Finding Causes in Traffic Behaviour

Theme Talk

SAMSI Workshop on Measurement and Modelling

Darryl Veitch
What is Burstiness?

Consider some stationary stochastic process $X(t), t \in \mathcal{R}$.

Two ‘orthogonal dimensions’

- Amplitude Burstiness – variability in marginal $\sigma/\mu$, subexponential family, tail exponent $\alpha$, ..... 
- Temporal Burstiness – dependencies covariance function, spectral density, ‘wavelet spectrum’,...

So how are they treated in traffic modelling?

- For theory - distinction is clear, eg counting measure $N$ captures all
- For modelling - tend to focus on one or the other
- But reality is both (nearly) always mixed!

Take discrete measured form $X_\tau(k), k \in \mathcal{Z}$.

Temporal and Amplitude burstiness mixed over scales $[0, \tau]$, result time-scale dependent.

Danger in modelling intrinsic scale dependent mix by only 1 paradigm
What are we Measuring?

- **Internet Protocol** (IP) packets, the unit of transport across networks.

- Data split into packets, with: header, payload.

- Payload carries higher layer protocols: TCP, UDP, ICMP.

- Protocols support services & applications:
  - **TCP**: HTTP, FTP, SNMP, ... (for reliable data)
  - **UDP**: VoIP, DNS, NTP,... (for real time)

- Passive measurement: watch complex mix pass by a fixed point.

- For each packet
  - Could capture all or part (eg just the header).
  - Must timestamp.

- Key concept, a flow (collection) of packets.

- ‘Spatio-temporal’ data - experiences of active probes.
Traffic Data Collection: Flows

**IP flow**: set of packets with the same 5-tuple

<table>
<thead>
<tr>
<th>IP protocol</th>
<th>Sources Address</th>
<th>Destination Address</th>
<th>Source Port</th>
<th>Destination Port</th>
</tr>
</thead>
</table>

![Graph showing IP flows over time]
Two Levels of Traffic Data: Packets & Flows

Flow arrivals: $Y(t)$

Packet arrivals: $X(t)$

A classic example of a flow is a TCP connection.
Burstiness of What?

Time series can be abstracted from myriad traffic ‘aspects’.

- Point process of packet arrivals.
- Packets per bin, bytes per bin.
- Link packet loss, path loss.
- End-to-end delay, delay variation, round-trip times.
- Vector versions of above over graph.
- Topological structure of network, address space, flow trajectories.
- Routing tables.

---

• .........
• ....
No shortage of sources of structure and therefore ‘burstiness’

1: User behaviour – diurnal cycle, think times, reading preferences, applications, attacks..

2: Application layer – function, protocols, methodology (caching)..

3: Nature of sources – real-time/not, reliable/not, volume, variability..

4: Session layer – dynamics of mixing 1, 2, 3, with influences from 5, 6, 7..

5: Connection level – protocol design (reliable/not), network cooperation..

6: Link layer – protocol design, wire/wireless, datagram size..

7: Physical – raw bandwidth, link failures, bit errors.

It is complex.. must let data speak.
The Idea: That ‘20%’ of packets account for ‘80%’ of impact.

- Usually applied to flow volume: distribution is heavy tailed.
- But, can also be applied to traffic rate.
- Is really time-scale dependent.
Weighing Burstiness: Mice & Elephants

Flow Density

Flow burstiness

(a)

(b)

(c)
Can Laboratory Experiments Establish Causation?

For many purposes, NO, too complex.

But for some purposes, an alternative - *Semi-Experiments*
A Link in the Chain: Semi-Experiments
Permutation of Flows [A-Perm]
Original Order Poisson Arrivals [A-Pord]

(a smooth form of internal shuffling)
Original Order Poisson Arrivals [A-Pord]
Permutated Poisson Arrivals [A-Pois]
From simple (Semi-)Experiments, we learn a lot

From these flow arrival *manipulations*:

- Correlations between flows can be **neglected**
  - No need for session level hierarchical models
  - TCP dynamics between flows can be **neglected**
- For IP modelling, flow arrivals can be modelled as **Poisson**
  - Justifies an assumption commonly used in traffic modelling.

**Note:** true flow arrival process is **LRD**.
The Semi-Experimental Method

Replacing selected aspects of data with model substitutes

Benefits

- Understand the impact of a particular feature on overall statistics
- Enables ‘convenient’, highly flexible virtual experimentation
- Avoids need to model all aspects simultaneously
- Enables physically meaningful parameters to be directly targeted

Example Manipulation Classes:

[A] on Flow Arrival process
[P] on Packet arrival process within a flow – can adjust entire process, scale rate, shuffle...
[S] Selection of flows according to number of packets, duration, rate, address, burstiness ....
[T] Truncation of flows according to number of packets, duration, silence duration ....

A Modelling Continuum:

observe → primitive SE → compound SE → class search SE → semi-model → model
(data)    (neutral model)    (multi-aspect)    (model choosing)    (1 data aspect)    (no data)
Return to Scientific Method

- Let the data speak.
- Aim to show what does Not cause, not what does.
- Find what matters → model suggests itself.
- Then can test if model performs well.
- Model is physically based by definition, avoid black magic.
Can it Capture Physics?

For packet arrival process on ‘lightly’ loaded links

Data and Model hard to distinguish at all aggregation levels