Traffic classification under sampling

Dario Rossi
dario.rossi@enst.fr
http://www.enst.fr/~drossi

Joint work with
Silvio Valenti
Davide Tammaro
Antonio Pescape'

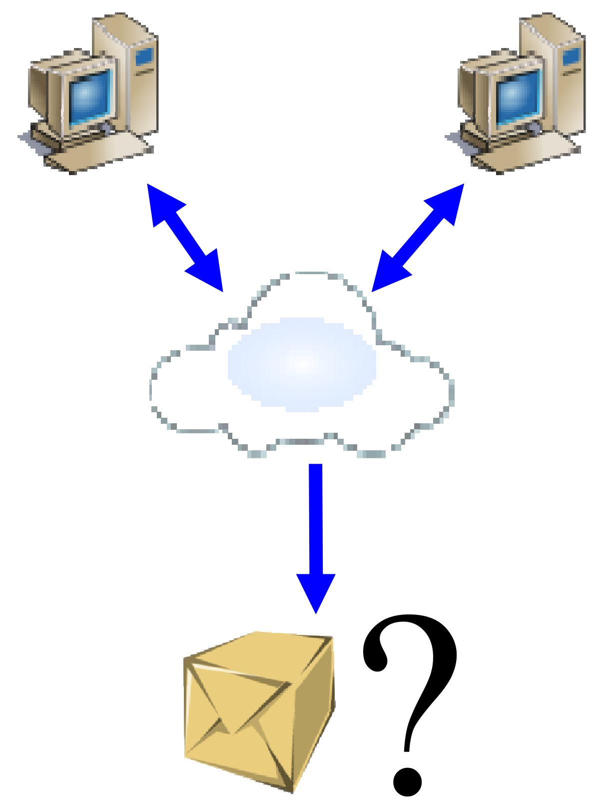
4th NMRG workshop, IETF83, 31 March 2012

Agenda

- Traffic classification taxonomy
- Sampling
- Methodology
  - Dataset, tools, workflow, metrics, etc.
  - Sampling strategies
- Experimental results
  - Feature distortion
  - Classification accuracy
- Conclusions
- Advertisement
- Further advertisement
  - if time allows and audience interested :)

Traffic classification

- Problem
  - Look at packets in the network and guess which application has generated them
- Applications
  - Intrusion Detection System
  - Quality of Service
  - Lawful interception
- Challanges
  - Applications try to cheat (well-known ports no longer reliable),
  - Applications evolve (proprietary protocols, encryption…)
  - Lightweight to keep up with modern network speed



Traffic classification taxonomy

Approach	Subcategory	Granularity	Timeliness	Complexity	Comment
Payload Based	[1,2] Deep Packet Inspection (DPI)	Fine-grained individual applications	Early (first few packets).	Access to packet payload of first few packets. Moderate cost	Deterministic technique;
	KISS[ToN'10] Stochastic Packet Inspection	Fine-grained individual applications	Online (100s packets windows)	Access to packet payload of several packets. High cost	Robust technique
Statistical Analysis	[4,5,6,7]	Coarse-grained, class of application	Late (after the flow end).	Access to flow-level information Lightweight cost	Post-mortem analysis
	[8,9]	Fine-grained individual applications	Early (first 5 packets)	Access to first few packets Lightweight cost	On the fly classification
Behavioral Analysis	[10,11]	Coarse-grained, class of application	Late (after the flow end).	Lightweight	Post-mortem analysis
	Abacus [ComNet'11]	Fine-grained, individual P2P applications	Online (1s-5s seconds windows)	Lightweight	Online classification Limited to P2P

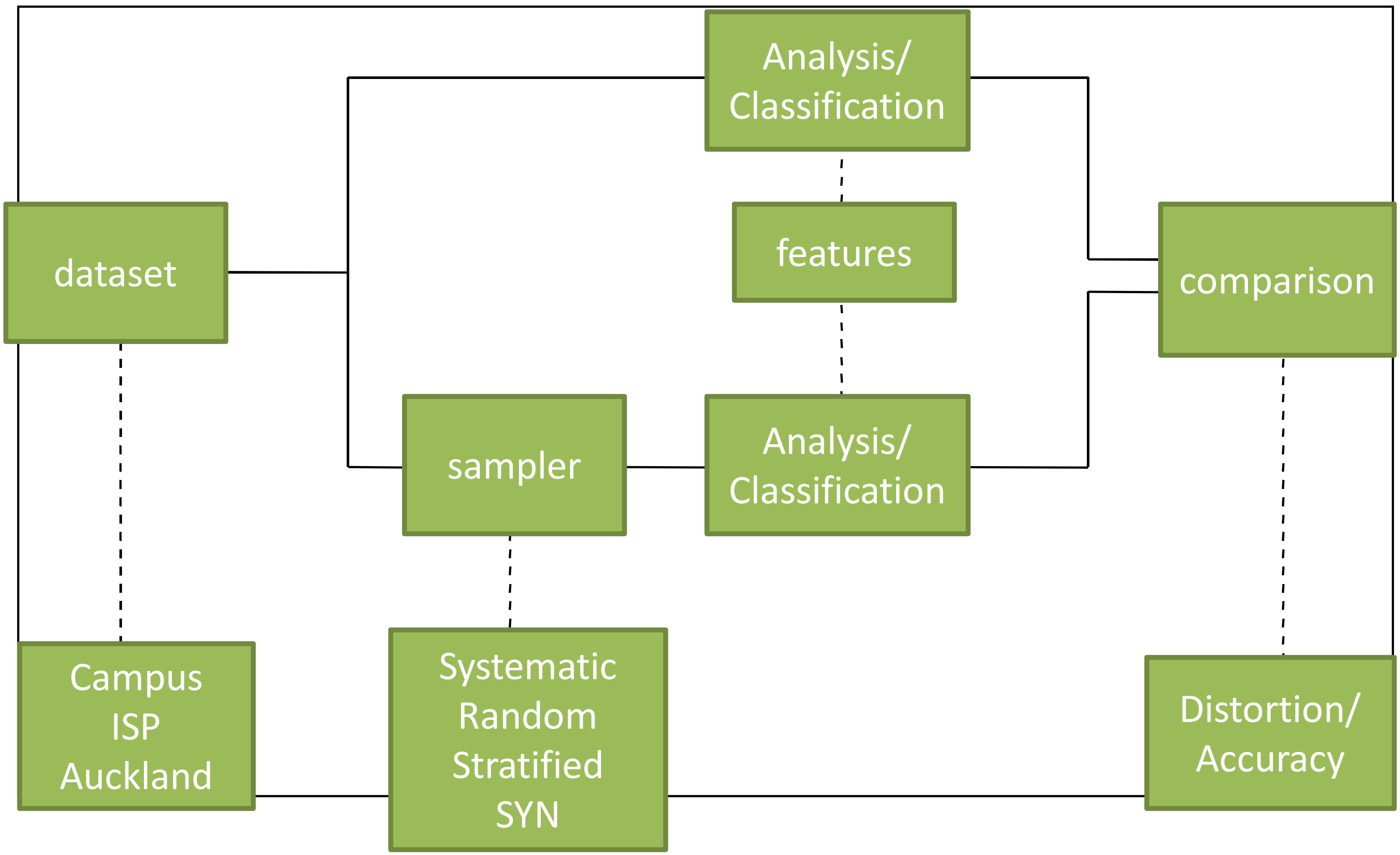
Traffic classification taxonomy (refs)

- [1] S. Sen, O. Spatscheck, D. Wang, "Accurate, Scalable In-Network Identification of P2P Traffic Using Application Signatures", 13th International Conference on World Wide Web (WWW'04), pp. 512-521, New York, NY, May 2004.
- [2] AW. Moore, K. Papagiannaki, "Toward the Accurate Identification of Network Applications", In Passive and Active Measurement (PAM'05), Boston, MA, USA, March/April 2005
- [3] J. Ma, K. Levchenko, C. Kreibich, S. Savage, G. M. Voelker, "Unexpected Means of Protocol Inference", 6th ACM SIGCOMM Internet Measurement Conference (IMC'06), pp. 313-326, Rio de Janeiro, BR, October 2006.
- [4] A. McGregor, M. Hall, P. Lorier, J. Brunskill, "Flow Clustering Using Machine Learning Techniques", PAM'04, Antibes Juan-les-Pins, Fr., pp. 205-214, April 2004.
- [5] M. Roughan, S. Sen, O. Spatscheck, N. Duffield, "Class-of-Service Mapping for QoS: a Statistical Signature-based Approach to IP Traffic Classification", 4th ACM SIGCOMM Internet Measurement Conference (IMC'04), Taormina, IT, pp. 135-148, October 2004.
- [6] A. W. Moore, D. Zuev, "Internet Traffic Classification Using Bayesian Analysis Techniques", ACM SIGMETRICS '05, Banff, Alberta, Canada, 2005
- [8] L. Bernaille, R. Teixeira, K. Salamatian, "Early Application Identification," Conference on Future Networking Technologies (CoNEXT'06), Lisboa, PT, December 2006.
- [9] M. Crotti, M. Dusi, F. Gringoli, L. Salgarelli, "Traffic Classification Through Simple Statistical Fingerprinting", ACM Computer Communication Review, Vol. 37, No. 1, pp.5-16, January 2007. Jan 2007
- [10] T. Karagiannis, K. Papagiannaki, M. Faloutsos "BLINC: Multilevel Traffic Classification in the Dark", ACM Communication Review, Vol. 35, No. 4, pp. 229 - 240, 2005
- [11] K. Xu, Z. Zhang, S. Bhattacharyya, "Profiling Internet Backbone Traffic: Behavior Models and Applications", ACM SIGCOMM'05, Philadelphia, PA, pp. 169-180, August 2005.

S mp i g

- Why sample Internet traffic ?
  - To reduce computation & storage
- How much information do we loose ?
  - Monitoring [ITC22]
  - Classification [IJNM’12,TRAC’11]
- In the reminder of this talk [IJNM’12]
  - (see advertisement)

Workflow



Analysis/
Classification
dataset
features
comparison
sampler
Analysis/
Classification
Campus
ISP
Auckland
Systematic
Random
Stratified
SYN
Distortion/
Accuracy

Dataset

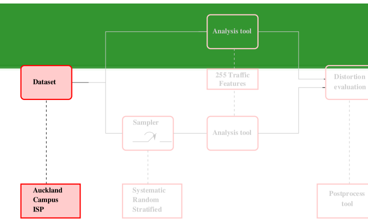
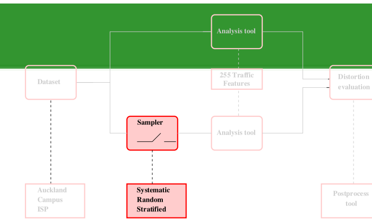


Table II. Subset of the dataset used for classification, and application breakdown.

	UniBS		Campus		Auckland	
	Flow	Byte	Flow	Byte	Flow	Byte
Protocol	%	%	%	%	%	%
HTTP	49.3	5.6	41.8	62.7	34.8	25.3
HTTPS	1.5	1.2	41.8	30.6	34.8	23.4
FTP	-	-	4.8	0.03	-	-
IMAPS	3.7	0.1	0.2	3.9	0.6	0.9
POP3	1	0.01	-	-	5.6	2.8
SMTP	-	-	-	-	23.9	47.5
Skype	1	0.7	11.1	2.6	-	-
eDonkey	40.1	87.2	-	-	-	-
BitTorrent	3.3	5.0	-	-	-	-

IPs	410K	61K	81K	6.59K
Available at	[30]	–	–	[31]
Ground truth	Port-based	–	DPI [32]	gt [33]



Sampling

Classification

SYN-sampling

= Systematic + SYN set

SYN needed to have a recall for all flows!

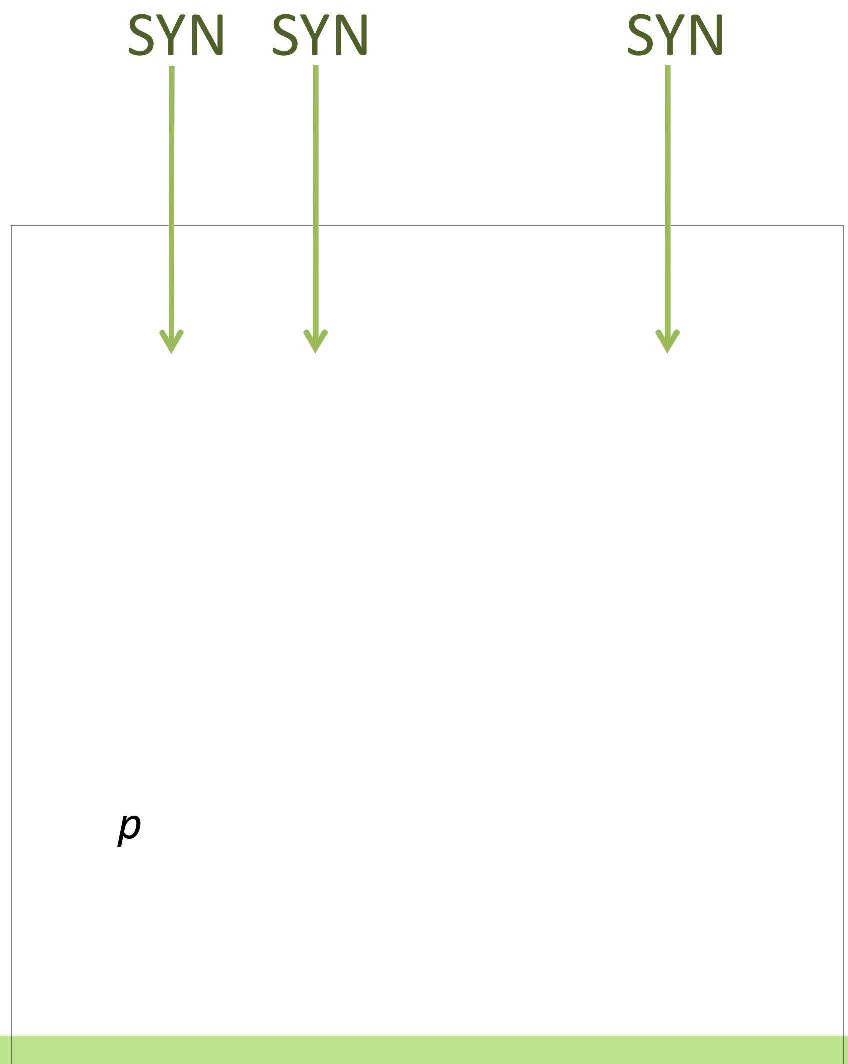
Monitoring

Systematic

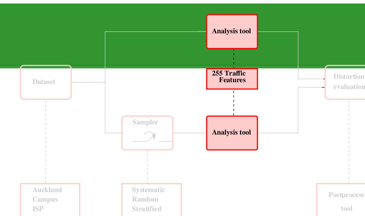
Stratified sampling

Random sampling

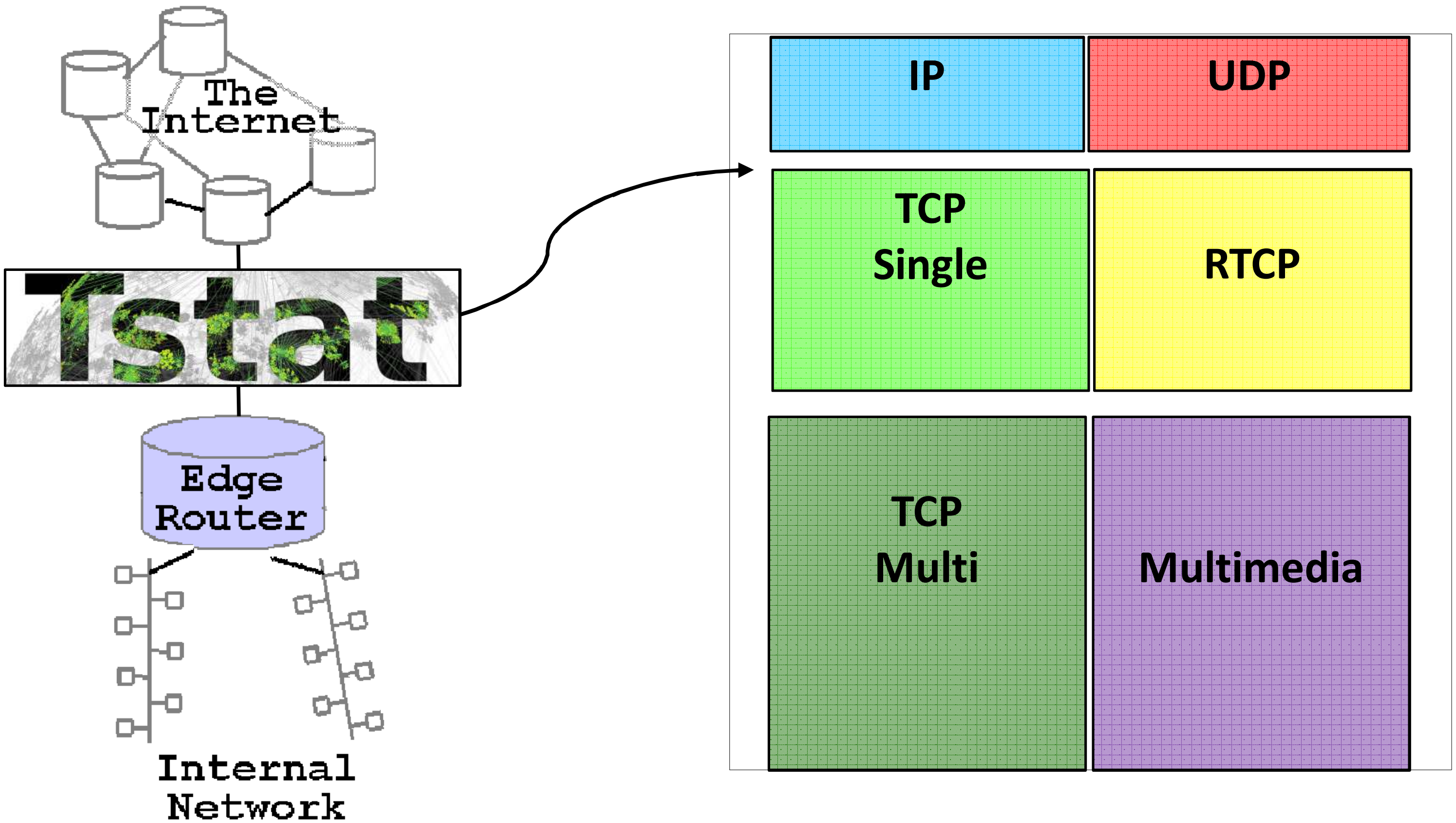
k=sampling period and $p=1/k$



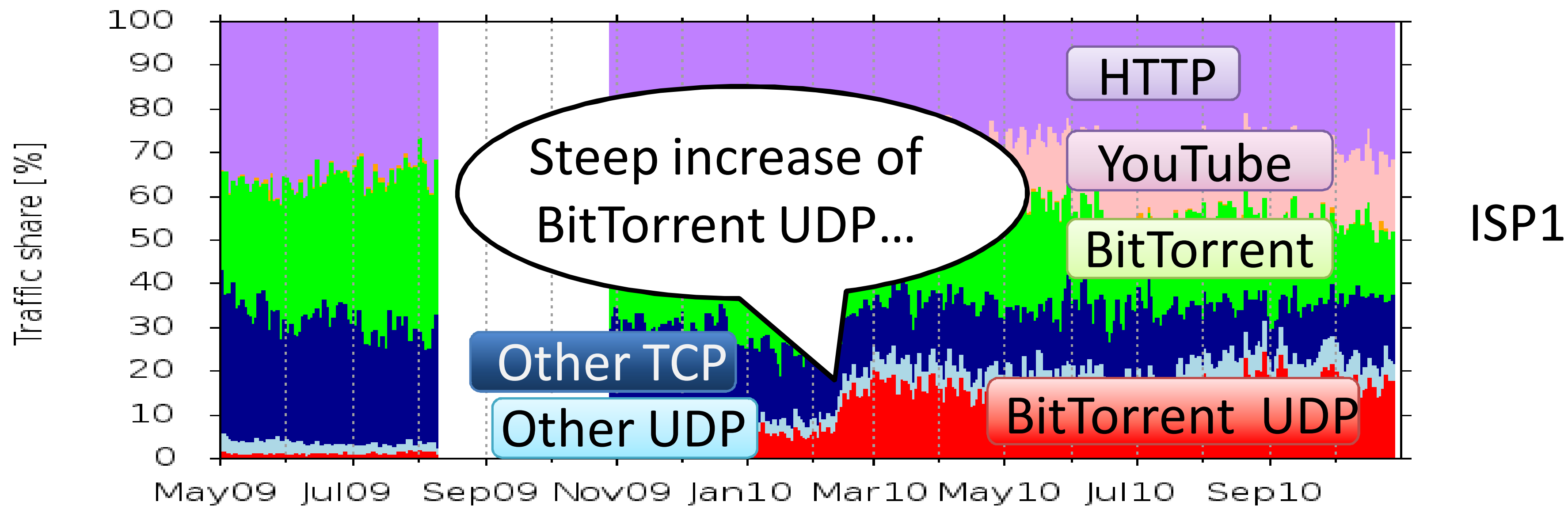
Features



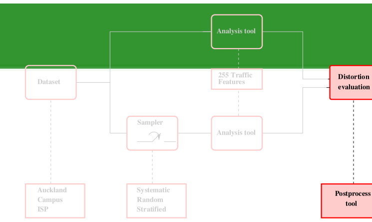
Tstat is a L4 traffic analyzer, produces >255 packet/flow level features



Unsampled Classification in Tstat



Analysis/Classification



- Traffic monitoring
  - Distance of feature distribution for aggregate [ITC22]
    - Hellinger Distance, Kullback-Leibler, Fleiss Chi-square
  - Distortion of individual flow features [IJNM’12]
    - Relative error, correlation coefficient
    - Instrumental for traffic classification

- Traffic classification
  - C45 trees (in this talk) and SVM
  - Accuracy w.r.t ground-truth

Feature distortion (1/2)

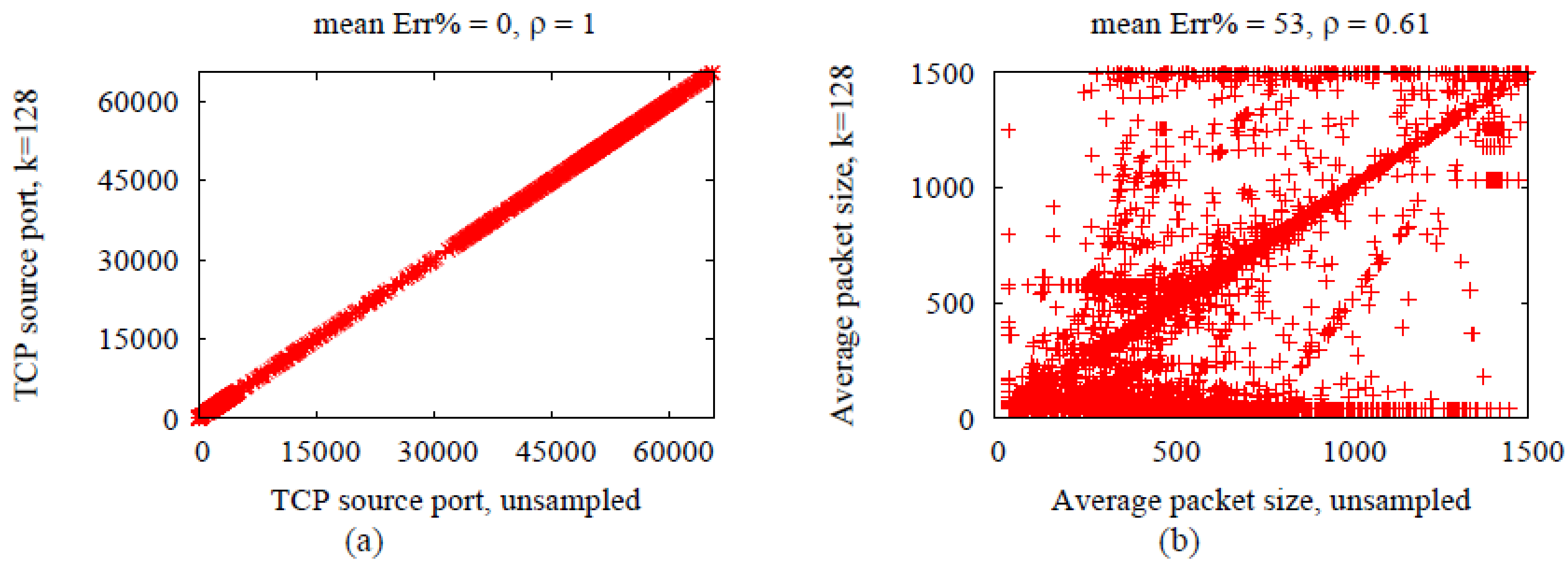
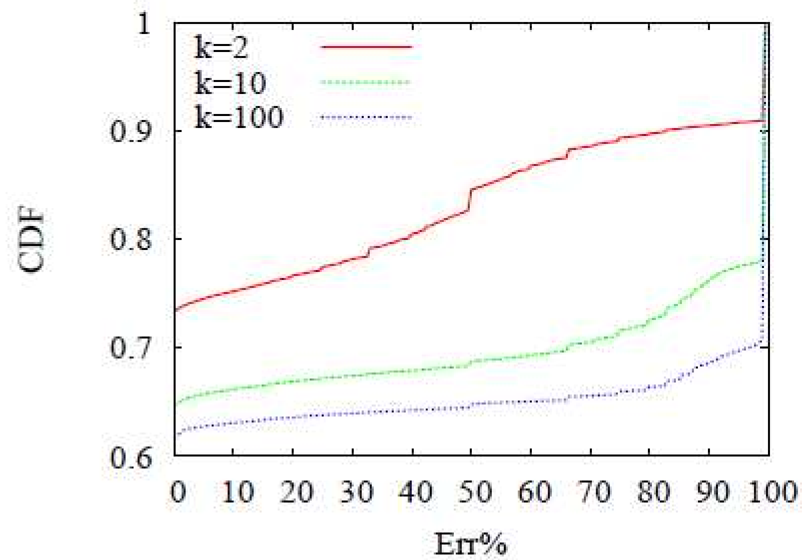


Figure 3. Example of distortion of per-flow features (Campus dataset): scatter plot of TCP source port (a) and average packet size (b) for unsampled vs sampled traffic, along with statistical indexes of correlation.

Feature distortion (2/2)



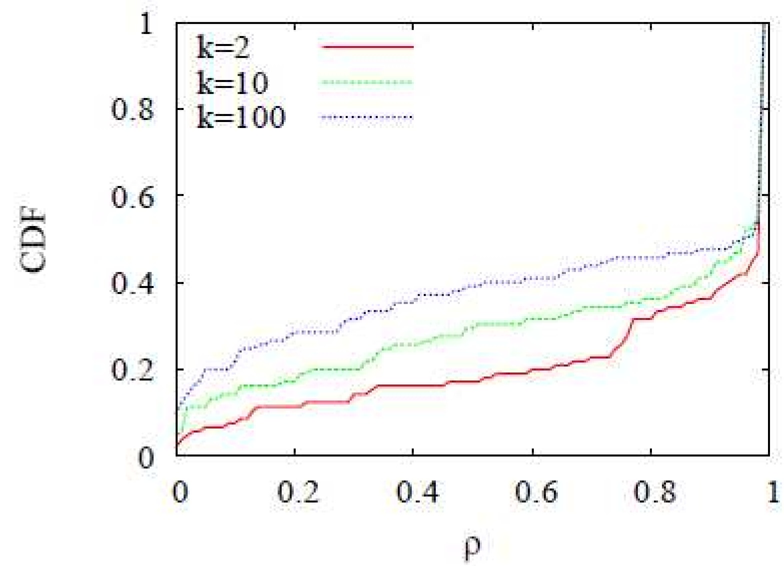


Figure 8. CDF of (left) Err% and (right) ρ for UniBS trace and different sampling step.

Expected impact on classification ?

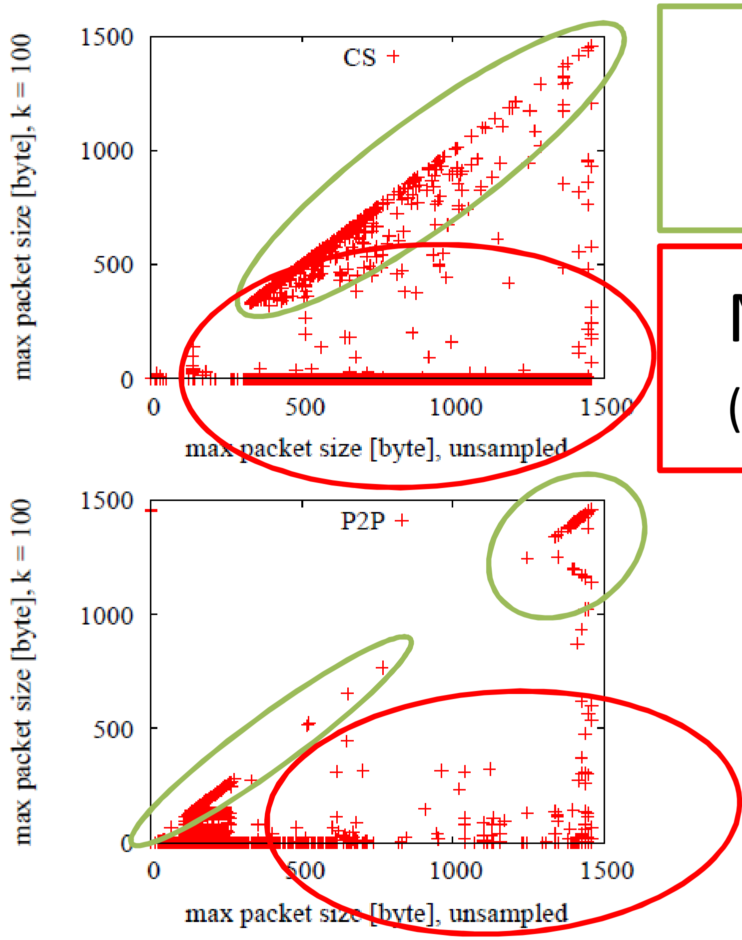


Figure 9. Scatter plot of features values for unsampled and SYN Sampling $k = 100$ for UniBS trace, contrasting peer-to-peer (P2P) and traditional client-server (CS) applications.

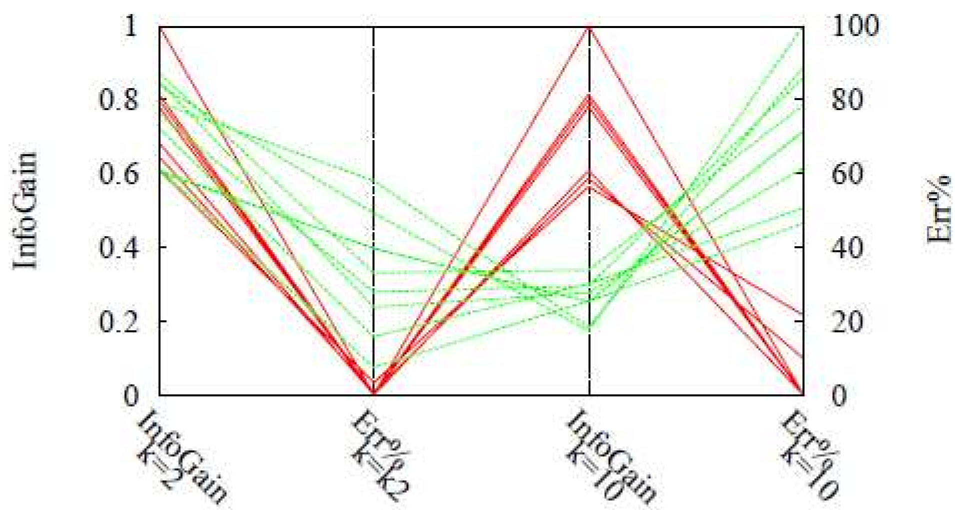
Some samples can still discriminate

Most samples can't
(pkt size==0 => only SYN seen)

Expect poor performance, unless… any guess?

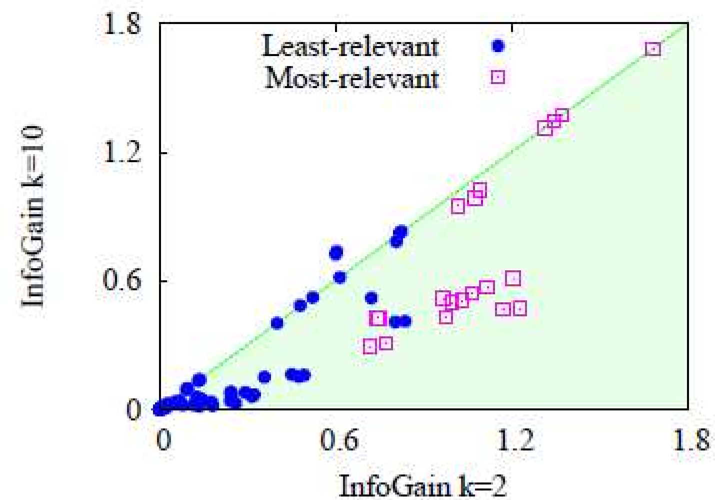
Features at different sampling step K

Single packet feature
Multiple packets features



(a)

Most relevant
Least relevant



(b)

Figure 10. (a) Parallel coordinates plot for most-relevant features and (b) scatter plot of information gain for all features with $k = 2, 10$.

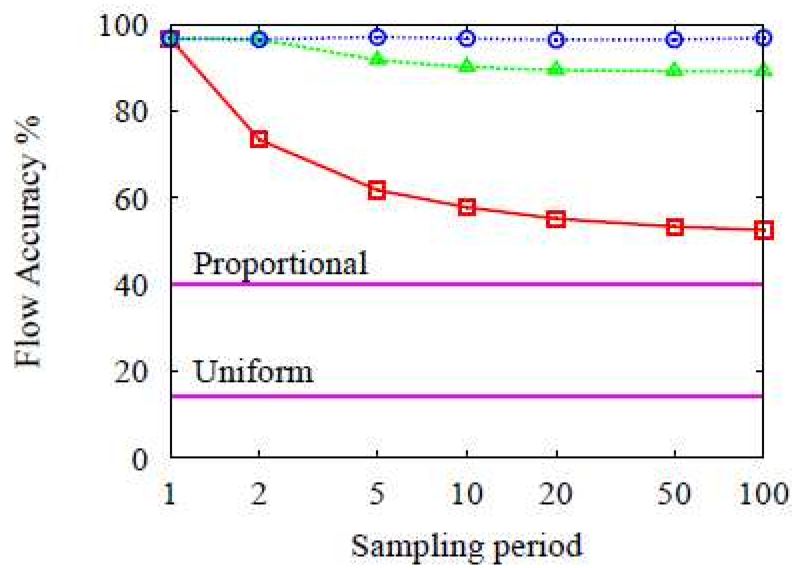
Most relevant features at different K

Table V. Feature Information gain for UniBS trace at different sampling rates.

Features	Unsampled		Sampled k=2		Sampled k=10	
	Score	Rank	Score	Rank	Score	Rank
Server-IP-address	**1.68**	1	**1.68**	1	**1.68**	1
cwin-min-c2s	**1.49**	2	**1.20**	6	0.60	14
min-seg-size-c2s	**1.48**	3	**1.22**	5	0.47	23
cwin-max-c2s	**1.47**	4	**1.11**	8	0.56	15
max-seg-size-c2s	**1.43**	5	**1.17**	7	0.46	24
initial-cwin-c2s	**1.41**	6	0.71	26	0.29	32
First-time	**1.37**	7	**1.37**	2	**1.37**	2
cwin-min-s2c	**1.35**	8	**1.06**	11	0.53	16
Server-TCP-port	**1.34**	9	**1.34**	3	**1.34**	3
initial-cwin-s2c	**1.33**	10	0.77	22	0.30	31
Client-IP-address	**1.31**	11	**1.31**	4	**1.31**	4
cwin-max-s2c	**1.28**	12	0.99	14	0.49	21
min-seg-size-s2c	**1.22**	13	0.96	16	0.51	19
max-seg-size-s2c	**1.21**	14	**1.03**	12	0.50	20
Last-time	**1.14**	15	**1.09**	9	**1.02**	5
win-max-s2c	**1.08**	16	**1.07**	10	0.98	6
Completion-time	**1.03**	17	0.97	15	0.42	25
win-min-s2c	**1.02**	18	**1.01**	13	0.94	7
unique-byte-s2c	**1.02**	19	0.74	23	0.42	27
data-byte-s2c	**1.01**	20	0.74	24	0.42	26

Training policy matters

- Classification accuracy remains if we train and classify with the same sampling step !



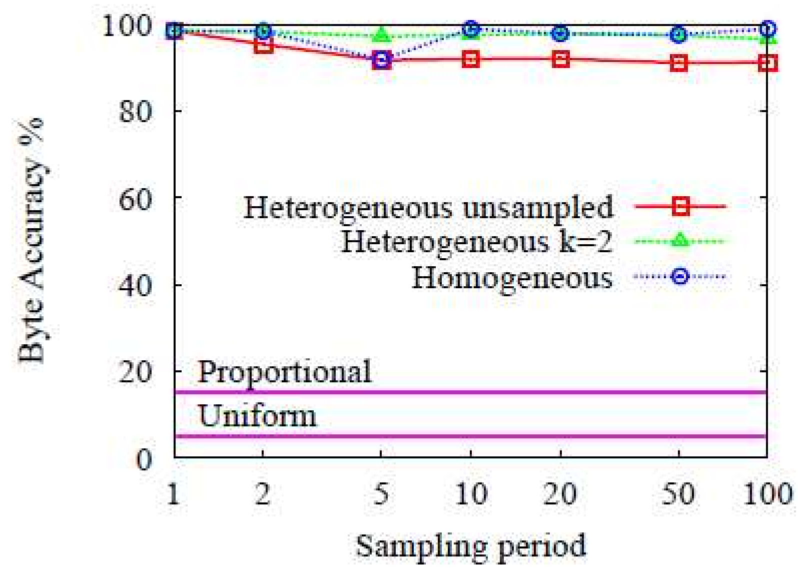


Figure 12. Impact of Homogeneous vs Heterogeneous training set policies at varying sampling rates in terms of flow and byte accuracy.

Conclusions

- Traffic sampling
  - Heavy distortion on aggregated features
    Details in [ITC22]
  - Heavy distortion on individual features
    Details in [IJNM'12]
  - Classification still accurate if training and testing at homogeneous sampling rates [IJNM'12]
    - Distorted features preserve distance in some non geometric space (e.g., gaussian SVM, or information gain amount for C45 trees)

Advertisement: selected publications

Classification

- **[Sigcomm'07]** D. Bonfiglio, M. Mellia, M. Meo, D. Rossi and P. Tofanelli, *Revealing Skype Traffic: When Randomness Plays with You* . ACM SIGCOMM Computer Communication Review, 37(4): 37-48, 2007.
- **[Ton'10]** Finamore, M. Mellia, M. Meo and D. Rossi**,** *KISS: Stochastic Packet Inspection Classifier for UDP Traffic* . IEEE Transactions on Networking, 18(5):1505 - 1515, October 2010.
- **[ComNet'11]** P. Bermolen, M.Mellia, M. Meo, D. Rossi and S. Valenti, *Abacus: Accurate, Fine-Grained Classification of P2P-TV Traffic* . Elsevier Computer Networks, April 2011.

Classification & Sampling

- **[Trac'11]** S. Valenti and D. Rossi*, Fine-grained behavioral classification in the core: the issue of flow sampling* . In IEEE TRAC'11 , Istanbul, Turkey, 5-9 July 2011.
- **[ITC'22]** Pescape, D. Rossi, D. Tammaro and S. Valenti**,** *On the impact of sampling on traffic monitoring and analysis* . In ITC22, Amsterdam, The Netherlands, September 7 - 9 2010
- **[IJNM'12]** Davide Tammaro, Silvio Valenti, Dario Rossi, Antonio Pescape*, Exploiting packet sampling measurements for traffic characterization and classification* . International Journal of Network Management, 2012.

?? || //

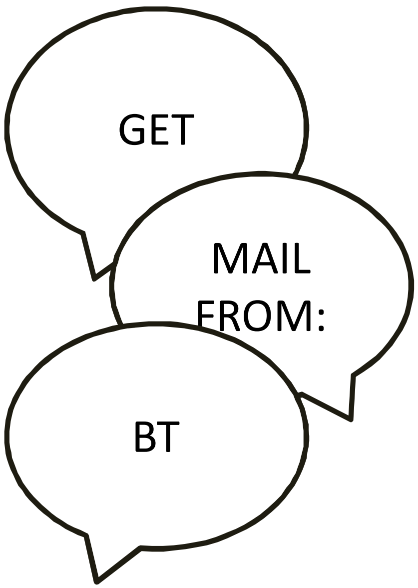
Further advertisement

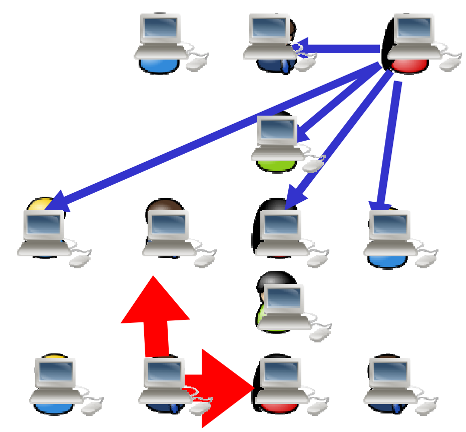
Traffic Classification Taxonomy

Approach	Subcategory	Granularity	Timeliness	Complexity	Comment
Payload Based	[1,2] Deep Packet Inspection (DPI)	Fine-grained individual applications	Early (first few packets).	Access to packet payload of first few packets. Moderate cost	Deterministic technique;
	KISS[Ton'10] Stochastic Packet Inspection	Fine-grained individual applications	Online (100s packets windows)	Access to packet payload of several packets. High cost	Robust technique
Statistical Analysis	[4,5,6,7]	Coarse-grained, class of application	Late (after the flow end).	Access to flow-level information Lightweight cost	Post-mortem analysis
	[8,9]	Fine-grained individual applications	Early (first 5 packets)	Access to first few packets Lightweight cost	On the fly classification
Behavioral Analysis	[10,11]	Coarse-grained, class of application	Late (after the flow end).	Lightweight	Post-mortem analysis
	Abacus [ComNet'11]	Fine-grained, individual P2P applications	Online (1s-5s seconds windows)	Lightweight	Online classification Limited to P2P

Overview

Deep Packet Inspection (DPI)	Stochastic Packet Inspection (KISS)	Behavior analysis (Abacus)
Specific Keyword	**Application syntax**	**Algorithm design**





KISS: Stochastic packet inspection

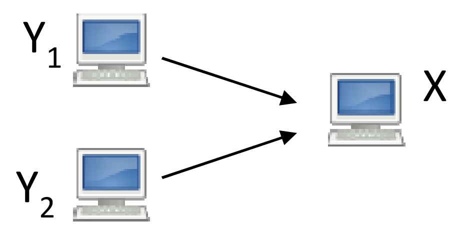
Header syntax is fixed, binary alphabet

1) Extract the **first N bytes** of the payload from a **window of W consecutive packets**
2) Divide each byte in **2 chunks of 4 bits**
3) Collect the **frequency distribution Oi** of the values assumed by each chunk
4) Compare the distribution to a **uniform distribution Ei=/2^4** with a **χ^2-like test**

$$X_g = \frac{\sum_{i=1}^{2^b} \left(O_g^i - E_g^i\right)^2}{\sum_{i=1}^{N/2} E_g^i} \sim \chi_g^2$$

measure the randomness of each chunk

KISS signature: $[X_1, X_2, \dots X_{2N}]$ *over W pkts*



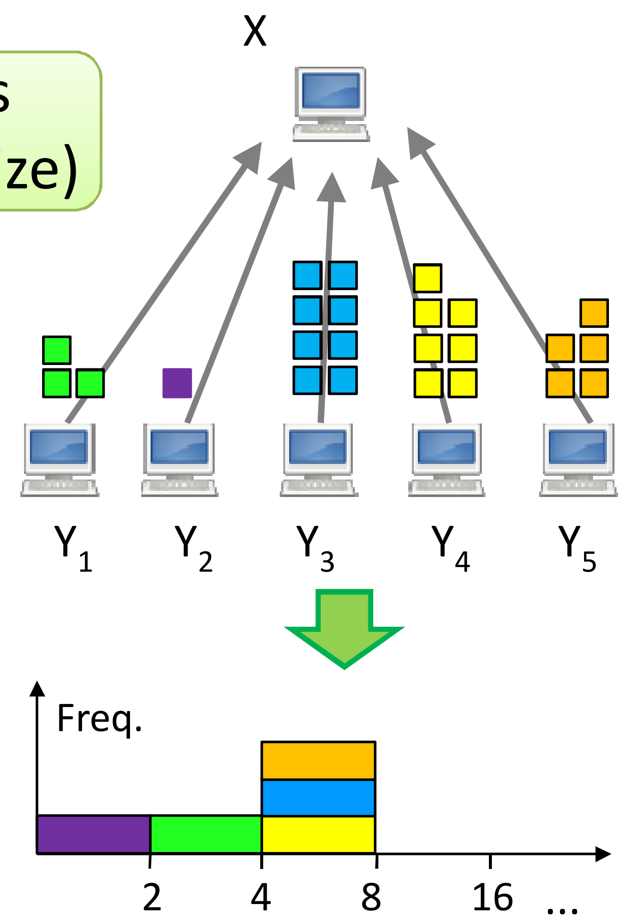
counters
C||D = 3 bit fixed
random
deterministic

Y1 pkt1	cb	d2	...	02	60
Y1 pkt2	cc	d5	...	02	08
Y2 pkt1	01	da	...	02	65
Y1 pkt3	cd	c0	...	02	d9
Y2 pkt2	02	c1	...	02	5c
Y2 pkt3	03	dc	...	02	11
Y1 pkt4	ce	cb	...	02	28
Y1 pkt5	cf	d1	...	02	8a
Y1 pkt6	d0	ca	...	02	3a
Y2 pkt4	04	c2	...	02	b7

Abacus: Behavioral signatures

Applications implement different activities (signaling, data chunks) and tuning (chunk size)

1) Count the **number of packets/bytes** received in a **fixed time window ΔT**
2) Count the **number of hosts** sending a given number of packets/bytes (exponential binning)
3) Normalize the packet/bytewise counts to gather two **probability mass functions**



Example using packets

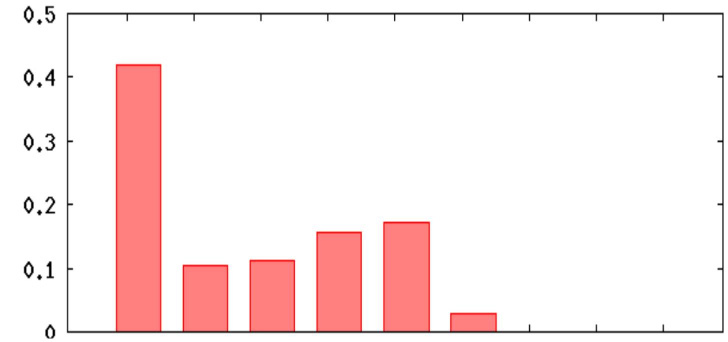
Distribution = [1, 1, 3, 0]
Signature = [0.2, 0.2, 0.6]

Kiss vs Abacus signatures

Normalized χ^2 (first 14 header bytes)

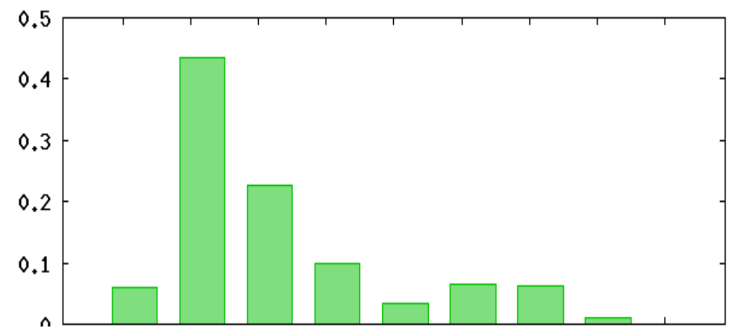
PPLive

0 1 2 3	4 5 6 7	8 9 10 11	12 13 14 15
0.80	0.77	0.73	0.74
0.89	0.79	0.20	0.11
0.69	0.68	0.70	0.64
0.28	0.14	0.16	0.68
0.69	0.68	0.70	0.69
0.69	0.68	0.67	0.65



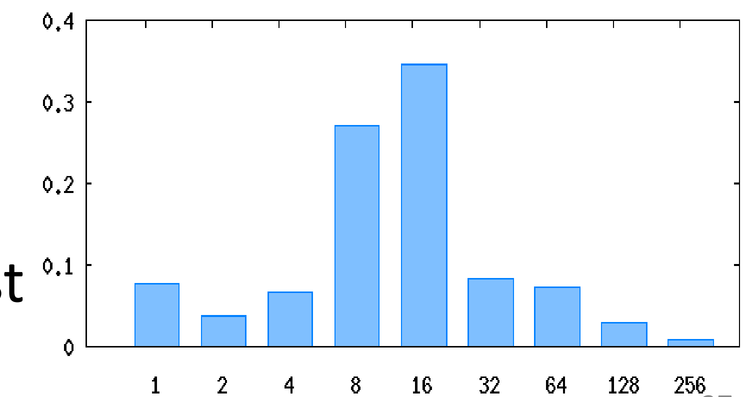
TVAnts

0 1 2 3	4 5 6 7	8 9 10 11	12 13 14 15
0.31	0.27	0.07	0.04
0.97	0.96	0.04	0.04
0.15	0.07	0.04	0.04
0.04	0.03	0.01	0.02
0.19	0.11	0.19	0.19
0.19	0.11	0.10	0.08



SopCast

0 1 2 3	4 5 6 7	8 9 10 11	12 13 14 15
0.98	0.98	0.51	0.29
1.00	0.96	0.30	0.30
0.77	0.31	0.30	0.23
0.30	0.25	0.21	0.22
0.92	0.55	0.95	0.95
1.00	0.57	0.55	0.36



Packets per sender peers pdf (5 sec intervals)

Oops!

- Sorry, wrong key