Time Signatures to detect multi-headed stealthy attack tools

Marc Dacier (EURECOM)
Guillaume Urvoy-Keller (EURECOM)
Fabien Pouget (CERTA)

Plan

- What we already have…
  - A world-wide project
  - Large amount of data
  - A classification
- On studying temporal evolution of malicious activities
- The SAX similarity detection method
- Applications to the Leurré.com dataset
- Conclusions

Observations

- There is a lack of valid and available data
- The understanding of what is going on in the Internet remains very limited
- This understanding might be useful in many situations:
  - To build efficient detection systems
  - To ease the alert correlation task
  - To tune security policies
  - To confirm or reject free assumptions

Consequences

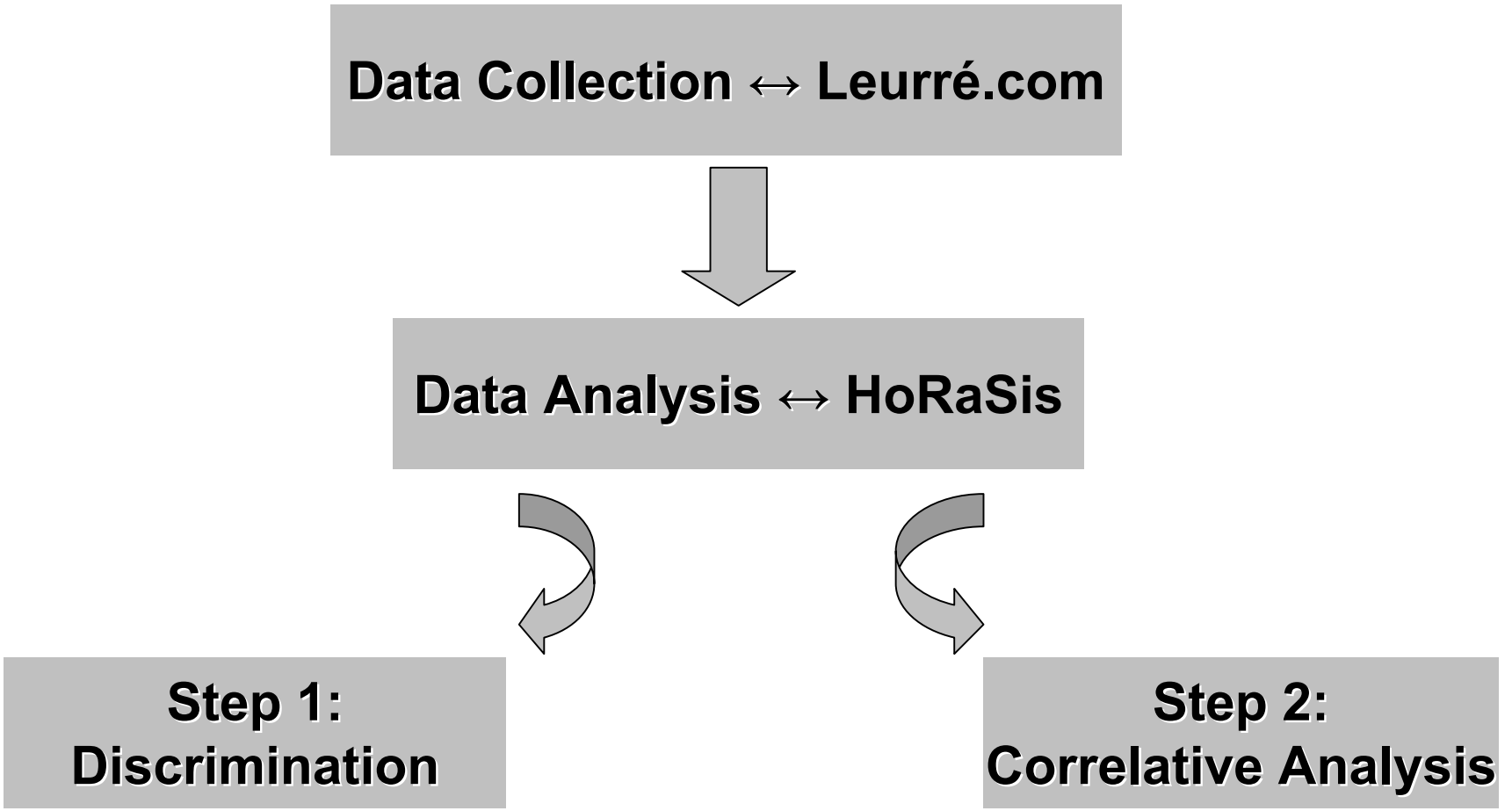
- We could consider an architecture of sensors deployed over the world

  … using few IP addresses

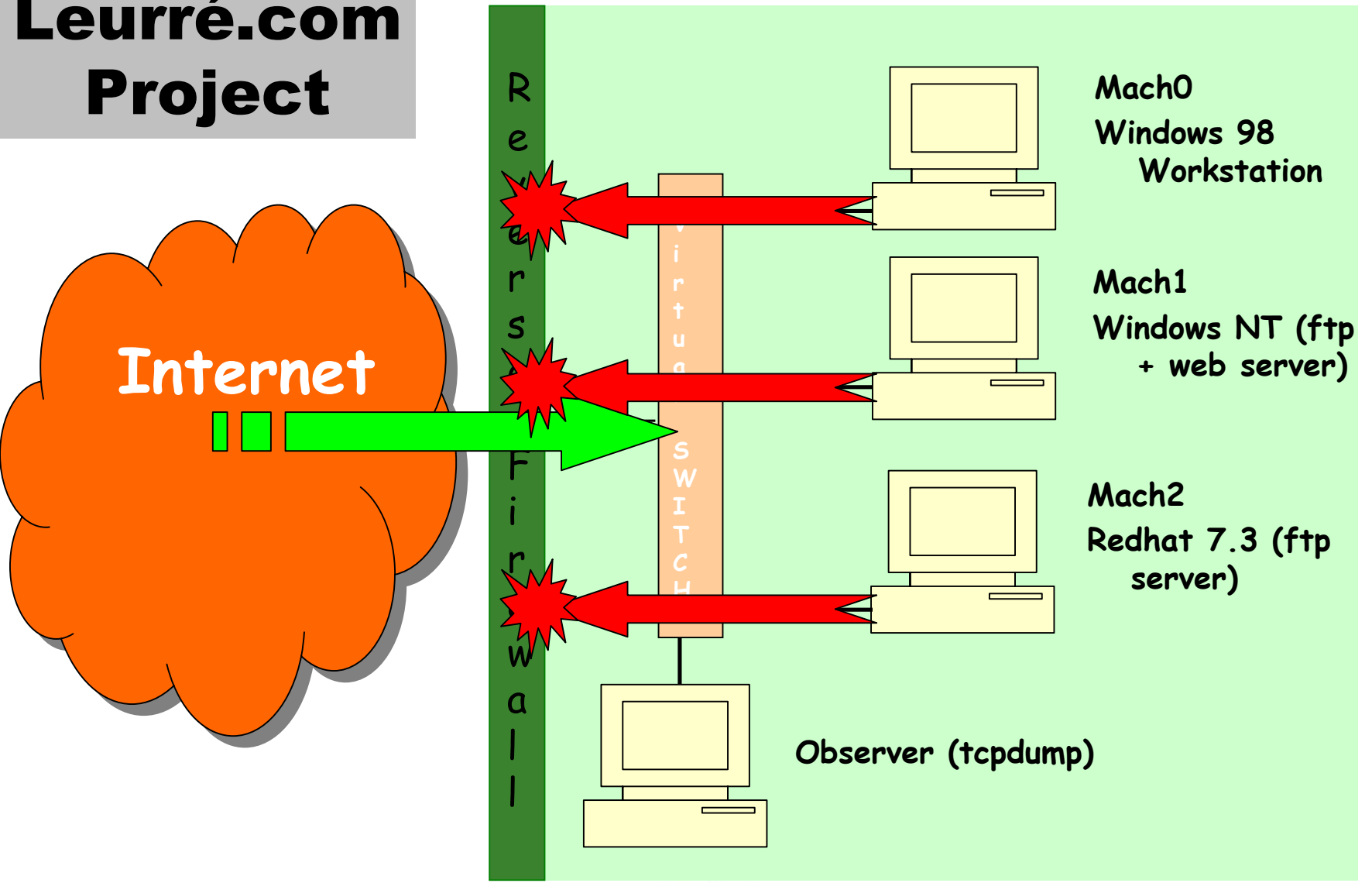
- Sensors should run a very same configuration to ease the data comparison

… and make use of the honeypot capabilities.

Our approach :



Leurré.com Project



45 sensors, 25 countries, 5 continents

Europe

Leurré.com Project

Sensor 1: logs(t')

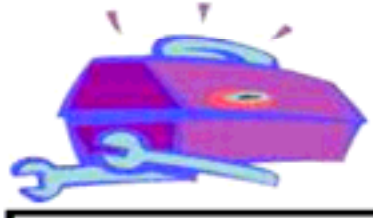
Sensor N: logs(t)

INTERNET

INSTITUT EURECOM

Events

IP headers
ICMP headers
TCP headers
UDP headers
payloads



TOOLS

IP geolocation
Name resolution (DNS, whois)
TCP stats
Passive OS fingerprinting
IDS alerts

[PDDP, NATO ARW'05]

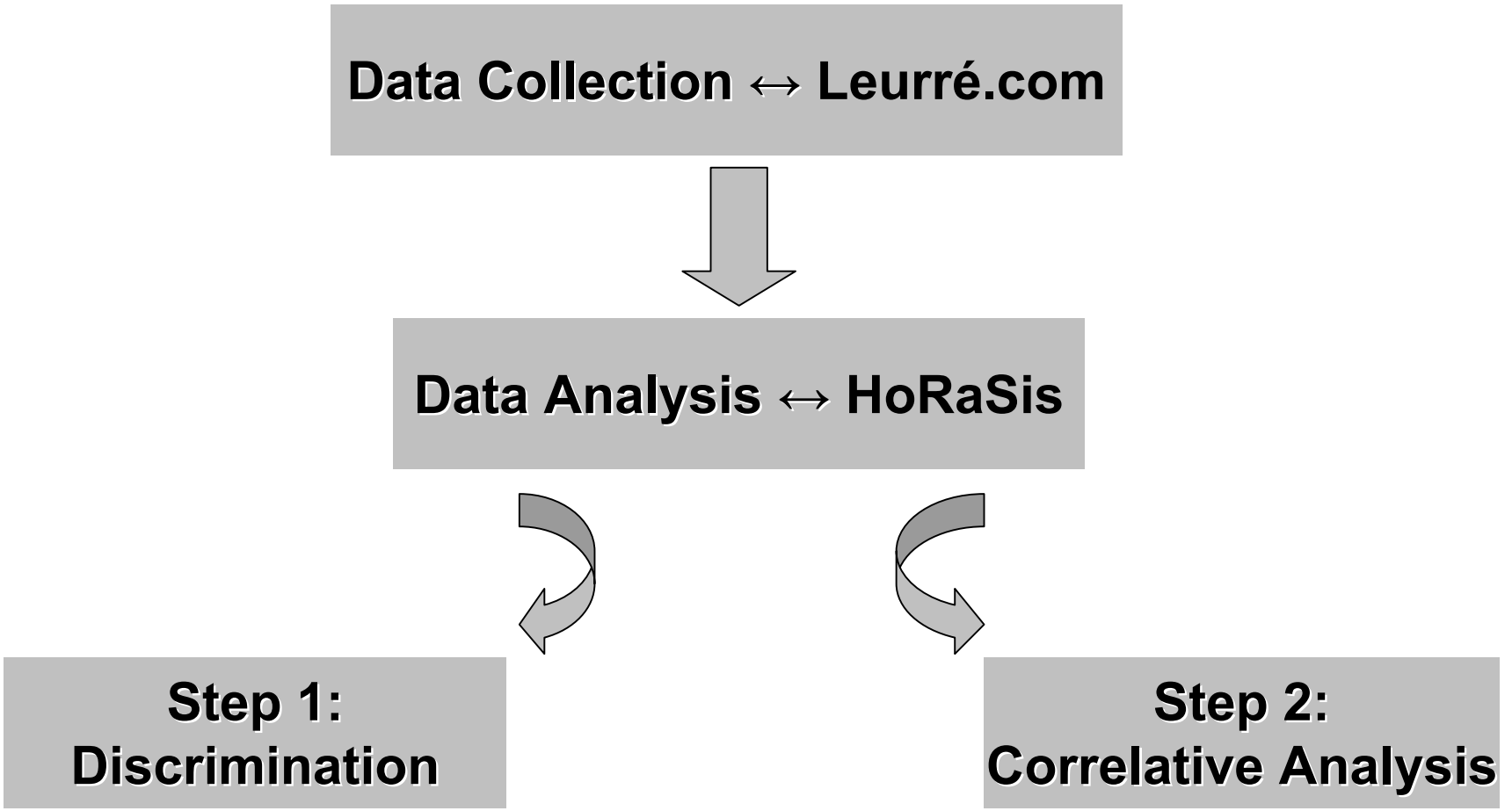
Big Picture

- Some sensors started running 3 years ago (30GB logs)
- 989,712 distinct IP addresses
- 41,937,600 received packets
- 90.9% TCP, 0.8% UDP, 5.2% ICMP, 3.1 others
- Top IP attacking countries
  (US, CN, DE, TW, YU…)
- Top operating systems
  (Windows: 91%, Undef.: 7%)
- Top domain names
  (.net, .com, .fr, not registered: 39%)

http://www.leurrecom.org

[DPD, NATO'04]

Considered approach :



***HoRaSis*: Honeypot tRaffic analySis**

- Our framework
- *Horasis*, from ancient Greek οραασις:
  “the act of seeing”
- Requirements
  - Validity
  - Knowledge Discovery
  - Modularity
  - Generality
  - Simplicity and intuitiveness

Identifying the activities

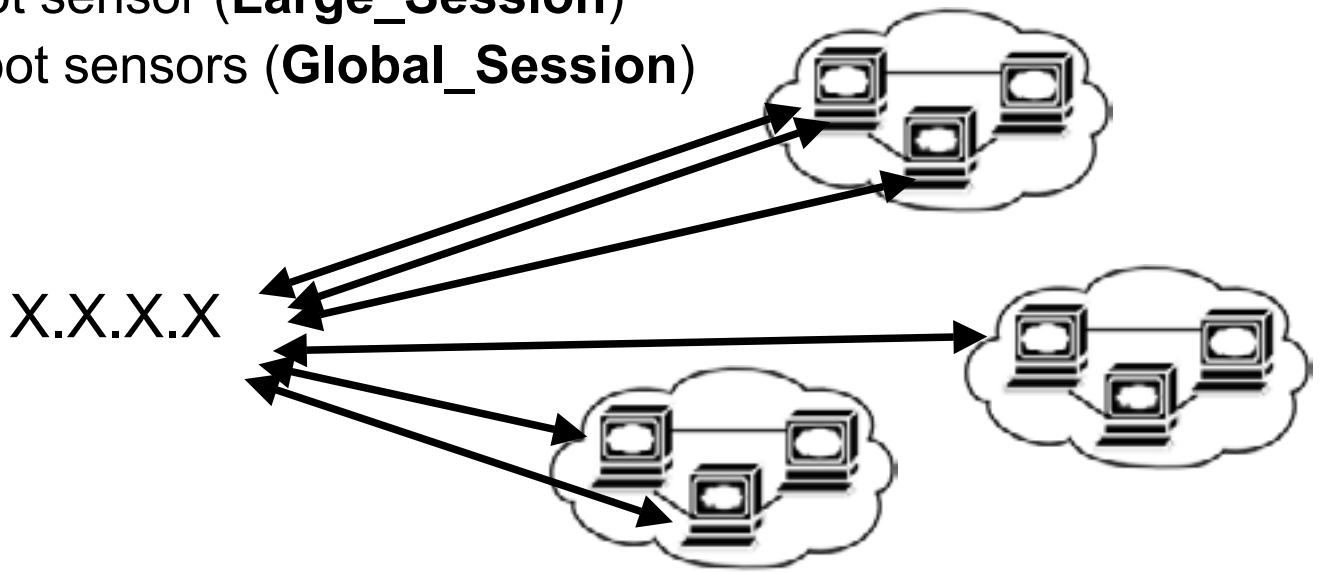
- Receiver side…
  - We only observe what the honeypots receive
- We observe several *activities*
- Intuitively, we have grouped packets in diverse ways for interpreting the activities
- What could be the analytical evidence (parameters) that could characterize such *activities?*

First effort of classification…

- **Source:** an IP address observed on one or many platforms and for which the inter-arrival time difference between consecutive received packets does not exceed a given threshold (25 hours).

We distinguish packets from an IP Source:

- To 1 virtual machine (**Tiny_Session**)
- To 1 honeypot sensor (**Large_Session**)
- To all honeypot sensors (**Global_Session**)



[PDP,IISW'05]

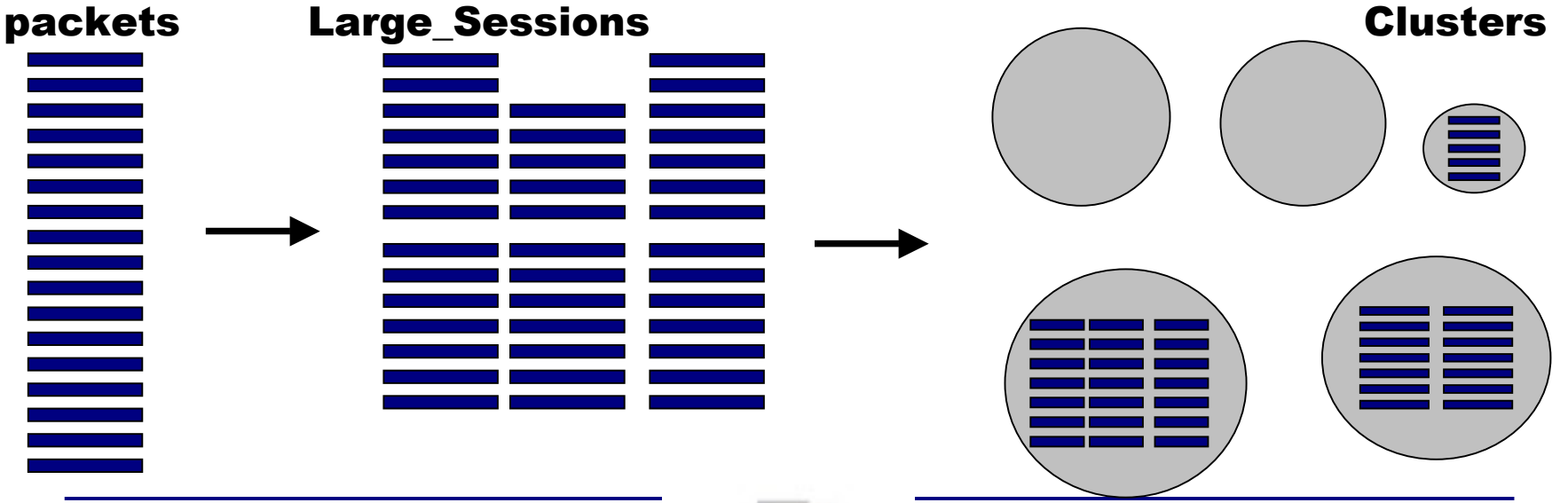
Fingerprinting the Activities

- **Clustering Parameters of Large_Sessions**:
  - Number of targeted VMs
  - The ordering of the attack against VMs
  - List of ports sequences
  - Duration
  - Number of packets sent to each VM
  - Average packets inter-arrival time

Discrimination step: summary

- A clustering algorithm
- An incremental version

Cluster = a set of IP Sources having the same activity fingerprint on a honeypot sensor



Cluster Signature

- A set of parameter values and intervals

CLUSTER ID:	IDENTIFICATION:
2145	

FINGERPRINT:

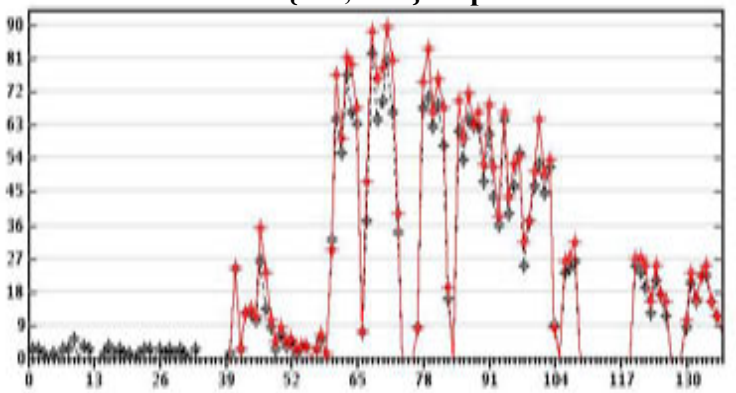
* Number Targeted Virtual Machines: 1
* Ports Sequence: 2745,2082,135,1025,445,3127,6129,139,1433,5000,80
* Number Packets sent VM: 33
* Global Duration: 7s < t < 11s
* Avg Inter Arrival Time: < 1s
* Payloads: yes (DCOM, Netbios, WebDav)

Plan

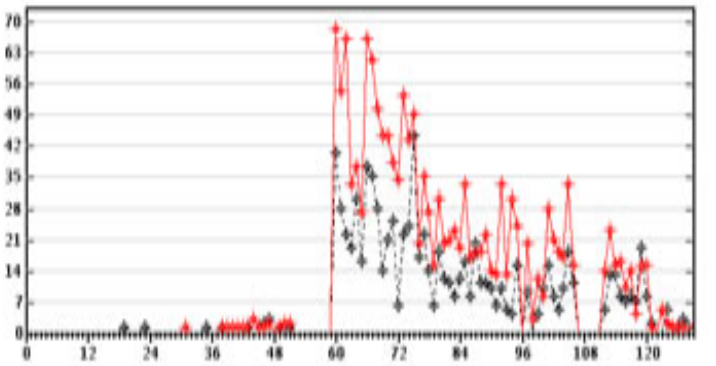
- What we already have…
  - A world-wide project
  - Large amount of data
  - A classification
- On studying temporal evolution of malicious activities
- The SAX similarity detection method
- Applications to the Leurré.com dataset
- Conclusions

On studying temporal evolution of activities… observation (1)

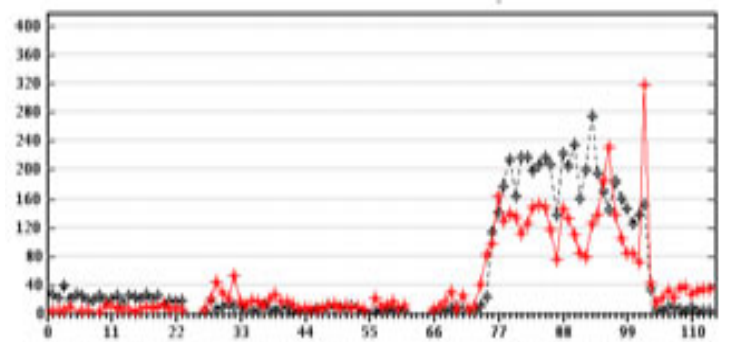
a) 2 attacks (clusters) targeting port {135} and ports {135,4444} resp.



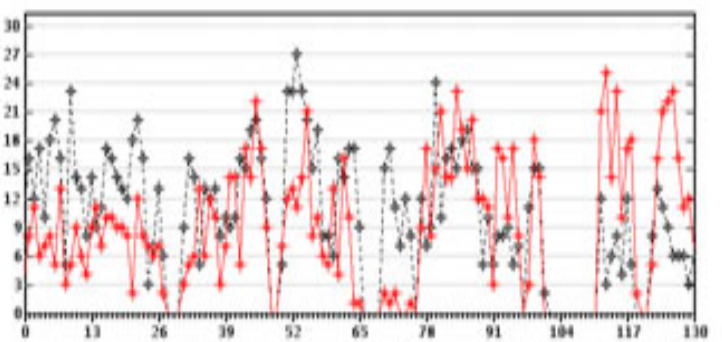
b) 2 attacks (clusters) targeting port {80} and port {135} resp.



c) 2 attacks (clusters) targeting port {1433} and port {139} resp.



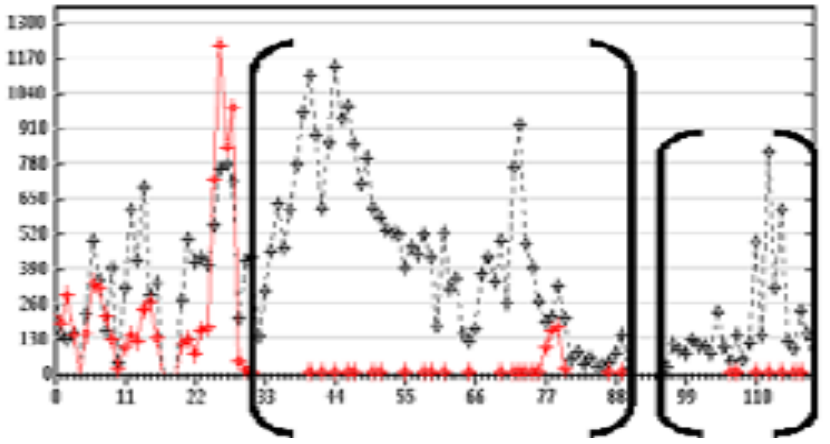
d) 2 attacks (clusters) targeting port {445} and ports {5554,1023,9898} resp.



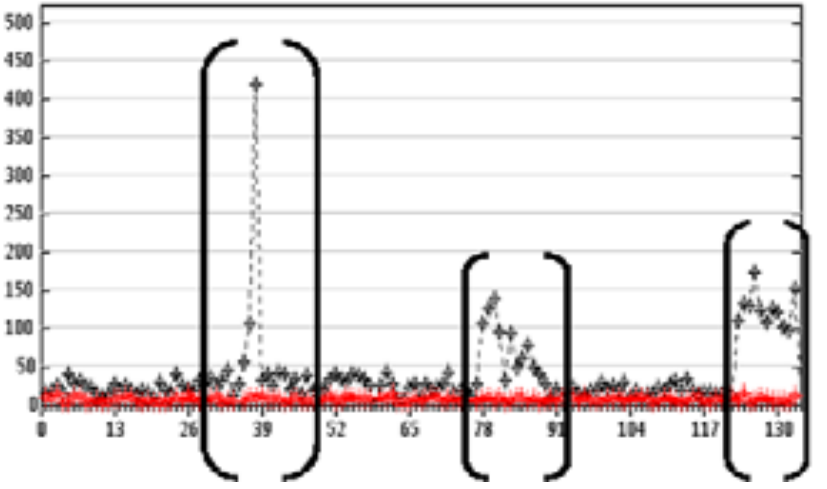
EURECOM

On studying temporal evolution of activities… observation (2)

a) Number of attacks having targeted port 80 or attacks having targeted port 135



b) Number of attacks having targeted port 139 or attacks having targeted port 1433



On studying temporal evolution

- Our Requirements…
  - Find an automatic method to find temporal similarities
- The method must be:
  - Incremental
  - Work at different granularity levels (day, week, month?)
  - Flexible: wipe out details but keep essential info

Plan

- What we already have…
  - A world-wide project
  - Large amount of data
  - A classification
- On studying temporal evolution of malicious activities
- The SAX similarity detection method
- Applications to the Leurré.com dataset
- Conclusions

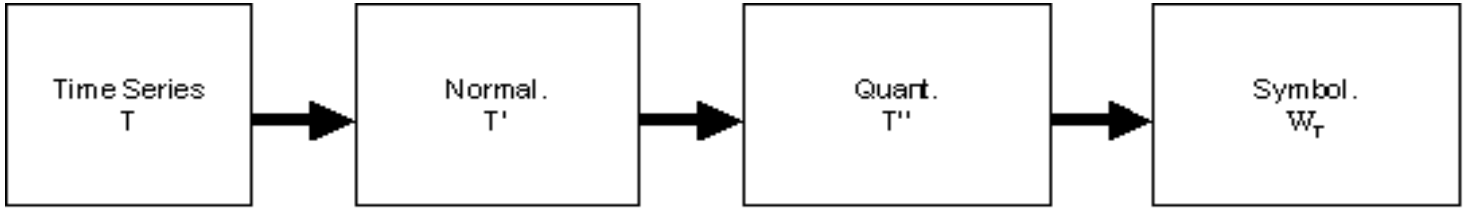
Symbolic Aggregate approXimation

- http://www.cs.ucr.edu/~jessica/sax.htm

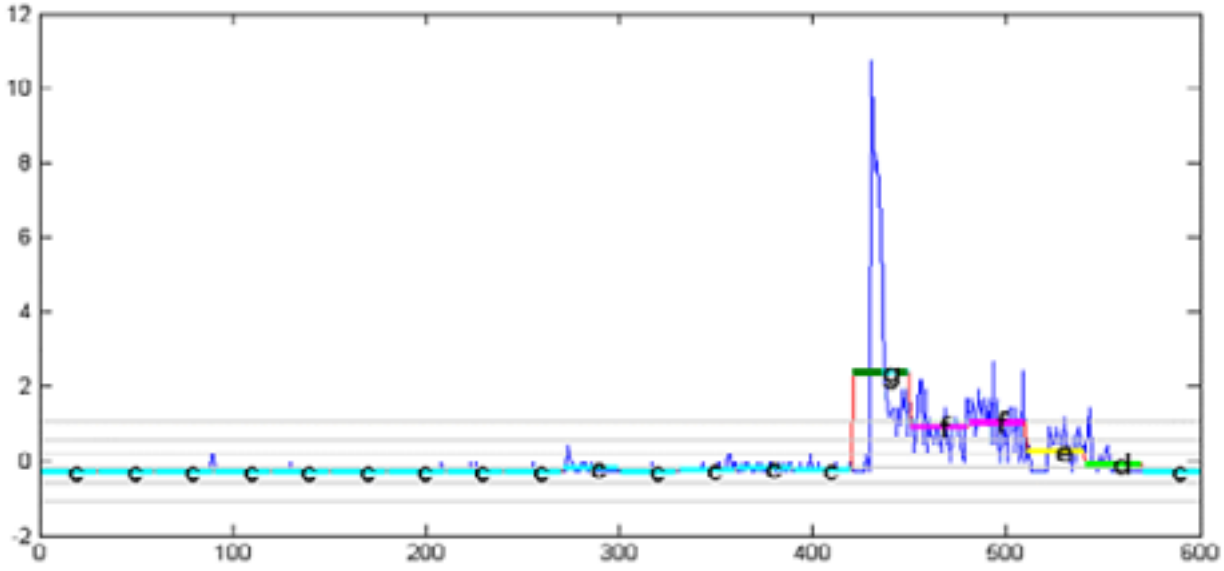
- J. Lin, E. Keogh, E. Lonardi, B. Chiu :

"A Symbolic Representation of Time Series, with Implications for Streaming Algorithms". *ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, 2003.

SAX principles



Three steps to get the SAX symbolic representation of T (PAA of initial time series)



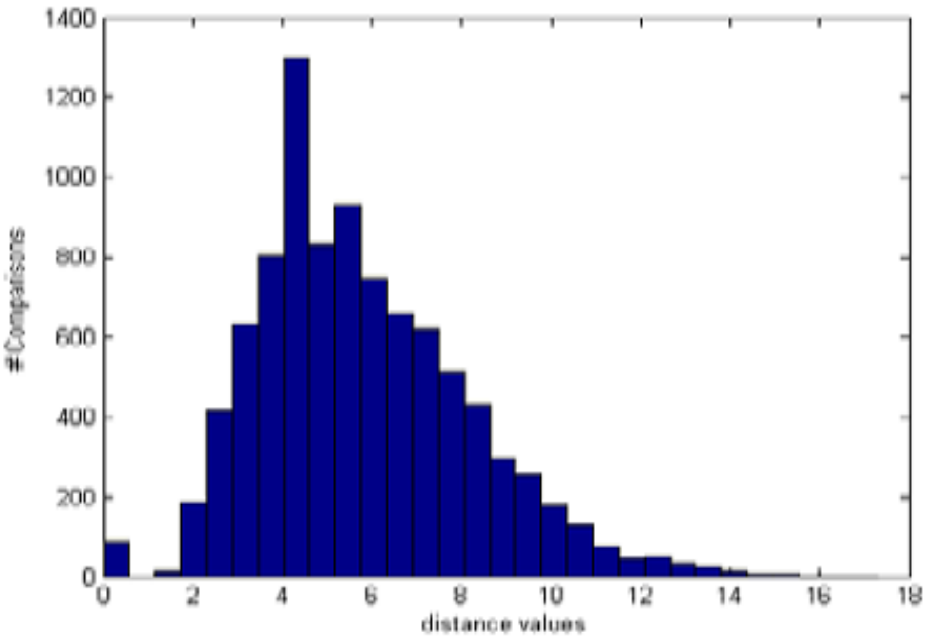
ccccccccccccgffedc

Similarity detection

- Distance between two SAX strings:

$$D(\mathrm{W}_{\mathrm{T}_1}, \mathrm{W}_{\mathrm{T}_2}) = \sqrt{\frac{N}{w}} \sqrt{\left(\sum_{i=1}^{w} TAB(\mathrm{W}_{\mathrm{T}_1}(\mathrm{i}), \mathrm{W}_{\mathrm{T}_2}(\mathrm{i}))\right)^2}$$

- Usefull feature:
  - If D>1, time series are visually dissimilar
  - If D==0, they are similar
- Remaining issue:
  - Choice of alphabet size
  - For our case:
    - 4 is too coarse
    - 5 is ok
    - 6 is too conservative



Plan

- What we already have…
  - A world-wide project
  - Large amount of data
  - A classification
- On studying temporal evolution of malicious activities
- The SAX similarity detection method
- Applications to the Leurré.com dataset
- Conclusions

SAX Analysis

- Input : the 137 largest clusters
- Output : 89 pairs of similar time series (a cluster might appear in several pairs)
- Parameter : 1-week = 1 symbol
- In terms of probabilities….
  - K = number of strings (Time Series)
  - w = string size

$$P = \frac{K(K-1)}{2} \times \left(\frac{13}{25}\right)^{w} < 10^{-13}$$

SAX Analysis : three categories of similarities (1)

- Malware targeting random IPs with sequential ports sequences

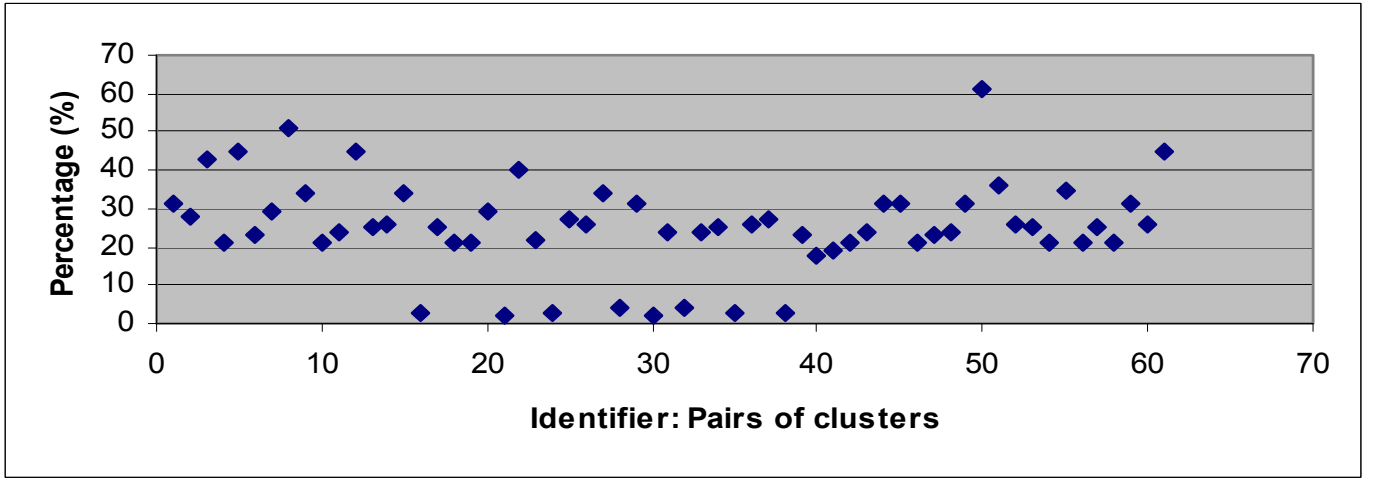
$$PS_a = (PS_b, *)$$

  Sophisticated tools that always target the same sequence of ports on a machine, but stop scanning if ever one of the ports is closed.
  - Typical example: MBlaster with 4 clusters
- Overlap (85 -100%) between source IPs

SAX Analysis : three categories of similarities (2)

- **Multi-headed :** Malware targeting different ports on each victim
  Strong domain similarities and common IPs

$$P(common\ \ domains\ : C_a\ and\ C_b) = \frac{card\ (Dom_a \cap Dom_b)}{card\ (Dom_a) + card\ (Dom_b) - card\ (Dom_a \cap Dom_b)}.100$$



Multi-Headed Worms

- Some identified malware :
  - Nachi (also called Welchia)
    Randomly chooses an IP address and then attacks it either against port 135 or port 445
  - Spybot.FCD
    Tries to exploit Windows vulnerabilities either on port 135, 445 or 443

SAX Analysis : three categories of similarities (3)

- Other cases…
  - No domain, network, IP clear similarity
  - No top domain, or country close distribution
  - Apparently more personal computers than the average (=> domain name including strings such as ‘%dial%’, ‘%dsl%’ or ‘%cable%’)
  - 8 cluster pairs, involving ports 21, 25, 80, 111, 135, 137, 139, 445, 554 and 27374.

Open Issue (capture and analysis)
- Stealthier multi-headed worms ?
- Other phenomena ?

Example :

- One pair :
  - cluster 1 : attacks targeting port 27374 (a port left open by some Trojans)
  - cluster 2 : attacks targeting port 21 (FTP).

C_a	C_b
CN: 24%	US: 47%
KR: 17%	KR: 11%
TW: 14%	FR: 10%
US: 10%	CA: 7%
DE: 7%	DE: 6%

C_a		C_b	
.net	31%	.net	32%
.com	4%	.com	40%
.it	3%	.fr	9%
others	28%	others	1%
undetermined	34%	undetermined	18%

Plan

- What we already have…
  - A world-wide project
  - Large amount of data
  - A classification
- On studying temporal evolution of malicious activities
- The interesting SAX method
- Applications to the Leurré.com dataset
- Conclusions

Conclusions

- We have highlighted the existence of so called multi-headed stealthy tools based on the similarity between their time signatures
  - difficult to identify except by reverse engineering their code (*a priori* knowledge)
- two distinct steps:
  1. we group attacks with a common fingerprint on a honeypot platform into the same cluster
  2. we compare the temporal evolution of these clusters to find out similarities

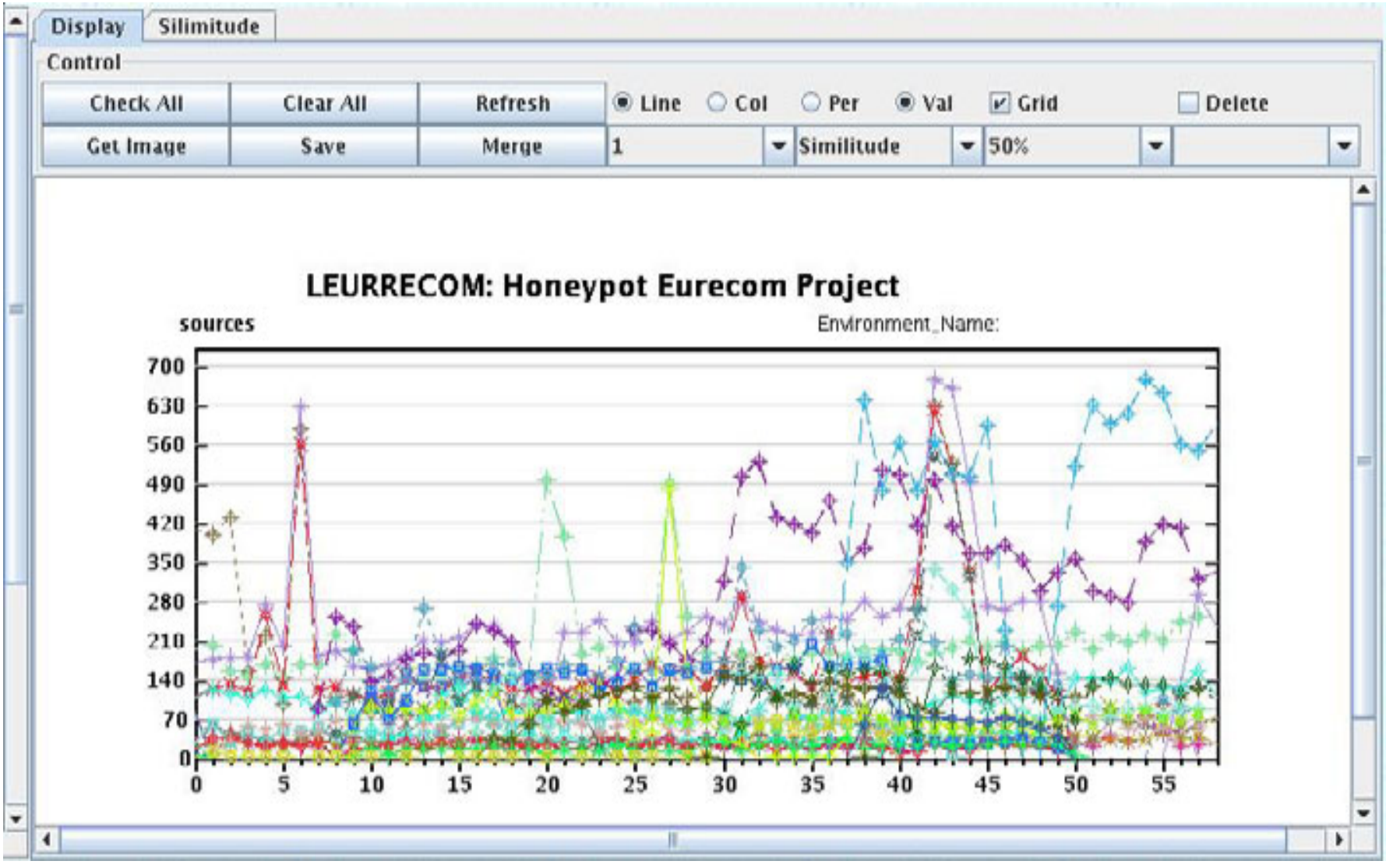
Conclusions and perspectives

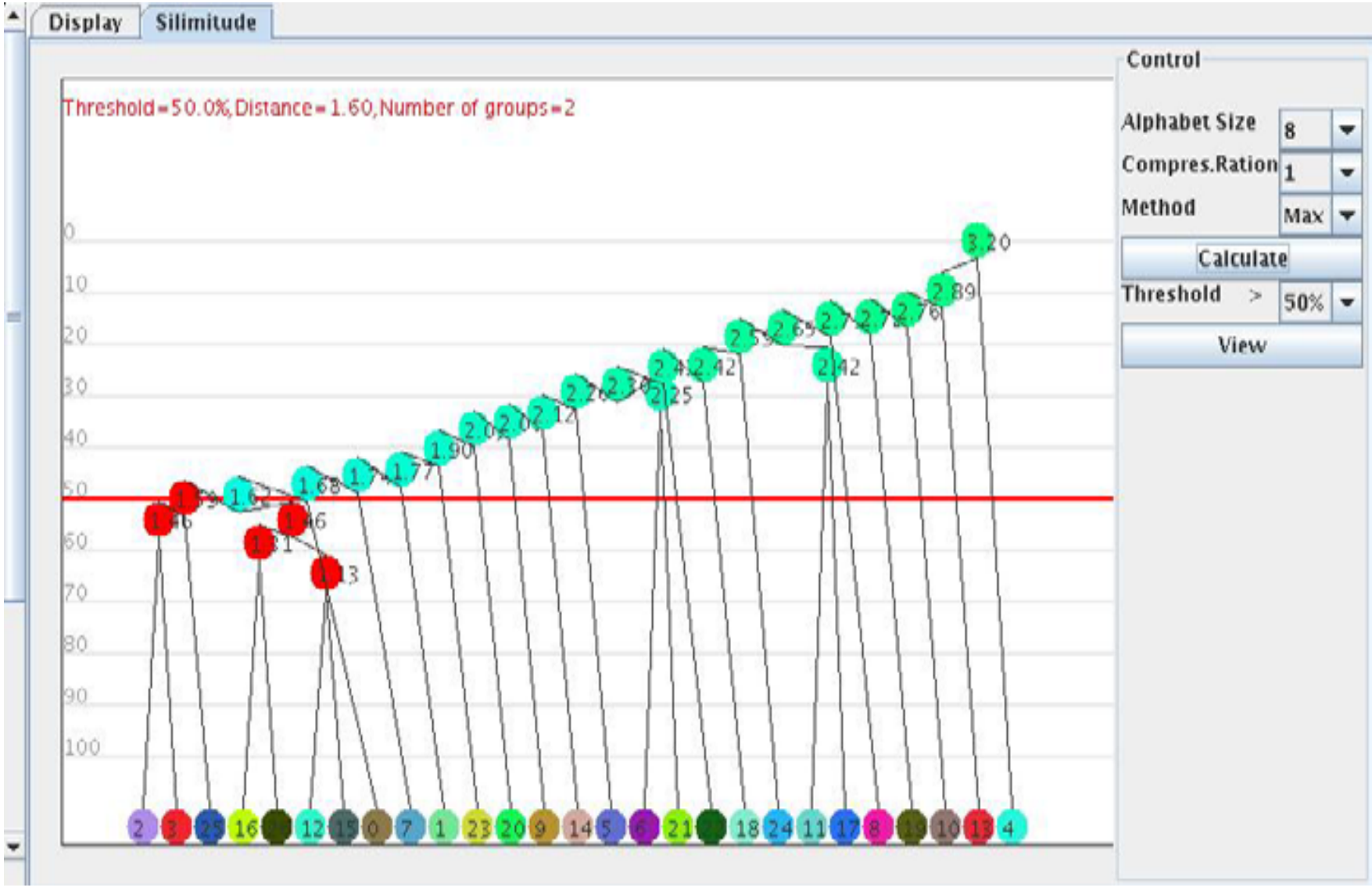
- SAX is a very interesting approach
- Results must be cross-correlated with other cluster-based analyses
  - *HoraSis Framework* (see TF-CSIRT Amsterdam, January 2006)
- Perspectives
  - Different time window granularities
  - Partial similarities

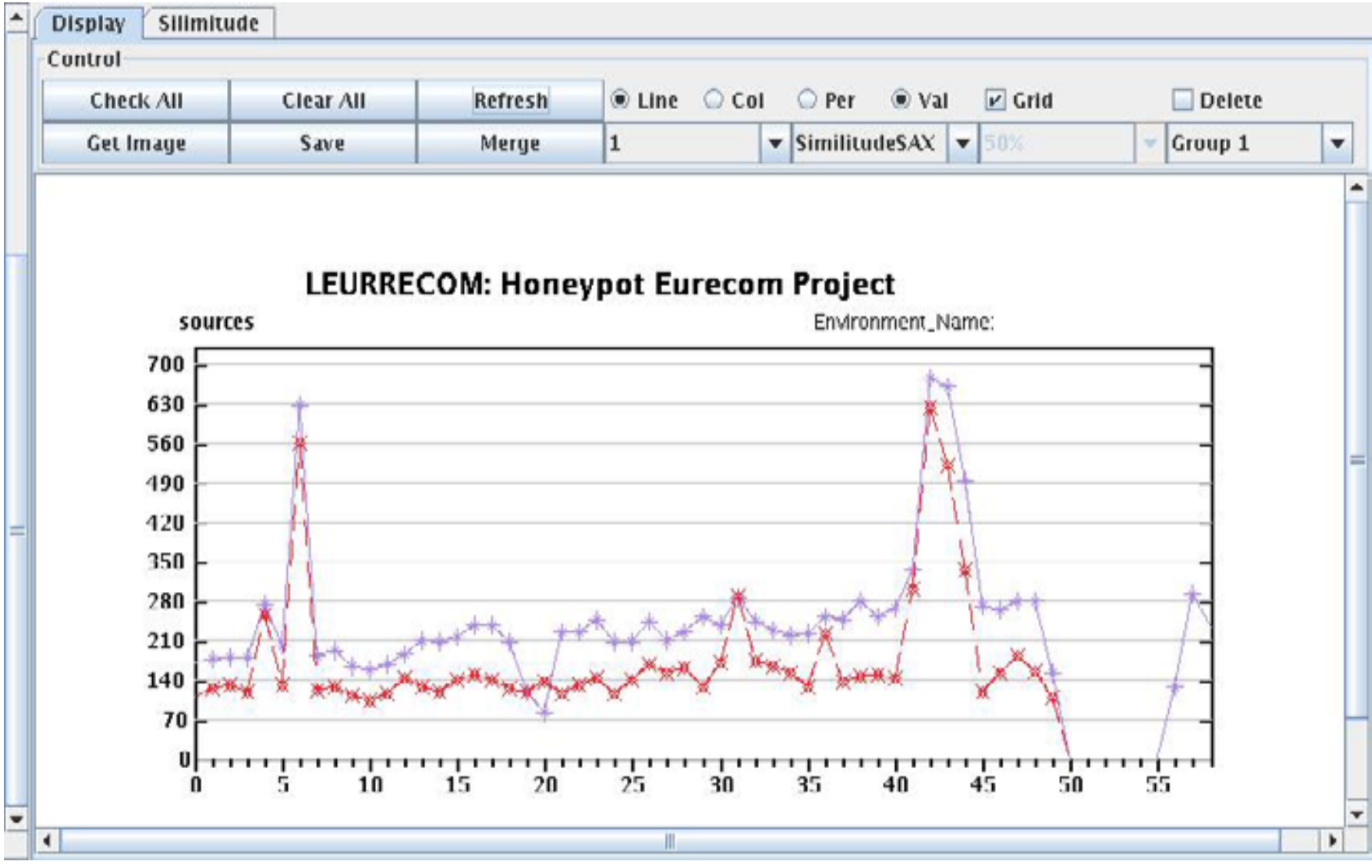
This method…

- … is available to all Leurré.com partners (see http://www.leurrecom.org)
- A Java applet

SCREENSHOT INTERFACE DEMO


Display
Silimitude
Control
Check All
Clear All
Refresh
Line
Col
Per
Val
Grid
Delete
Get Image
Save
Merge
1
Similitude
50%
LEURRECOM: Honeypot Eurecom Project
sources
Environment_Name:
700
630
560
490
420
350
280
210
140
70
0
0
5
10
15
20
25
30
35
40
45
50
55


Display
Silimitude
Threshold=50.0%,Distance=1.60,Number of groups=2
Control
Alphabet Size
8
Compres.Ration
1
Method
Max
Calculate
Threshold
>
50%
View


Display
Silimitude
Control
Check All
Clear All
Refresh
Line
Col
Per
Val
Grid
Delete
Get Image
Save
Merge
1
SimilitudeSAX
50%
Group 1
LEURRECOM: Honeypot Eurecom Project
sources
Environment_Name:
700
630
560
490
420
350
280
210
140
70
0
0
5
10
15
20
25
30
35
40
45
50
55

Thank You for your Attention!

Questions