The Web: Source of Big Data with a Measurement Perspective

Marco Mellia
Politecnico di Torino

7th Annual IMDEA Networks Workshop
Big Data and Cloud Computing
June 11, 2015, Madrid

mPlane

Agenda

- Internet (passive) traffic measurements
  - A good source of big data
- And applications to
  - Knowledge of (benign) Internet traffic
  - Understanding of (malicious) traffic
  - Implications on your privacy

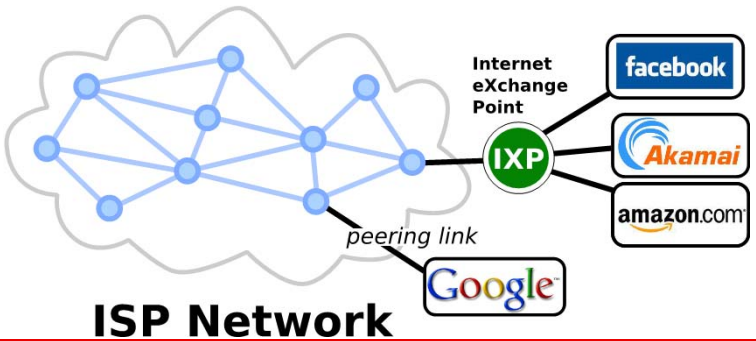
I'm by far not a big data expert…
… I may be a measurement expert…

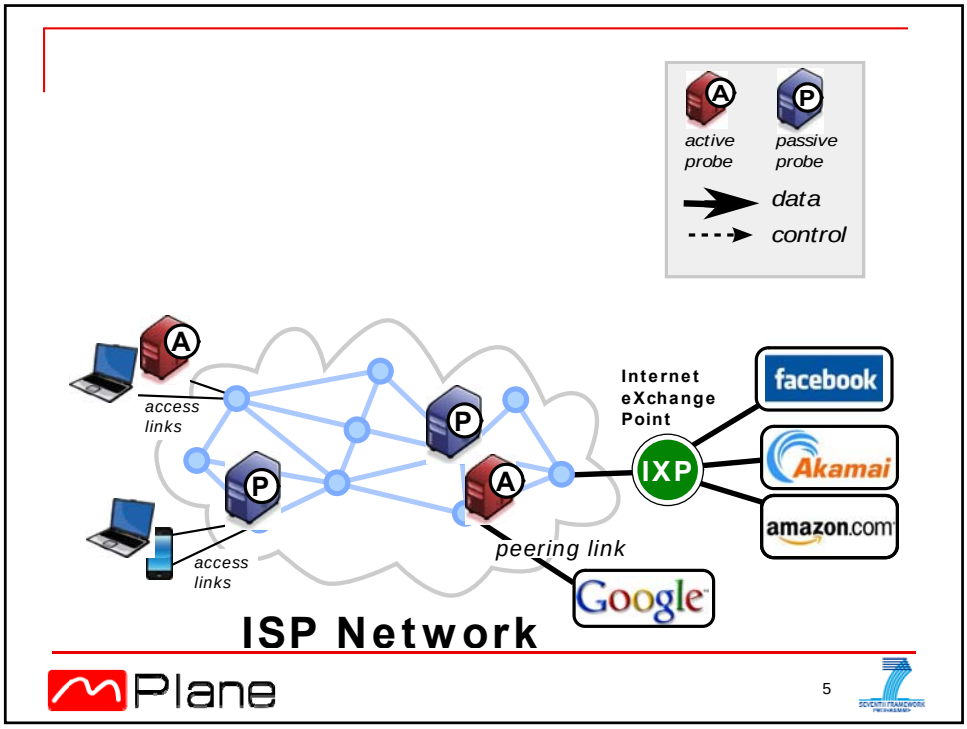
The nowadays Internet

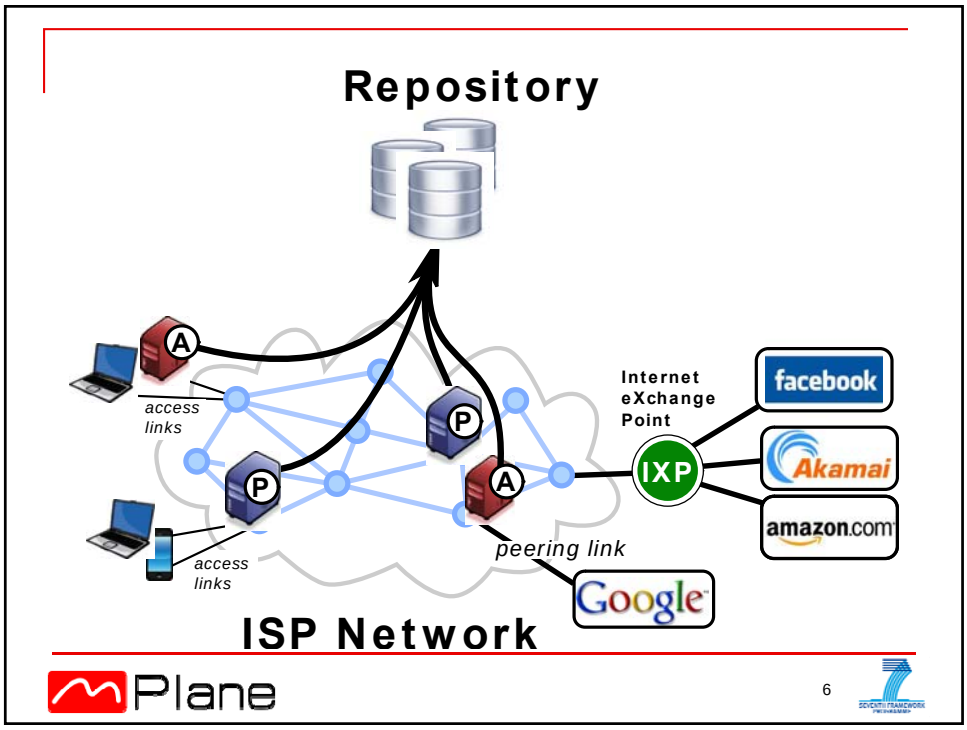
*"The Internet is the first thing that humanity has built that humanity doesn't **understand**, the largest experiment in **anarchy** that we have ever had."*

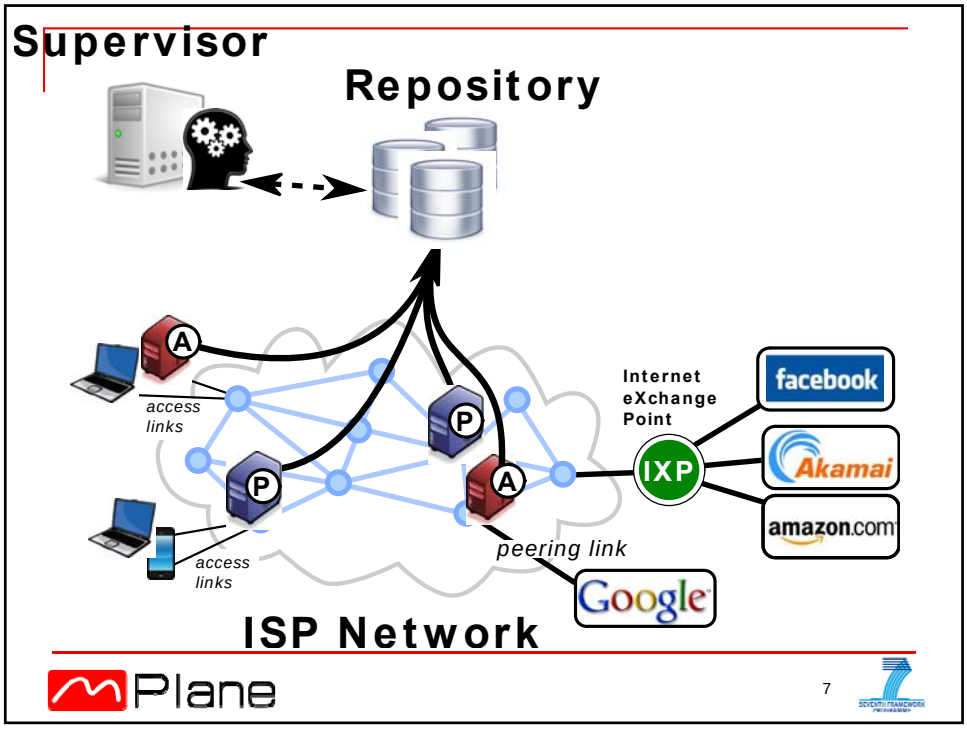
Eric Schmidt – ex Google Exec. Chairman

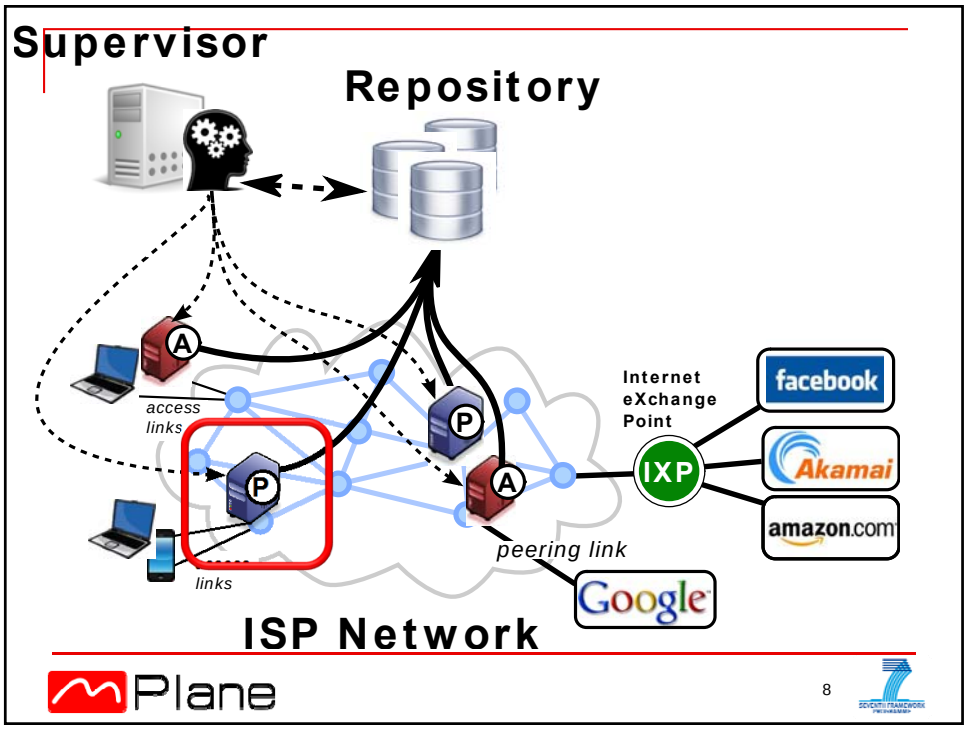
mPlane architecture




active probe
passive probe
data
control
access links
access links
Internet eXchange Point
IXP
facebook
Akamai
amazon.com
peering link
Google
ISP Network
mPlane
5


Repository
access links
access links
Internet eXchange Point
IXP
facebook
Akamai
amazon.com
peering link
Google
ISP Network
mPlane
6

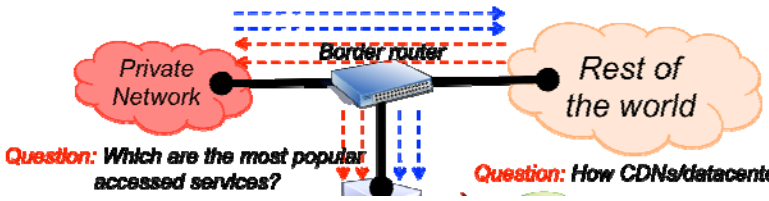

Supervisor
Repository
A
access links
P
P
A
Internet eXchange Point
IXP
facebook
Akamai
amazon.com
peering link
Google
access links
ISP Network
mPlane
7


Supervisor
Repository
A
access links
P
P
A
Internet eXchange Point
IXP
facebook
Akamai
amazon.com
peering link
Google
links
ISP Network
mPlane
8

Passive data analysis

mPlane

Measurement scenario

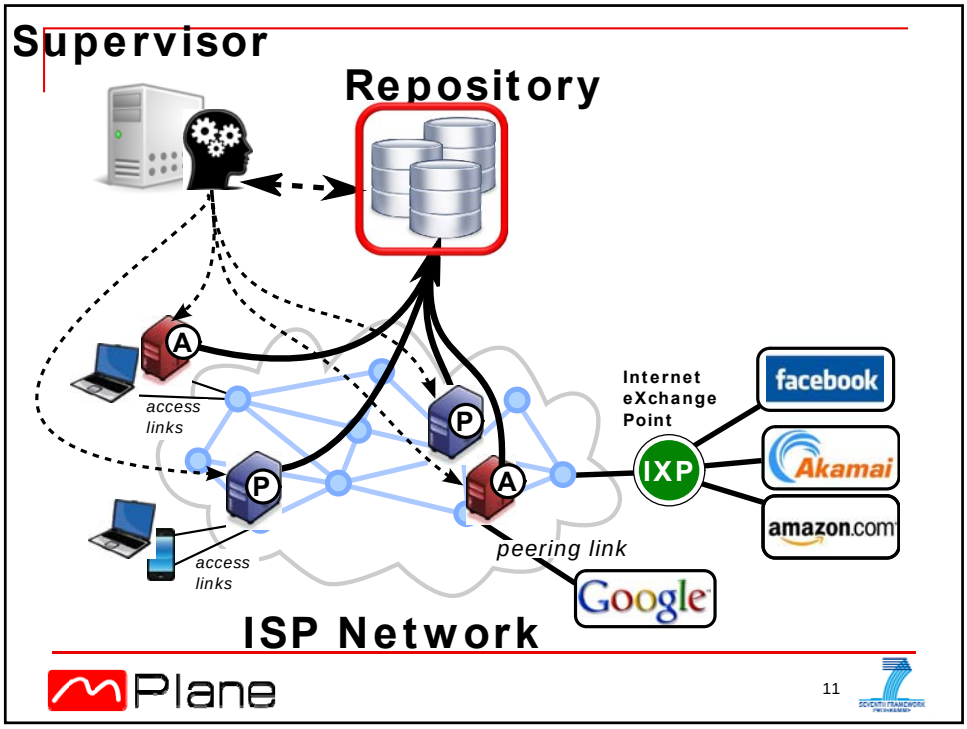


log_tcp_complete - log_tcp_nocomplete

C2S	S2C	Short description	Unit	Long description
1	45	Client/Server IP addr	-	IP addresses of the client/server
2	46	Client/Server TCP port	-	TCP port addresses for the client/server
3	47	packets	-	total number of packets observed form the client/server
4	48	RST sent	0/1	0 = no RST segment has been sent by the client/server
5	49	ACK sent	-	number of segments with the ACK field set to 1
6	50	PURE ACK sent	-	number of segments with ACK field set to 1 and no data
7	51	unique bytes	bytes	number of bytes sent in the payload
8	52	data pkts	-	number of segments with payload
9	53	data bytes	bytes	number of bytes transmitted in the payload, including retransmissions
10	54	rexmit pkts	-	number of retransmitted segments
11	55	rexmit bytes	bytes	number of retransmitted bytes
12	56	out seq pkts	-	number of segments observed out of sequence

Open-source: http://tstat.polito.it

mPlane



Dataset

- 1 day of traffic from an European ISP
  - Vantage point, about 20,000 households
- Approx 4Gb/s => 43 TB/day

Class	Users (%)	Records (%)
HTTP	16,217 (79.1)	39.7 M (11.8)
Email	3,640 (17.7)	880.7 k (0.2)
Chat	3,045 (14.8)	100.8 k (0.03)
P2P	3,163 (15.4)	17.1 M (5.05)
Other TCP	18,806 (91.8)	22.7 M (6.7)
DNS	15,164 (74.1)	30.7 M (9.3)
VoIP	8,371 (40.8)	80.5 k (0.02)
Other UDP	17,664 (86.2)	224.6 M (66.8)
total	*20,486*	*336.1 M*

mPlane

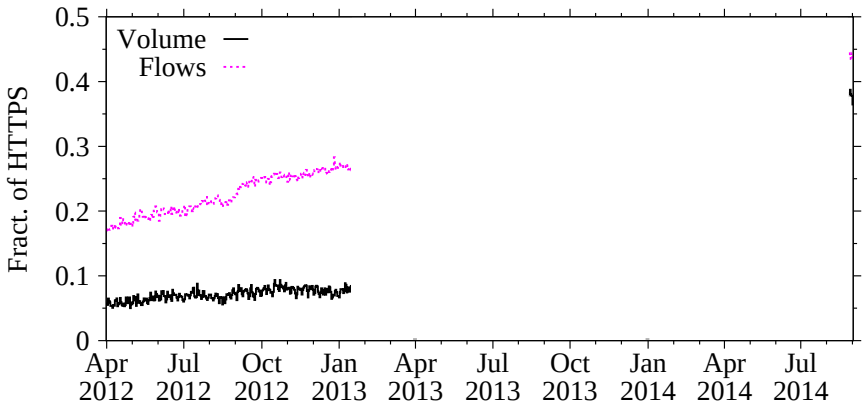
12

Questions you may answer

- How much traffic is TCP or UDP?
- How much traffic is HTTP or P2P?
- What is the best goodput observed by a user?
- What is the CDN being used to serve YouTube videos?
- **How much traffic is HTTPS?**
  - **And what are the "cost of the S"?**

mPlane

"S" usage trends

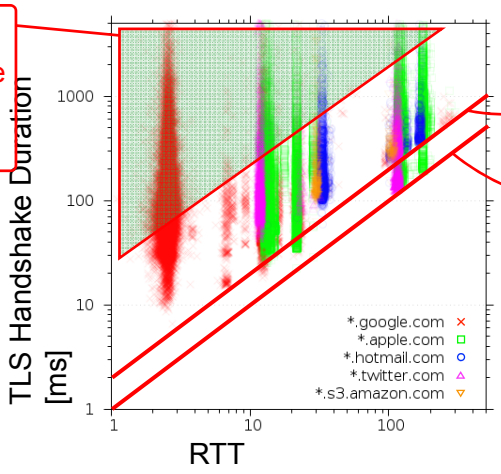


mPlane

D. Naylor, et al, "The Cost of the "S" in HTTPS", CoNEXT '14

Handshake Latency

Stragglers
Long handshake no matter how close server is



y = 2x
Full handshake

y = x
Session resumption

mPlane

Questions you may answer

- How much traffic is TCP or UDP?
- How much traffic is HTTP or P2P?
- What is the best goodput observed by a user?
- What is the CDN being used to serve YouTube videos?
- **How much traffic is HTTPS?**

- **How much traffic is "malicious"**
  - **And how can you detect it**

mPlane

THE RISE OF CYBERWARFARE

+261% - Stealing personal information

+23% - Threatening visitors' security during web surfing

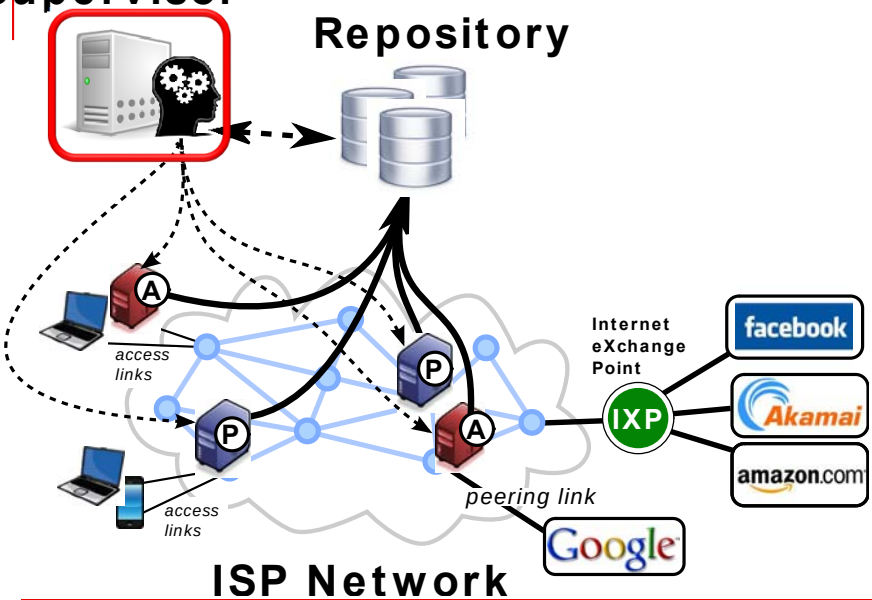
+ 28% - Exploiting legitimate websites

Source: Symantec Internet Security Threat Report http://www.symantec.com/security_response/publications/threatreport.jsp

mPlane

17

Supervisor



mPlane

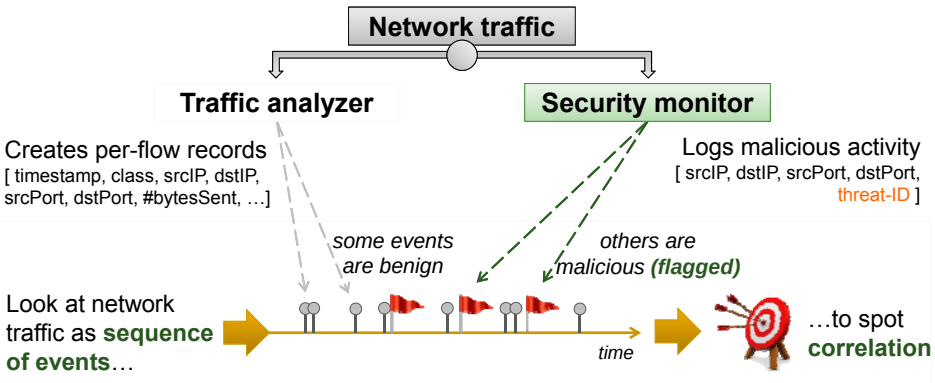
18

Goal

- ***Provide a characterization of malicious traffic in a real network***
  - Identify **common malicious patterns** in network traffic
  - **Augment the information** coming from Intrusion Detection System (IDS) software
  - **Generalize** the IDS knowledge to unknown threats
- With a "Big Data" approach
  - Directly extract knowledge from data

A. Finamore, et al, "Macroscopic View of Malware in Home Networks", IEEE CCNC'15

Methodology



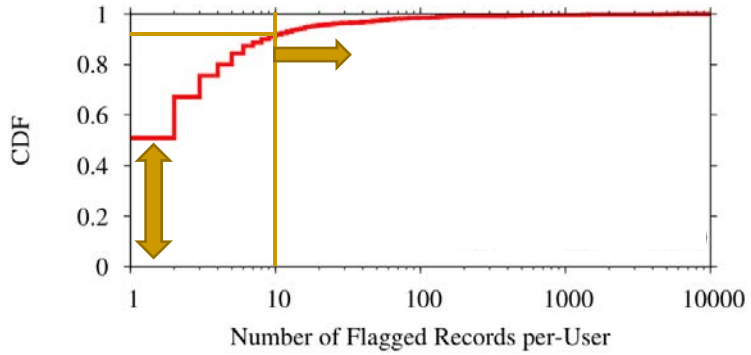
Dataset

- 1 day of traffic from an European ISP
  - Vantage point, about 20,485 households
  - 1,321 (6.4%) households have ≥ 1 connection flagged over a total of 151 threat-IDs

	Traffic Analyzer	
Class	Users (%)	Records (%)
HTTP	16,217 (79.1)	39.7 M (11.8)
Email	3,640 (17.7)	880.7 k (0.2)
Chat	3,045 (14.8)	100.8 k (0.03)
P2P	3,163 (15.4)	17.1 M (5.05)
Other TCP	18,806 (91.8)	22.7 M (6.7)
DNS	15,164 (74.1)	30.7 M (9.3)
VoIP	8,371 (40.8)	80.5 k (0.02)
Other UDP	17,664 (86.2)	224.6 M (66.8)
total	*20,486*	*336.1 M*

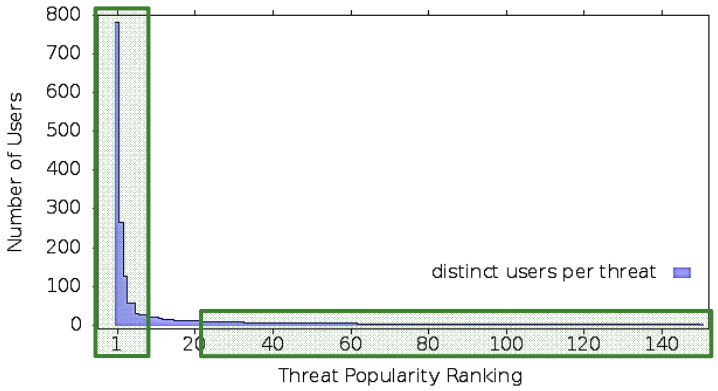
High-level characterization

Malware characteristics – Flags per User



- Malware is generally stealthy
  - ~51% of flagged users have just 1 flag (in a day!)
  - Only 8% of flagged users shows more than 10 flags

Malware characteristics – Threat popularity



- Threat popularity highlights a diversified scenario
  - The most popular infects 800 users
  - 121 threats have less than 10 users

Getting your hand into the data

mPlane

Top -15 most diffused threats

		Users ≥ 1 flag		Users > 1 flag	
#	Name	Users	Flags	Users	AVG Flags
1	Drive-by download [type 1]	781	1,427	265	3
2	DynDNS activity [type 1]	266	26,270	181	144
3	Blackhole EK [type 1]	127	158	20	2
4	Skintrim [type 2]	56	301	46	6
5	Skintrim [type 3]	56	301	46	6
6	Facebook plugin attack	30	31	1	2
7	Threat-A	25	27	2	2
8	Blackhole EK [type 2]	25	25	-	-
9	Toolbar activity [type 1]	21	105	19	5
10	Threat-B	21	23	2	2
11	Threat-C	21	22	1	2
12	Toolbar activity [type 2]	17	19	2	2
13	Drive-by download [type 2]	15	33	5	4
14	Tidserv	14	228	12	18
15	Threat-D	14	470	7	66

mPlane

27

Threat-D – Flagged HTTP objects

HTTP objects requested	Users	# HTTP events
ads.staticyonkis.com/www/delivery/afr.php	5	223
bloggasaurus.com/wp-content/instal/file.php	2	23
www.clublhsnowboards.com/blog/	1	1
www.clublhsnowboards.com/blog/%3Cahref=http://www.snowboardipendente.it/forum.php%3E%3Cimgsrc=http://www.snowboardipendente.it/allegati/member.jpgalt=forumsnowboardborder=1%3E%3C/a%3E	1	1
www.clublhsnowboards.com/blog/divisa-del-club/%3Cahref=http://www.snowboardipendente.it/forum.php%3E%3Cimgsrc=http://www.snowboardipendente.it/allegati/member.jpgalt=forumsnowboardborder=1%3E%3C/a%3E	1	1
www.electricguitarlearn.com/	1	1
www.sportrenoteam.it/	1	1
casariposo.org/banner//jquerymini.js	1	1
www.calcolostipendionetto.it/	1	1
www.clublhsnowboards.com/blog/webcam-neve/%3Cahref=http://www.snowboardipendente.it/forum.php%3E%3Cim//www.snowboardipendente.it/allegati/member.jpgalt=forumsnowboardborder=1%3E%3C/a%3E	1	1
total	*14*	*263*

Most recurrent objects

afr.php	count
User 1	66
User 2	64
User 3	38
User 4	37
User 5	18
total	*223*

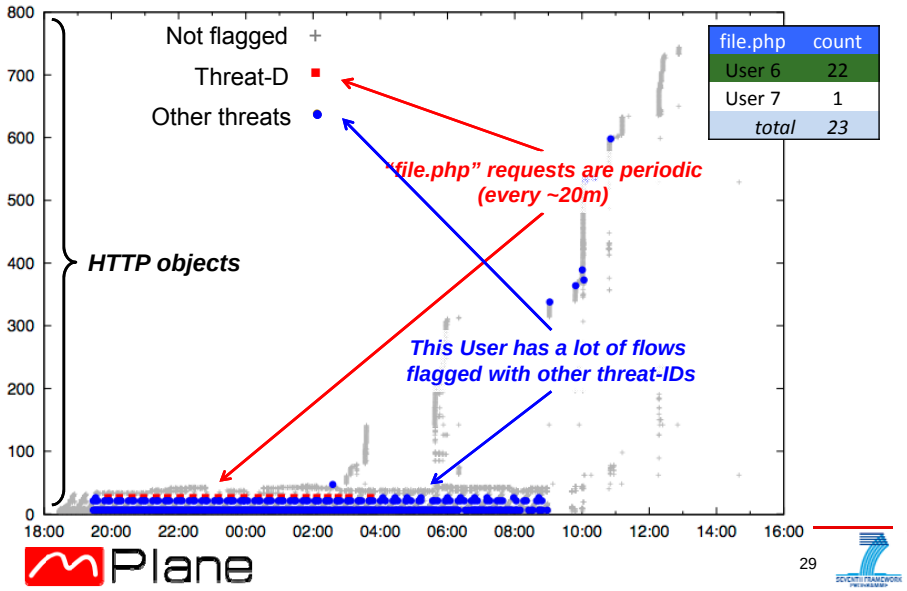
Focus

file.php	count
User 6	22
User 7	1
total	*23*

mPlane

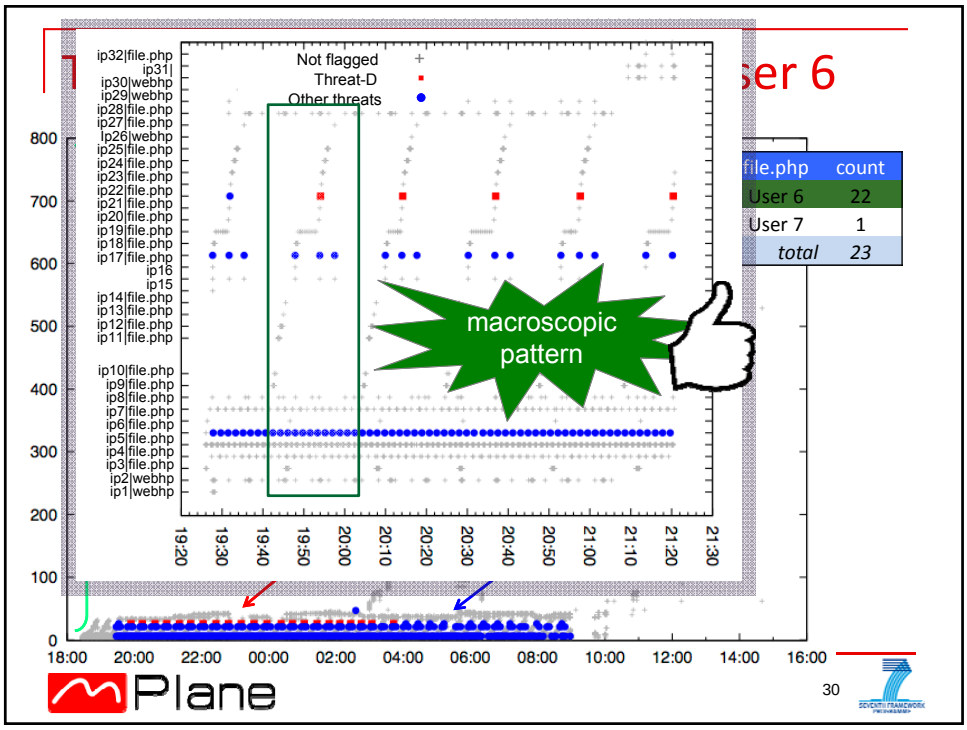
28

Temporal patterns – file.php, User 6



mPlane

29



mPlane

30

Threat-D – Flagged HTTP objects

HTTP objects requested	Users	# HTTP events
ads.staticyonkis.com/www/delivery/afr.php	5	223
bloggasaurus.com/wp-content/instal/file.php	2	23
www.clublhsnowboards.com/blog/	1	1
www.clublhsnowboards.com/blog/%3Cahref=http://www.snowboardipendente.it/forum.php%3E%3Cimgsrc=http://www.snowboardipendente.it/allegati/member=forumsnowboardborder=1%3E%3C/a%3E	1	1
www.clublhsnowboards.com/blog/divisa-del-club/%3Cahref=http://www.snowboardipendente.it/forum.php%3E%3Cimgsrc=http://www.snowboardipendente.it/allegati/member.jpgalt=forumsnowboardborder=1%3E%3C/a%3E	1	1
www.electricguitarlearn.com/	1	1
www.sportrenoteam.it/	1	1
casariposo.org/banner//jquerymini.js	1	1
www.calcolostipendionetto.it/	1	1
www.clublhsnowboards.com/blog/webcam-neve/%3Cahref=http://www.snowboardipendente.it/forum.php%3E%3Cim//www.snowboardipendente.it/allegati/member.jpgalt=forumsnowboardborder=1%3E%3C/a%3E	1	1
total	*14*	*263*

Most recurrent objects

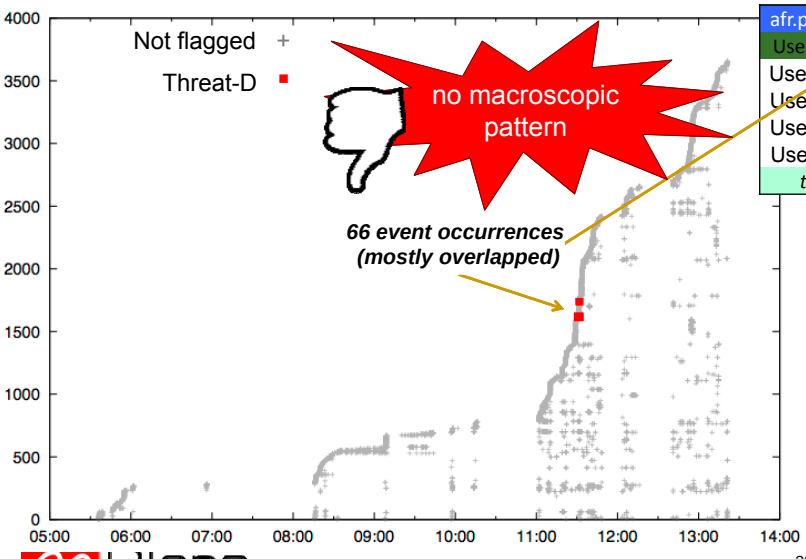
Focus

afr.php	count
User 1	66
User 2	64
User 3	38
User 4	37
User 5	18
total	*223*

file.php	count
User 6	22
User 7	1
total	*23*

mPlane

Temporal patterns – file.php, User 1



afr.php	count
User 1	66
User 2	64
User 3	38
User 4	37
User 5	18
total	*223*

mPlane

Lesson learned

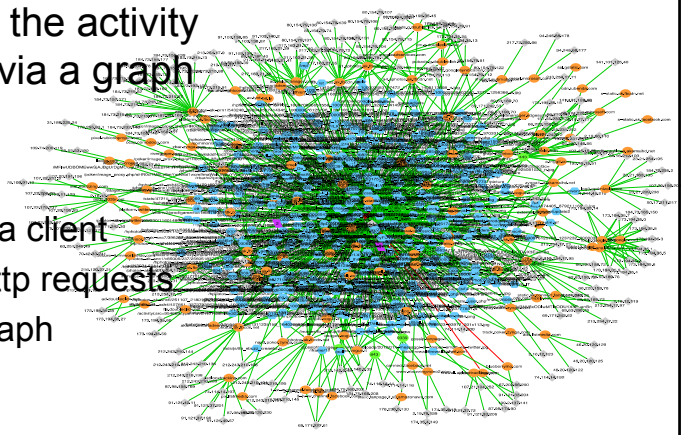
- ***The picture is very fuzzy and complicated***
  - *Volume, velocity,* ***variety***, ***variability***, *…*
- Repeating the experiment on other threats does not help
  - **Most of the times** there is not a clear pattern
  - **In few cases** there is a pattern recurrent over time

Going toward automated detection

Via a connectivity graph

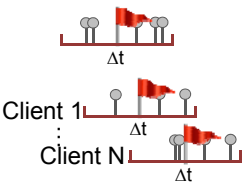
Connectivity graph

- Represent the activity of a client via a graph
  - Consider a client
  - Look at http requests
  - Build a graph

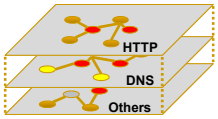


Building the Malicious Graph

- One HTTP request plays the role of **seed** and triggers the analysis
  - Malicious seeds, flagged by the IDS
  - Benign seed, no suspicious activities



1. Define an **analysis snapshot** surrounding the event
2. Collect multiple snapshots coming from different clients to find a common pattern



3. Leverage activities relationships detected on multiple **layers** (i.e., HTTP, DNS, TCP, …)

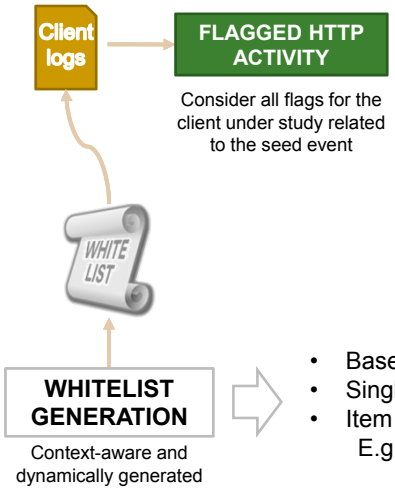
4. Join layers to have a unified **connectivity graph**

Building the Connectivity Graph

1. Select a seed, either malicious or benign
2. Process analysis snapshots coming from involved clients
3. Find a common pattern among snapshots and clients

4. Plus tons of “minor details”

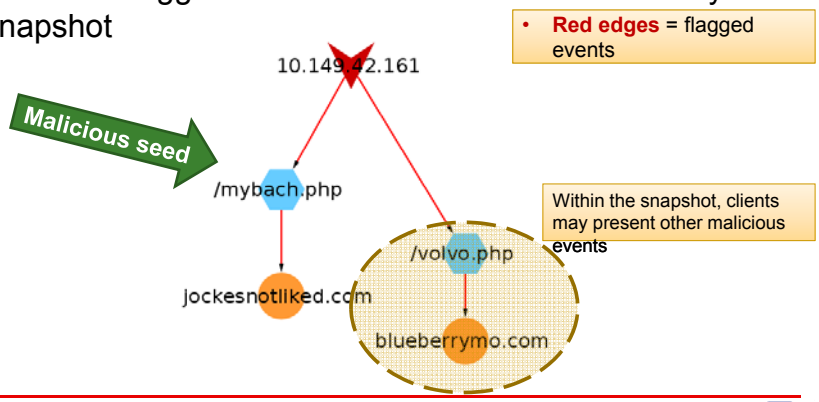
Building the Connectivity Graph



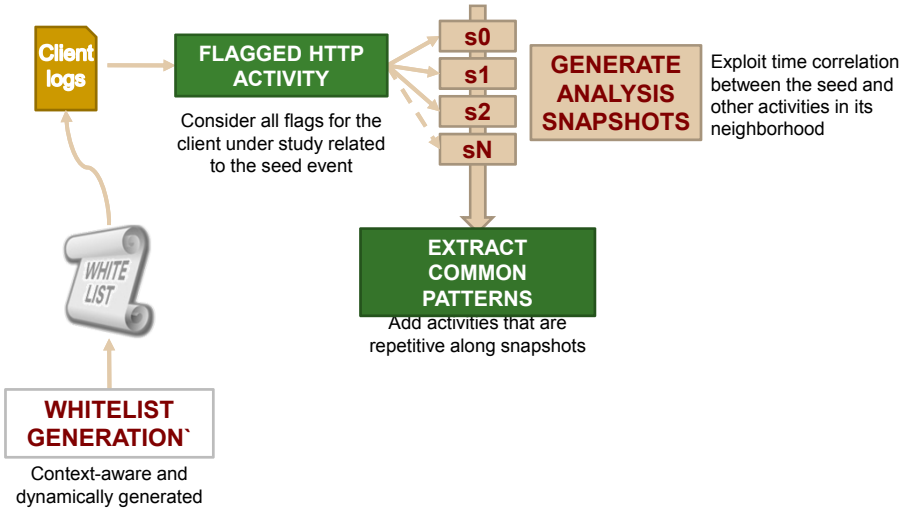
- Based on popular items
- Single HTTP request as white list entry
- Item = an HTTP request, domain + object path
  E.g.: www.facebook.com+/plugins/activity.php

Step 1 – Flagged activity

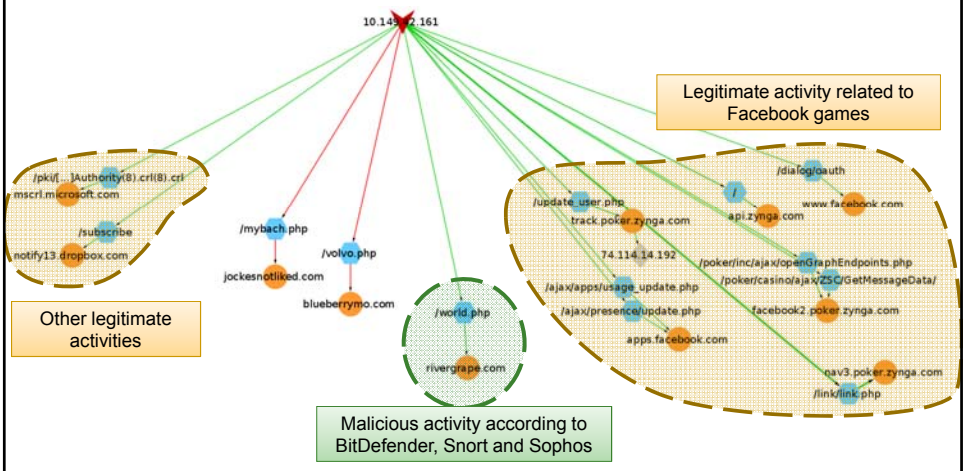
- **Client 1** being analyzed: 3 snapshots
- Focus on flagged activities detected in the first analysis snapshot



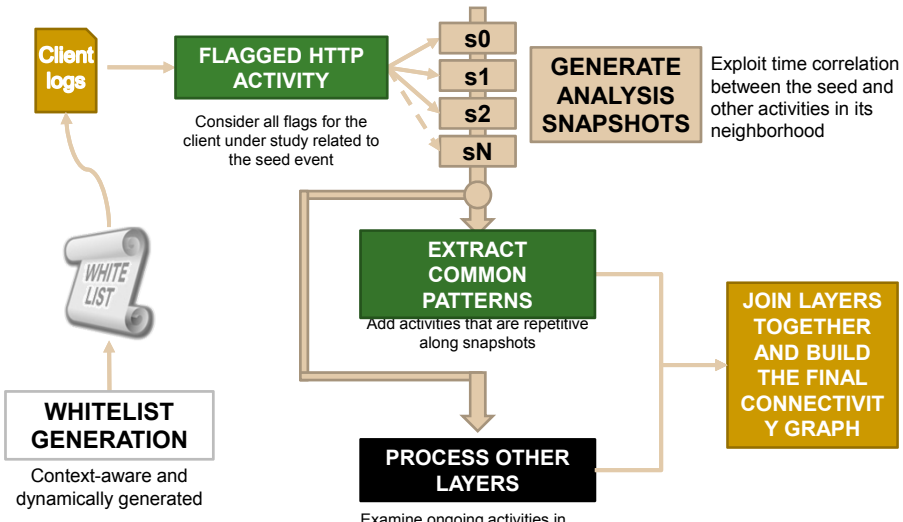
Building the Connectivity Graph



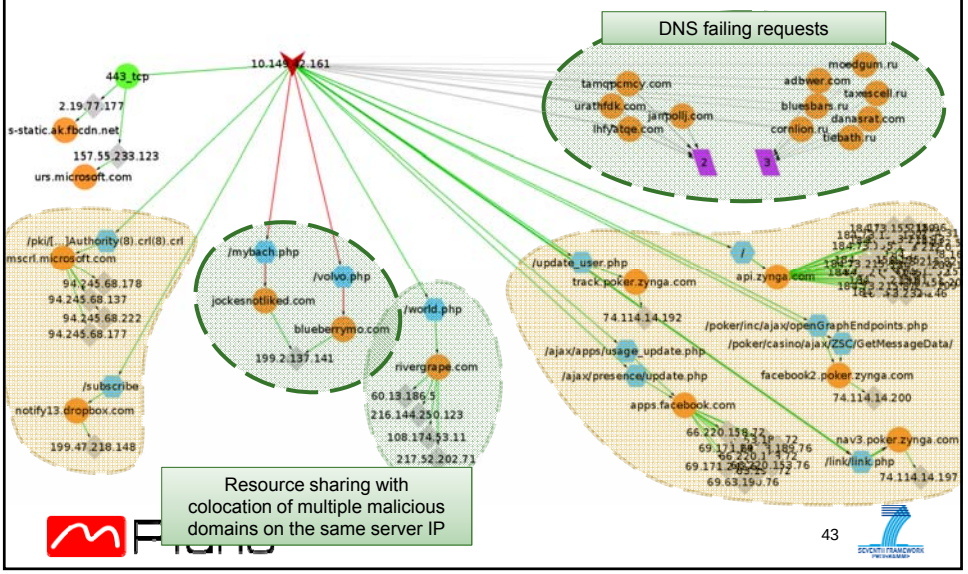
Step 2 - common patterns



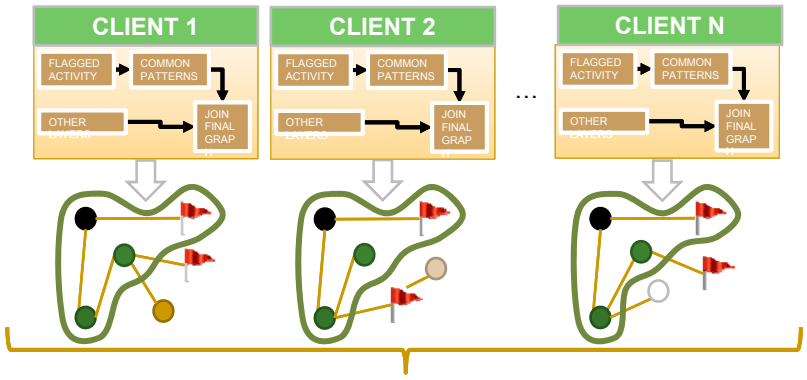
Building the Connectivity Graph



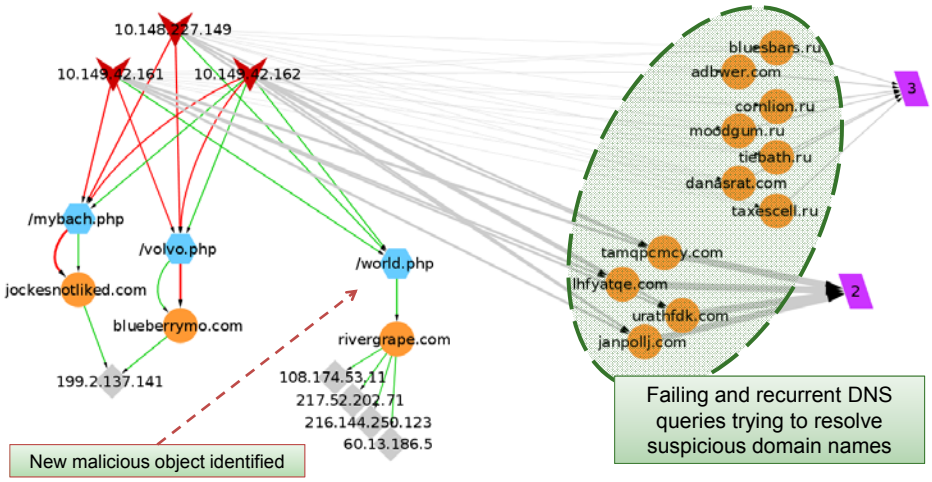
Step 3 – Per-client connectivity graph



Aggregating multiple clients

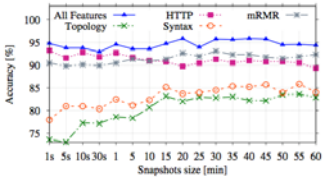


Step 4 – Final connectivity graph



Conclusions

- It is possible to augment the knowledge of the malicious activity thanks to big data analysis
- Correlation and filtering are key
- Next step:
  - Extract signature from malicious and benign seeds
  - Instruct a machine learning classifier with the two classes
  - Design a self update system to automatically augment its knowledge



Agenda

- Passive traffic measurements
  - Source of big data
- And applications to
  - Knowledge of (benign) Internet traffic
  - Understanding of (malicious) traffic
  - Implications on your privacy

One more example…
… about you…

Word count

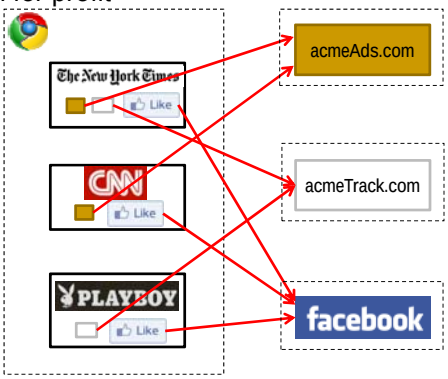
Take 100 users, and observe which services they contact while on the Internet…

… most of those are so called "third-party sites" - aka tracking services…

Third-Party Trackers

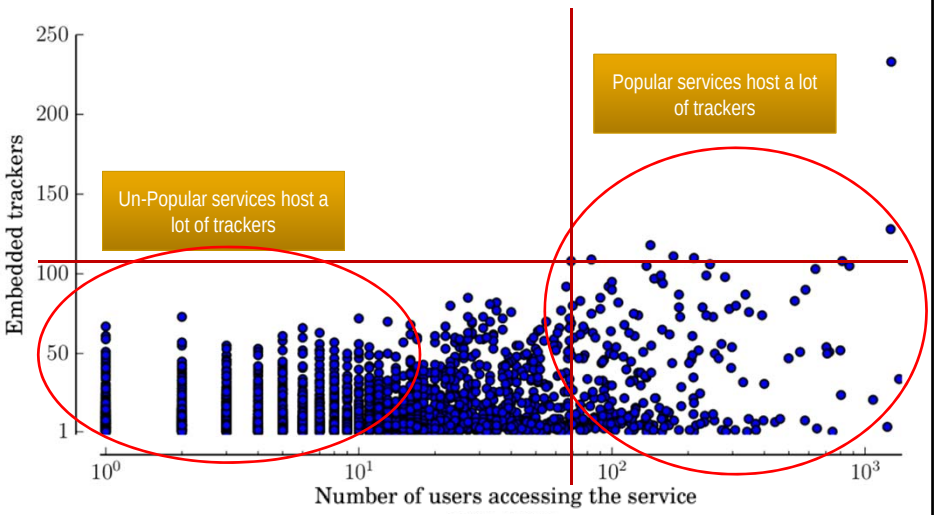
- **Third-party web tracking** refers to the practice by which a service records user web activities often for profit
- Many techniques
  - Cookies
  - HTML5 LocalStorage
  - Finger printing (browser/OS/IP)
- Some are visible

  Like
- Others are not



Metwalley, et al. "The Online Tracking Horde: A View from Passive Measurements." *TMA'15*

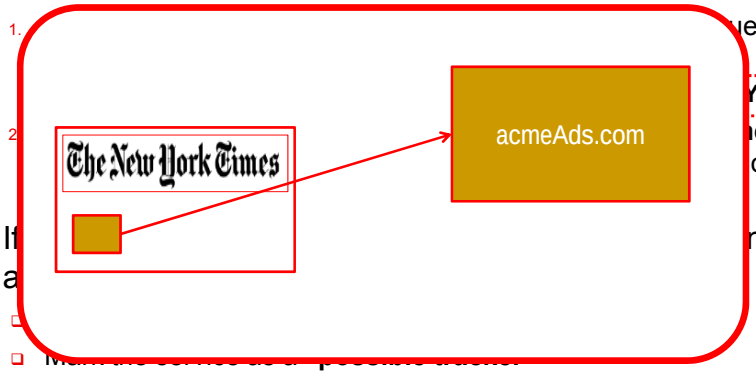
Trackers per Service



Going toward automated detection

Automatic Tracker Detection

- For a set of events in which hostname is different from referrer field



Automatic Tracker Detection

Algorithm 1 Automatic third party tracker identifier.

#HTTP request log and target website
Input: **HS**, $target$
#List of possible third party trackers and their user-tracking keys
Output: **TS**
#Init hashtable of user identifiers
$\mathbf{H_u} \leftarrow$ init_hash_table()
#Init hashtable of key-value pairs
$\mathbf{H_{kv}} \leftarrow$ init_hash_table()
#Read HTTP request logs
while h in **HS** **do**
  #Extract fields of interest
  $h \leftarrow$ csUid, host, path, referer
  #Check target is third party
  if $target$ not in h.host and $target$ in h.referer **then**
    #Extract key, value pair from the URL path
    $K, V \leftarrow$ extract_keys(h.path)
    #Iterate all key names and values
    while k, v in K, V **do**
      #Create hash for $\mathbf{H_u}$
      $host_key_csUid \leftarrow$create_hash(h.host,k,h.csUid)
      #Create hash for $\mathbf{H_{kv}}$
      $host_key_value \leftarrow$create_hash(h.host,k,v)
      #Insert all key-value pairs in $\mathbf{H_u}$
      ADD_DISTINCT($\mathbf{H_u}[host\_key\_csUid]$,v)
      #Insert CS-UID in $\mathbf{H_{kv}}$
      ADD_DISTINCT($\mathbf{H_{kv}}[host\_key\_value]$,h.uid)
    end while
  end if
end while

#Iterate over $\mathbf{H_u}$
while hash in $\mathbf{H_u}$ **do**
  #Iterate over values mapped to current hash
  while value in $\mathbf{H_u}$[hash] **do**
    #Check current hash refers to one value only
    if LEN($\mathbf{H_u}$[hash]) == 1 **then**
      #Decode hash into host, key and CS-UID
      $host, key, csUid \leftarrow$ decode_hash(hash)
      #Create an auxiliary hash using host, key and value
      $hash_{aux} \leftarrow$ create_hash($host, key, value$)
      #Check the auxiliary hash in $\mathbf{H_{kv}}$ contains only one CS-UID and
      #check this corresponds to the one in $\mathbf{H_u}$
      if LEN($\mathbf{H_{kv}}[hash_{aux}]$) == 1 and $\mathbf{H_{kv}}[hash_{aux}] == csUid$ **then**
        #Add host and key to the output list
        ADD(**TS**, $host, key$)
      end if
    end if
  end while
end while
return **TS**

Automatic Tracker Detection - Results

- Third party sites in
  - Repubblica.it
  - YouTube
  - Facebook

Website	Third party sites	Keys
Repubblica		
YouTube		
Facebook		

mPlane

Automatic Tracker Detection - Results

- Third party sites in
  - Repubblica.it
  - YouTube
  - Facebook
- Third party hostnames identified

Website	Third party sites	Keys
Repubblica	pix04.revsci.net	
	su.addthis.com	
	track.adform.net	
YouTube	bh.ams.contextweb.com	
	eu-jet-01.sociomantic.com	
	ib.adnxs.com	
	www.wajam.com	
	uip.semasio.net	
Facebook	adadvisor.net	
	data.bncnt.com	
	go.flx1.com	
	ira.spysomeone.com	
	tags.bluekai.com	
	ww1.collserve.com	
	www.skyscanner.com	

mPlane

Automatic Tracker Detection - Results

- Third party sites in
  - Repubblica.it
  - YouTube
  - Facebook
- Third party hostnames identified
- Third party hostnames confirmed ☺

Website	Third party sites	Keys
Repubblica	pix04.revsci.net	
	su.addthis.com	
	track.adform.net	
YouTube	bh.ams.contextweb.com	
	eu-jet-01.sociomantic.com	
	ib.adnxs.com	
	www.wajam.com	
	uip.semasio.net	
Facebook	adadvisor.net	
	data.bncnt.com	
	go.flx1.com	
	ira.spysomeone.com	
	tags.bluekai.com	
	ww1.collserve.com	
	www.skyscanner.com	

mPlane

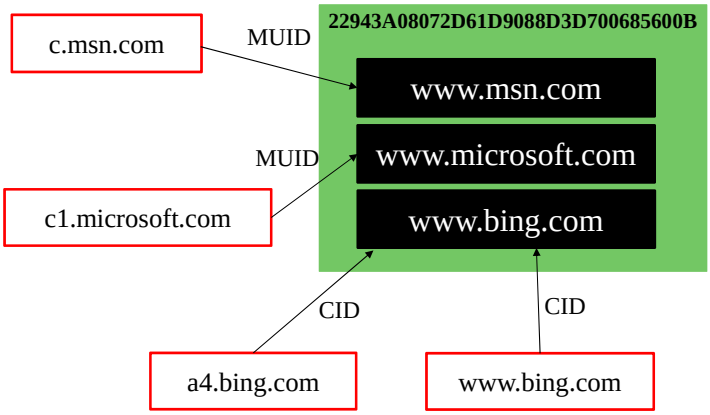
Automatic Tracker Detection - Results

- Third party sites in
  - Repubblica.it
  - YouTube
  - Facebook
- Third party hostnames identified
- Third party hostnames confirmed ☺
- Keys suggest the exchange of **client identifiers**

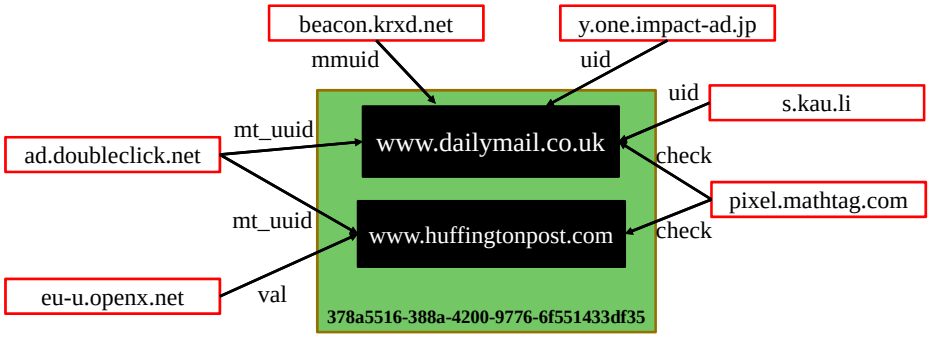
Website	Third party sites	Keys
Repubblica	pix04.revsci.net	id
	su.addthis.com	puid
	track.adform.net	icid
YouTube	bh.ams.contextweb.com	vgd
	eu-jet-01.sociomantic.com	fpc
	ib.adnxs.com	uuid
	www.wajam.com	sExtCookieId
	uip.semasio.net	install_timestamp
Facebook	adadvisor.net	bk_uuid
	data.bncnt.com	uid
	go.flx1.com	anuid, euid
	ira.spysomeone.com	s
	tags.bluekai.com	google_gid
	ww1.collserve.com	bk_uuid
	www.skyscanner.com	ksh_id

mPlane

Collusions



Collusion ?!??!



Conclusions

- Internet is possibly the “best” source for big data
  - Volume, Variety, Velocity, Variability, Veracity, …
- With multiple applications
  - Characterization
  - Security
  - Privacy
  - User behaviour
  - …
- Lot of domain knowledge, and lot of ingenuity to better understand the “*largest experiment in anarchy that we have ever had*”