Adding rigor to the comparison of anomaly detector outputs

Romain Fontugne, National Institute of Informatics / SOKENDAI, Tokyo
Pierre Borgnat, Physics Lab, CNRS, ENS Lyon
Patrice Abry, Physics Lab, CNRS, ENS Lyon
Kensuke Fukuda, National Institute of Informatics / PRESTO JST, Tokyo

April 25, 2010

Motivation

Anomaly detection in backbone traffic

- Active research domain
  - Wavelet [IMC 02], PCA [SIGCOMM 05, SIGMETRICS 07], gamma law [LSAD 07], association rule [IMC 09]...
- Tricky evaluation, lack of common ground truth:
  - Manual inspection
  - Synthetic traffic
  - Comparison with other methods

Similar problems arise in traffic classification

Goal

Long term goal: Provide common “ground truth data”

- Labeling MAWI archive
- Combining several anomaly detector results
- Ground truth relative to the state of the art

Goal of this work: Find relations between outputs of different classifiers

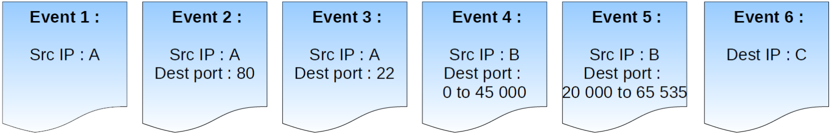
Problem statement: Eventx=Eventy??

Event (= anomaly detector's alarm)

Set of traffic feature containing at least 2 timestamps and one traffic feature.
i.e. one flow, one IP address, a set of flows, a set of packets...

Main difficulties

- Different granularities: Event1=Event2?=Event3?
- Overlapping: Event4=Event5?
- Different points of view: Event1=Event6?



Event 1 :
Src IP : A

Event 2 :
Src IP : A
Dest port : 80

Event 3 :
Src IP : A
Dest port : 22

Event 4 :
Src IP : B
Dest port :
0 to 45 000

Event 5 :
Src IP : B
Dest port :
20 000 to 65 535

Event 6 :
Dest IP : C

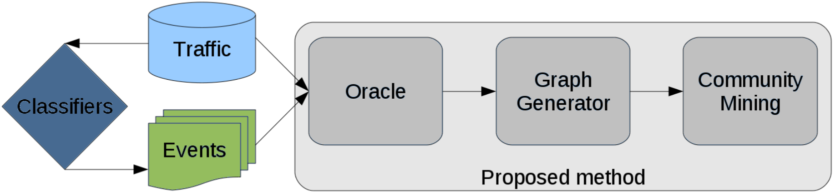
Proposed method

Approach

Identify similar events by using community mining on graph

Overview

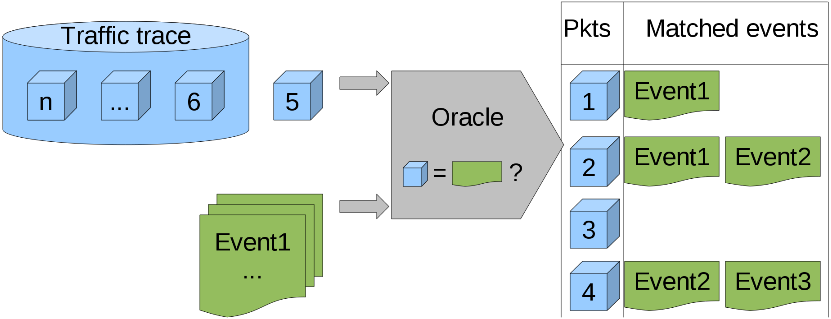
- Oracle: Uncover relations between traffic and events
- Graph gen.: Represent events and their relations in a graph
- Community Mining: Find similar events by looking at dense components



Oracle

Uncover relations between original traffic and events

- List the events that match each packet of the original traffic
- i.e. pkt1:$\{IP1 : 80 \rightarrow IP2 : 12345\} = $ Event1:$\{srcIP = IP1\}$

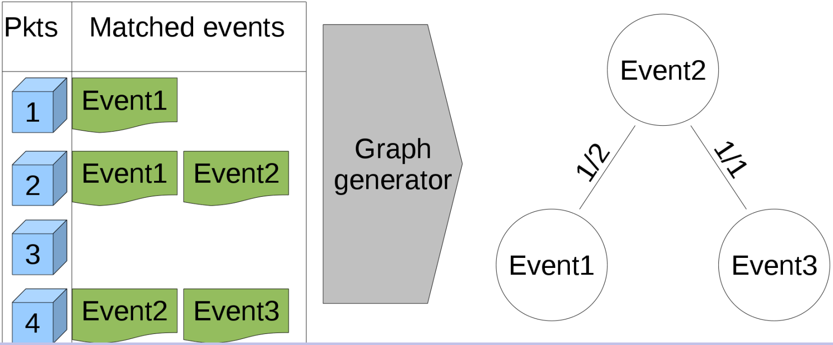


Pkts	Matched events	
1	Event1	
2	Event1	Event2
3		
4	Event2	Event3

Graph generator

Build a non-directed weighted graph from the Oracle output

- Nodes are events and edges are shared packets
- Weight on each edge: similarity measure, Simpson index, $|E_1 \cap E_2| / \min(|E_1|, |E_2|)$, E_i: packets matching event i

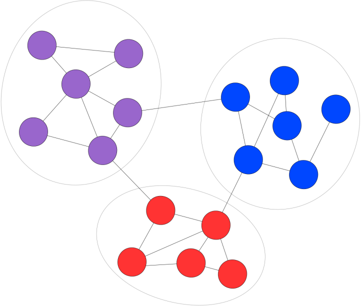


Pkts	Matched events
1	Event1
2	Event1 Event2
3	
4	Event2 Event3

Community mining

Identify community (= dense component) in the graph

- Louvain algorithm[1]: based on Modularity[2]
- Take into account node connectivity and edge weight



[1] Blondel et al.: Fast unfolding of communities in large networks. J.STAT.MECH. (2008)

[2] Newman, Girvan: Finding and evaluating community structure in networks. Phys. Rev.E (Feb 2004)

Data and anomaly detectors

Data set

- MAWI archive (trans-Pacific link)
- During the outbreak of the Sasser worm (08/2004)

Anomaly detectors

- Sketches and multiresolution gamma modeling [3]
  Report source or destination IP
- Image processing: Hough transform [4]
  Report set of packets

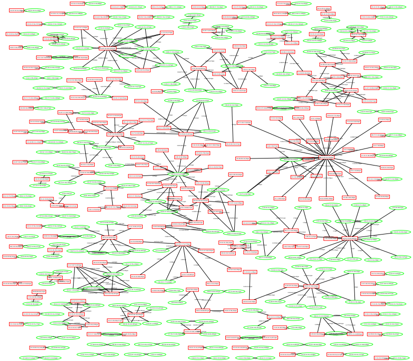
[3] Dewaele, G., Fukuda, K., Borgnat, P., Abry, P., Cho, K.: Extracting hidden anomalies using sketch and non gaussian multiresolution statistical detection procedures. SIGCOMM LSAD 07

[4] Fontugne, R., Himura, Y., Fukuda, K.: Evaluation of anomaly detection method based on pattern recognition. IEICE Trans. on Commun. E93-B(2) (February 2010)

Results

Graph

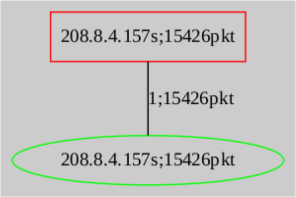
- Reported events; Gamma-based: 332, Hough-based: 873
- Intersection 235 and 247 events: 124 connected components
- Biggest component: 47 events (G.34, H.13), 8 communities



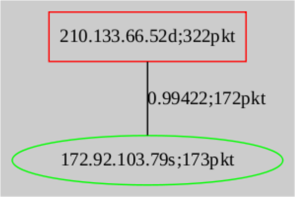
Simple connected components

Two event component

- 86 small components, mainly Sasser
- Gamma-based = red; Hough-based = green



(1) Sasser infected host.

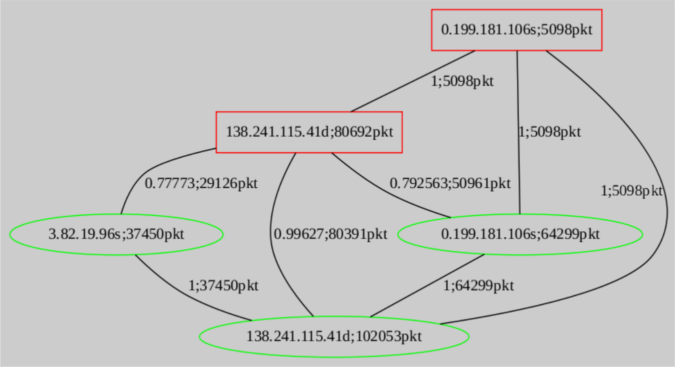


(2) Different src.IP and dest.IP.

Large connected components I

Large component with one community

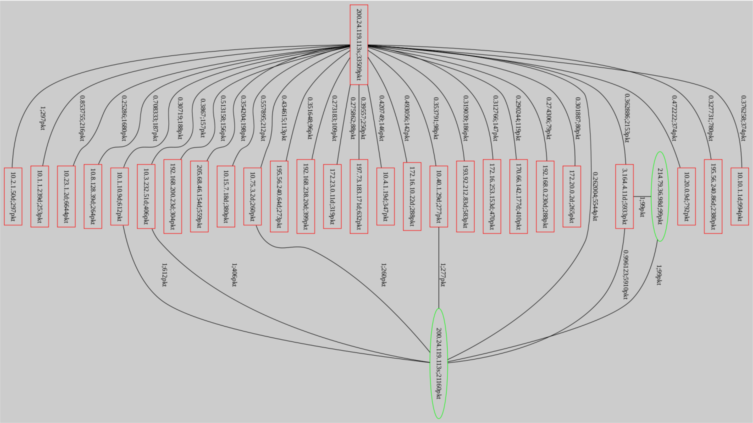
- 38 components having more than two events
- RSync traffic identified by 5 events



Large connected components II

DNS traffic

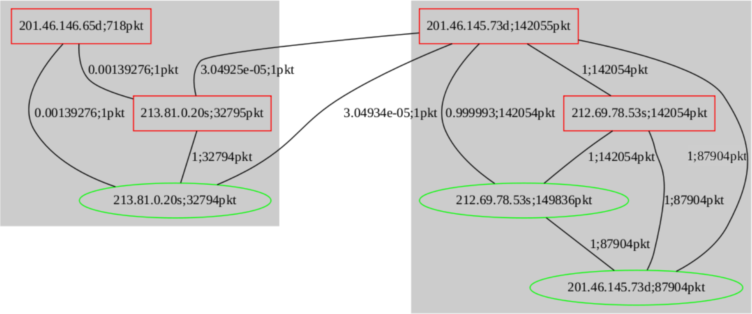
29 events in which 27 are from the gamma-based detector



Communities in components

Distinct traffics

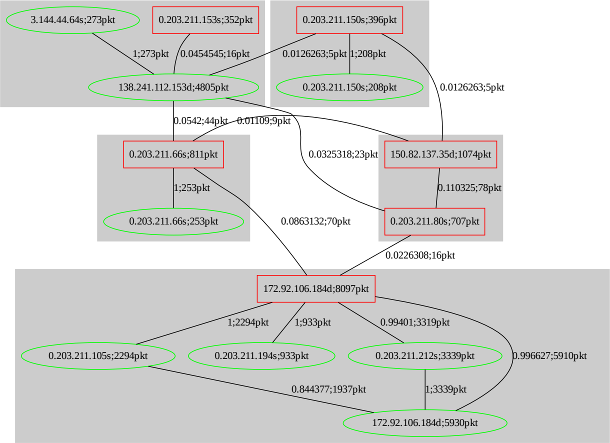
Network scan on port 3128 and nntp traffic



Communities in components

Same kind of traffic

14 events reporting HTTP traffic



Discussion

Advantages

- Uncover relations between classifier outputs
- Able to compare outputs of different kinds of classifiers

Applications

- Comparing/combining anomaly detectors
- Clarifying output of a single detector
- Understanding detector sensitivity to parameter tuning

Conclusion and future work

Conclusion

- Uncover relations between classifiers outputs
- Graph theory
- General and rigorous method

Future work

- Deeper analysis of the method
- Combining anomaly detectors
- Labelling MAWI

Thank you!

Questions?

romain@nii.ac.jp

[1]

Fontugne, R., Borgnat, P., Abry, P., Fukuda, K.:
Uncovering relations between traffic classifiers and anomaly detectors via graph theory.
TMA (2010) 101–114