Towards Informative Statistical Flow Inversion

Hamed Haddadi

Networks & Services Research Group

MSN07, 12th July 2007

Introduction to sampling and NetFlow

Packet Sampling: Pick 1 in N, relaxes router by~80%

Flow Records : [IP protocol, source address, source port, destination address, destination port]

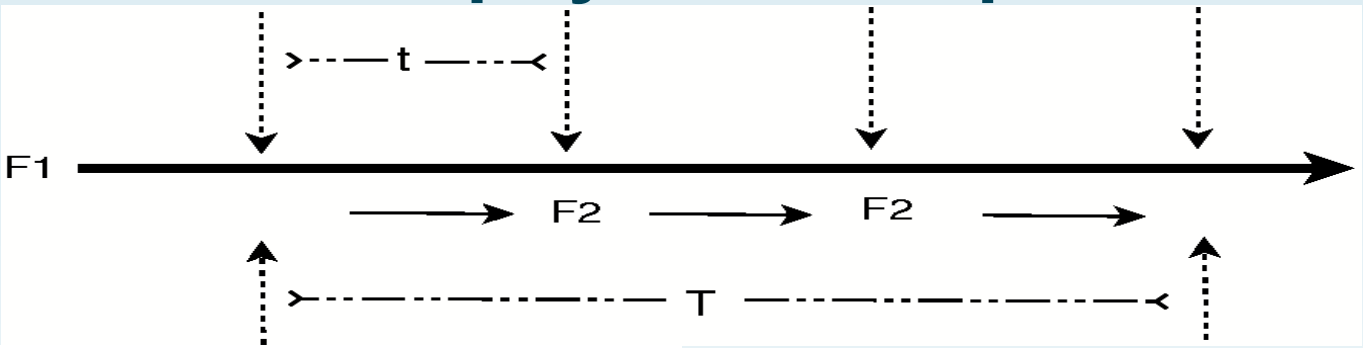
criteria for expiring flows in the cache:

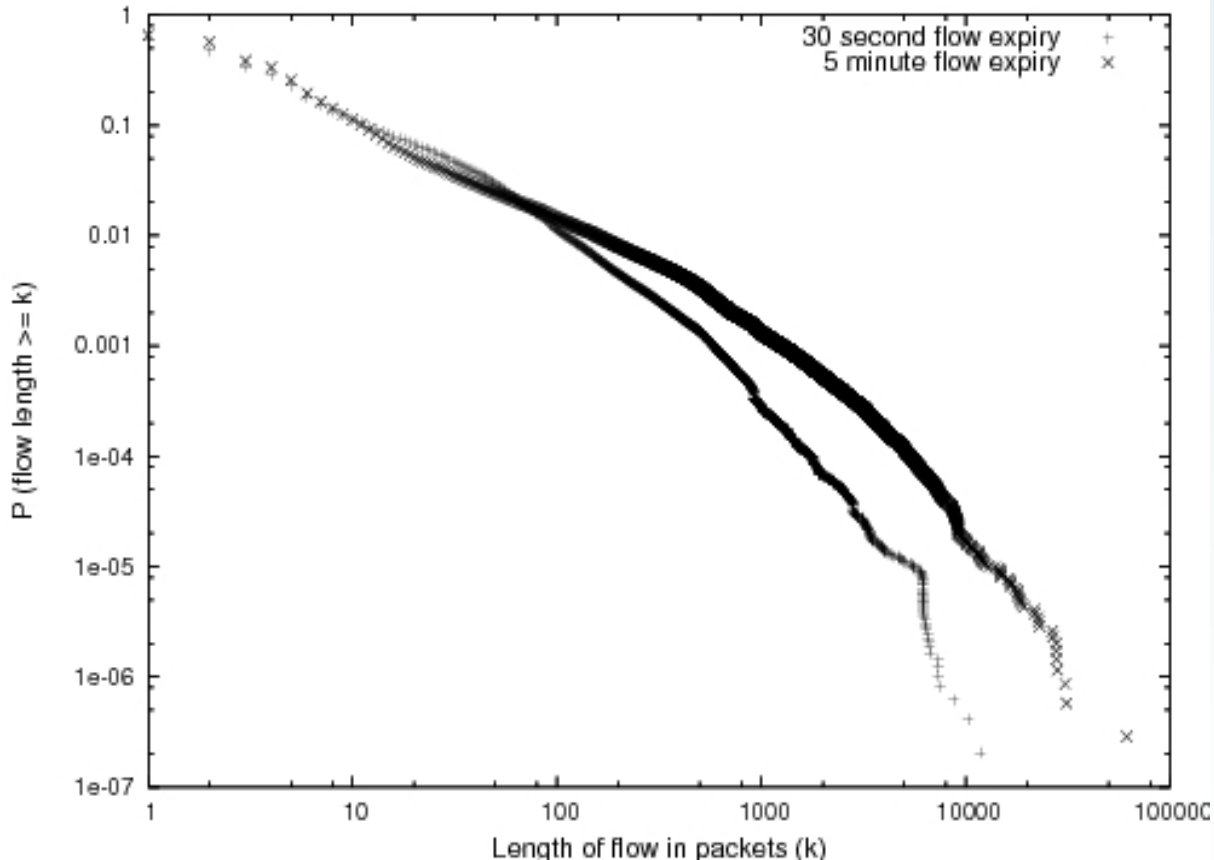
- Idle time-out (15 seconds is default)
- Long lived flows (30 minutes is default)
- Heuristics expiration (cache is full!)
- Protocol flags

....add sampling rate and export frequency...
it is complex stuff!

Effects of flow expiry, an example

Two flows:
F1: YouTube
F2: Port-Scan





- Missing short flows
- Mis-ranking flows
- Slicing long flows

Sampling techniques

CAIDA data, April 2003, 30 mins, 47,047,240 packets f
83% TCP, 7% UDP

Sampling strategies used in this paper:

- Packet sampling [Widely used]
- Flow sampling
  - sample-and-hold (by byte) [Estan & Varghese 2002],
  - sample-and-hold (by packet)
  - sample-and-hold (by SYN)

Creation of flow records

```
while PACKETSLEFT(trace)
do { P ← READPACKET(trace)
     (φ, t, N_b) ← DECODEPACKET(P)

     if FLOWISBEINGTRACKED(φ)
       then { comment: Has the flow expired?
              if (t_s(φ) > t_t)
                { ψ ← GETFLOWID(φ)
                  TERMINATEFLOW(ψ)
                  ψ ← CREATEFLOW(φ) }

              t_s(φ) ← t
              T_p(ψ) ← T_p(ψ) + 1
              T_b(ψ) ← T_b(ψ) + N_b }
       else { comment: Is the flow going to be sampled?
              if FLOWSELECTEDFORSAMPLING(p, N_b)
                then { t_s(φ) ← t
                       ψ ← CREATEFLOW(φ)
                       Ψ ← ψ
                       T_p(ψ) ← 1
                       T_b(ψ) ← N_b }

     if FLOWBUFFERFULL(|Ψ|, N_f)
      or FLOWEXPORTTIMEREXPIRED(t, t_w)
       then { EXPORTFLOWBUFFER()
              RESETFLOWBUFFER() }
```

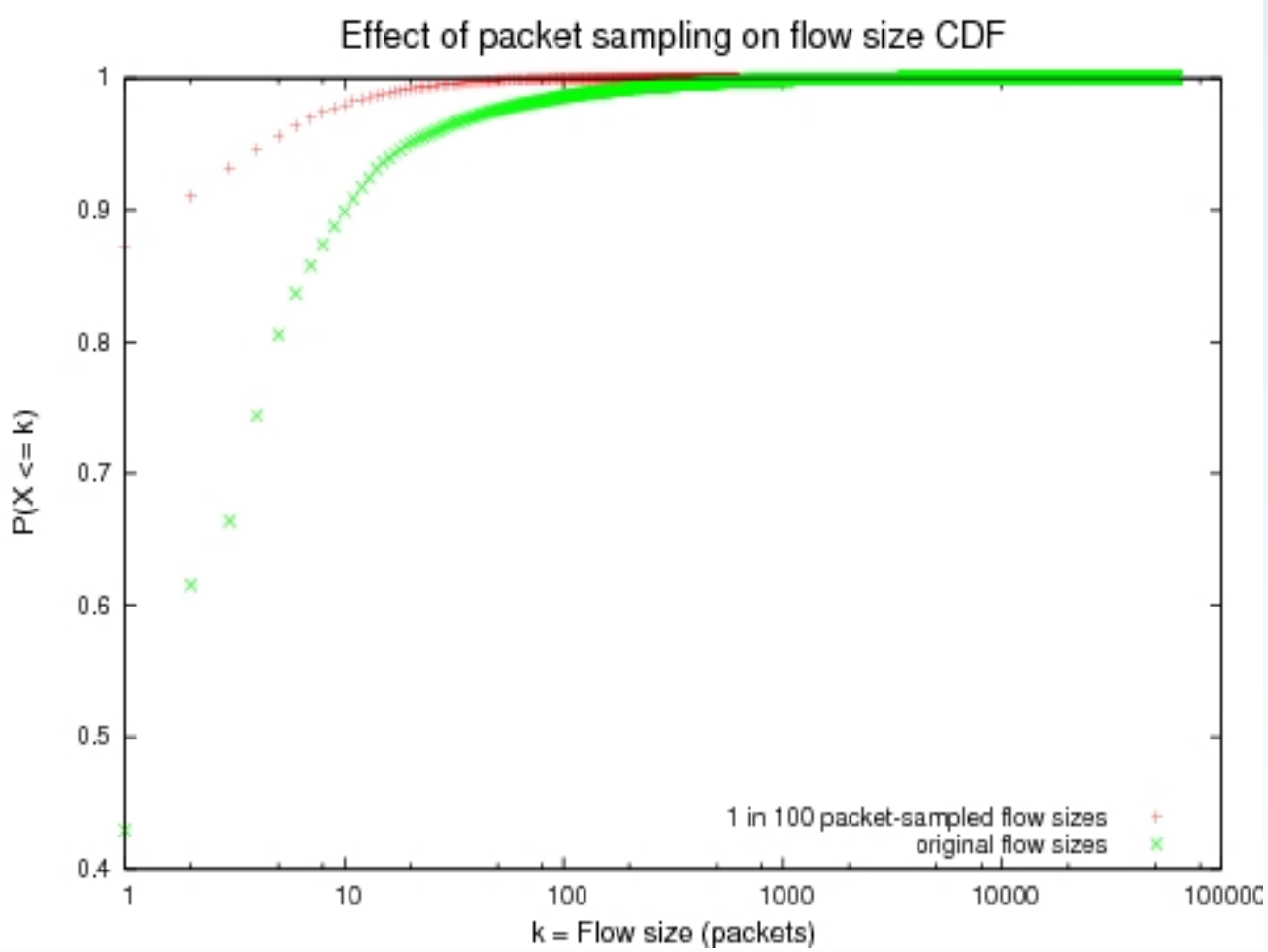
ϕ	5-tuple corresponding to packet P
t	Capture time of packet P
N_b	Number of bytes in packet P
$t_s(\phi)$	Trace time since the last packet on 5-tuple ϕ was seen
t_t	Flow expiry timeout
ψ	Flow identification number
Ψ	Set of all ψ
$T_p(\psi)$	Total number of packets in flow ψ
$T_b(\psi)$	Total number of bytes in flow ψ
p	Probability of starting to follow flow ϕ
t_w	Flow buffer export timeout
N_f	Flow buffer size in records

Impact of packet sampling

worst problem:

can not get the right distribution of the traffic...

Important e.g., Billing, provisioning, management,..



Inverting SYN-sampled flows

SYN-based sampling gives an unbiased estimator of the number of flows

Challenges:

- Need to inspect every packet
- Some flows have no SYN
- Only applicable to TCP
- Some flows have 2 SYNs

Inverting the sampled flows

Packet sampling inversion proven to be impossible! [Hohn & Veitch 2003]

Sample and hold inversion a new problem.

θ'_i = probability of sampling *i* packets from a stream.

$X_i \in N$ = Distribution of observable flow lengths

after loads of Maths, final estimate becomes:

$$\theta_i = \frac{X_i - qX_{i+1}}{C}$$
$$= \frac{X_i - qX_{i+1}}{1 - q + qX_1}.$$

Weaknesses of estimation

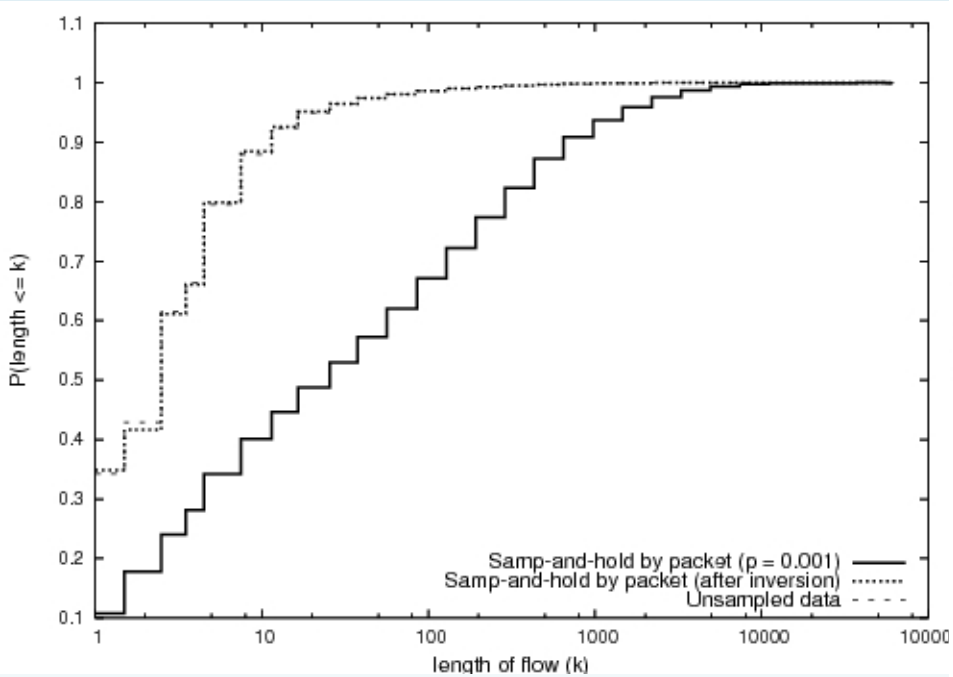
method wholly relies on the difference between X_i and X_{i+1}.

At large flows this creates problems.

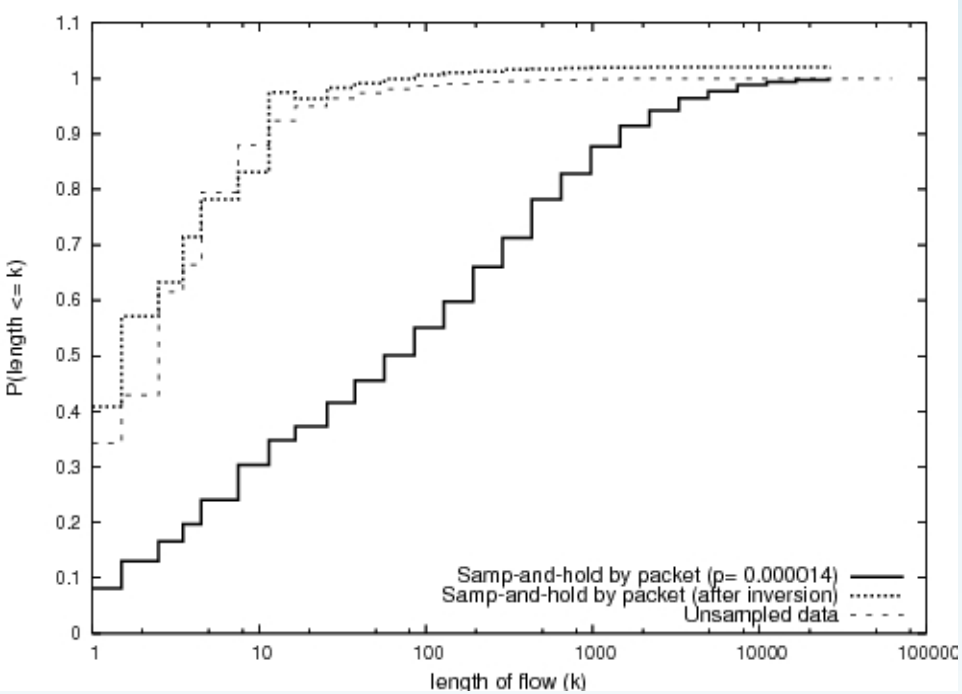
In particular, if $X_{i+1} > X_i$, the method will produce a negative estimate for the probability.

Estimate the probability that a flow has a length in the range *i, i+1, . . . , i+n*

Results

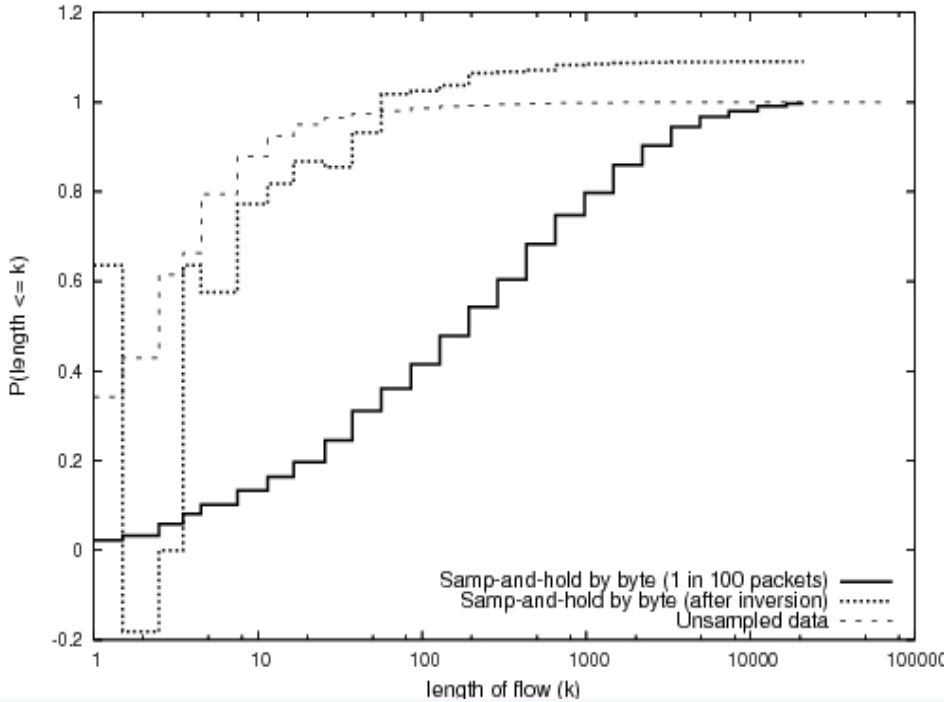


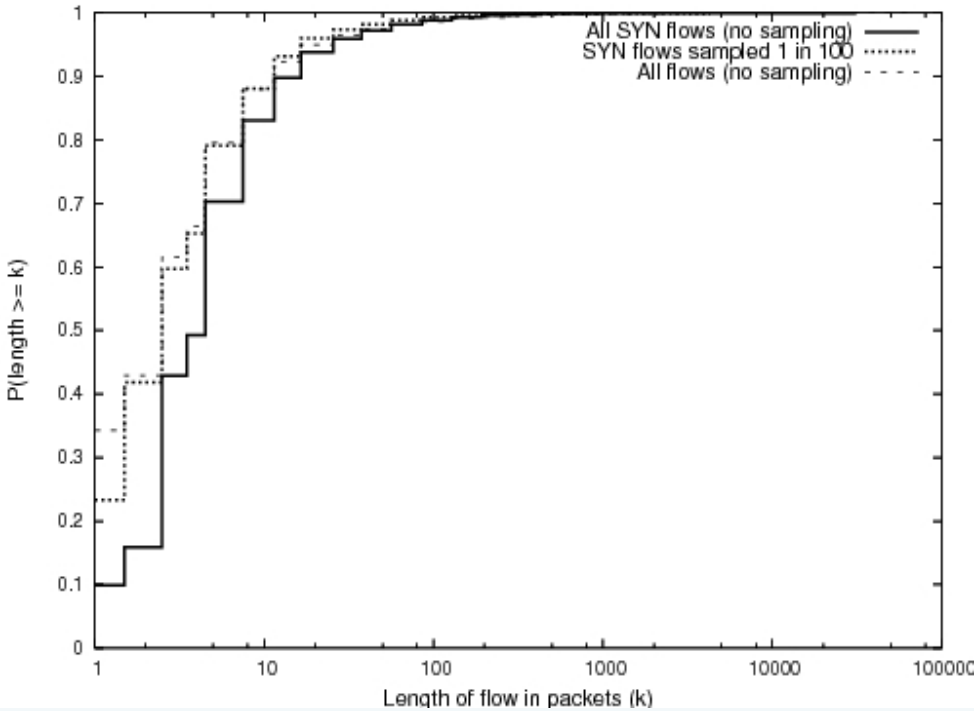
Sample & Hold, 1 in 1000 FLOWS



Sample & Hold, 1 in 1000 PACKETS

Results



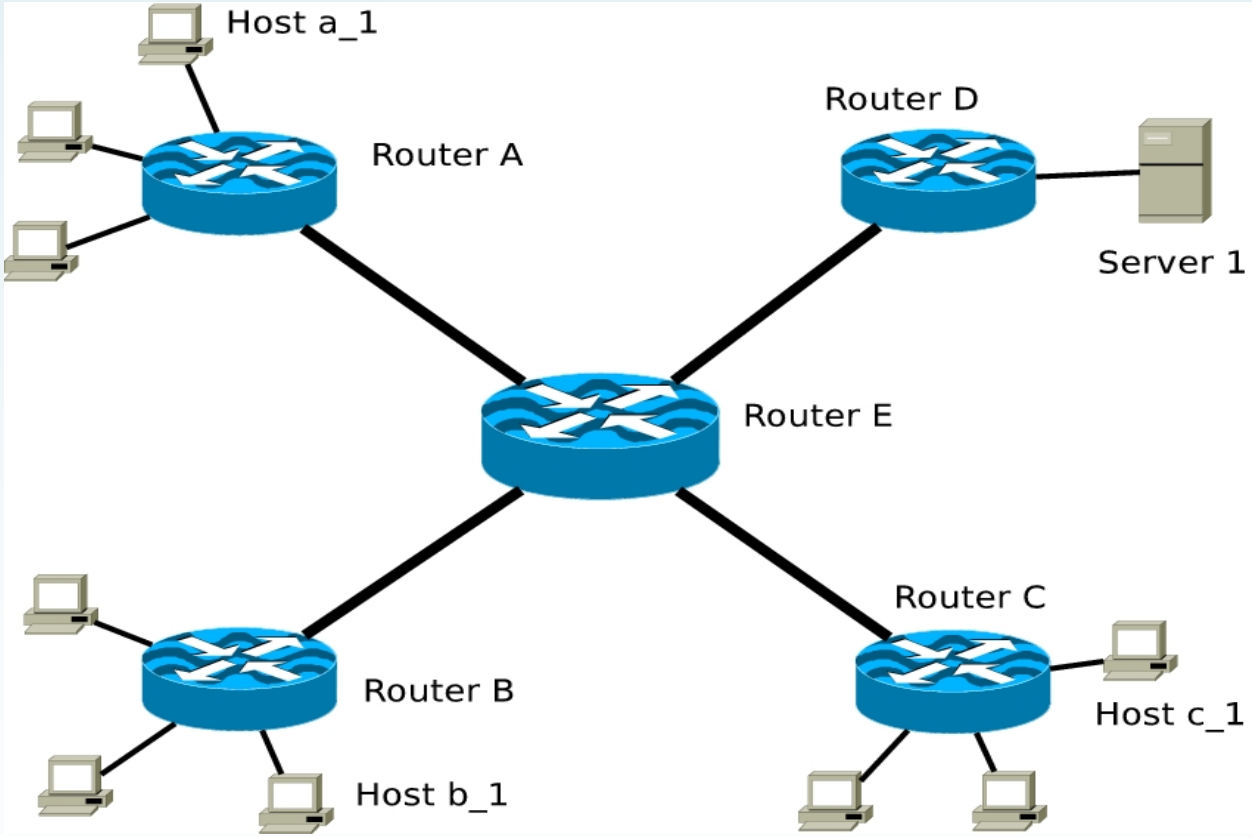


Sampled & Hold by byte inversion

SYN flows sampled 1 in 100

Future work

- Use More sources of information : TCP flags, SEQ Numbers, ACKs,..
- Topology-aware sampling, selective flow selection
- Inverting the traffic distribution across all the links to form TM



Thank you!!

Richard G. Clegg, Hamed Haddadi, Raul Landa, Miguel Rio, “Towards Informative Statistical Flow Inversion” May 2007 http://arxiv.org/abs/0705.1939

Hamed Haddadi, Raul Landa, Miguel Rio, Saleem Bhatti, “Revisiting the Issues On Netflow Sample and Export Performance” December 2006 http://arxiv.org/abs/0704.0730

http://www.ee.ucl.ac.uk/~hamed/