# TIE Traffic Identification Engine



Alberto Dainotti - *alberto@unina.it* COMICS Research Group University of Napoli "Federico II"

## TRAFFIC CLASSIFICATION

# To associate **flows** to the **applications** that generate them

# {UDP, IPsrc: 10.0.0.1, PORTsrc: 31215, IPdst: 212.48.72.19, PORTdst: 80}

# {TCP, IPsrc: 10.0.0.1, PORTsrc: 2233, IPdst: 13.29.10.199, PORTdst: 25}



## MOTIVATIONS Why classify traffic?

### • To understand what our links carry

- -How are people using the Internet?
- -What's the killer application?
- -Does it really matter to model this or that?
- -Is something "strange" happening and we don't know it?

#### • To operate networks

- -Resource allocation and QoS
- -Enforcement of security policies (e.g. Firewalling)
- -Billing based on typology of traffic
- -Network provisioning
- -Diagnostics: retracing phenomena (e.g. congestion) to specific applications and protocols



## APPROACHES

### an evolving complex scenario

#### Port-based

✓ Fast and Simple-Unreliable (e.g.TCP:80 ≠ HTTP)

### Payload inspection

- ✓ Often reliable
- -Privacy concerns
- -Computationally heavy
- -Can be tricked by protocol encapsulation, encryption, ...

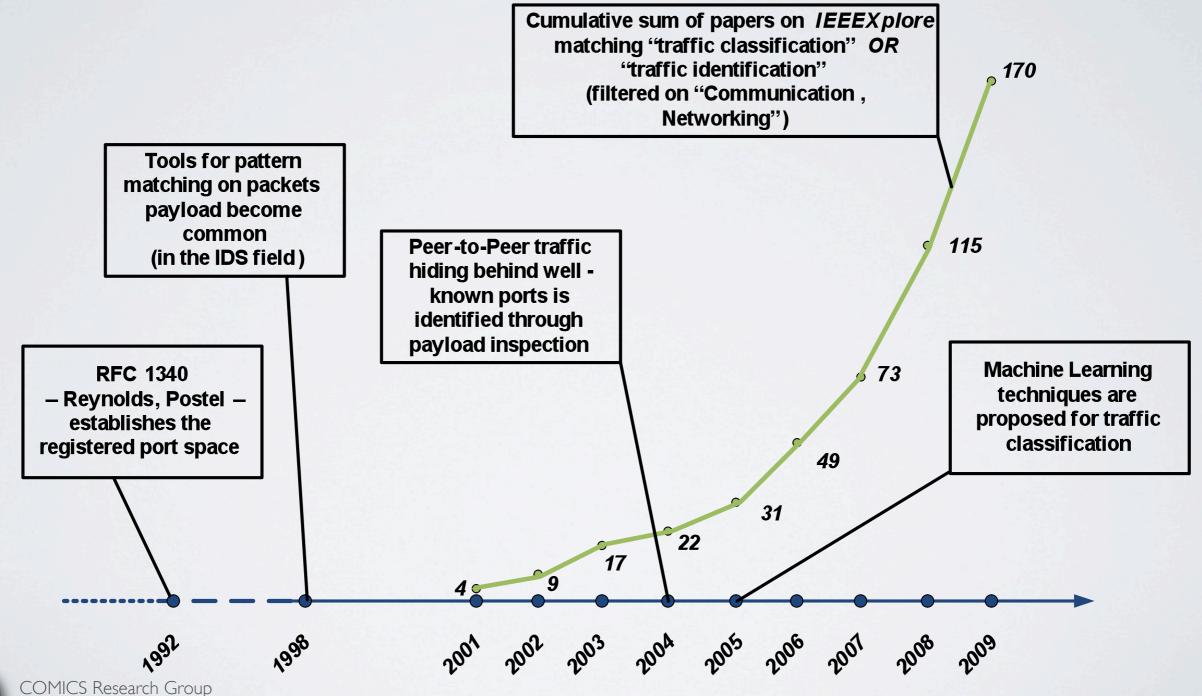
### Pattern Recognition & Behavioral

Promising with respect to current trends (encryption, obfuscation, novel applications, ...)
 -Experimental

-Reliable?

Mellia et al., "Traffic classification and its applications to modern networks", Elsevier Computer Networks, Dec. 2008
 Callado et al., "A survey on internet traffic identification", IEEE Communications Surveys & Tutorials, July 2009.

## SCIENCE EFFORTS dramatically increased in past years





## WHERE WE ARE difficulties...

- A lot of work is still in experimental stage
- Scarce availability of real implementations
- Sharing traffic data in scientific community
- Lack of benchmarks
- Lack of standard formats



# WHERE WE ARE

... and opportunities

- Large interest of different communities -Scientists
  - -Providers
  - -Industry
  - -Society
- Several approaches and code proved to be effective
- Increasing complexity of Internet applications and traffic will continue to keep this topic *hot*!



## TIE Traffic Identification Engine

A software **platform** for **building** traffic classifiers and for **experimenting** with them

- Multi-approach Framework
- Open source
- Fast (Clanguage, Libpcap, Endace DAG support, ...)
- Modular
- Supports multiclassification
- Supports online traffic classification



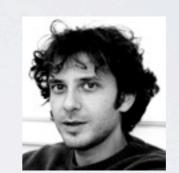
•

## TIE HISTORY the genesis

# Started in **2007** by researchers of the "TRAFFIC" project inside COMICS

		TRAFFIC IDENTIFICATION	ENG
You are here: start		Q. Inc	dex 👰 Lo
Search	Home	Table of Contents	
Home     News	Intro	-Home -Intro - Main features - Write us at	
<ul> <li>Documentation </li> <li>Download </li> <li>How to Help</li> </ul>	TIE is a project for application Traffic Classification, Traffic Identi	identification through network traffic analys fication, etc.).	s <b>is</b> (aka
Support     People		platform for the study and the development on ng collaboration among researchers and practition	
<ul><li>Friends</li><li>Projects</li></ul>		orm working as a multiple classifier system techniques (implemented as separate plug cision combination.	
<ul><li>Publications</li><li>Links</li></ul>		in the documentation pages (see links on the i	left) and

http://tie.comics.unina.it













## TIE HISTORY opening to the world

During these 4 years has been/is the subject of

- Graduate and undergraduate **students** theses
- Collaborations with other research groups



• Collaborations with the **Industry** (manufacturing, customer service assurance consultancy, ...)

National and European Research Projects





# TIE HISTORY

### publications/inventions

#### Papers

- A. Dainotti, F. Gargiulo, L. Kuncheva, A. Pescapè, C. Sansone, Identification of traffic flows hiding behind TCP port 80, IEEE ICC 2010 -May 2010, Capetown (South Africa)
- G. Aceto, A. Dainotti, W. de Donato, A. Pescapè, PortLoad: taking the best of two worlds in traffic classification, IEEE INFOCOM 2010 -WIP Track - March 2010, San Diego (CA, USA)
- V. Carela-Espanol, P. Barlet-Ros, M. Solé-Simò, A. Dainotti, W. de Donato, A. Pescapè, K-dimensional trees for continuous traffic classification, International Workshop on Traffic Monitoring and Analysis (TMA'10) @ PAM 2010 April 2010, Zurich (Switzerland)
- A. Dainotti, W. De Donato, A. Pescapè, "TIE: a Community-Oriented Traffic Classification Platform", International Workshop on Traffic Monitoring and Analysis (TMA'09) @ IFIP Networking 2009 - May 2009, Aachen (Germany)
- Marco Mellia, Antonio Pescapè, Luca Salgarelli, "Traffic classification and its applications to modern networks", Computer Networks, Volume 53, Issue 6, 23 April 2009, Pages 759-760.
- A. Dainotti, W. De Donato, A. Pescapè, P. Salvo Rossi, "Classification of Network Traffic via Packet-Level Hidden Markov Models", IEEE GLOBECOM 2008 - Dec 2008, New Orleans (LA, USA)

#### **Book Chapters**

 G. Aceto, A. Dainotti, W. de Donato, F. Gargiulo, A. Pescapè C. Sansone, "Combining Multiple Traffic Classification Techniques within a Single Platform", *RECIPE Robust and Efficient traffic Classification in IP nEtworks, Fridericiana Editrice Universitaria*, pp.1-16, ISBN: 978-88-833-8081-5, Napoli, Italy, 2009

#### **Technical Reports**

 A. Dainotti, W. de Donato, A. Pescape', Giorgio Ventre, "TIE: a community-oriented traffic classification platform", Technical Report TR-DIS-10-2008, Dipartimento di Informatica e Sistemistica, University of Napoli "Federico II", Italy <u>tr-dis-10-2008-tie.pdf</u>

#### Patents

 A. Dainotti, G. Aceto, W. de Donato, A. Pescapè, "Method and system for traffic classification in communication networks using contentbased signatures". 9th March 2010 - code NA2010A000011#



## TIE OVERVIEW operating modes

### •Offline

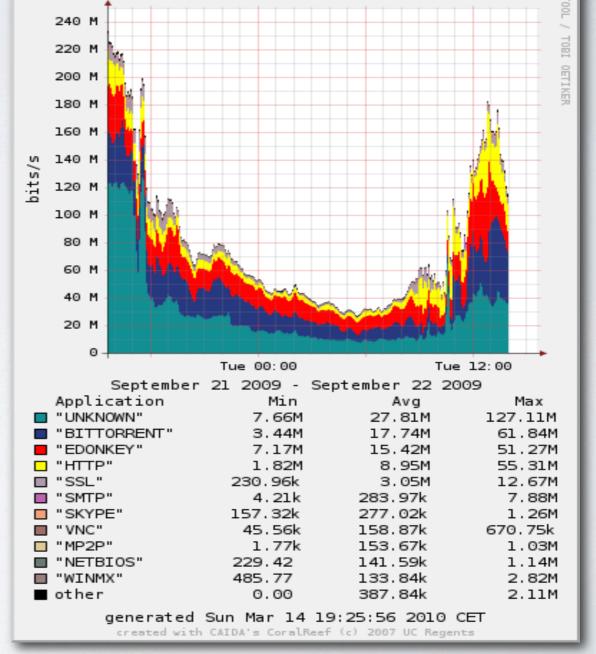
- a session is *classified* only when it ends or at the end of TIE execution

### • Realtime

- a session is classified as soon as possible and output is immediately available

### • Cyclic

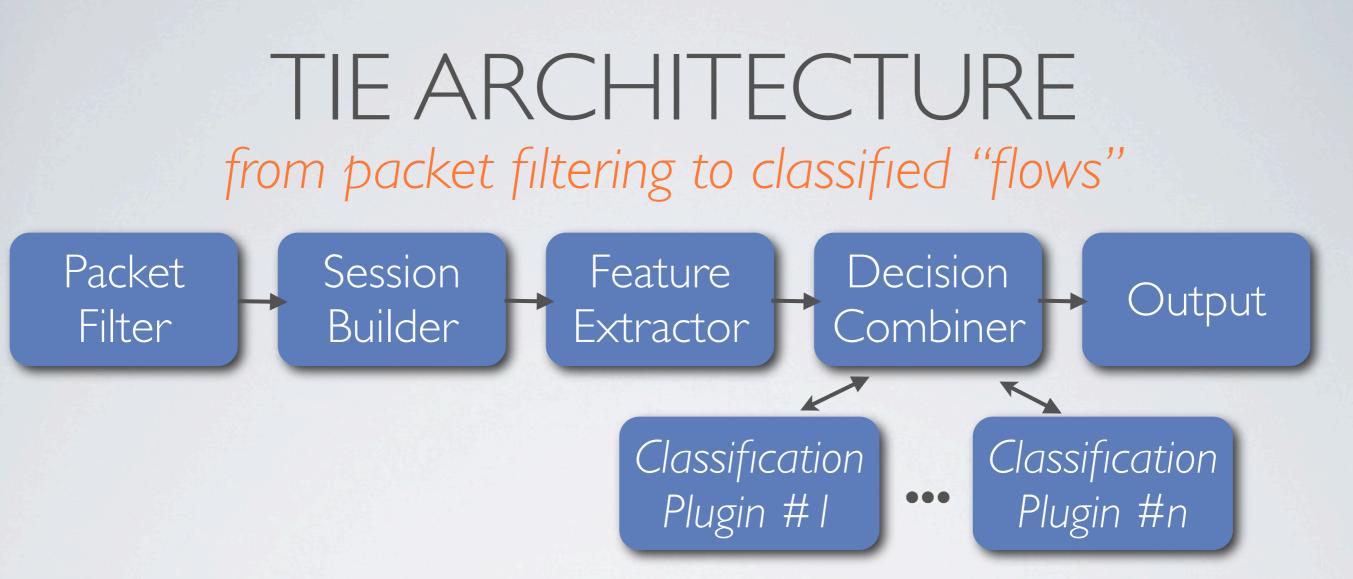
-the classification of all live sessions is generated at regular intervals (e.g. each 5 min.)



Application bits/s - 1 day

In Cyclic mode, automated Web Reports can be built using CAIDA's CoralReef tools.





• It can work with configurable definitions of sessions

#### -Flows

-<L4Proto, IPsrc, Portsrc, IPdst, Portdst> + timeout

#### -Biflows

-Same as above but src and dst are swappable

-Support for TCP connections through simple heuristics based on TCP flags

-Hosts

-Under development



# CLASSIFICATION PLUGINS

Name	<b>Based on</b>	Status	Contributor
Port	L4 ports	Available	UNINA(signatures from CAIDA)
L7	Deep Payload Inspection	Available	UNINA(signatures/code from Linux L7- filter)
PortLoad	Lightweight Payload Insp.	Licensable	UNINA
GMM-PS	Statistical Approach: PS	Under Test	UNINA
нмм	Statistical Approach: PS, IPT	Under Test	UNINA
FPT	Statistical Approach: PS, IPT	Under Dev	UNIBS
Joint	Machine Learning: PS, IPT	Under Test	UNINA-CENS
GT	Information from hosts	Under Dev	UNINA-UNIBS
OpenDPI	Deep Payload Inspection	Beta	OpenDPI, UNINA, TUM
WEKA	Imports the output of a WEKA classifier	Available	UNINA



## OUTPUT sample ASCII output

#### # tie output version: 1.0 (text format)

# generated by: ./tie -r traffic.pcap -S 2048

# Working Mode: off-line # l plug-ins enabled: l7filter

# begin trace interval: 1222078328

# begin TIE Table

# id	src_ip	dst_ip	proto	sport	dport	dwpkts	uppkts	dwby	tes upbytes	t_start		t_last	app_id	sub_id	confidence
844	143.225.22	29.169 89.96.63.82	6	33837	29867	1	1	4	15	1222078300	. 965969	1222078300.984039	Θ	0	0
843	143.225.22	29.169 213.140.17.96	6	33837	29014	1	1	4	14	1222078300	. 965951	1222078300.983139	Θ	Θ	0
225	# id	src ip	dst i	p		proto	spor	٠t	dport	dwpkts	uppkts	1222078278.674796	163	0	100
503		143.225.229.169				-	3383		I	1	1	1222078317.672792	Θ	0	0
589						6	3303	)/	29867	T	T	1222078290.640406	163	0	100
661	843	143.225.229.169	213.1	40.17.	96	6	3383	37	29014	1	1	1222078294.110945	Θ	Θ	0
134	225	143.225.229.169	87 5	180 25	6	17	3383	27	13604	1	1	1222078279.994987	163	Θ	100
327										- -	-	1222078281.557751	163	0	100
	503	143.225.229.169	151.8	.66.21	.0	6	3383	37	48781	2	2				
	589	143.225.229.169	87.3.	228.23	34	17	3383	37	34930	1	1				
	661	143.225.229.169	85.34	.207.1	.0	6	3383	37	16508	1	1				
	134	143.225.229.169	96.20	.21.10	)8	17	3383	37	8056	1	1				and the second
	327	143.225.229.169	74.72	.218.2	9	17	3383	37	11788	1	1				

A set of utilities is distributed with TIE for the post-processing of the output
In *realtime* mode, the output can also be sent through network sockets to another application



## A CASE STUDY PortLoad\*

• TIE's modular framework allows to easily **implement a new** classification technique and run it on real traffic

• By using a unified framework and standard definitions and formats it is easy to **compare and benchmark** three different classification techniques

\*G. Aceto, A. Dainotti, W. de Donato, A. Pescapè, PortLoad: taking the best of two worlds in traffic classification, IEEE INFOCOM 2010 - WIP Track - March 2010, San Diego (CA, USA)

Patent pending "Method and system for traffic classification in communication networks using content-based signatures". 9th March 2010 - code NA2010A000011#



# PORTLOAD

merging two "worlds" in traffic classification

### **Port-based** approach

- Very inaccurate
- + Simple & Fast
- + Privacy-friendly

		_			_	_			_				
Ver.	Head Leng		T S	yp Gerv	e c vic	of e		Total Length					
	Identific	atior				Flags Offset							
Time To Live	D	F	Protocol					Checksum					
				Sc	our	се	Ad	dress					
Destination Address													
			0	ptic	ons	a	nd I	Padding					
	Source	Po	rt				Destination Port						
			Se	equ	ler	nce	e N	umber					
	Ac	kno	wle	dg	en	ne	nt N	Number (AC	CK)	)			
Offset Re	served	U	٩P	R	s	F	F Window						
	Check	sum	ı					Urge	nt	Pointer			
			Op	tio	ns	ar	nd F	Padding					

COMICS Research Group University of Napoli ''Federico II'' - Italy

### **Deep Packet Inspection**

- + Accurate
- CPU intensive
- Doesn't care about Privacy

file	:/hor	me/mł	neffne	er/e90	)7/pe	desta	al/bui	ld/mi	ipp01	L0444	.000	).raw	- KH	exEd	it				
<u>F</u> ile	<u>E</u> dit	<u>V</u> iew	<u>B</u> 00	kmar	ks <u>T</u>	ools	<u>D</u> οcι	imen	ts <u>S</u>	etting	ls <u>H</u> e	elp							
گ 🖄	26	5 🛱		14				Þ	P	ĥ	9	୍ୱ	8	?					
0014	4:c	e10	00	83	00	7f	00	81	00	80	00	7f	00	7a	00	7e	00	80	z.~
		e20	00	85	00	83	00	83	00	82	00	82	00	81	00	82	00	84	
0014			00	85	00	84	00	85	00	82	00	83	00	80	00	81	00	83	
0014			00	83	00	7e	00	80	00	7f	00	81	00	7f	00	81	00	82	
0014			00	85	00	81	00	7e	00	7d	e9	07	03	d2	00	00	00	00	·····~·}□··□····
0014			00	14	CC	0c	78	00	00	00	00	03	00	10	00	00	02	88	••••••••••••••••••••••••••••••••••••••
0014			00	00	00	01	00	01	00	00	00	03	00	44	00	44	06	e8	
0014			07	3b 00	06	ff 00	07	19	09	86	08	2C	08	9a 00	00	10	00	00	.;
			00		••	10	10	80	• •		00 ££	00	00 ££		01 ff		00 ff	01 ff	
0014 0014			00 7f	00 ff	00 ££	T0	II ££	11	ff	ff	ff	II EE	ff	ff	II ff	ff	II ff	ff	
0014			ff	ff	fh	ff	f f	£1 66	11 ff	11 44	II ff	EE EE	II ff	ff	II ff	ff	II ff	ff	.fffffffffffffffff
0014			ff	ff	ID ff	ff	II ff	1 I F F	ff	ff	II ff	1 I 6 6	ff	ff	ff	ff	ff	ff	ffffffffffffffffff
0014			ff	ff	fe	ff	19	49	04	26	e0	08	40	00	28	00	00	02	2222.I.&D.@.(
0014			80	00	04	80	00	02	00	00	20	80	00	44	40	10	30	60	
0014			00	00	0.0	80	40	C9	00	00	00	40	00	0.0	02	80	80	10	
0014			00	00	00	10	08	0.4	00	00	00	82	08	00	00	10	00	40	a
0014			00	00	50	02	ff	ff	ff	ff	ff	ff	ff	ff	ff	ff	ff	ff	
0014			ff	ff	ff	ff	ff	ff	ff	ff	ff	ff	ff	ff	ff	ef	fd	ff	ffffffffffffff
0014			ff	ff	ff	ff	ff	ff	ff	ff	ff	ff	ff	ff	bf	ff	ff	ff	ffffffffffffffffff
0014	4:c	£50	ff	ff	ff	ff	ff	ff	ff	ff	ff	ff	ff	ff	ff	ff	ff	ff	ffffffffffffffffff
0014	4:c	£60	ff	ff	ff	ff	80	80	02	90	08	00	01	00	08	00	01	22	2222
0014	4:c	£70	00	16	40	00	84	04	00	60	12	10	2c	00	05	00	10	00	
0014	4:c	£80	0a	00	4c	00	05	00	00	04	01	01	10	00	00	00	00	00	L
0014	4:c	£90	00	00	00	00	30	04	00	20	00	00	08	00	60	40	00	20	`@.
0014	4:c	fa0	00	00	00	00	ff	ff	ff	ff	ff	ff	ff	ff	ff	ff	ff	ff	fffffffffff
Hex	<b>-</b>	e907													Fi	ind		Ba	ckwards 🗌 Ignore case
	Sig	ned 8	bit		120	D	Signe	ed 32	bit		2013	2659	20	Hexa	decin	nal			78
ι	Jnsig	ned 8	bit		0x78	B U	nsigne	ed 32	bit [		0x78	0000	00		Oc	tal			170
	Sign	ed 16	bit	1	30720	D	32	bit fl	oat	1.	0384	59E+	34		Bina	ary			01111000
Ur	nsign	ed 16	bit	0	<7800	)	64	bit fl	oat	4.	17236	3E-3	09		Te	ext			x
	Show	v little (	endia	n dec	oding	<b>v</b>	Sho	w un	signe	d as	hexa	decim	ial S	Stream	n leng	gth [	Fixed	8 bit	<b>±</b>
						Sel	ectior	1: 001	4:ce	58 00	00:00	00c	0		ze: 1	27946	5752	0	ffset: 0014:ce64-7 Hex H

## PORTLOAD do we need all that payload?

• Experiments on sample traces with TIE-L7 (L7-Filter DPI based on regular expressions)

-Evaluated where the matches happen

-Packet position inside flow

-Bytes in payload

• L ·	Site	Date	Size	Pkts	biflows
• E.g.	Univ. Napoli	Oct 3rd 2009	59 GB	80M	1M

-87% of the matches start at the first packet -Almost all matching strings start (99.98%) and finish (90.77%) in the first 32 bytes of payload of a packet



## PORTLOAD taking the benefits of both approaches

### • **Port-based** is fast and privacy-friendly because:

- -It needs the 1st packet only
- -It uses fixed fields (protocol and port)
- -It uses few data
- -It can be considered as a special case of packet-classification techniques developed for routers, flow-monitors, etc.

• **Payload-based** is *accurate* because it relies on applicationlevel headers and other information from the payload -Payload-based signatures



## PORTLOAD Port + Payload = PortLoad

• A single packet (1st one with payload), fixed fields, limited data (e.g. 32B of payload)

Payload-based signatures

			direction		fields							
н.	App_ID	TCP/UDP	UP/DW/BOTH	offset	1	2	3	4	5	6	7	8
	34	UDP	BOTH	0	Ι	С	Y	\x20	$\odot$	$\odot$	$\odot$	\x20

Example of signature for the Shoutcast MP3 streaming application

Packet-classification matching approach

-Indipendent field searches

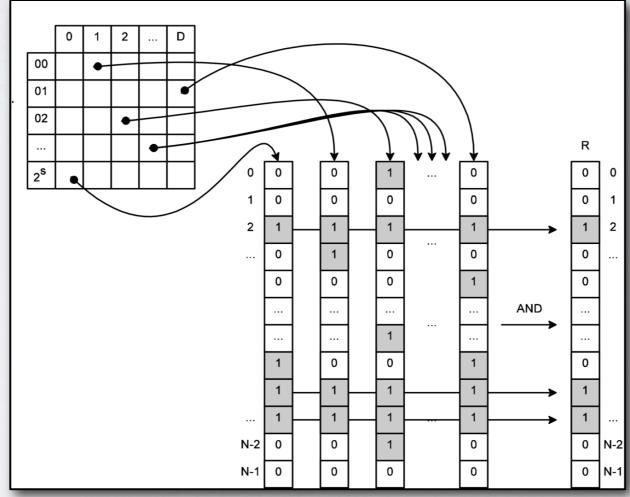
-E.g. bitmap intersection (Lakshman and Stiliadis, SIGCOMM Computer Communication Review, 1998)



Ver.	Head Lenç		Type of Service		Total L	ength			
I	Identific	cation	1		Flags	Offset			
Time To Live	Time To Live Protocol				Chec	ksum			
			Source	) 	Address				
Destination Address									
Options and Padding									
	Source	e Por	rt		Destination Port				
Sequence Number									
	Ad	ckno	wledgeme	en	t Number (ACK	)			
Offset Re	serveo				Win	dow			
	Check	sum	1		Urgent	Pointer			
			Options a	n	d Padding				
Payload									

## PORTLOAD Bitmap Intersection

- A **bitmap** is assigned to each Field-Value pair
- I's in a bitmap indicate signatures compatible with that pair
- AND-ing the bitmaps corresponding to packet content will return the matching signatures





## PORTLOAD evaluation of classification accuracy

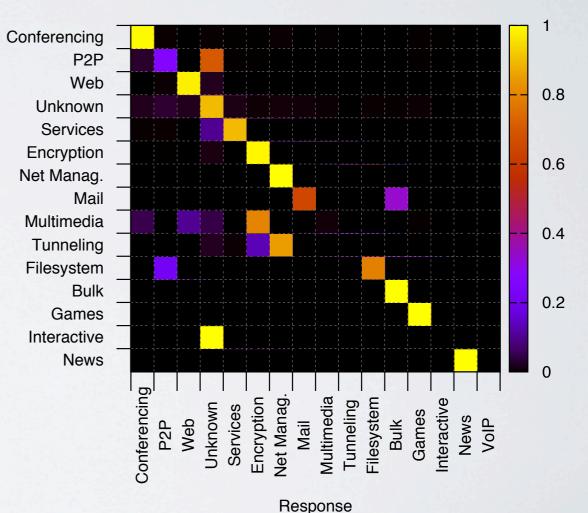
• Evaluation (accuracy against TIE-L7) on UNINA trace from Oct. 2009, with a preliminary set of signatures

Reference

• We compared results on the same traffic trace obtained with

- TIE-L7
- TIE-PortLoad
- TIE-Port

	Accuracy on applications							
Classifier	sessions	bytes						
PortLoad	74.24%	97.83%						
Port-based	19.57%	25.12%						

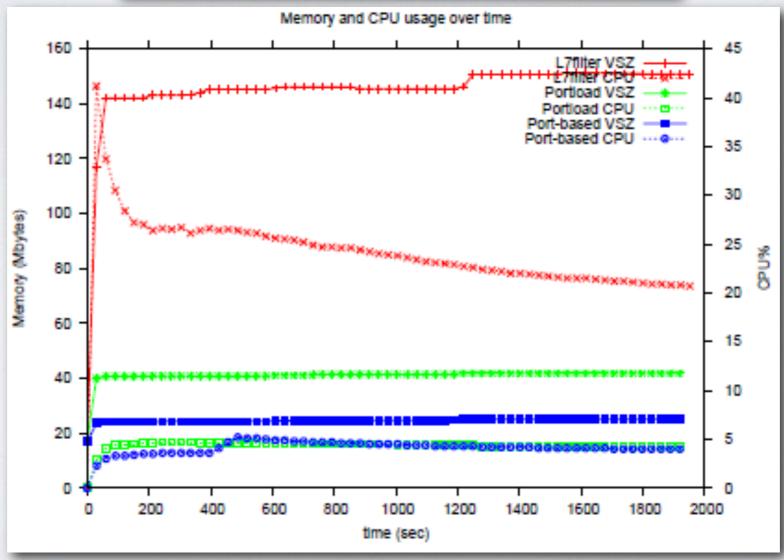




# PORTLOAD

### evaluation of performance

	Mean Time	Mean Time	Variance
Classifier	$(\mu sec)$	(vs Port-based)	$(\mu sec^2)$
Port-based	2.48	1.0	0.88
PortLoad	6.99	2.8	11.15
L7-Filter	211.4	85.2	47057.88





# TIE DEPLOYMENT

what do you need at least

A Linux/FreeBSD box

• An optical splitter or switch/router doing port mirroring

- A spare network adapter or an ENDACE DAG card
- The **pcap** library
- The CAIDA's CoralReef library for live web reports

E.g. we live monitor a 200Mbps link with a Xeon box / FreeBSD 6.3 and a ~\$800 DAG card.



## RIPE MEETING TIE and Internet Service Providers

• We are always seeking for **collaborations** TIE can be used by ISPs for:

- Deploying traffic classification with **low costs**
- Developing traffic classifiers targeted to specific needs and **operating problems** (novel/custom network protocols and encapsulations, specific classes of traffic and applications, etc.)
- Helping in monitoring and diagnosing network events
- Deploy differentiated **QoS** or **security** policies
- Forecasting users-traffic trends

•We are particularly interested in **listening to ISPs needs** and unsolved technical problems and to discuss them



- ....

## THANKS feedback is very welcome

#### http://www.grid.unina.it/Traffic

This work has been funded by the European Project "OneLab2" (ICT FP7 IP 224263).

