Packet Sampling for Flow Accounting: Challenges and Limitations

Tanja Zseby (Fraunhofer FOKUS)
Thomas Hirsch (Fraunhofer FOKUS)
Benoit Claise (Cisco Systems)

April 29, 2008

FOKUS

Fraunhofer Institute for Open Communication Systems

CISCO SYSTEMS

This work was funded by Cisco Systems as part of the VEGAS project.

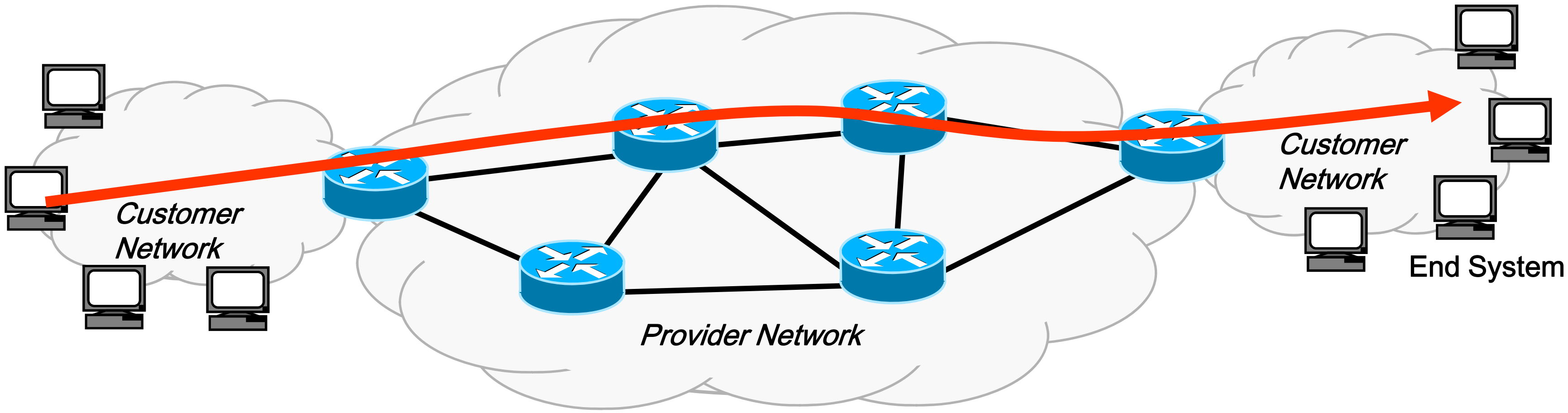
Outline

- Problem statement
- Packet selection techniques
- Accuracy assessment in theory
- Accuracy assessment in practice
- Experimental results
- Standardization (IPFIX/PSAMP)
- Conclusion

Usage-based Accounting

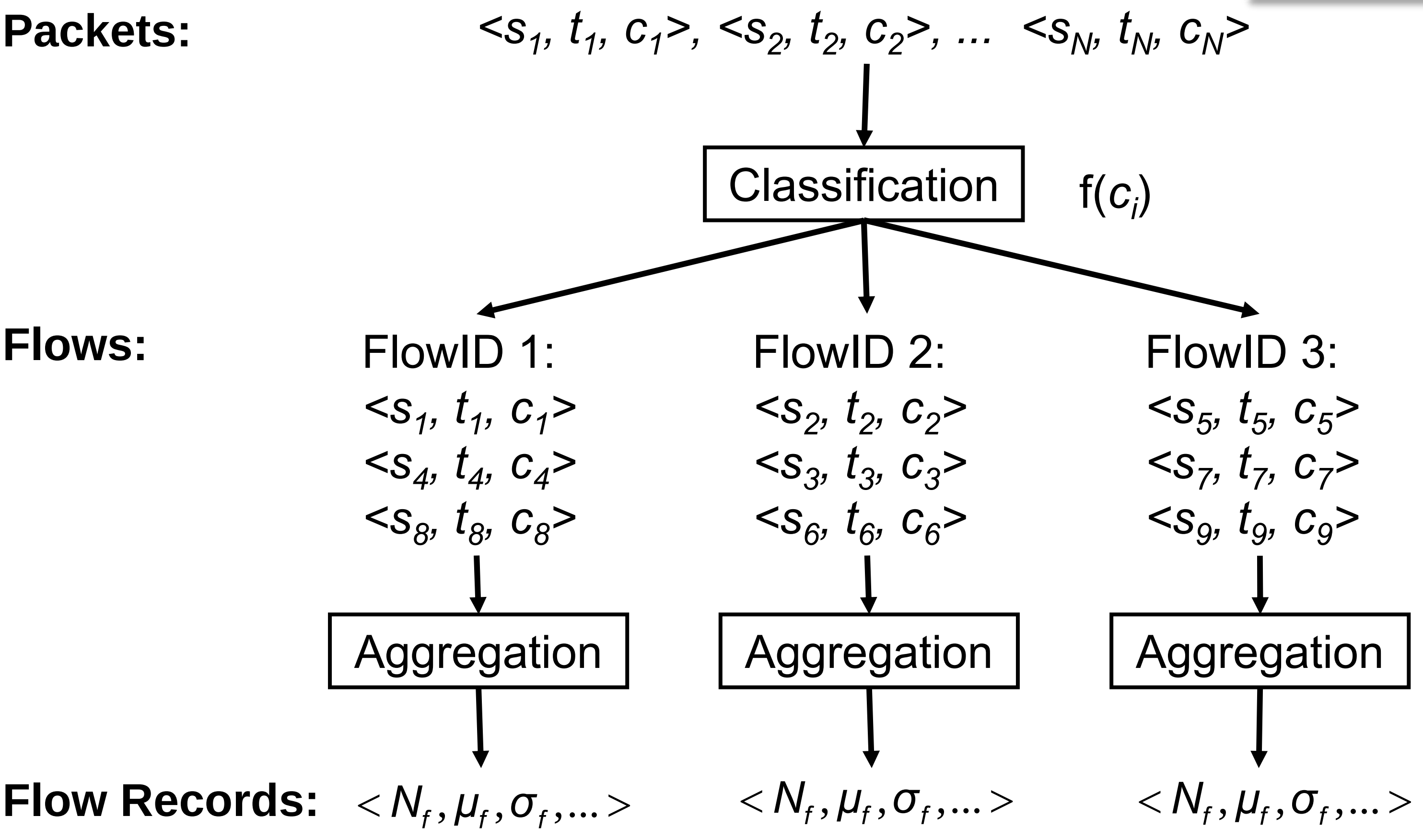
- Accounting based on flow volume (transferred bytes)
- Requires flow measurements
  - all packets from network A
  - all packets with DSCP=x
  - all VoIP packets
  - …

Flow:= *packets with common properties*



Flow Measurements

Packet Attributes at Observation Point
s_i, –sequence number
t_i – arrival time
c_i – content (header, payload)



Packets: $<s_1, t_1, c_1>, <s_2, t_2, c_2>, \ldots <s_N, t_N, c_N>$

Classification f(c_i)

Flows:

FlowID 1:
$<s_1, t_1, c_1>$
$<s_4, t_4, c_4>$
$<s_8, t_8, c_8>$

FlowID 2:
$<s_2, t_2, c_2>$
$<s_3, t_3, c_3>$
$<s_6, t_6, c_6>$

FlowID 3:
$<s_5, t_5, c_5>$
$<s_7, t_7, c_7>$
$<s_9, t_9, c_9>$

Aggregation

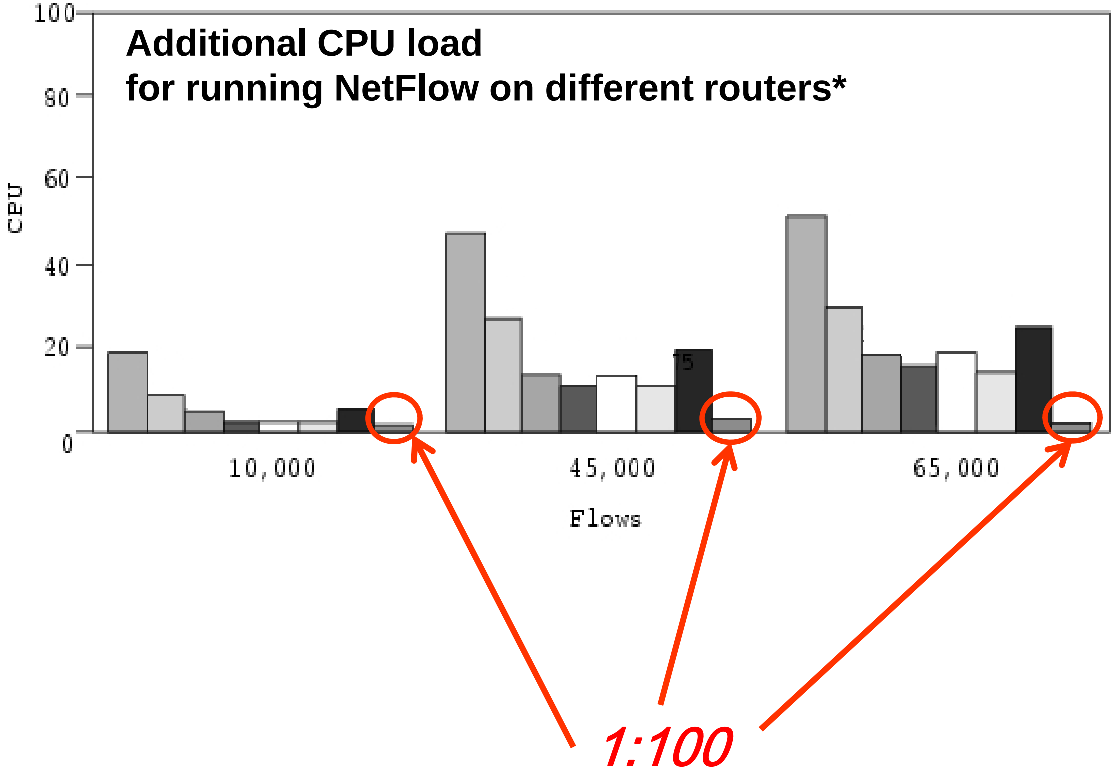
Aggregation

Aggregation

Flow Records: $< N_f, \mu_f, \sigma_f, \ldots >$ $< N_f, \mu_f, \sigma_f, \ldots >$ $< N_f, \mu_f, \sigma_f, \ldots >$

Problem: Resource Consumption

- Resource Limitations
  - Processing power
  - Transmission
  - Storage
- Demand depends on
  - Data rates
  - Required granularity
- Solutions
  - Dedicated Hardware
  - Improved Algorithms
  - Data Selection

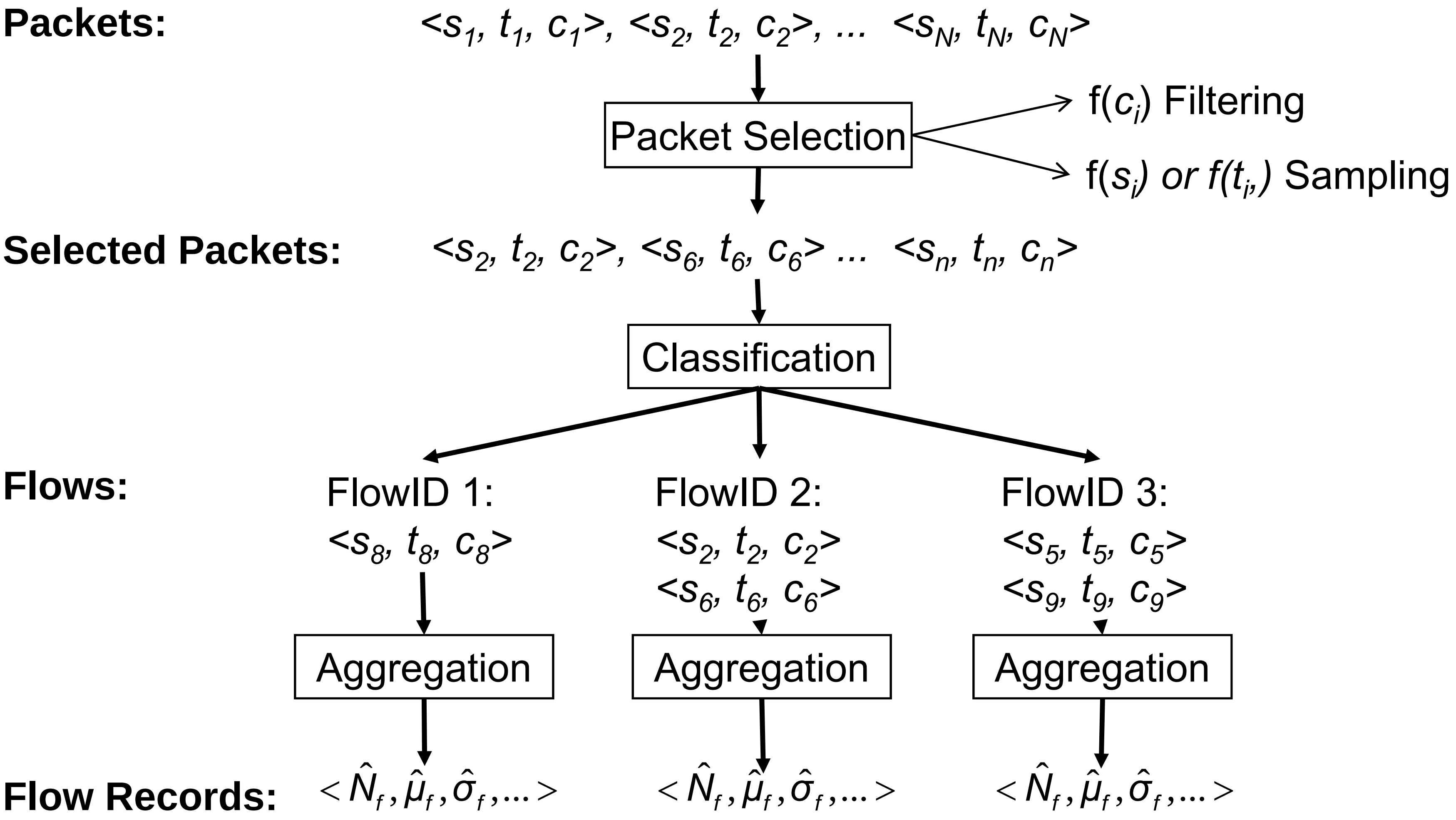


Additional CPU load
for running NetFlow on different routers*

1:100

*source: NetFlow Performance Analysis, Cisco white paper

Packet Selection



Packets: $<s_1, t_1, c_1>, <s_2, t_2, c_2>, \dots <s_N, t_N, c_N>$

Packet Selection

f(c_i) Filtering

f(s_i) *or f(t_i,)* Sampling

Selected Packets: $<s_2, t_2, c_2>, <s_6, t_6, c_6> \dots <s_n, t_n, c_n>$

Classification

Flows:

FlowID 1:
$<s_8, t_8, c_8>$

FlowID 2:
$<s_2, t_2, c_2>$
$<s_6, t_6, c_6>$

FlowID 3:
$<s_5, t_5, c_5>$
$<s_9, t_9, c_9>$

Aggregation

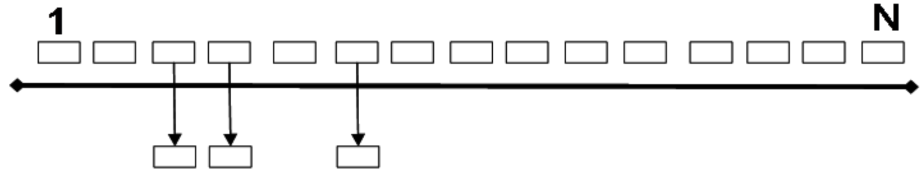
Aggregation

Aggregation

Flow Records: $< \hat{N}_f, \hat{\mu}_f, \hat{\sigma}_f, \dots >$ $< \hat{N}_f, \hat{\mu}_f, \hat{\sigma}_f, \dots >$ $< \hat{N}_f, \hat{\mu}_f, \hat{\sigma}_f, \dots >$

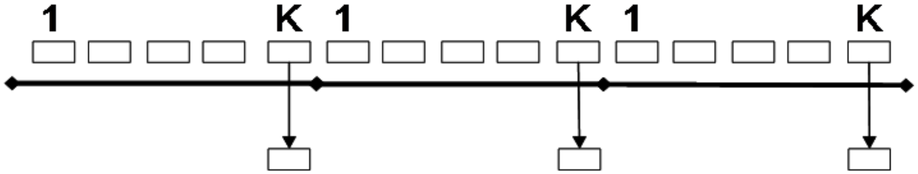
Packet Selection Techniques (Examples)

Random n-out-of-N:



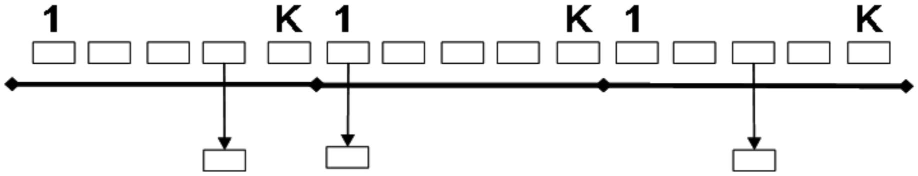
MI with N elements
n-of-N selected
N=15, n=3

Systematic:



MI with N elements
every Kth selected
N=15, K=5, n=N/K=3

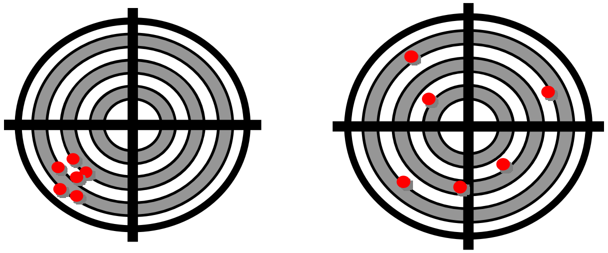
Random 1-in-K:



L subintervals with K elements
L x 1-of-K selected
N=15, K=5, L=N/K=3, n=L=3

Problem: Accuracy Assessment

Accuracy Assessment required



- Achievable accuracy depends on
  - Sampling and estimation method
  - Sampling parameters
  - Population characteristics 🡐 **unknown and highly dynamic**
- Accuracy assessment **during** measurement
  - For each measurement interval
  - For each flow
  - Based on sampled data

Theoretical Model

Case: n-out-of-N, sampling ***before*** classification

N	Packets in MI
n	Packets in sample
Sum_f	Volume in flow f
N_f	Number of packets in flow f
n_f	Sampled packets from flow f
μ_f	Mean packet size in flow f
$\sigma^2_{x_f}$	Packet size variance in flow f
$x_{f,i}$	Bytes in i-th packet

Estimation of Flow Volume: $\hat{S}um_f = \frac{N}{n} \cdot \sum_{i=1}^{n_f} x_{f,i}$

Variance of Estimate: $V[\hat{S}um_f] = V[\frac{N}{n} \cdot \sum_{i=1}^{n_f} x_{f,i}] = \frac{N^2}{n^2} \cdot V[\sum_{i=1}^{n_f} x_{f,i}]$

random variables

Estimation accuracy for flow f :

$$StdErr_{rel} = \frac{1}{N_f \cdot \mu_f} \cdot \sqrt{\frac{N\, N_f \cdot (\sigma_f^2 + \mu_f^2)}{n} - \frac{N_f^2 \cdot \mu_f^2}{N}}$$

Flow Characteristics:

Number of packets from flow f in MI

Packet size variance in flow f

Packet size mean in flow f

Sampling Parameters:

Number of all packets in MI (population size)

Number of selected packets in MI (sample size)

Accuracy Assessment in Practice

- Flow characteristics unknown
  - Estimation from sampled data
- Storing per-packet information too costly
  - Storing aggregates
- NetFlow Records
  - Number of packets stored
  - Sum of packet sizes stored
  - Calculation/estimation of mean packet size possible
  - **BUT:** calculation/estimation of packet size variance not possible

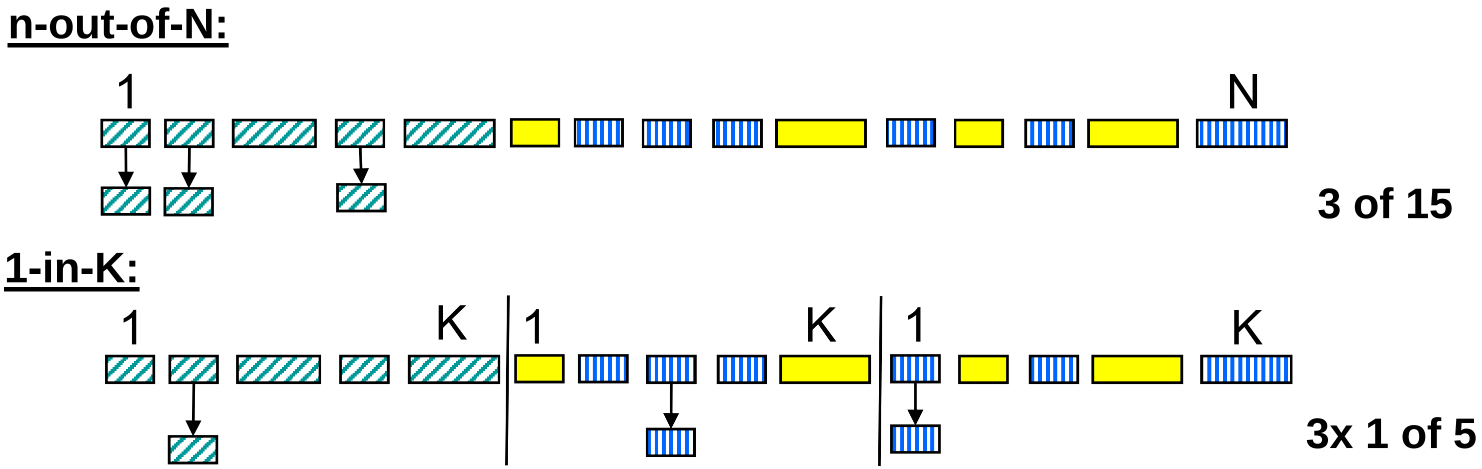
NetFlow Records: $N_f, \sum x_f \; , \sum x_f^2$

Store sum of squares !

1-in-K Sampling (Cisco)

- 1-in-K: Count-based **stratification** with **equal allocation**
  - Packet selection limited to 1 packet per subinterval
  - Theoretical Model ➔ see paper
- Stratification gain
  - Depends on variance of packet sizes from flow f in strata
  - 1 packet per sub-interval selected

  ➔ not sufficient to estimate variance in sub-interval



n-out-of-N:

1

N

3 of 15

1-in-K:

1

K

1

K

1

K

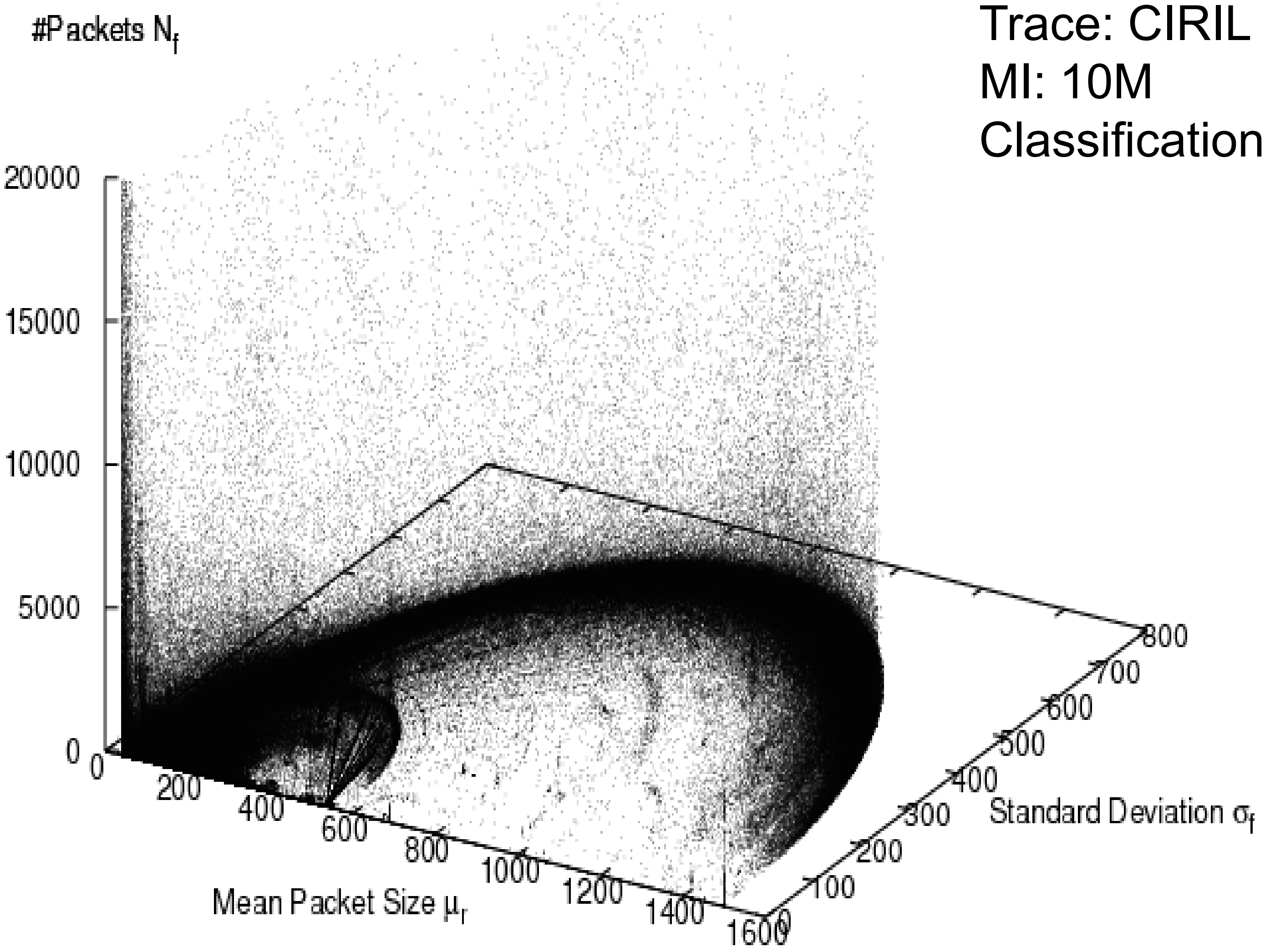
3x 1 of 5

Experiments

- Setup
  - Traces from three different networks
  - Different sampling schemes
  - Different classification schemes
  - Different measurement interval lengths
  - Sampling before and after classification

- Accuracy Calculation
  - based on theoretical model
  - using real flow characteristics

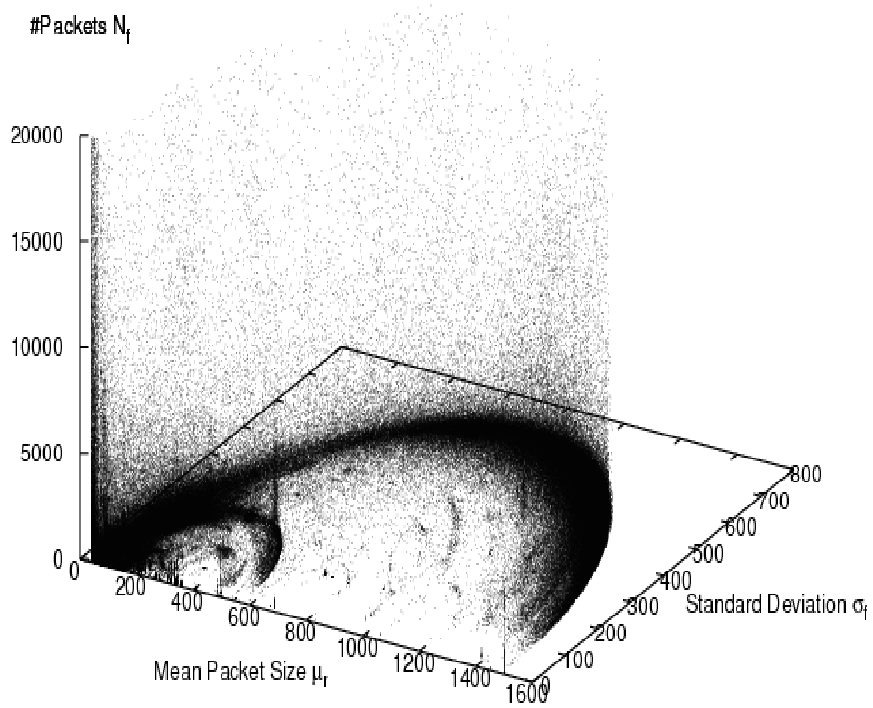
Flow Characterstics



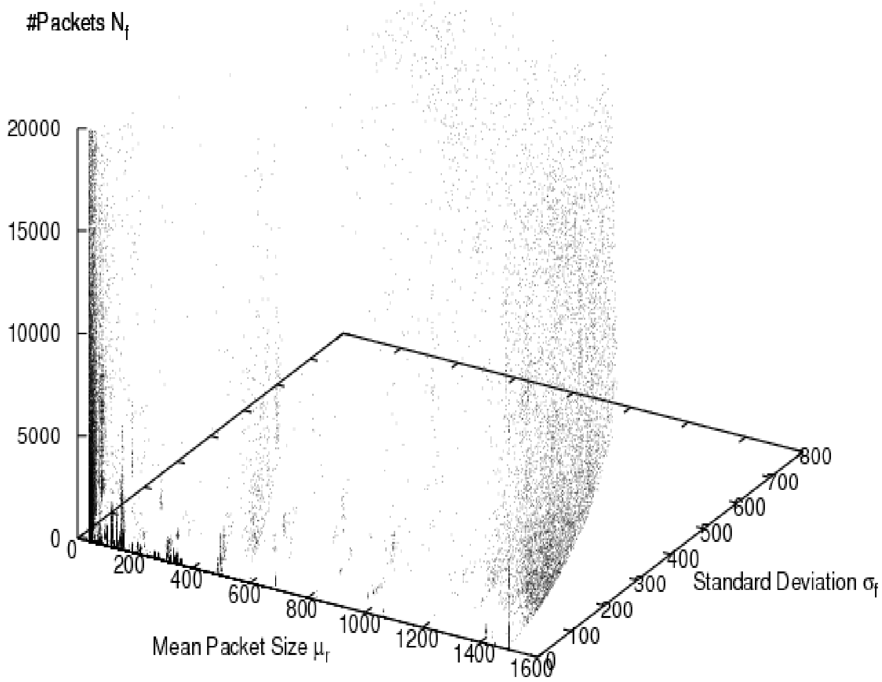
Trace: CIRIL
MI: 10M
Classification S24D00

Conformant Flows

Sampling fraction: 5%, StdErr ≤0.05



Sampling ***after*** classification



Sampling ***before*** classification

Sampling Experiments

- 1000 sampling runs per experiment
- Different sampling rates
- Calculation of bias and standard error
- Comparison of schemes
  - n-out-of-N
  - 1-in-K
  - systematic

Conformant Flows

Trace: NZIX
MI: 1M
Classification S24D00
Sampling fraction =5%

Max rel. StdErr	Error/CL	n-of-N	1-in-K	Systematic
0.003876	0.01/99%	0	0	0
0.005102	0.01/95%	0	0	0
0.019380	0.05/99%	64	64	62
0.025510	0.05/95%	72	72	83
0.051020	0.1/95%	473	475	567
0.076531	0.15/95%	1406	1425	1580
0.102041	0.2/95%	2316	2568	2860
0.1531	0.3/95%	5146	5397	5799
>0.1531	-	79383	79383	79383

Results

- Comparison of schemes
  - n-out-of-N close to n-out-of-N model
  - 1-in-K close to n-out-of-N model
  - Systematic sampling
    - Better results for some flows
    - But unpredictable (high variance of results)
    - Differs from model
- Higher accuracy achievable with
  - Larger sample fraction
  - Longer observation periods (if flow characteristics remain)
  - Coarse grained classification
  - Aggregation of flows

IPFIX/PSAMP IEs

- IP Flow Information Export (IPFIX)
  - Standard for flow information export (RFC5101)
  - Information elements (RFC5102)
- Packet Sampling (PSAMP)
  - Packet selection techniques (filtering, sampling)
  - Packet export using IPFIX

Parameter	**IPFIX/PSAMP IEs**
Number *N* of packets in measurement interval	`samplingPopulation`
Number *n* of packets in sample	`samplingSize`
Number of packets from flow *f* in sample	`packetTotalCount`
Sum (bytes in sampled packets)	`octetTotalCount`
Sum of squares (bytes in sampled packets)	`octetTotalSumOfSquares`

Conclusion

- Accuracy Assessment in theory and practice
  - n-out-of-N (before/after classification) ➔ store sum of squares
  - 1-in-K ➔ not possible in practice (although model exists)
- Experiments
  - Small flows ➔ poor accuracy for sampling before classification
  - 1-in-K close to n-out-of-N
  - Accuracy depends on settings (obs. period, classification)
  - Alternative: Flow selection based on expected accuracy
- IPFIX provides required information elements
- Work in progress:
  - Sampling for other metrics (e.g. for anomaly detection)
  - Hash-based selection

Thank you!

tanja.zseby@fokus.fraunhofer.de

FOKUS Open Source IPFIX Library:

http://net.fokus.fraunhofer.de/libipfix/

Measurement data always welcome at:

http://www.ist-mome.org/

FOKUS

Fraunhofer Institute for Open
Communication Systems