WOMBAT: towards a Worldwide Observatory of Malicious Behaviors and Attack Threats

Fabien Pouget

Institut Eurécom

January 24th 2006

EURECOM
Sophia Antipolis

TF-CSIRT 2006

Observations

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

Statement

It is possible to build a framework that helps better identifying and understanding of malicious activities in the Internet.

Data Collection

↓

Data Analysis

Research in this Direction…
… Capturing/Collecting Data (1)

A **Honeypot** is an information system resource whose value lies in unauthorized or illicit use of that resource

- **Darknets, Telescopes, Blackholes:** CAIDA Telescope, IMS, iSink, Minos, Team Cymru, Honeytank
  - ☒ Generally good for seeing explosions, not small events
  - ☒ Assumption that observation can be extrapolated to the whole Internet
  - ☒ Can be blacklisted and bypassed
- **Other Honeypots, Honeytokens:** mwcollect, nepenthes, honeytank
  - ☒ Interesting but quite specific collection techniques

Research in this Direction…
… Capturing/Collecting Data (2)

- **Log Sharing:**
  Dshield, Internet Storm Center (ISC) from SANS Institute, MyNetWatchman, Symantec DeepSight Analyzer, Worm Radar, Talisker Defense Operational Picture
  - ☒ Mixing various things
  - ☒ No information about the log sources

Research in this Direction…
… Analyzing Data

- Netflow flow level aggregation
  - ☒ Not always fine grained analysis
  - ☒ Information often limited to netflow recorded fields

- Intrusion Detection System alerts and derived tools (Monitoring Consoles)
  - ☒ Analysis as accurate as alerts…

- Modeling
  - ☒ Validation Process and specificity
  - ☒ *A priori* knowledge

Conclusions

- We should 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.

Refined Statement

It is possible to build a framework that helps better identifying and understanding of malicious activities in the Internet.

1. By collecting data from simple honeypot sensors (few IPs) placed in various locations.

2. By building a technique adapted to this data in order to automate knowledge discovery.

Our Approach

Data Collection ↔ Leurré.com

Data Analysis ↔ HoRaSis

Step 1: Discrimination

Step 2: Correlative Analysis

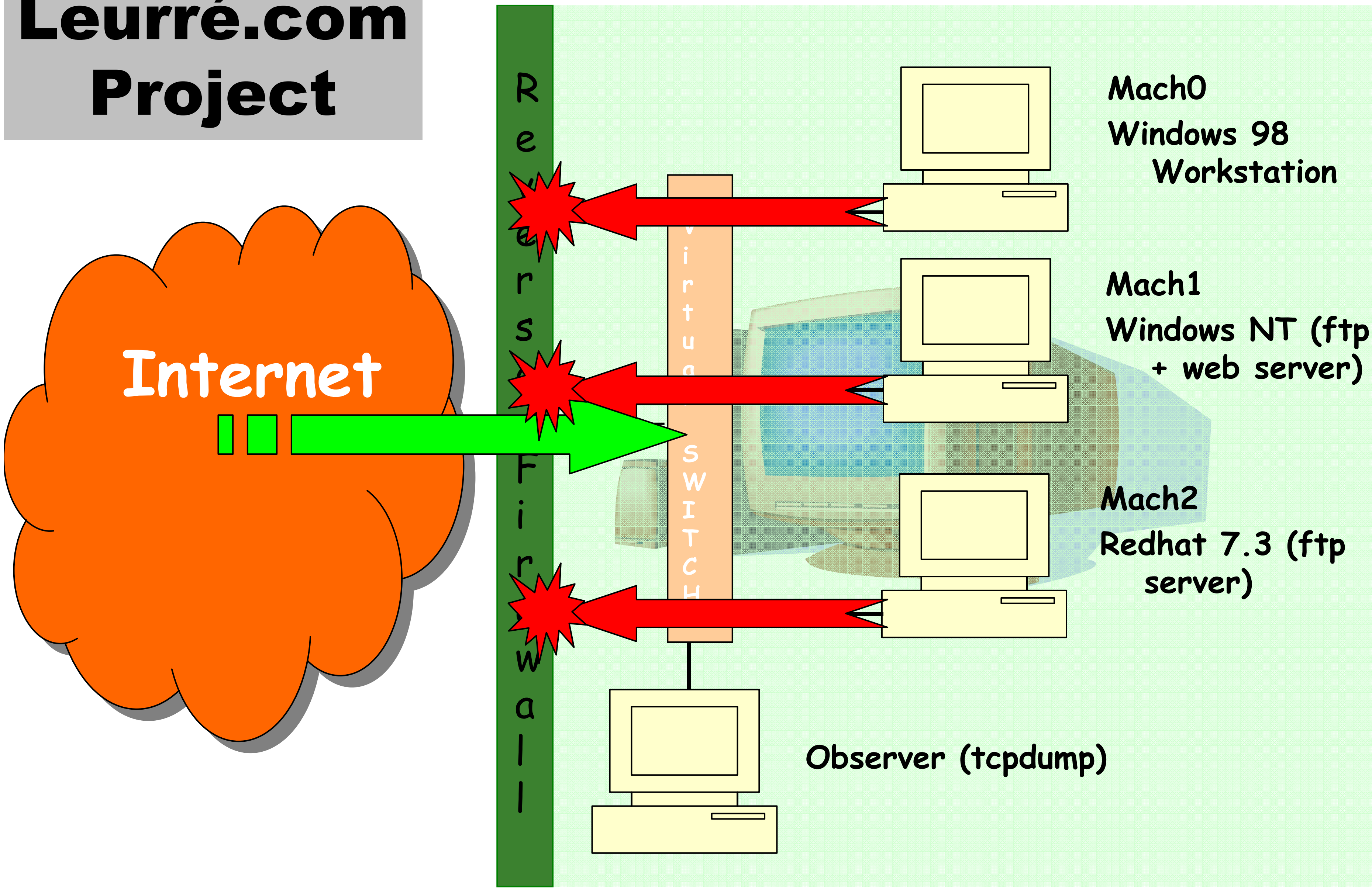
Win-Win Partnership

- The interested partner provides …
  - One old PC (pentiumII, 128M RAM, 233 MHz…),
  - 4 routable IP addresses,
- EURECOM offers …
  - Installation CD Rom
  - Remote logs collection and integrity check.
  - Access to the whole SQL database by means of a secure web access.

- Partially funded by the French ACI Security named CADHO (CERT Renater and CNRS LAAS)

- Joint Research with France Telecom R&D

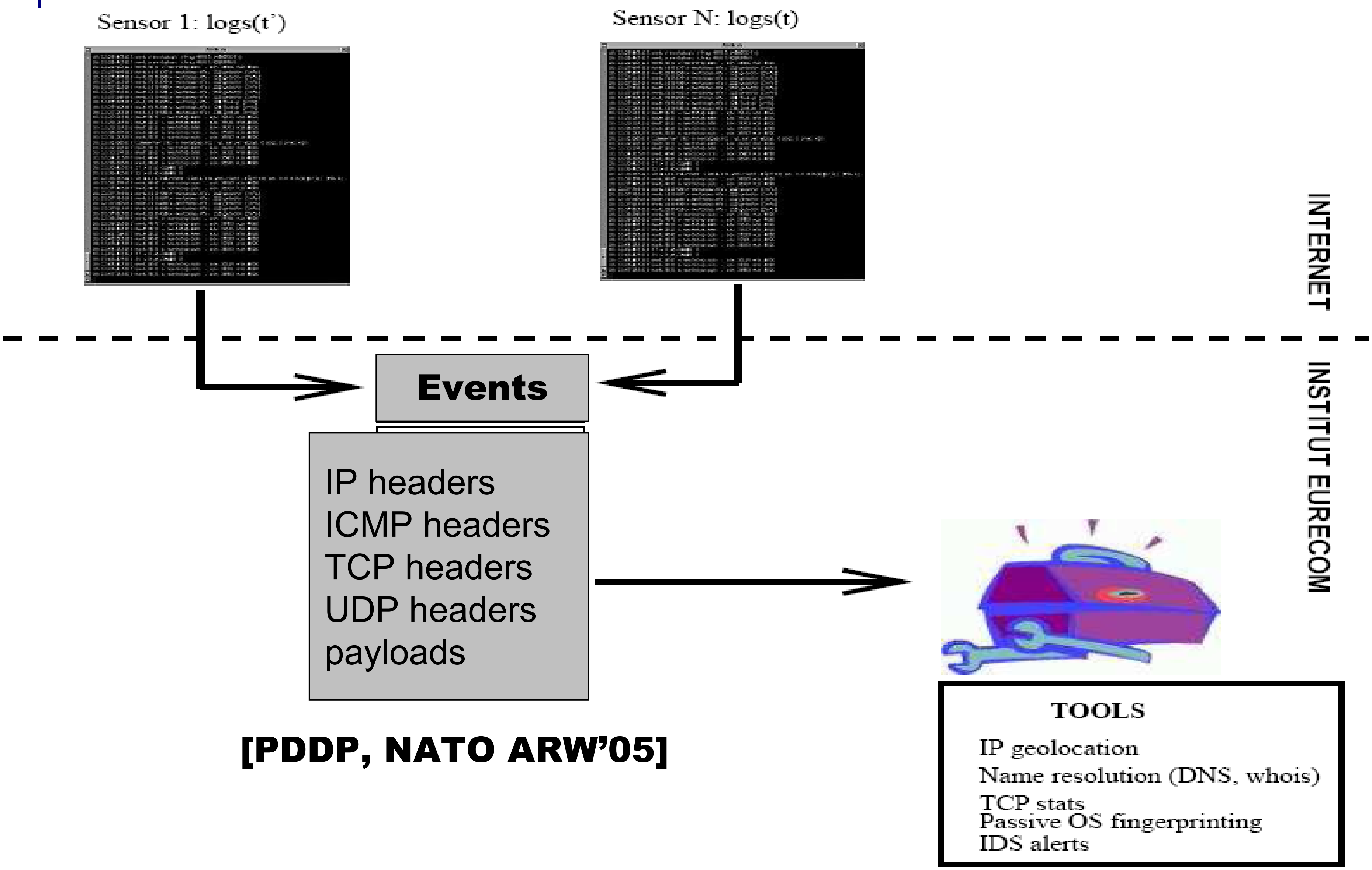
Leurré.com Project



40 sensors, 25 countries, 5 continents

Leurré.com Project

Europe



[PDDP, NATO ARW'05]

Some Relevant Details

What is the bias introduced by using honeypots with *low interaction* instead of real systems for the analysis?

- High Interaction Honeypots as 'Etalon Systems': reference for checking port interactivity

[PH, DIMVA'05] For each port:

$$I(H_1) = \sum_p P_p . f_p$$

$$I(H_2) = \sum_k P_k . f_k$$

$$\frac{I(H_1)}{I(H_2)} = \eta$$

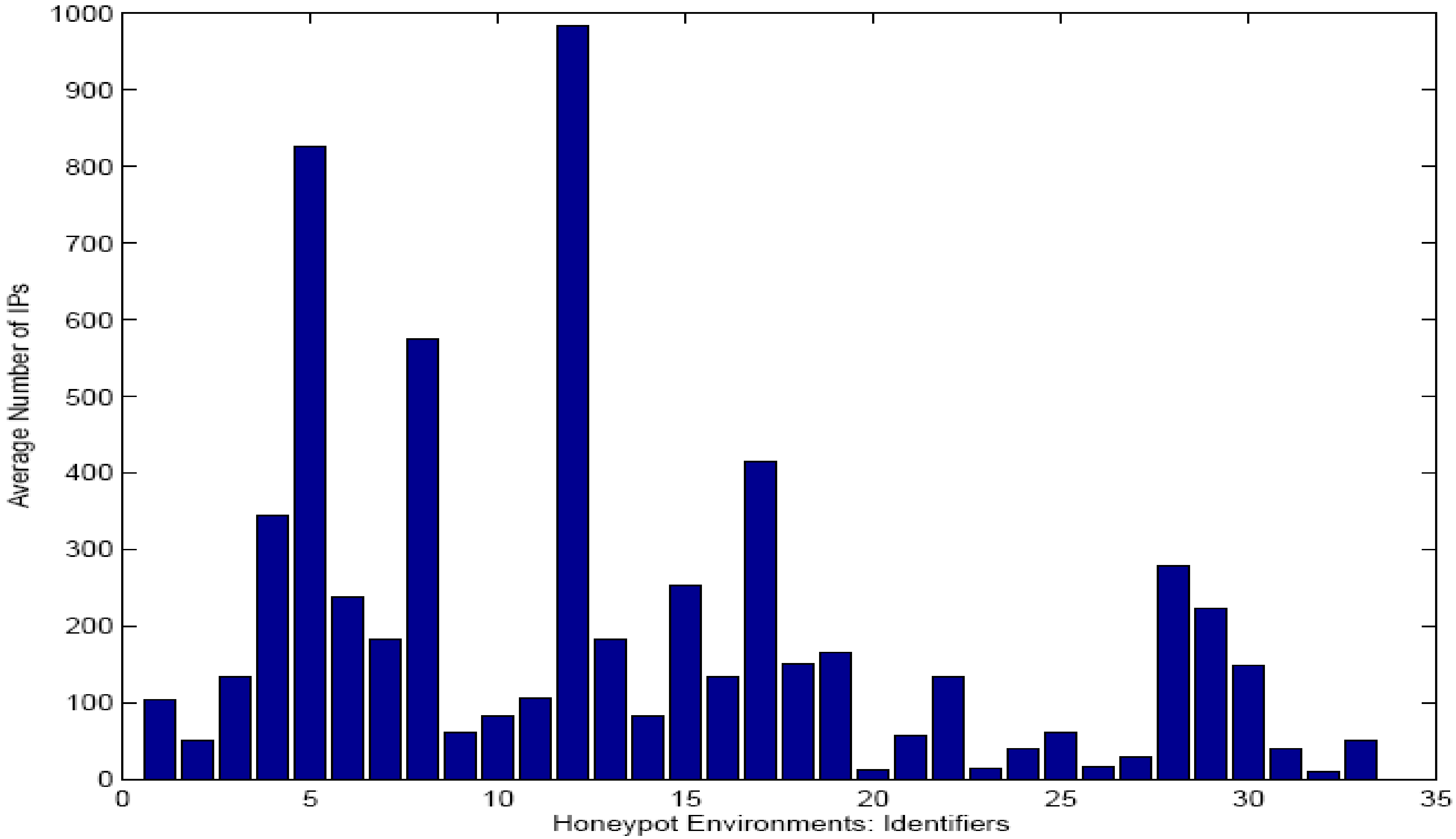
Principle:

- To check basic statistics
- To check the interaction relevance

Big Picture

- Some sensors started running 2 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 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]**



[CLPD, SADFE'05]

[PDP, ECCE'05]

IP addresses observed per sensor per day

Our Approach

Data Collection ↔ Leurré.com

↓

Data Analysis ↔ HoRaSis

Step 1: Discrimination

Step 2: Correlative Analysis

HoRaSis: Honeypot tRaffic analySis

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

HoRaSis

First step: Discrimination of attack processes

1. Remove network influences
2. Identify parameters characterizing activities (fingerprint)
3. Cluster the dataset according to chosen parameters
4. Check consistency of clusters

Identifying the activities

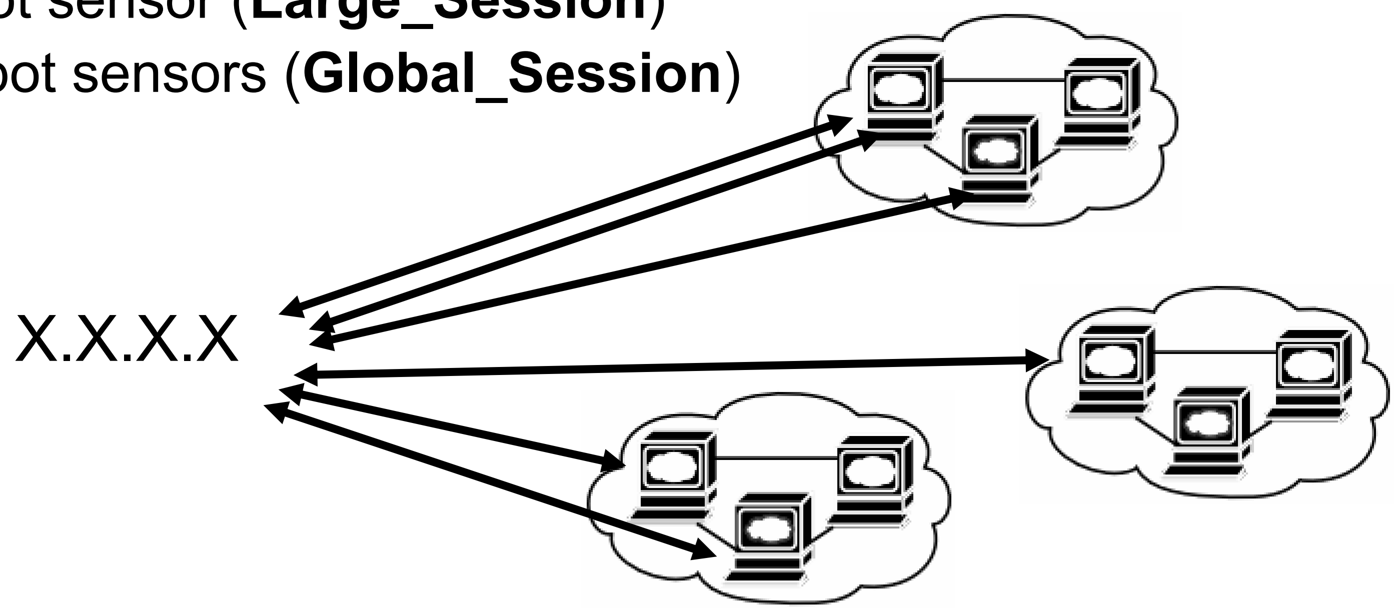
- Receiver side…
  - We only observe what the honeypots receive
- We observe several *activities*
- Intuitively, we have grouped packets in diverse ways for interpreting the activities
- What could be the analytical evidence (parameters) that could characterize such *activities?*

First effort of classification…

- **<u>Source:</u>** 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

- **<u>Clustering Parameters of Large_Sessions</u>**:
  - 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

Parameters

- Discrete values
  - Resistant to network influences
  - Ex: Ports Sequence

Clustering function:

Exact n-tuplet match

- Generalized values
  - Modal properties
  - Ex: Nb rx packets

Clustering function:

Peak picking strategy
Bins creation

Parameters relevance estimated by the entropy-based Information Gain Ratio (IGR)

$$IGR(Class, Attribute) = \frac{(H(Class) - H(Class \langle Attribute \rangle))}{H(Attribute)}$$

[DPD, PRDC'04]

Clusters Consistency

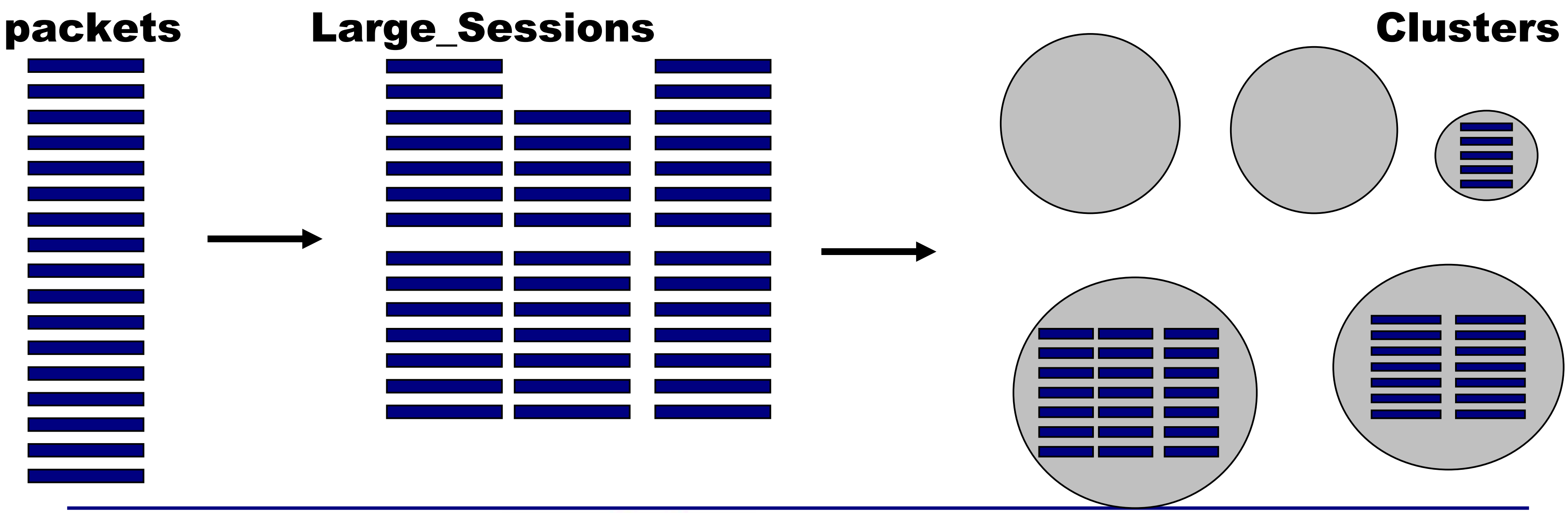
- Unsupervised classification
- Levenshtein-based distance function
  - Concatenated payloads => activity sentences
  - Count *deletions*, *insertions*, *substitutions* btw sentences
  - Pyramidal agglomerative bottom-up algorithm

- Payload Homogeneity **[PD, AusCERT'04]**
- Splitting Ratio:

$$\gamma_d = \frac{\#\ \text{Obtained Subclusters}}{\#\ \text{Sources grouped in the initial Cluster}}$$

Discrimination step: summary

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)

Our Approach

Data Collection ↔ Leurré.com

Data Analysis ↔ HoRaSis

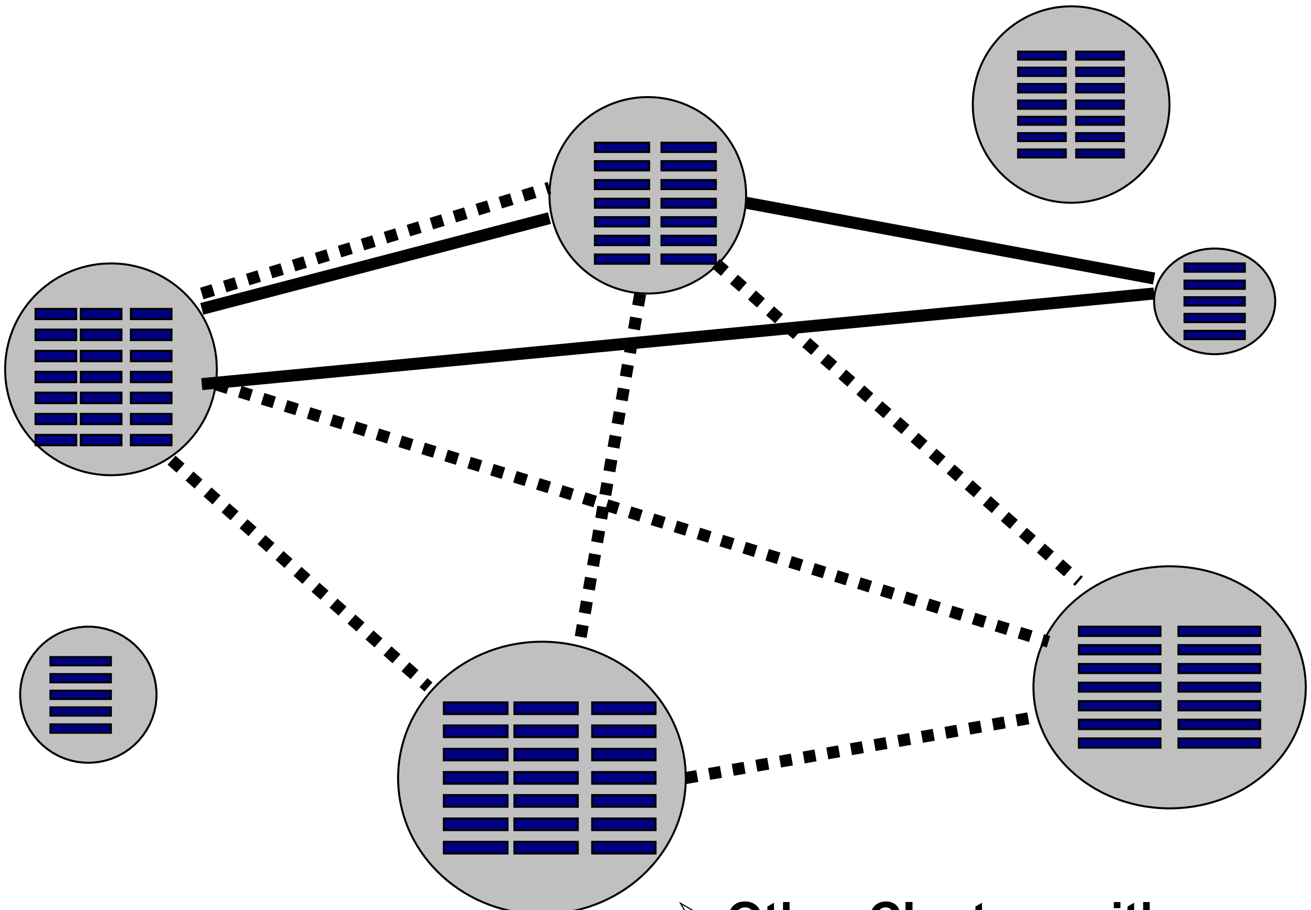
Step 1: Discrimination

Step 2: Correlative analysis

HoRaSis

Second step:
Correlative Analysis of the Clusters

Correlative Analysis of Clusters



Clusters containing Sources from Countries A and B only

Clusters having been observed on Sensor X only

- Other Clusters with same properties?
- Other relationships from previous analyses?

► **Recurrent Questions**

► **Need to automate this analysis**

Dominant Sets Extraction (1)

- *Similar* characteristics between clusters
- Clusters as Nodes: graph
- For each analysis, construct several edge-weighted graphs
- a Graphic Theoretic problem of finding *maximal cliques* in *edge-weighted* graphs.

[PUD, RR-05]

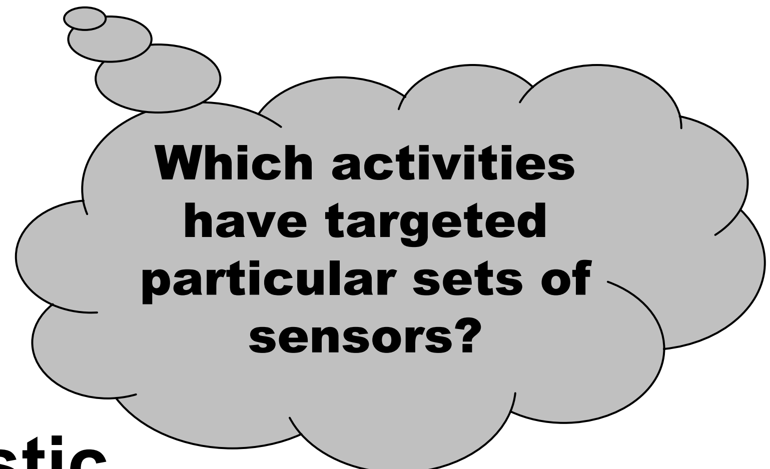
Dominant Set Extraction (2)

- Maximal Clique problem:
  NP-hard (even for unweighted graphs)
- Dominant Set Extraction approach
- Based on the solution from Pelillo & Pavan(2003):
  - Dominant set extracted by replicator dynamics
  - Fast convergence to one solution

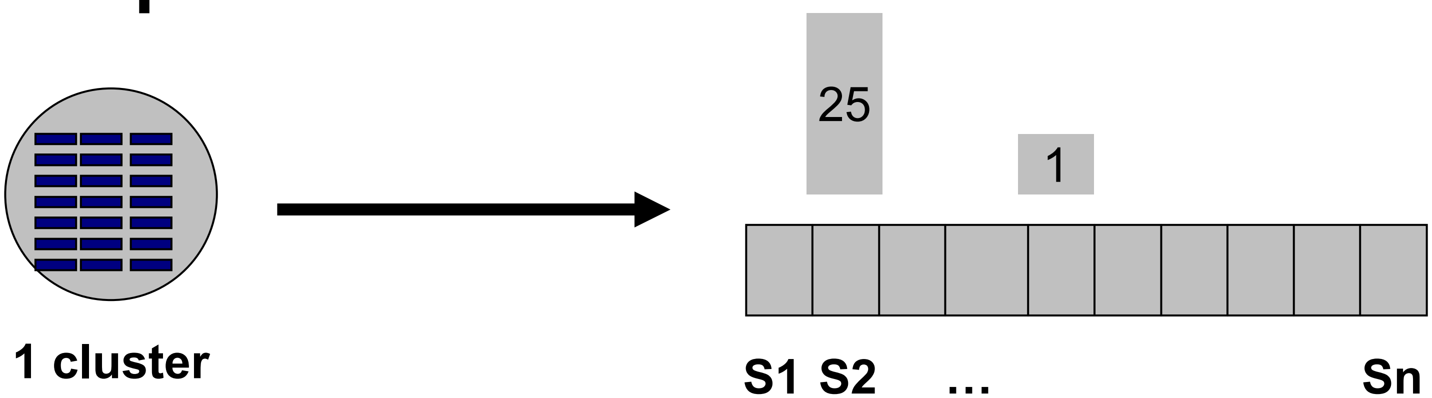
$$x_i(t+1) = x_i(t)\frac{(Ax(t))_i}{x(t)^T Ax(t)}$$

Our Algorithm
Step 1 – Define a correlation analysis

1. **Consider a characteristic**

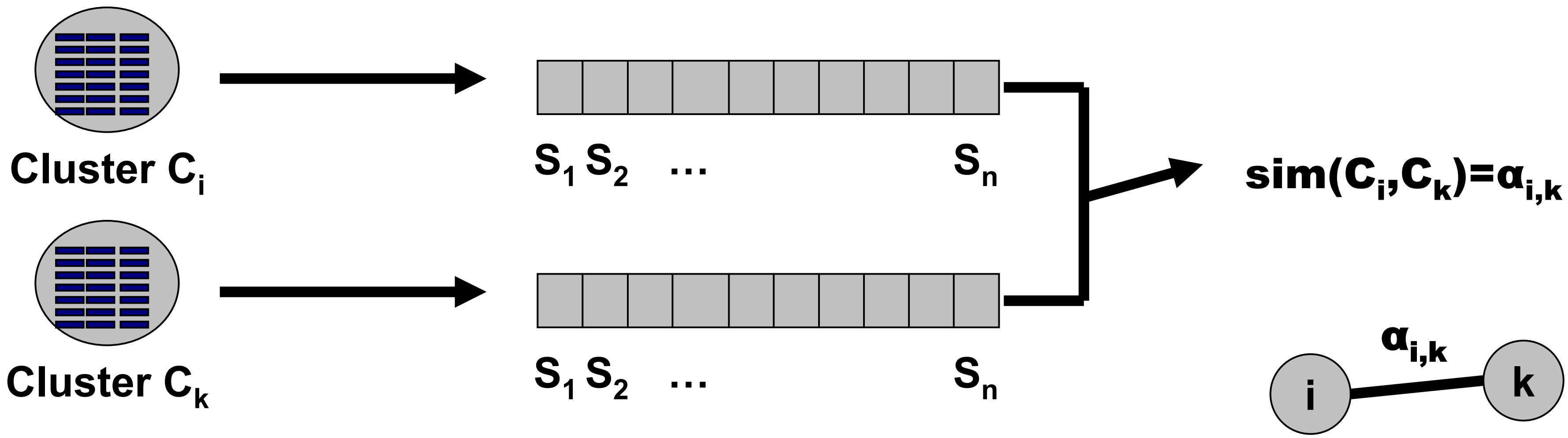


2. **Represent this characteristic**

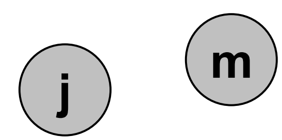


Our Algorithm
Step 2 – Build the edge-weighted graph

3. **Define a similarity function that compares values**

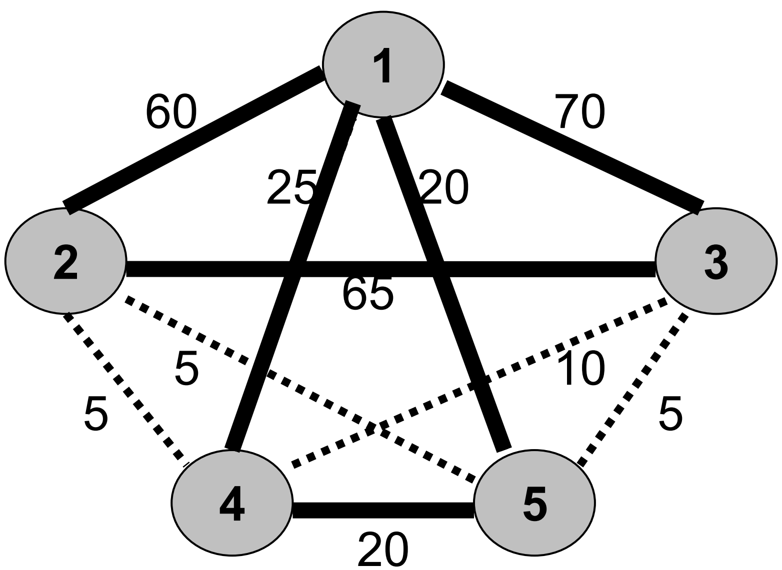


4. **Insert the values in a similarity matrix (edge-weighted graph)**



Our Algorithm
Step 3 – Extract Relevant Dominant Sets

5. **Apply recursively Pelillo&Pavan technique**



{1,2,3}

{1,4,5}

$$A_1 \begin{pmatrix} 0 & & a_{i,k} \\ & .. \quad 0 & \\ a_{k,i} & .. & 0 \end{pmatrix} \Rightarrow \begin{matrix} DS_{1,1} \\ DS_{1,2} \\ \ldots \\ DS_{1,N1} \end{matrix}$$

$$A_2 \begin{pmatrix} 0 & & b_{i,k} \\ & .. \quad 0 & \\ b_{k,i} & .. & 0 \end{pmatrix} \Rightarrow \begin{matrix} DS_{2,1} \\ DS_{2,2} \\ \ldots \\ DS_{2,N2} \end{matrix}$$

$$A_3 \begin{pmatrix} 0 & & c_{i,k} \\ & .. \quad 0 & \\ c_{k,i} & .. & 0 \end{pmatrix} \Rightarrow \begin{matrix} DS_{3,1} \\ DS_{3,2} \\ \ldots \\ DS_{3,N3} \end{matrix}$$

$\cap$	$DS_{1,1}$	$DS_{1,2}$	…	$DS_{1,N1}$
$DS_{2,1}$				
$DS_{2,1}$				
…				
$DS_{2,N2}$				

Intersection $DS_{1,2}$ with $DS_{2,1}$:

- List of Common Clusters
- Weight (%) of this new set of Clusters

$$W(\%) = \frac{card(new\_set\_of\_clusters)}{\min(card(DS_{1,2}); card(DS_{2,1}))}$$

Matrices in use

- 8 distinct matrices having developed.
- 3 distinct similarity functions have been defined

Matrix Name	Similarity Meaning btw Clusters
A_Geo	Distribution of attacking countries
A_Env	Distribution of targeted environments
A_OSs	Distribution of attacking OSs
A_IPprox	IP proximity of attacking sources
A_TLDs	Distribution of attacking Top-Level Domains
A_Hostnames	Attacking machine types
A_ComIPs	Shared attacking IPv4 addresses
A_SAX	Temporal evolution over weeks

Results (1): A_Geo

Dominant Set ID	# Clusters	Corresp. Peaks
ID 1	20	{CN}
ID 2	14	{CN,US}
ID 3	12	{YU}
ID 4	11	{YU,GR}
ID 5	10	{CN,US,JP}
ID 6	6	{CN,KR}
ID 7	10	{CN,CA}
ID 8	4	{CN,KR,JP}
ID9	9	{CN,US,TW}

12 distinct activities have been launched by Sources coming from YU only.

Results (2): A_Env

Dominant Set ID	# Clusters	Corresp. Peaks
ID 1	30	{20}
ID 2	28	{6}
ID 3	20	{20,8}
ID 4	18	{32}
ID 5	14	{20,25}
ID 6	26	{25}
ID 7	43	{6,31}
ID 8	10	{8,6}
ID 9	8	{6,8}
ID 10	14	{23}
ID 11	12	{10}
ID 12	5	{25,20,36}

28 distinct activities have been observed against Sensor 6 only.

Results (3): A_Env & A_Geo

	1	2	3	4	5	6	7	8	9	10	11	12
1	0	0	0	0	0	4	0	0	0	0	0	1
2	0	0	0	0	0	0	0	0	0	0	1	1
3	0	7	0	1	0	0	0	0	0	0	0	0
4	0	7	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	2	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	0	0

7 distinct activities coming from YU Sources only have targeted the sole Sensor 6.

Results (4): A_SAX

- Symbolic Aggregate approXimation (SAX)
- Alphabet size=5 , Compression Ratio=8

Dominant Set ID	# Clusters
ID 1	9
ID 2	5
ID 3	7
ID 4	4
ID 5	5
ID 6	3
ID 7	4
ID 8	3
ID 9	3
ID 38	3

Clique ID	Ports Lists
ID 1	{80},{139}
ID 2	{139},{1433}
ID 3	{1434_udp},{445, 135}
ID 4	{1433},{1434_udp},{445, 135}
ID 8	{80},{22}
ID 9	{9898},{5554},{5554, 9898}

Intersection A_SAX	# Common Clusters	% initial clusters
with $A\_commonIPs$	7	6.1%
with $A\_Hostnames$	35	30.7%
with A_OSs	102	86.5%

[PUD, RR-05]

Correlative Analysis: summary

- We obtain all dominant sets for all similarity combined matrices we have developed
- All groups are interesting case studies
- Each cluster is labeled according to the sets identifiers it belongs to
- Reasoning based on the association and non-association of clusters within sets
- Potential validation by means of Telescopes

CLUSTER ID:

1931

IDENTIFICATION:

FINGERPRINT:

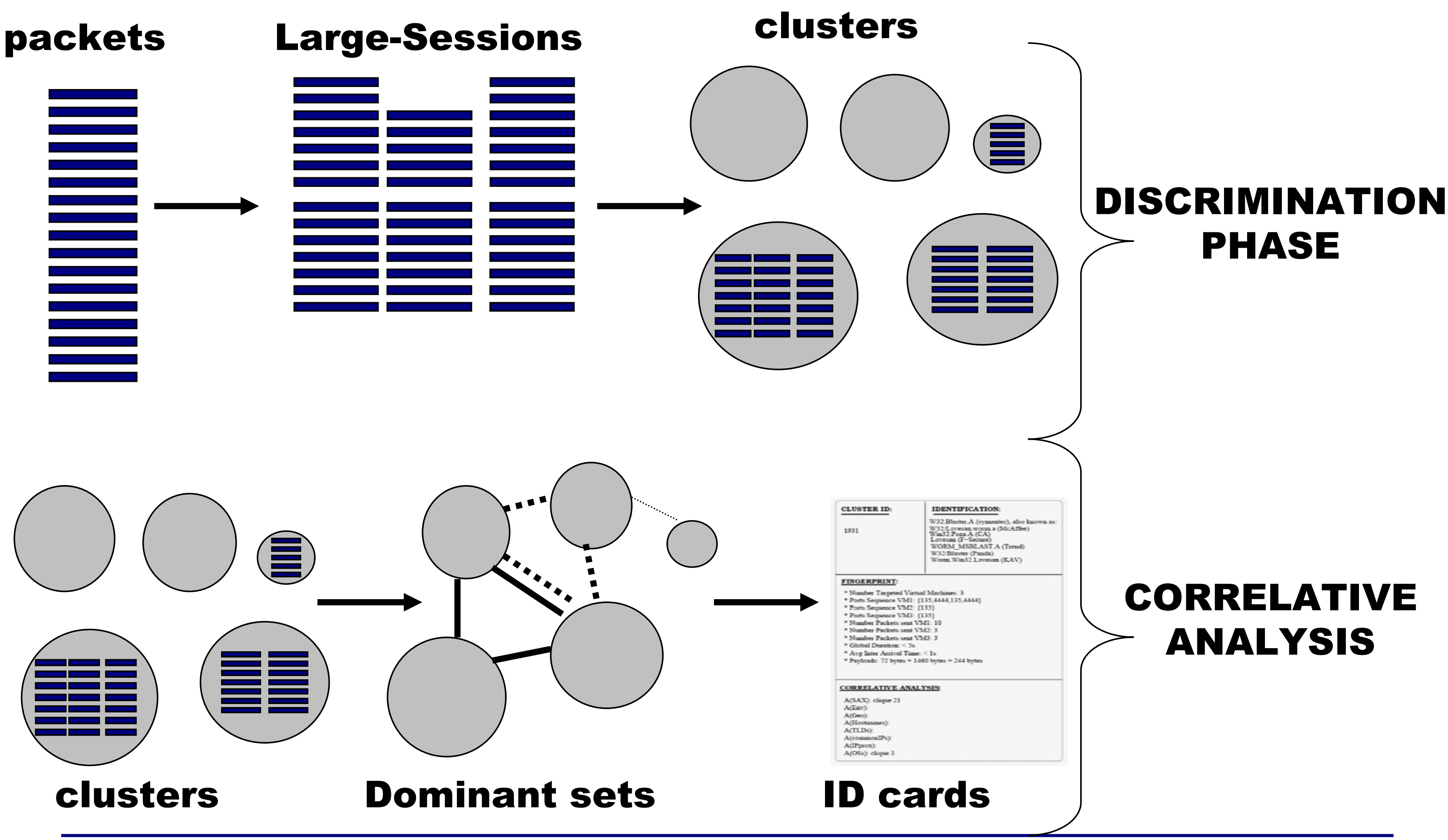
- Number Targeted Machines: 3
- Ports Sequence VM1: {135,4444}
- Ports Sequence VM2: {135}
- Ports Sequence VM3: {135}
- Number Packets sent to VM1: 10
- Number Packets sent to VM2: 3
- Number Packets sent to VM3: 3
- Global Duration: < 5s
- Avg Inter Arrival Time: < 1s
- Payloads:

72 bytes + 1460 bytes + 244 bytes

CORRELATIVE ANALYSIS:

A(SAX): DS 21
A(Env):
A(Geo):
A(Hostnames):
A(TLDs):
A(commonIPs):
A(IPprox):
A(OSs): DS 3

HoRaSis: Brief Summary



Conclusions (1)

We have demonstrated that it is possible to build a framework which helps better identifying and understanding of malicious activities in the Internet.

1. By collecting data from simple honeypot sensors (few IPs) placed in various locations.

2. By building a technique adapted to this data in order to automate knowledge discovery.

Conclusions (2)

Help feeding the WOMBAT!!

References

- More information on the French ACI Security available at acisi.loria.fr
- Exhaustive and up to date list of publications available at

http://www.leurrecom.org

- F. Pouget, M. Dacier, V.H. Pham, **Leurre.Com: On the Advantages of Deploying a Large Scale Distributed Honeypot Platform**. Proc. Of the E-Crime and Computer Conference 2005. ECCE'05), Monaco, March 2005.
- F. Pouget, M. Dacier, H. Debar, V.H. Pham, **Honeynets: Foundations For the Development of Early Warning Information Systems**. NATO Advanced Research Workshop, Gdansk 2004. Cyberspace Security and Defense: Research Issues. Publisher Springler-Verlag, LNCS, NATO ARW Series, 2005.
- E. Alata, M. Dacier, Y. Deswarte, M. Kaaniche, K. Kortchinsky, V. Nicomette, V.H. Pham, F. Pouget, **CADHo: Collection and Analysis of Data from Honeypots**. In Proc. Of the Fifth European Dependable Computing Conference. (EDCC-5), Budapest, Hungary, April 2005.
- F. Pouget, T. Holz, **A Pointillist Approach for Comparing Honeypots**. Proc. Of the Conference on Detection of Intrusions and Malware & Vulnerability Assessment. (DIMVA 2005), Vienna, Austria, July 2005.
- J. Zimmermann, A. Clark, G. Mohay, F. Pouget, M. Dacier, **The Use of Packet Inter-Arrival Times for Investigating Unsolicited Internet Traffic**. In Proc. Of the First International Workshop on Sytematic Approaches to Digital Forensic Engineering. (SADFE'05), Taipei, Taiwan, November 2005.
- P.T. Chen, C.S. Laih, F. Pouget, M. Dacier, **Comparative Survey of Local Honeypot Sensors to Assist Network Forensics**. In Proc. Of the First International Workshop on Sytematic Approaches to Digital Forensic Engineering. (SADFE'05), Taipei, Taiwan, November 2005.

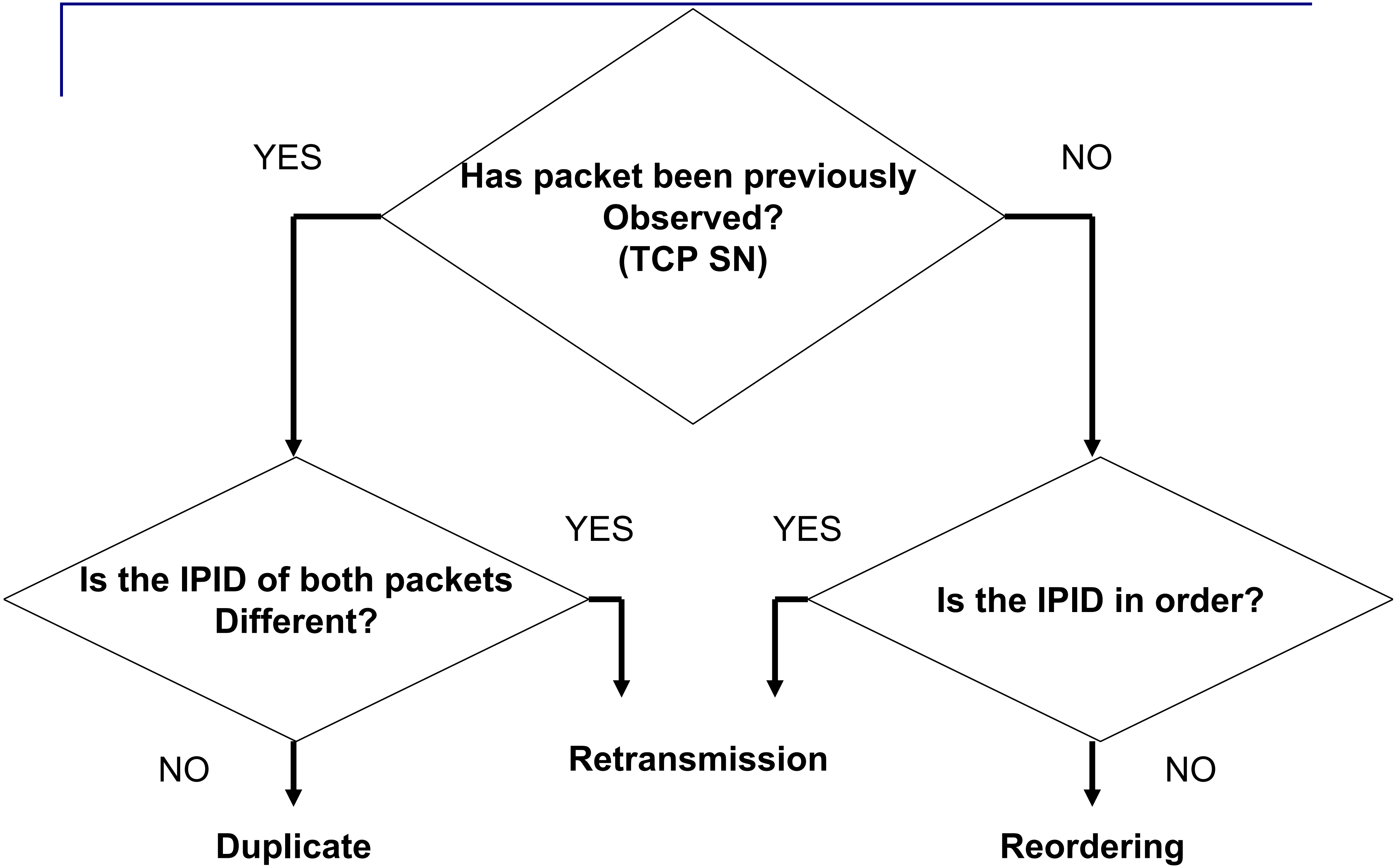
Removing Network Influences

- Examples:
  - Duplicates, retransmission, losses, delays, jitter, reordering,etc
- Network and transport layers can address these phenomena…
- … which can also be part of an attack process
- Hard to discriminate both cases

Solution: **[PUD, RR-05]**

Exploit the IP Identifier implementation (RFC 791)

We have addressed this way the following influences:


Has packet been previously Observed? (TCP SN)
YES
NO
Is the IPID of both packets Different?
Is the IPID in order?
YES
YES
NO
NO
Retransmission
Duplicate
Reordering