Inverting Sampled Traffic

Nicolas Hohn, Darryl Veitch

Australian Research Council
Special Research Center for
Ultra-Broadband Information Networks
THE UNIVERSITY OF MELBOURNE
Inverting Sampled Traffic

- Motivation

- Sampling Techniques
  - Packet Sampling
  - Flow Sampling

- Comparison of sampling techniques
  - Distribution of the number of packets per flows
  - Spectral density of packet arrival process

- Application to traffic modelling
Introduction

Motivation

- Traffic statistics collected by routers don’t scale well with link speed: exact traffic logging is **impossible** for backbone links
- Need to **sample** the traffic, export **partial** statistics
- Aim: **infer** statistics of **original** traffic from partial measurements
Introduction

Motivation

- Traffic statistics collected by routers don’t scale well with link speed: exact traffic logging is impossible for backbone links.
- Need to sample the traffic, export partial statistics.
- Aim: infer statistics of original traffic from partial measurements.

Short history

- 1993: Claffy et al. advocate sampling techniques at the packet level to reduce the load on measuring infrastructure.
- 2002-2003: Duffield et al. give estimates of first order quantities from packet level sampled traffic: average rate, mean number of packets per flows.
Inverting Sampled Traffic

- Motivation

- Sampling Techniques
  - Packet Sampling
  - Flow Sampling

- Comparison of sampling techniques
  - Distribution of the number of packets per flows
  - Spectral density of packet arrival process

- Application to traffic modelling
Packet Sampling

Original traffic

- Sample packets with probability $q$.
- Measure rate of sampled traffic: $\lambda(q)$.
- Infer rate of original traffic: $\lambda(q)/q$. 

Time
Packet Sampling

Original traffic

i.i.d. sampling with probability $q$

Simple example: recover original packet rate
- Sample packets with probability $q$.
- Measure rate of sampled traffic: $\lambda(q)$.
- Infer rate of original traffic: $\lambda(q)/q$. 
Packet Sampling

Original traffic

i.i.d. sampling with probability $q$

Sampled traffic

Simple example: recover original packet rate
- Sample packets with probability $q$.
- Measure rate of sampled traffic: $\lambda(q)$.
- Infer rate of original traffic: $\lambda(q)/q$. 

Time

Time
Packet Sampling

Simple example: recover original packet rate
- Sample packets with probability $q$,
- Measure rate of sampled traffic $\lambda^{(q)}$,
- Infer rate of original traffic $\lambda^{(q)}/q$.
**Terminology**

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

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

![Time Diagram](image-url)
**Terminology**

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

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

**Flow Level**

**Packet Level**
Original Traffic

Recovering original flow sizes not straightforward
Flow Sampling
Original Traffic

Recovering original flow sizes not straightforward
Packet Sampling

Recovering original flow sizes not straightforward
Inverting Sampled Traffic

- Motivation

- Sampling Techniques
  - Packet Sampling
  - Flow Sampling

- Comparison of sampling techniques
  - Distribution of the number of packets per flows
  - Spectral density of packet arrival process

- Application to traffic modelling
Distribution of number of packets per flow

Original traffic

Packet sampling

Flow Sampling

Potential inversion problems

No ‘inversion’ problems
Distribution of number of packets per flow

Packet sampling

\( p_j \): Probability that a flow had \( j \) packets before sampling.

\( p^{(q)}_k \): Probability that a flow has \( k \) packets after sampling,

\[
p^{(q)}_k = \sum_{j=k}^{\infty} \Pr\{k \text{ packets after thinning} | j \text{ packets before thinning}\} p_j
\]

\[
p^{(q)}_k = \sum_{j=k}^{\infty} \binom{j}{k} q^k (1 - q)^{j-k} p_j
\]  \hspace{1cm} (1)

- Aim: express \( p_j \) as a function of \( p^{(q)}_k \) by inverting (1)
Inverting (1) with generating functions

Definition: \( G_{P}(z) = \sum_{j=0}^{\infty} p_{j} z^{j}, z \in \mathcal{D}(0,1). \)

\( \mathcal{D}(z, r) \): open disc centered at \( z \) with radius \( r \)

**Singularity** at \( z = 1 \) if heavy tailed distribution.

From (1): \( G_{P}^{(q)}(z) = \sum_{k} p_{k}^{(q)} z^{k} = G_{P}(1 - q + qz), z \in \mathcal{D}(0,1) \)

\[ G_{P}(z) = G_{P}^{(q)} \left( \frac{z - (1 - q)}{q} \right), z \in \mathcal{D}(1 - q, q) \]

Aim: Find power series expansion of \( G_{P} \) at \( z = 0 \)

Methods:
- Analytic Continuation
- Cauchy Integral
Scheme 1: Analytic Continuation

$q = 0.6$

\[ p_j = \sum_{n=j}^{\infty} \binom{n}{j} \frac{(-1)^{n-j}}{q^n} (1 - q)^{n-j} p_n^{(q)} \]  

(2)
Scheme 1: Analytic Continuation

$q = 0.1$

$p_j = \ldots$

Diagram showing points labeled $z_0$, $z_1$, $z_2$, $z_3$, $z_4$, and $z_5$. The diagram includes concentric circles and points marked with different symbols.
Scheme 2: Cauchy Integral

\[ p_j = \oint_S \frac{G_P(z)}{z^{j+1}} \, dz, \quad (3) \]

- \( S \): any closed contour containing the origin, for instance \( \mathcal{D}(0, 1) \).
- Inversion methods work well when \( G_P \) can be directly evaluated on \( S \).
- Values of \( G_P \) on \( \mathcal{D}(0, 1) \) are unknown: obtained with Padé Approximants.
Distribution of number of packets per flow

$q = 0.6$

Theoretical original density
Flow thinning
Packet thinning: scheme 1
Packet thinning: scheme 2
Distribution of number of packets per flow

Packet sampling

- Easy to implement,
- Need for consistent flow definition for sampled traffic (new timeout $T_0$),
- Problems to estimate $p_0^{(q)}$ from sampled data,
- Severe numerical issues to recover the packet distribution ("impossible" for $q < 0.5$ !),

Flow Sampling

- Need on-line processing to create flows.
- No need to change flow definition,
- No inversion to recover packet distribution,
- $q$ plays no theoretical role. Only the remaining number of flows matters for the estimation,
Spectral density of packet arrival process

Original traffic

Packet sampling

Potential inversion problems

Flow Sampling

Potential inversion problems
Spectral density of packet arrival process

$\Gamma_X(\omega)$: spectral density of original traffic

$\Gamma^{(q)}_X(\omega)$: spectral density of sampled traffic

Packet sampling

Results from theory of thinned point processes give direct inversion

$$\Gamma_X(\omega) = \frac{1}{q^2} \left( \Gamma^{(q)}_X(\omega) - (1 - q) \lambda^{(q)} \right)$$

Flow sampling

Assumptions needed:

- Flow arrivals follow a **Poisson process**,
- Flows are **uncorrelated**.

$$\Gamma_X(\omega) = \frac{1}{q} \Gamma^{(q)}_X(\omega)$$
Study Second Order Structure

Analysis tools: **Discrete Wavelet Transform**

**Definition:**
Comparison of a signal $X(t)$ with a family of functions $\psi_{j,k}$ by means of inner products $d_X(j,k) = \langle X, \psi_{j,k} \rangle$, where $\psi_{j,k} = 2^{-j/2} \psi(2^{-j}t - k)$, and $\psi$ is the mother wavelet, localised both in time and frequency.

**Properties:**
- $\{d_X(j,k), k \in \mathbb{Z}\}$ is **stationary** and **short range dependent** for $j$ fixed,
- $\text{variance}(j) = \mathbb{E}|d_X(j,k)|^2$
- For scaling processes: $\mathbb{E}|d_X(j,k)|^2 = 2^{j\alpha} \mathbb{E}|d_X(0,k)|^2$
- For LRD processes: $\mathbb{E}|d_X(j,k)|^2 \sim 2^{j\alpha} \mathbb{E}|d_X(0,k)|^2$ for large $j$.

**Wavelet Spectrum Estimate:** $\log_2 \left[ \frac{1}{n_j} \sum_k |d_X(j,k)|^2 \right]$ vs $j$

**Link with power spectral density:** $\mathbb{E}|d_X(j,k)|^2 = \int \Gamma_X(\nu) 2^j |\Psi(2^j \nu)|^2 d\nu$
Spectral density: \( q = 0.1 \)
Spectral density: $q = 0.001$
## Conclusions

### Packet Sampling

- Easy to implement,
- Need for consistent flow definition for sampled traffic (new timeout $T_0$),
- Problems to estimate $p_0^{(q)}$ from sampled data,
- Severe numerical issues to recover the packet distribution ("impossible" for $q < 0.5$ !),
- Inaccurate estimation of the spectrum from sampled traffic for small $q$.  

### Flow Sampling

- Need on-line processing to create flows.
- No need to change flow definition,
- No inversion to recover packet distribution,
- $q$ plays no theoretical role. Only the remaining number of flows matters for the estimation,
- Accurate spectrum estimation,
Inverting Sampled Traffic

- Motivation

- Sampling Techniques
  - Packet Sampling
  - Flow Sampling

- Comparison of sampling techniques
  - Distribution of the number of packets per flows
  - Spectral density of packet arrival process

- Application to traffic modelling
Application to traffic modelling

Aim
- Fit model to sampled traffic,
- Infer model parameters for unsampled traffic.

Theory
- Closure properties of the Bartlett-Lewis Point Process under both packet and flow sampling.

Practice
- Only flow thinning is applicable.
Sampling the Bartlett-Lewis Point Process

\[ j = \log_2(a) \]

\[ \log_2 \text{Var}(d_j) \]

Original

BLPP matched to Original

Flow Thinned

BLPP matched to Flow Thinned

BLPP reconstructed from Thinned
Conclusions

- Packet sampling
  - Easy to implement but hard to infer original statistics beyond first order.

+ Flow sampling
  - Harder to implement but leads useful information about original traffic, for both flow and packet level statistics.