and predicts some key features in a natural way. In particular, the 'knee' separating small and large scale behaviour, a feature which has been found by numerous researchers at around the 1 [sec] time scale but which has eluded a definitive explanation, can be explained in this model as the competition between the two source of variance, P and A. The knee is seen as the cross-over point of two effects, one which controls the small times scales, and the other the large.

Conclusion: flows rule, but invisibly

We are left with the following picture: flows are not only real, their existence has an impact on the observed statistics which is clear, and a model which has insight into this flow structure is capable of explaining and predicting key features of that impact.

and yet...

If the existence of flows were not known in advance, they would not announce themselves, their very existence may remain unsuspected.

Furthermore, even if it were known they existed, it may not be possible to even roughly reconstruct them.

An analogy could be made with physical laws:

It is as if we lived in a universe where laws existed which, if they could be found, would enable an entire technology based on them in detail to be built, but which were so 'faint' in another sense, that they gave no sign of their presence and could not be detected even if a genius (married to a psychic) deduced what they might be and searched for them explicitly.

Someone at the workshop later suggested that perhaps the known physical laws we now take for granted are in fact in this category already! but we managed to find them eventually! Certainly, as yet undiscovered laws may be.... Note however that even the current, very challenging problem of the detection of gravitational waves it not 'difficult' in principle, one simply needs a large enough detector.

A Challenge for SP

The challenge for SP in all this is to develop new ways to detect ‘signal in noise’ which are more powerful, which can somehow test for the presence of subtle *structure*, even when the form of the structure is unknown. In terms of the art of model choice, this is akin to re-emphasizing the importance of prediction (and in particular sample path ‘deterministic’ prediction), over ensemble distributional agreement.

Bibliography

1

Nicolas Hohn, Darryl Veitch, and Patrice Abry, “Cluster processes, a natural language for network traffic,” *IEEE Transactions on Signal Processing, special issue “Signal Processing in Networking”*, vol. 51, no. 8, pp. 2229–2244, August 2003.