Multidimensional Aggregation for DNS monitoring

Jérôme François, Lautaro Dolberg, Thomas Engel

jerome.francois@inria.fr

03/11/15

Outline

Outline

DNS monitoring

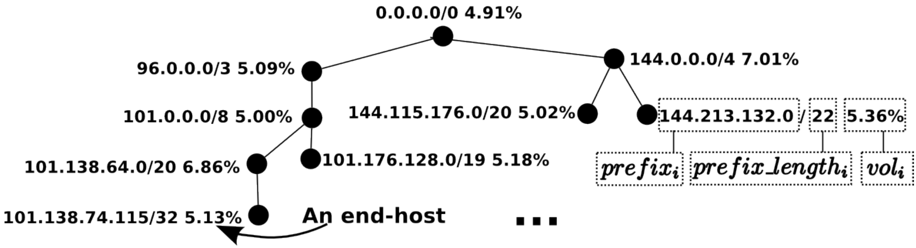
DNS traffic reflects host activities and behaviors

- Internet threats growing: Phishing, Malware, botnet, Spoofed Domains, data ex-filtration, etc.
- Identify malicious domains behavior by assessing associations between names and IP subnets (and how this evolves)
- Passive DNS analysis: easy to collect, reflect user activities without tracking individually them
- $\rightarrow$ from all collected DNS answers collected over multiple weeks, is it possible to detect divergent behaviors?

Outline

State of the art

- Spatio temporal aggregation:
  - Aguri QofIS 2001: subnetwork prefix based aggreagation
  - Danak NSS 2011: Aguri applied to anomaly detection
- TreeTop Usenix Sec 2010: DNS domain based aggregation



Aggregation

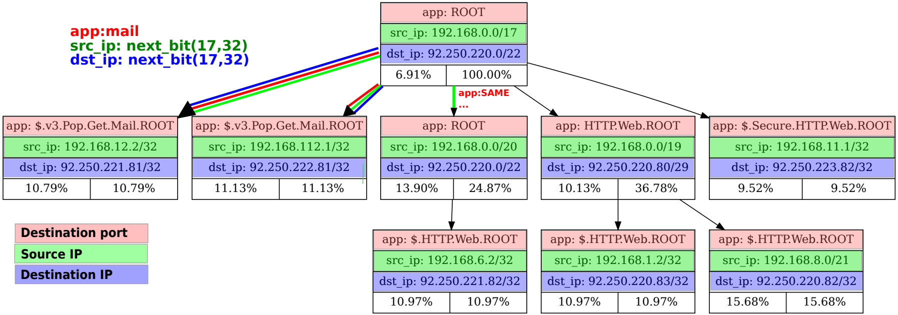
Aggregation

- Scalable way to represent information
  - Outline relevant correlated facts
  - reduce storage needs and post processing time
- Temporal and Spatial aggregation
  - temporal: time windows split (β)
  - spatial: keep nodes with activity $> \alpha$ *e.g. traffic volume*, aggregate the others into their parents $\rightarrow$ needs hierarchical relationships
- Heterogeneous Data
  - No specific order
    - ~~1st Source IP@, 2nd Destination IP@~~
  - Auto adjust to Information Granularity
    - ~~/18 /24 /27 subnetworks...~~

Mutidimensional Aggregation Example

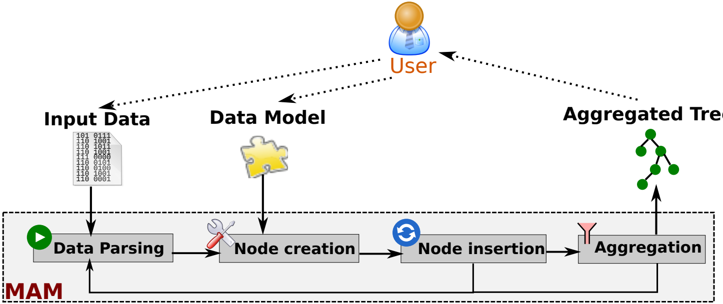
PORT	PROTO	KB	TIME	SOURCE	DEST
80	TCP	1491	2010−02−24 02:20:15	192.168.6.2	92.250.221.82
110	TCP	988	2010−02−24 02:20:19	192.168.8.2	92.250.223.87
443	TCP	902	2010−02−24 02:20:27	192.168.11.2	92.250.220.82
110	TCP	1513	2010−02−24 02:20:29	192.168.112.1	92.250.222.81
80	TCP	1205	2010−02−24 02:20:29	192.168.11.1	92.250.220.82
80	TCP	1491	2010−02−24 02:20:31	192.168.1.2	92.250.220.83
110	TCP	1467	2010−02−24 02:20:39	192.168.12.2	92.250.221.81
80	TCP	927	2010−02−24 02:20:39	192.168.12.2	92.250.220.82
443	TCP	1294	2010−02−24 02:20:39	192.168.11.1	92.250.223.82
110	TCP	940	2010−02−24 02:20:49	192.168.21.2	92.250.221.81
80	TCP	917	2010−02−24 02:20:49	192.168.23.1	92.250.220.82
443	TCP	460	2010−02−24 02:20:59	192.168.26.2	92.250.220.85

Mutidimensional Aggregation Example



Outline

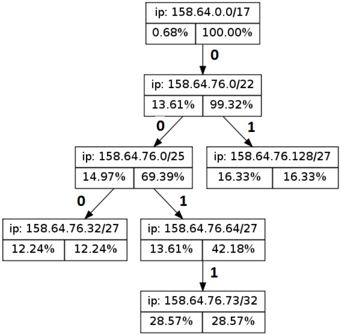
Data processing cycle

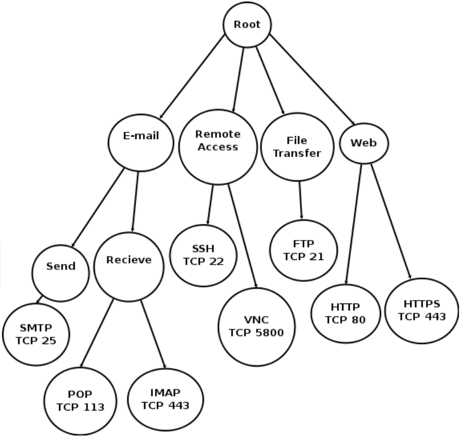


- Nodes constructed based on input data and continuously included in the tree
- Aggregation: at the final step *vs.* when the tree size is too large

Data Model

Underlying Data Model





Data Structure

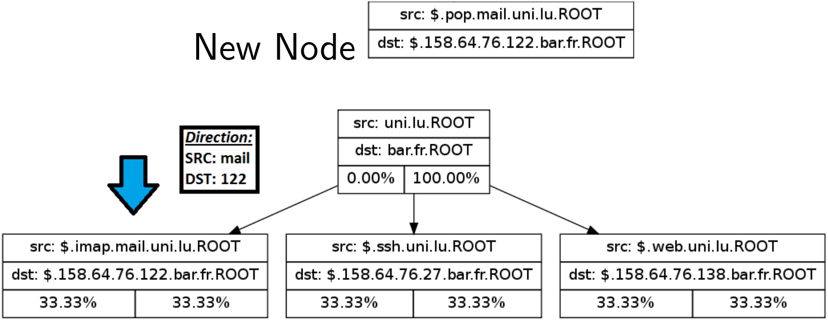
Tree based structure: Root node and multiple children

Directions

- How to find the right path to insert a node within a tree?
- Every hierarchical data can be implemented (MaM can be easily extended)
  - common ancestor between two nodes
  - direction function
- IP@ binary function (0,1) as next bit value
- DNS: every level name is a direction
- ports: service taxonomy

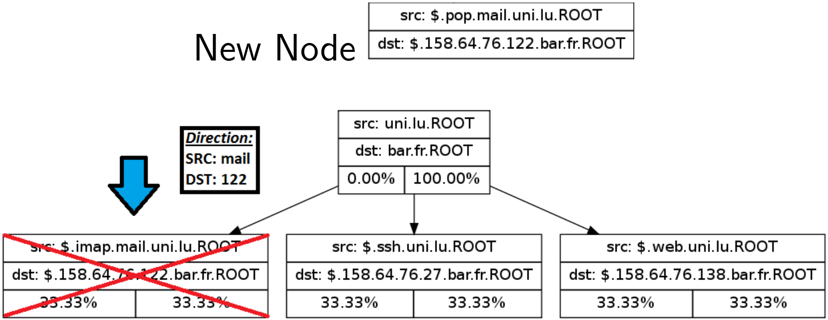
Data Structure

Node Insertion (Branching Point)



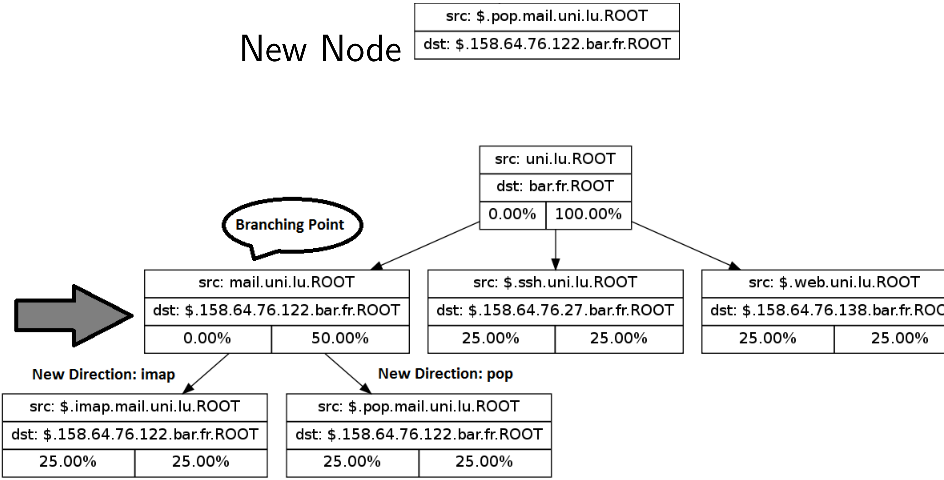
Data Structure

Node Insertion (Branching Point)



Data Structure

Node Insertion (Branching Point)



Optimization

Aggregation

- From leafs to root node
- On a complete tree of a time window
- $\rightarrow$ Large data structures in memory before aggregation

Online Strategies (before the end of the time window)

- Tree size $>$ MAX_NODES $\rightarrow$ aggregation

	Root	LRU
	Aggregation is triggered from root node	Aggregation is triggered in the least recently used node
RAM	+	+
Performance	- -	-

Outline

Applications

- Output of MaM = sequence of trees
- $\rightarrow$ monitoring the network using these trees
  - trees are well known data structure $\rightarrow$ distance metrics, kernel functions, homomorphisms,...
  - manual vs automated analysis
  - visual inspection

User inputs

- Data + parsing function
- List of attributes to extract + dimensions
- (definition of dimensions if not supported by default)
- parameters: aggregation threshold (α), time window size (β), max nodes (2000), strategy (LRU)
- $\rightarrow$ monitoring the network using these trees

Outline

Contributions

Malicious domains names are usually changing IP association. How can this be exploited?

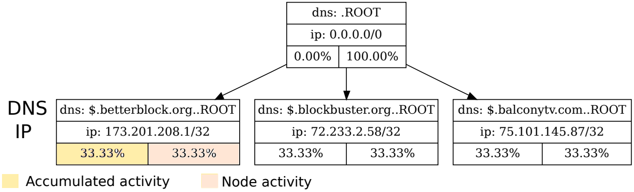
- Large Scale Aggregation: DNS and IP addresses, into single data structure.
- Steadiness Metrics: Formal measure of DNS and Subnetwork address association over time.
- Metric Validation: Long term experiments using Passive DNS Database.

Data sample

DATE	NAME	IP ADDRESS	TLD	TTL	TYPE
2012-07-07	twistedblood.co.uk	72.233.2.58	uk	20691609.0	A
2012-07-07	besttraintravel.com	69.43.161.181	com	1e-18	A
2012-07-07	besttraintravel.com	82.98.86.167	com	84428.0	A
2012-07-07	thedigitour.com	67.195.140.36	com	14161531.0	A
2012-07-07	thedigitour.com	67.195.145.141	com	6557703.0	A
2012-07-07	thedigitour.com	98.138.19.88	com	1158108.0	A
2012-07-07	thedigitour.com	98.139.135.21	com	17369531.0	A
2012-07-07	thegcblog.com	72.233.2.58	com	24044547.0	A
2012-07-07	equestriadaily.com	216.239.32.21	com	32253581.0	A
2012-07-07	livehoods.org	75.101.145.87	org	1e-18	A

With MAM is possible to generate aggregated views combining multiple dimensions at the same time.

- Hierarchically derived from data model
- Provides different levels of granularity
- Accelerates Post processing



Experiments & Data set

The objectives of the experiments are:

- Discriminate between malicious and normal domains
- Attack detection ability
- Performance decay

Experiments & Data set

The objectives of the experiments are:

- Discriminate between malicious and normal domains
- Attack detection ability
- Performance decay

Passive DNS + Blacklist

	Domains	IP Address
Name Servers	661968	164559
Blacklist	173066	174619
Total	835034	339178

Monitoring

Logs to Time Series of Trees

- An aggregation process outputs a series of trees
- Monitoring aggregated series of trees
- i.e $T_1 \dots T_m$

Metrics $\rightarrow$ *correlate*

- IP subnets
- Domain names
- Volume of Traffic

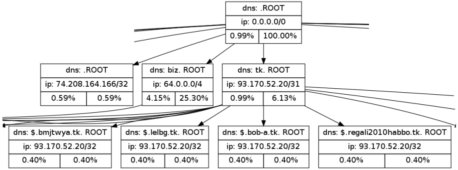
Monitoring

Logs to Time Series of Trees

- An aggregation process outputs a series of trees
- Monitoring aggregated series of trees
- i.e $T_1 \dots T_m$

Metrics $\rightarrow$ *correlate*

- IP subnets
- Domain names
- Volume of Traffic



Metrics I

Tree comparison, how to establish a similarity criteria?

Metrics I

Tree comparison, how to establish a similarity criteria?

$$sim(n1, n2) = \alpha \times IP\_sim(n1, n2) + \beta \times DNS\_sim(n1, n2) + \gamma \times vol\_sim(n1, n2)$$

Metrics I

Tree comparison, how to establish a similarity criteria?

$$sim(n1, n2) = \alpha \times IP\_sim(n1, n2) + \beta \times DNS\_sim(n1, n2) + \gamma \times vol\_sim(n1, n2)$$

$$IP\_sim(n1, n2) = 1 - \frac{|n1_{prefix\_len} - n2_{prefix\_len}|}{32}$$

Metrics I

Tree comparison, how to establish a similarity criteria?

$$sim(n1, n2) = \alpha \times IP\_sim(n1, n2) + \beta \times DNS\_sim(n1, n2) + \gamma \times vol\_sim(n1, n2)$$

$$IP\_sim(n1, n2) = 1 - \frac{|n1_{prefix\_len} - n2_{prefix\_len}|}{32}$$

$$DNS\_sim(n1, n2) = \frac{|n1_{dns} \cap n2.dns|}{|n1_{dns} \cup n2_{dns}|}$$

Metrics I

Tree comparison, how to establish a similarity criteria?

$$sim(n1, n2) = \alpha \times IP\_sim(n1, n2) + \beta \times DNS\_sim(n1, n2) + \gamma \times vol\_sim(n1, n2)$$

$$IP\_sim(n1, n2) = 1 - \frac{|n1_{prefix\_len} - n2_{prefix\_len}|}{32}$$

$$DNS\_sim(n1, n2) = \frac{|n1_{dns} \cap n2.dns|}{|n1_{dns} \cup n2_{dns}|}$$

$$vol\_sim(n1, n2) = 1 - 0.01 \times |n1_{acc\_vol} - n2_{acc\_vol}|$$

Metrics II

- Two goals at different levels
  1. Detecting the presence of an anomaly in the traffic:
     - *sim* metric is between two nodes $\rightarrow$ maximise this metric for each node

$$\begin{aligned} & n1 \in T_i, n2 \in T_{i-1}, n2 = most\_sim(n1) \\ & stead(n1) = sim(n1, n2) + \mu \times stead(n2) \\ & pers(T_i) = \frac{\sum_{n \in T_i} stead(n)}{|\{n \in T_i\}|} \end{aligned} \tag{1}$$

  2. Identifying the anomaly, i.e. the domains and IP addresses $\rightarrow$ look for nodes with the smallest *stead* values

Experiments

Aggregation Window

Time Length

- Macro: Up to 52 weeks
- Micro: 10 weeks maximum

Malicious data

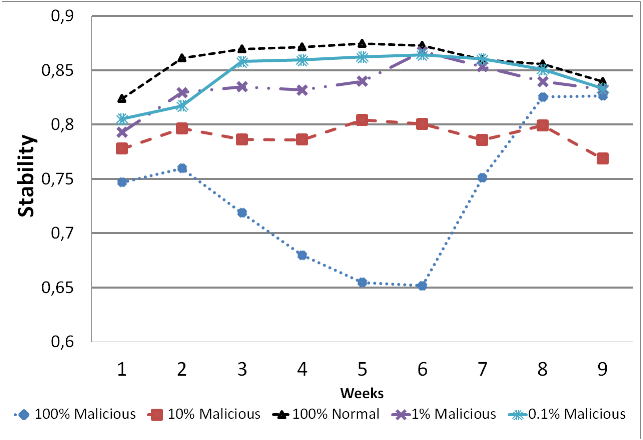
- Time: Periodically, Steady
- Proportion

Aggregation Granularity

Outline

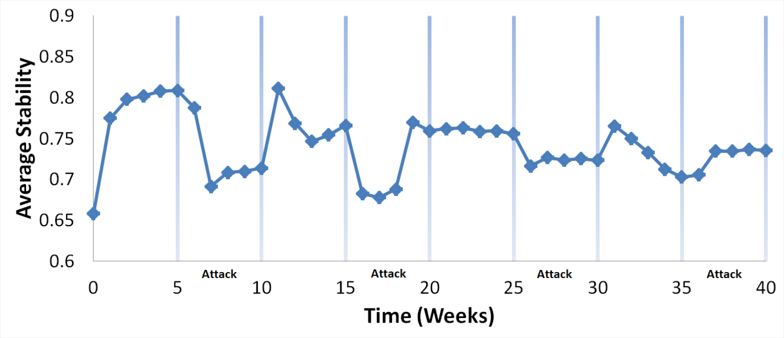
Results

Malicious domains causes a drop on average steadiness: Micro



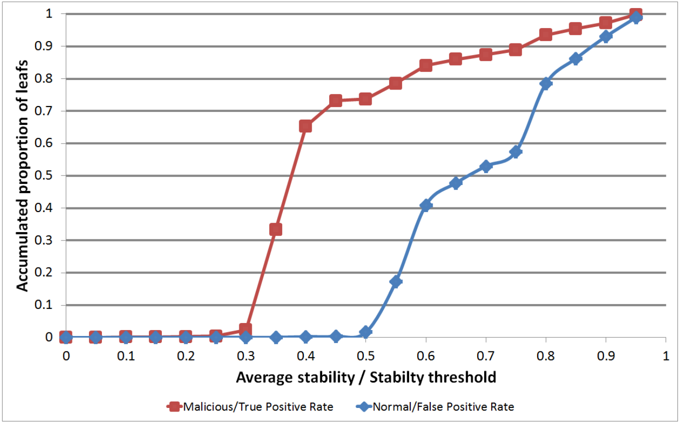
Results

Malicious domains causes a drop on average steadiness: Macro



Results II

Accuracy: Steadiness as metric for filtering malicious domains



Outline

MAM extensions

- define any hierarchical dimension
- successfully applied to different domains: vehicular networks, Netflow monitoring
- again MAM is only producing trees = aggregation
  - metrics / feature engineering
  - methods / machine learning
- but data to handle are squeezed to a smaller scale

Performances

- Number of nodes
  - main performance parameter when computing metrics
  - depends on the aggregation threshold (α) = minimum of activity to not be aggregated
- DNS monitoring
  - $\alpha = 2\%$
  - avg. = 2200 nodes / weekly tree
  - 13000 IP addresses / week
  - 5300 domain names / week

Other use case

- Dataset from major ISP in Luxembourg
  - Capture: 26 Days, 60,000 flows/sec at peak hours
  - IP Address: 279815 unique IP addesses using 64470 different UDP and TCP Ports
  - Extracting: Timestamp, IP Source and Destination Addresses, TCP/UDP source and destination ports, traffic Volume in bytes
- Anomaly detection
  - Raw output
  - Visually enhanced output
  - Automated analysis

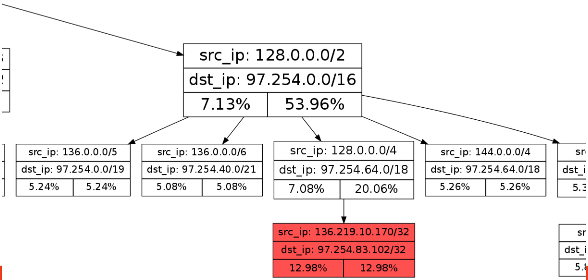
Raw output

- Trees as text with indentation

```
[src_ip-->0.0.0.0/0 dst_ip-->0.0.0.0/0 ] 92 (0.19% / 100.00%)
    [src_ip-->0.0.0.0/1 dst_ip-->0.0.0.0/1 ] 3104 (6.34% / 19.30%)
       [src_ip-->32.0.0.0/3 dst_ip-->96.0.0.0/3 ] 3868 (7.91% / 12.95%)
          [src_ip-->43.160.0.0/11 dst_ip-->120.194.118.20/32 ] 2470 (5.05% /
5.05%)
    [src_ip-->97.254.47.254/32 dst_ip-->138.146.47.197/32 ] 3581 (7.32% / 7.32%)

    [src_ip-->128.0.0.0/1 dst_ip-->0.0.0.0/1 ] 4182 (8.55% / 47.08%)
       [src_ip-->128.0.0.0/3 dst_ip-->97.254.0.0/16 ] 3734 (7.63% / 19.32%)
          [src_ip-->128.0.0.0/4 dst_ip-->97.254.64.0/18 ] 3012 (6.16% / 6.16%)

          [src_ip-->137.57.71.255/32 dst_ip-->97.254.131.93/32 ] 2706 (5.53% /
5.53%)
       [src_ip-->128.0.0.0/2 dst_ip-->0.0.0.0/1 ] 3223 (6.59% / 19.22%)
          [src_ip-->135.251.160.3/32 dst_ip-->97.254.23.33/32 ] 3438 (7.03% /
7.03%)
          [src_ip-->128.0.0.0/5 dst_ip-->97.254.128.0/21 ] 2740 (5.60% / 5.60%)

    [src_ip-->0.0.0.0/0 dst_ip-->0.0.0.0/1 ] 2504 (5.12% / 26.11%)
       [src_ip-->138.146.47.197/32 dst_ip-->97.254.47.254/32 ] 7030 (14.37% /
14.37%)
```

Visually enhanced output

- pictures (integrated in GUI)
- improvement
  - node size: importance of the represented attributes (feature space usage)
  - node color: instability of the represented attributes ($\sim$ new events)
  - needs to be user-defined $\rightarrow$ semantics can be freely chosen



Outline

Conclusion

- MaM
  - Scalable aggregation of heterogeneous data
  - Easily extensible to new features (geolocated IP flows, vehicular networks
- DNS monitoring
  - MaM only performs aggregation
  - Needs to define: hierarchical order, metrics and methods to analyze
- References
  - General description + theoretical foundations + network traffic monitoring
    - Dolberg L., François J., Engel T., Efficient Multidimensional Aggregation for Large Scale Monitoring, USENIX LISA 2012
  - DNS trafic monitoring
    - Dolberg L., François J., Engel T., Multi-dimensional Aggregation for DNS Monitoring, to appear in IEEE LCN 2013

Multidimensional Aggregation for DNS monitoring

Jérôme François, Lautaro Dolberg, Thomas Engel

jerome.francois@inria.fr

03/11/15