MaM: Multidimensional aggregation Monitoring

Lautaro Dolberg, Jérôme François, and Thomas Engel

UNIVERSITÉ DU LUXEMBOURG

SnT

NETLAB

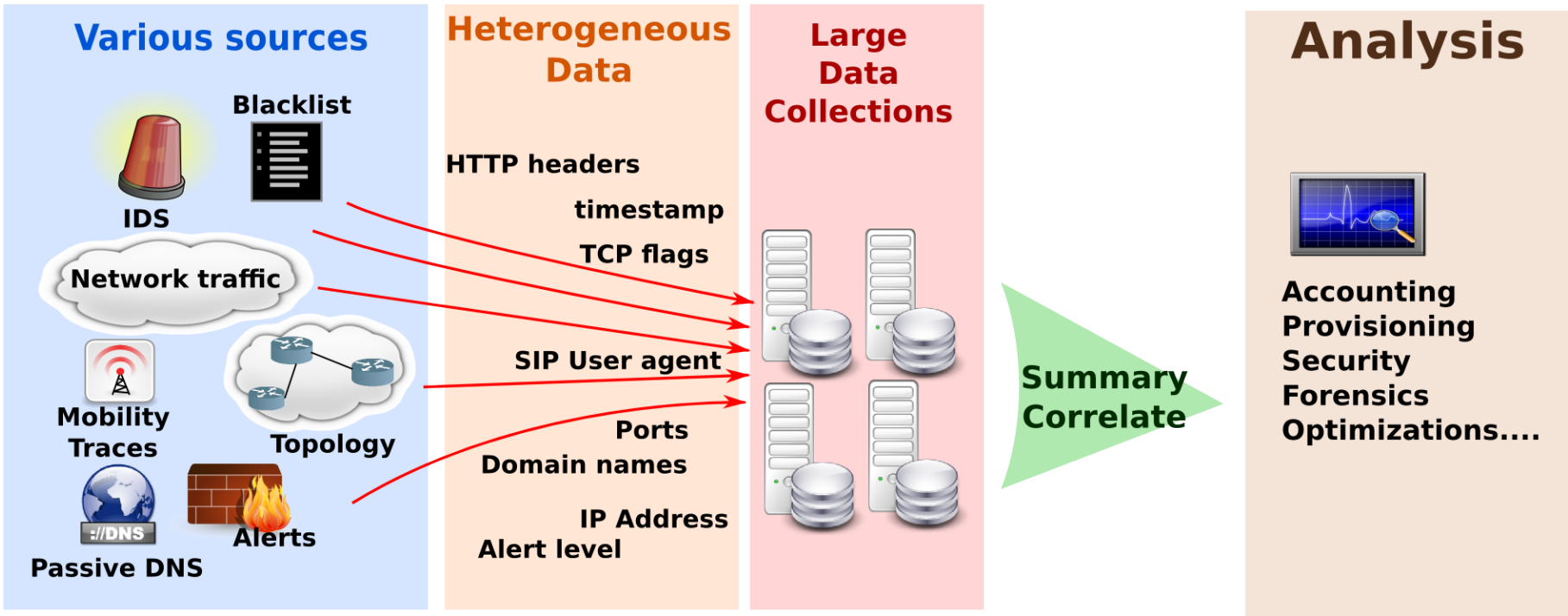
SNT

Outline

SMT

Outline

Motivation



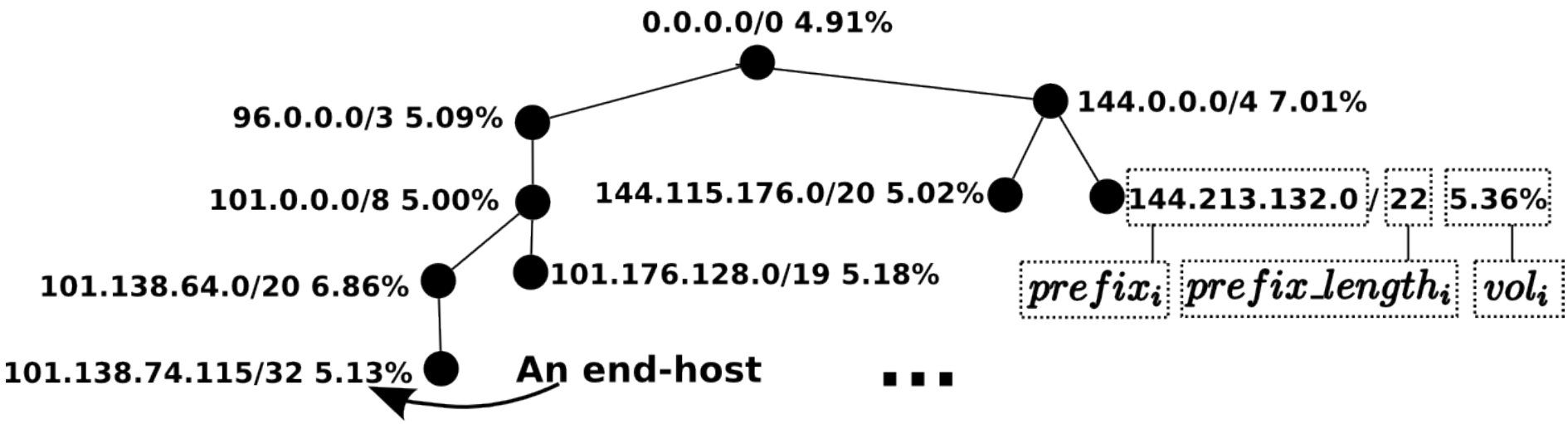
- How to combine large volume of heterogneous data in order to extract summarized relevant data?

SMT

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



SNT

Aggregation

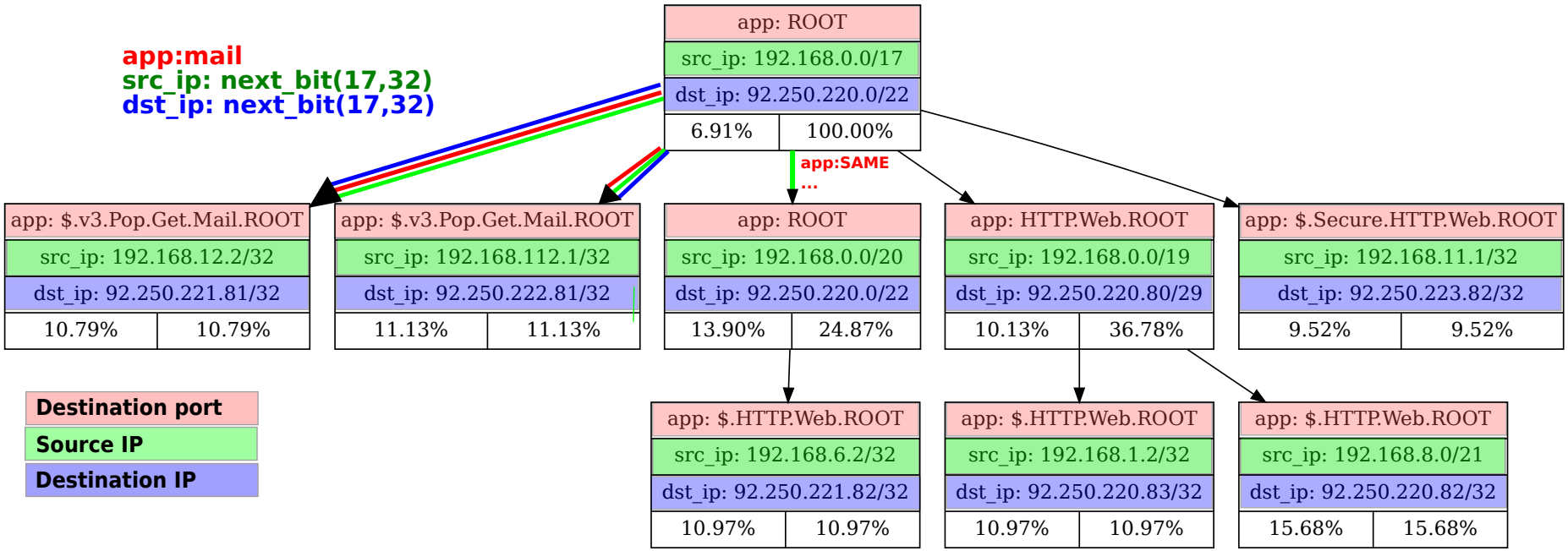
Aggregation

- 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

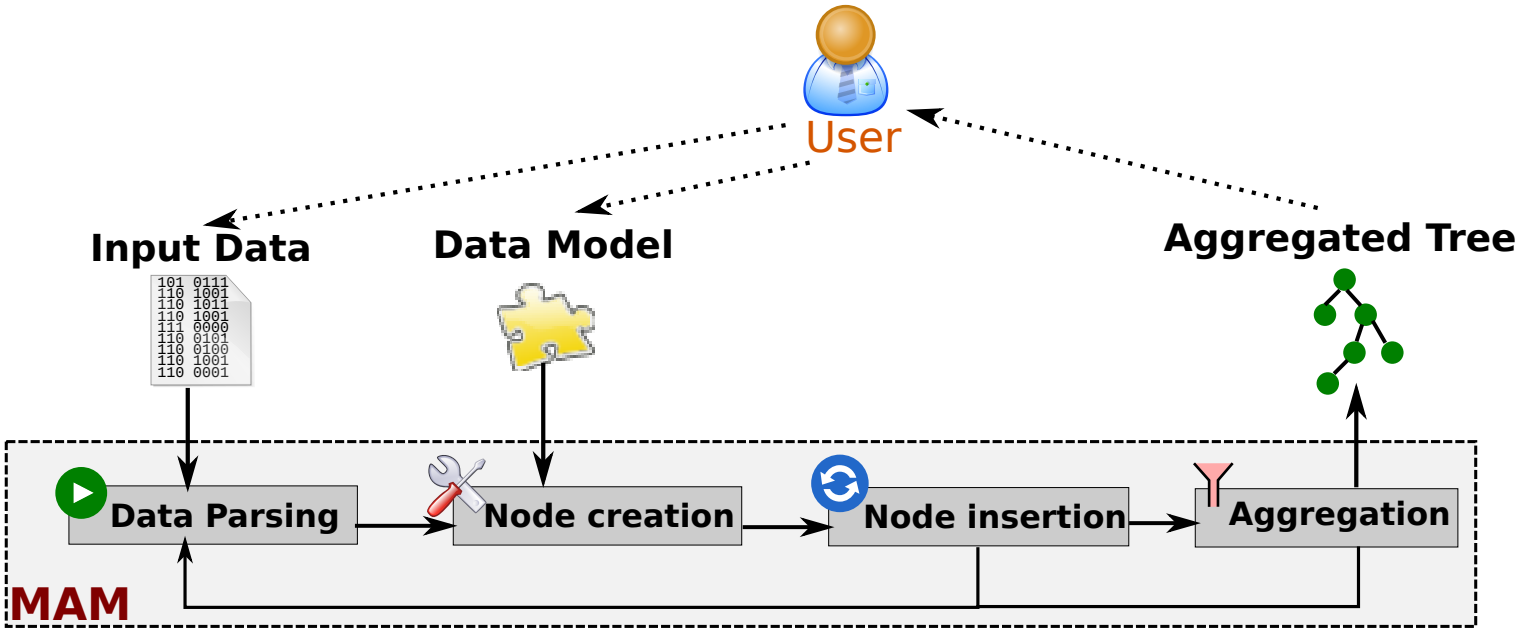
Mutidimensional Aggregation Example



SNT

Outline

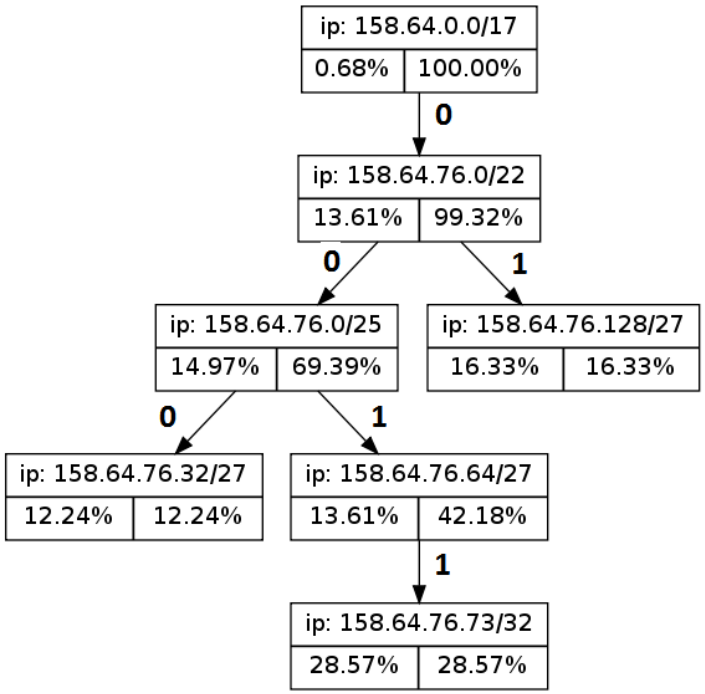
Data processing cycle

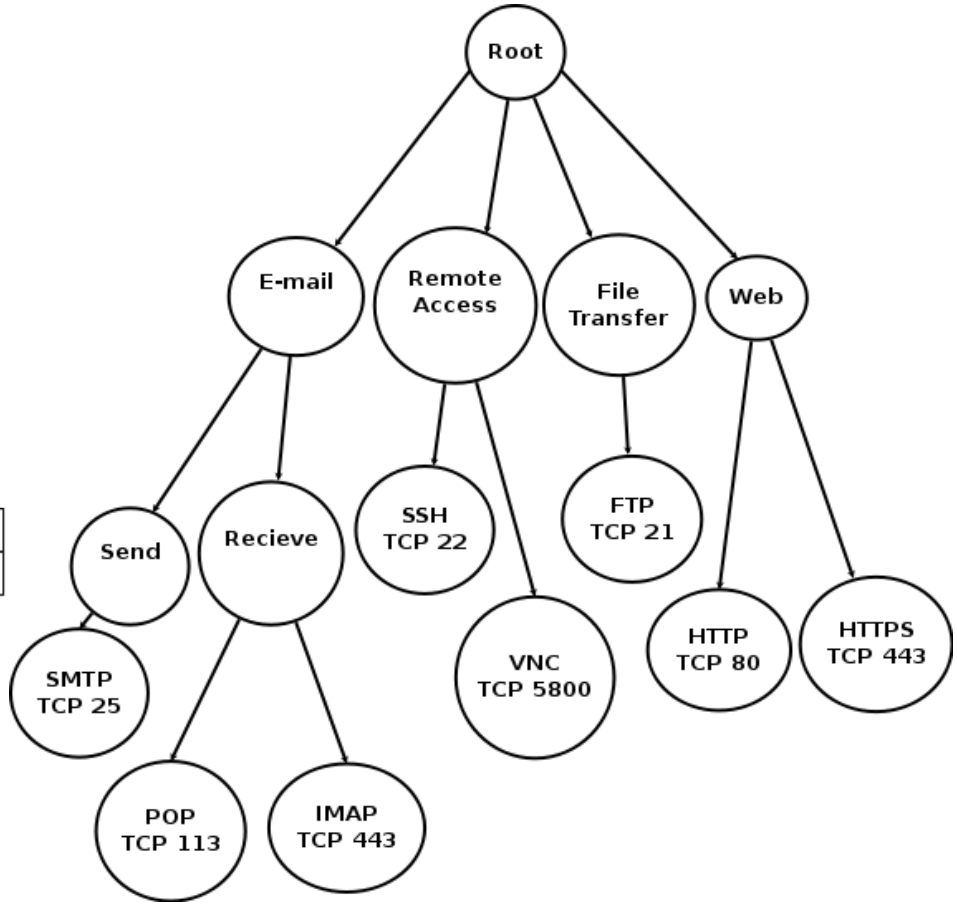


- Nodes constructed based on input data and continuously included in the tree
- Aggregation: at the final step *vs.* when the tree size is too large

Data Model

Underlying Data Model





SMT

Data Structure

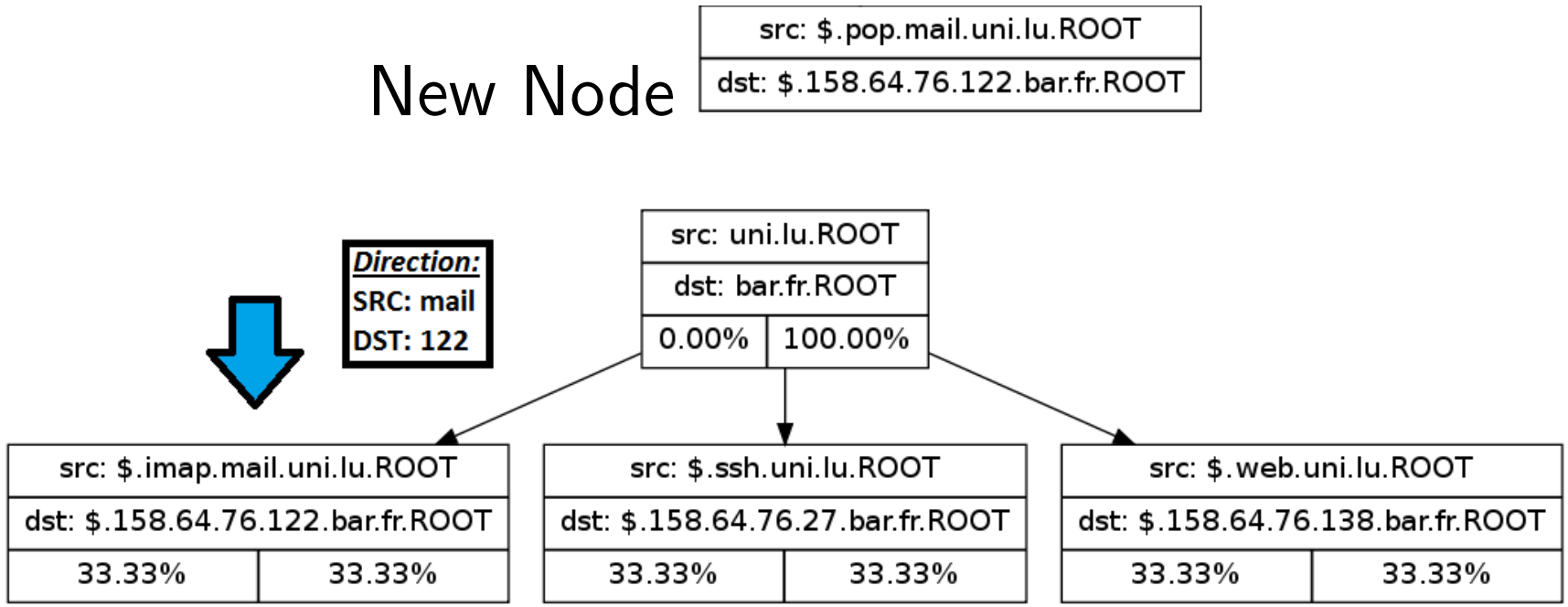
Tree based structure: Root node and multiple children

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

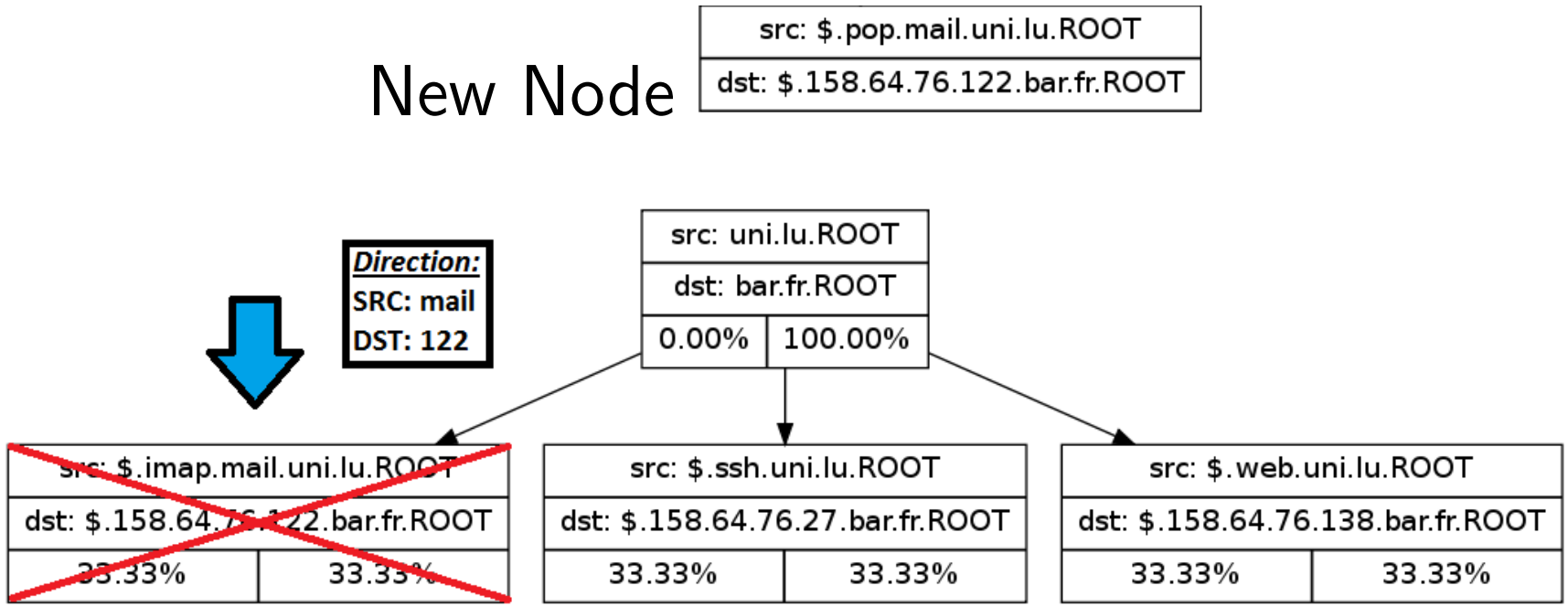
Data Structure

Node Insertion (Branching Point)



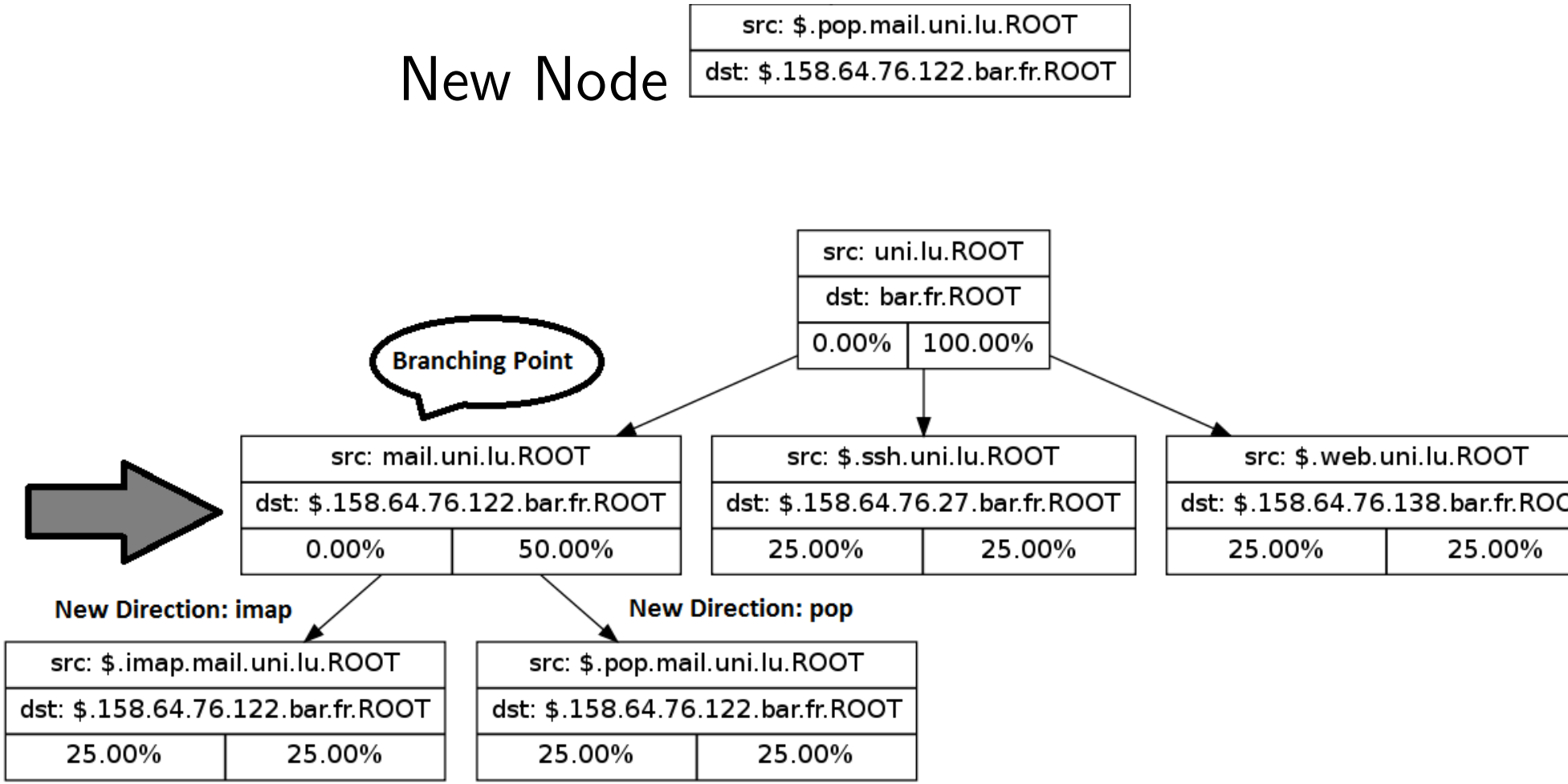
Data Structure

Node Insertion (Branching Point)



Data Structure

Node Insertion (Branching Point)



SNT

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	- -	-

SMT

Outline

SnT

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

SNT

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

Netflow monitoring with MaM

- 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

SNT

Raw output

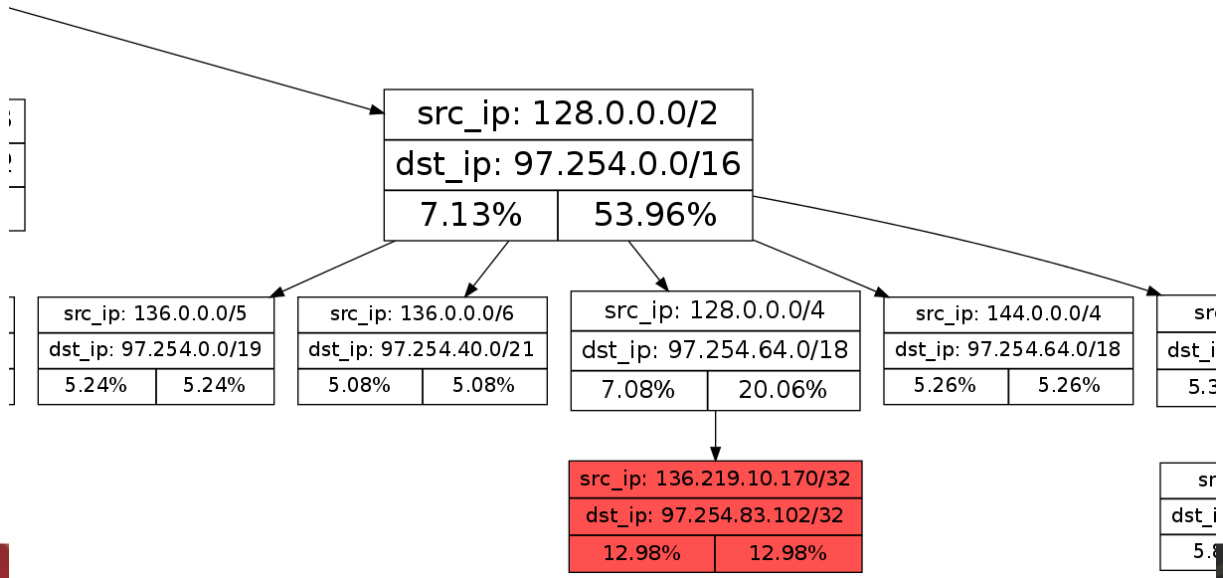
- Trees as text with indentation

```
[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%
       [src_ip-->158.200.136.60/32 dst_ip-->97.254.16.47/32 ] 3240 (6.62% / 6.62%)
```

SNT

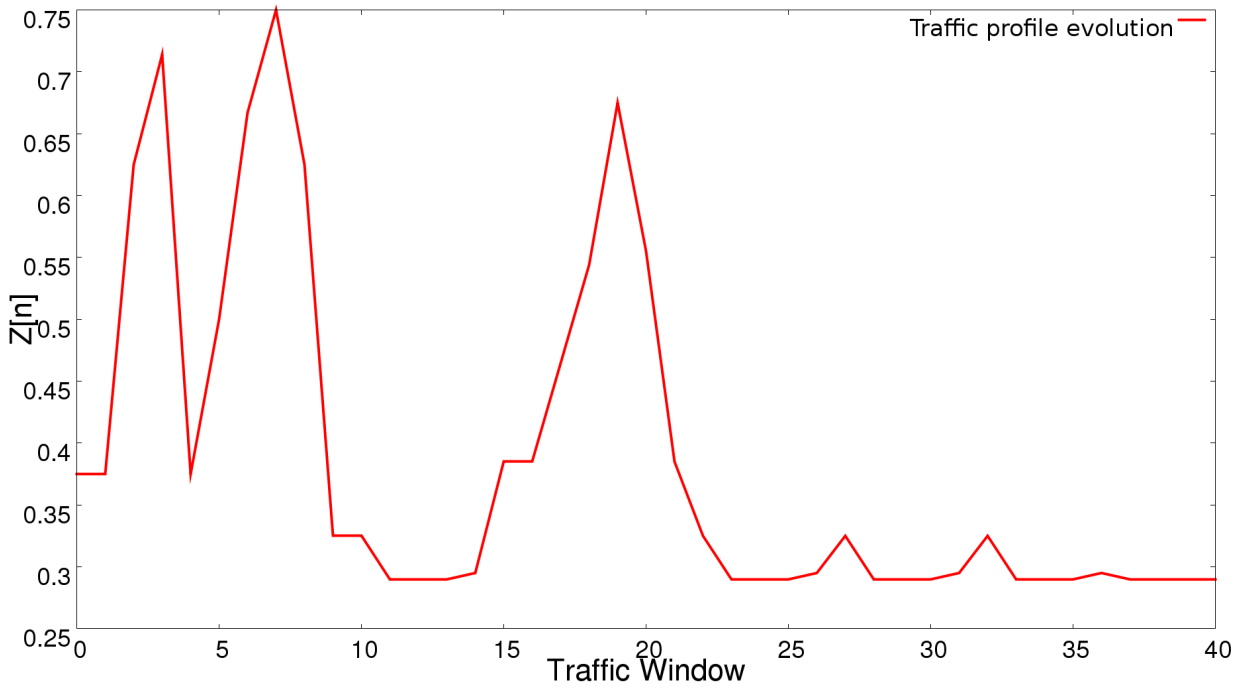
Visually enhanced output

- 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



Automated analysis

- TCP flood example
  - detect deviations in a time series of tree
  - edit-distance between following trees
  - 90 nodes (avg.) with $\alpha = 0.05$

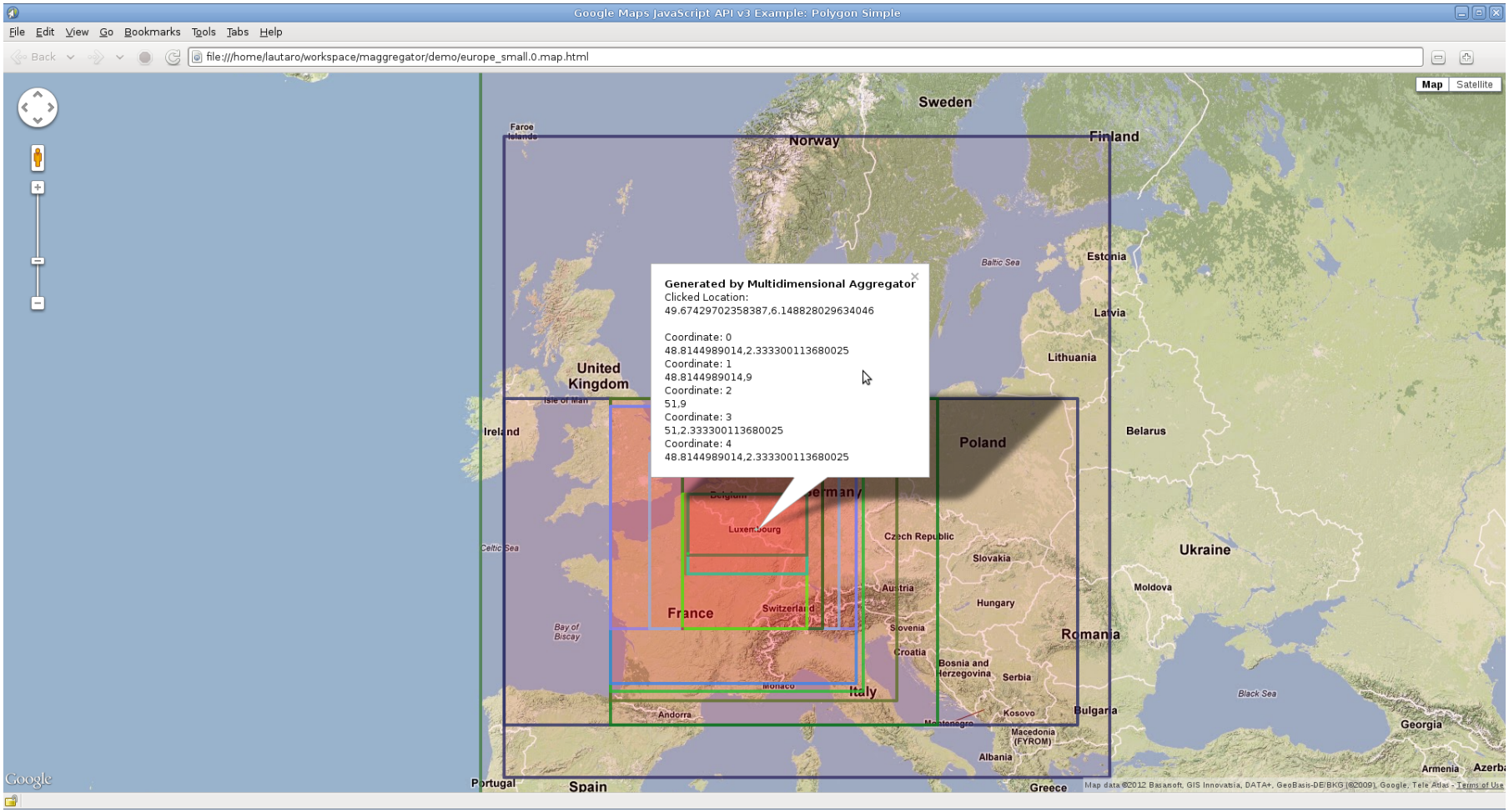


SNT

Geolocated IP Flows

- Same Dataset
- Geolocated source and destination IP Addresses
- Google Maps API

Geolocated IP Flows (Example)



SNT

Outline

SNT

Conclusion

- Scalable aggregation of heterogeneous data
- Realtime Strategies
- Easily extensible to new features
- Visual Tool with a Graphical User Interface (GUI)
- `https://github.com/jfrancois/mam`
- Supported by
  - IoT6, a European FP7 funded project under the grant agreement 288445
  - MOVE, a CORE project funded by FNR in Luxembourg

SNT

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

MaM: Multidimensional aggregation Monitoring

Lautaro Dolberg, Jérôme François, and Thomas Engel

UNIVERSITÉ DU LUXEMBOURG

SnT

NETLAB