On the Use of Data Mining Techniques for the Clustering of URLs Extracted from Network-based Malware Traces

Partner

Orange Labs

Auteurs

Anthony Verez
Dingqi Yang

Dataset

Collection

Provided by our partner, URLs are extracted from pcap files generated by executing malware in a sandbox. Legitimate communications are filtered out to reduce the noise. The original malware samples were collected by a third party

Pre-processing

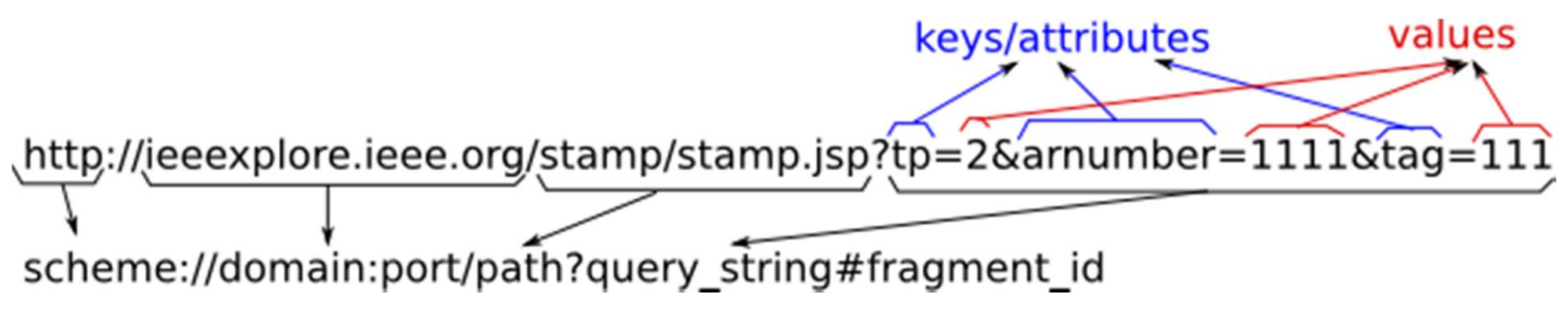
The original dataset contains more than 3.5M samples. Removing duplicates leaves a little less than 2M samples. Domain names are then removed to thwart obfuscation, and only GET requests are retained for privacy concerns, leaving around 1.2M samples

Typing

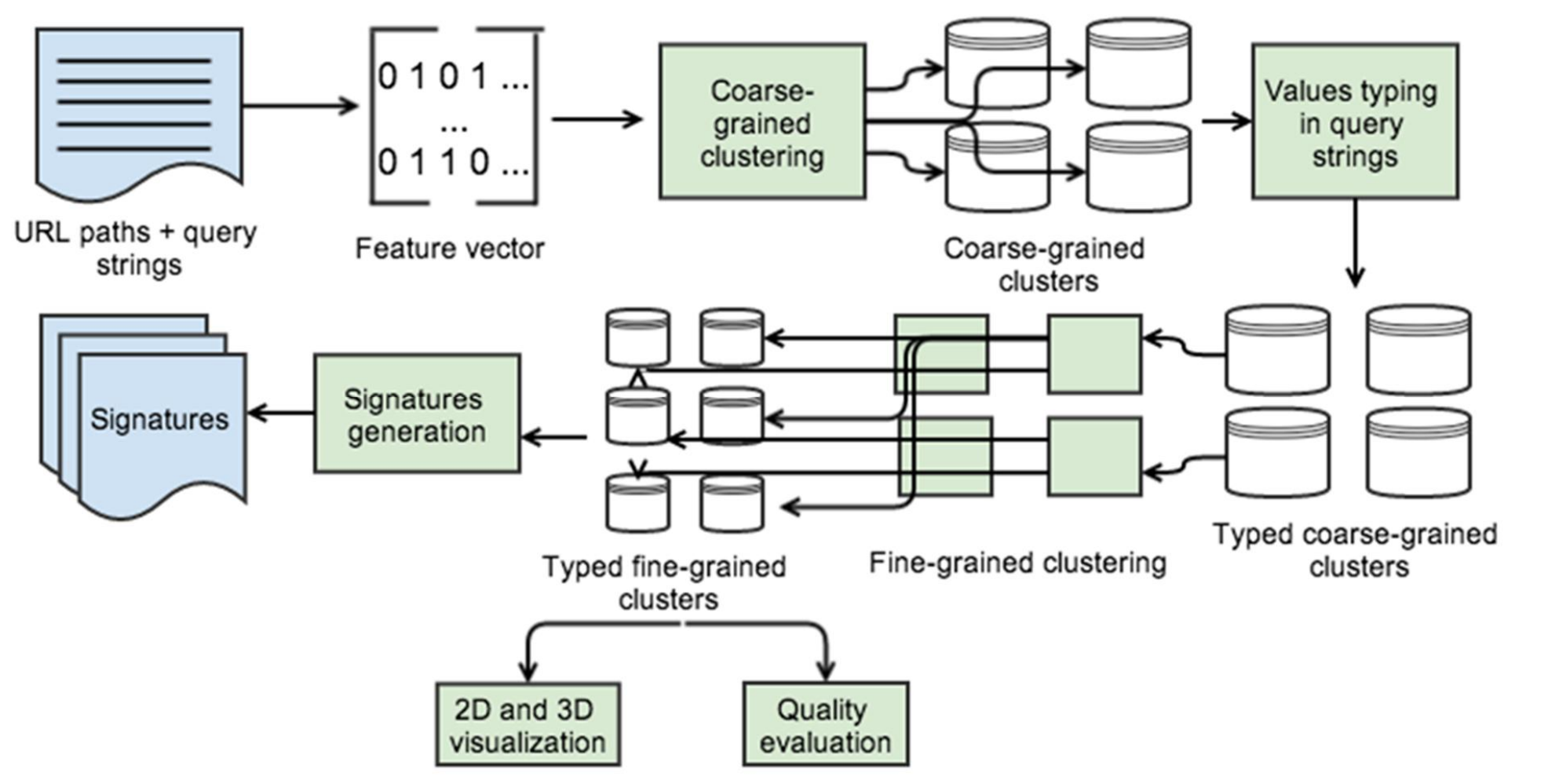
Intermediate step between the two clustering stages, typing allows to replace values in query string by types. Such abstraction offers better performance during fine-grained clustering. More than 70% of values match with one of the 13 types we defined

A ML-based 2-stage URL Clustering Framework

Overview


keys/attributes
values
http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=2&arnumber=1111&tag=111
scheme://domain:port/path?query_string#fragment_id

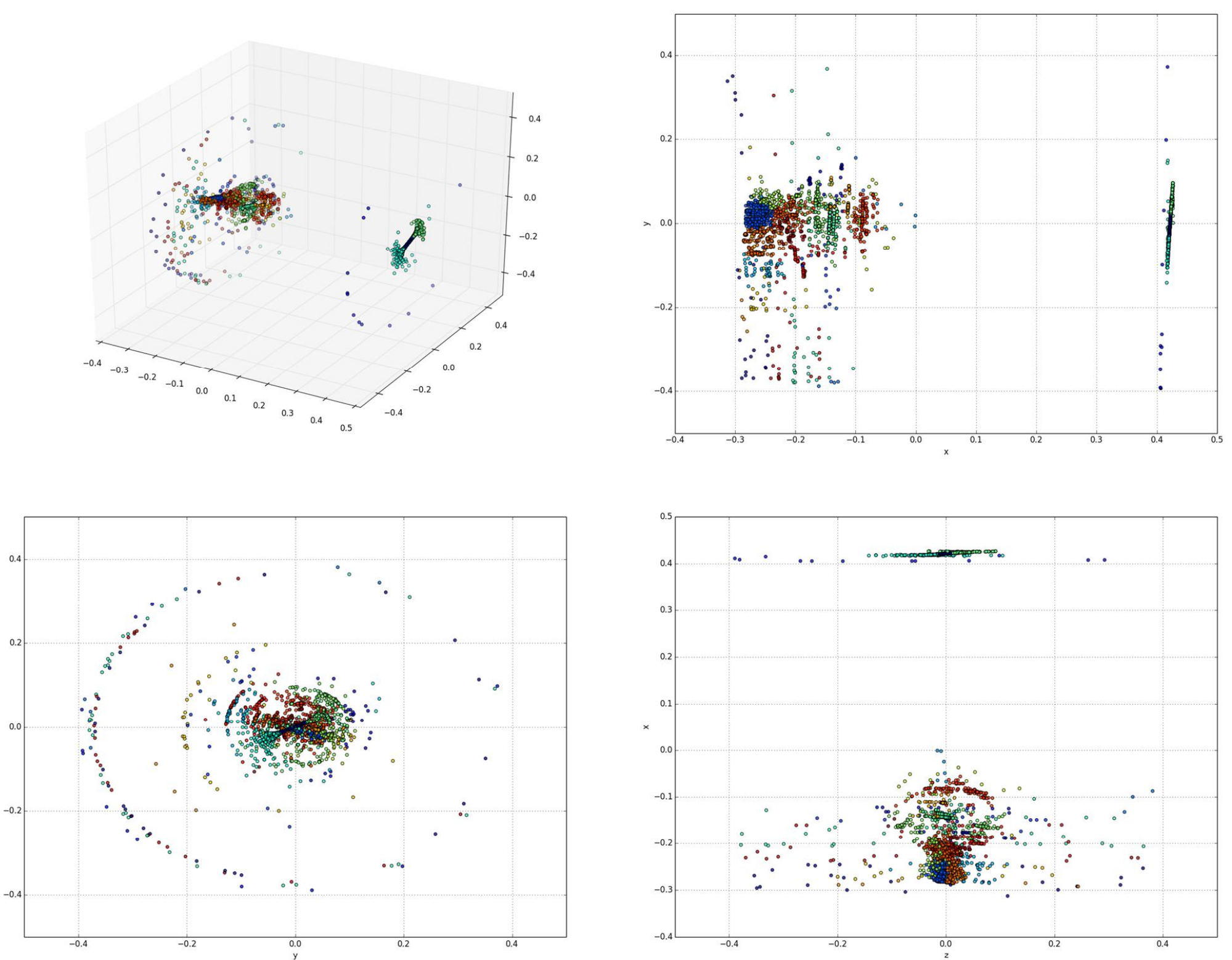
- ***Goal:*** identify families of malware by grouping issued URLs
- ***Assumption:*** common patterns hint at variants or code reuse
- ***Dataset collection:*** URLs extracted from network capture of sandboxed malware communications with possible peers and C&C



➲ Proposed architecture with two steps clustering and typing after the first clustering algorithm

Contributions

- New framework with a typing step and a DBSCAN step to create clusters from a dataset of 1.2M URLs
- Features and distances based only on paths and query strings, not on domains or HTTP headers
- Centralized web platform to monitor and execute machine learning experiments
- Development of generic tools to visualize and navigate through huge numbers of points in 2D and 3D



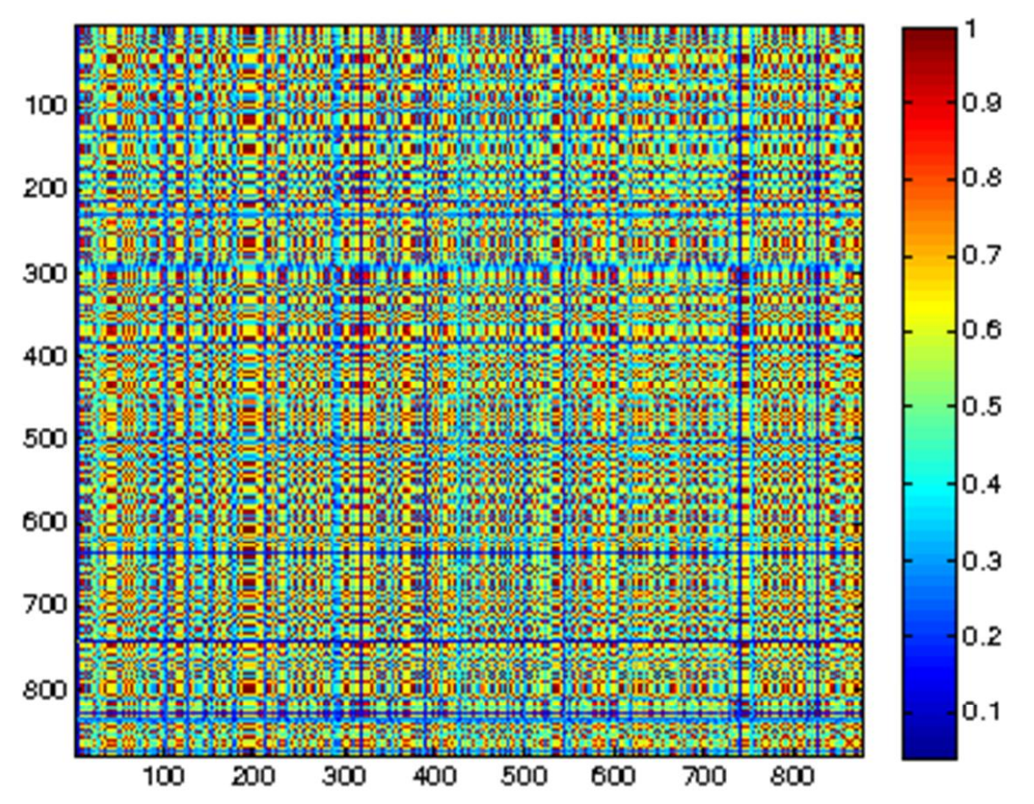
Coarse-grained Clustering

- ***Goal:*** reduce performance overhead of fine-grained clustering by providing smaller input
- ***Method:*** k-means (with k=30) based on ASCII character frequency
- ***Advantages:*** unsupervised learning to automate malware discovery, low complexity and ability to specify number of clusters

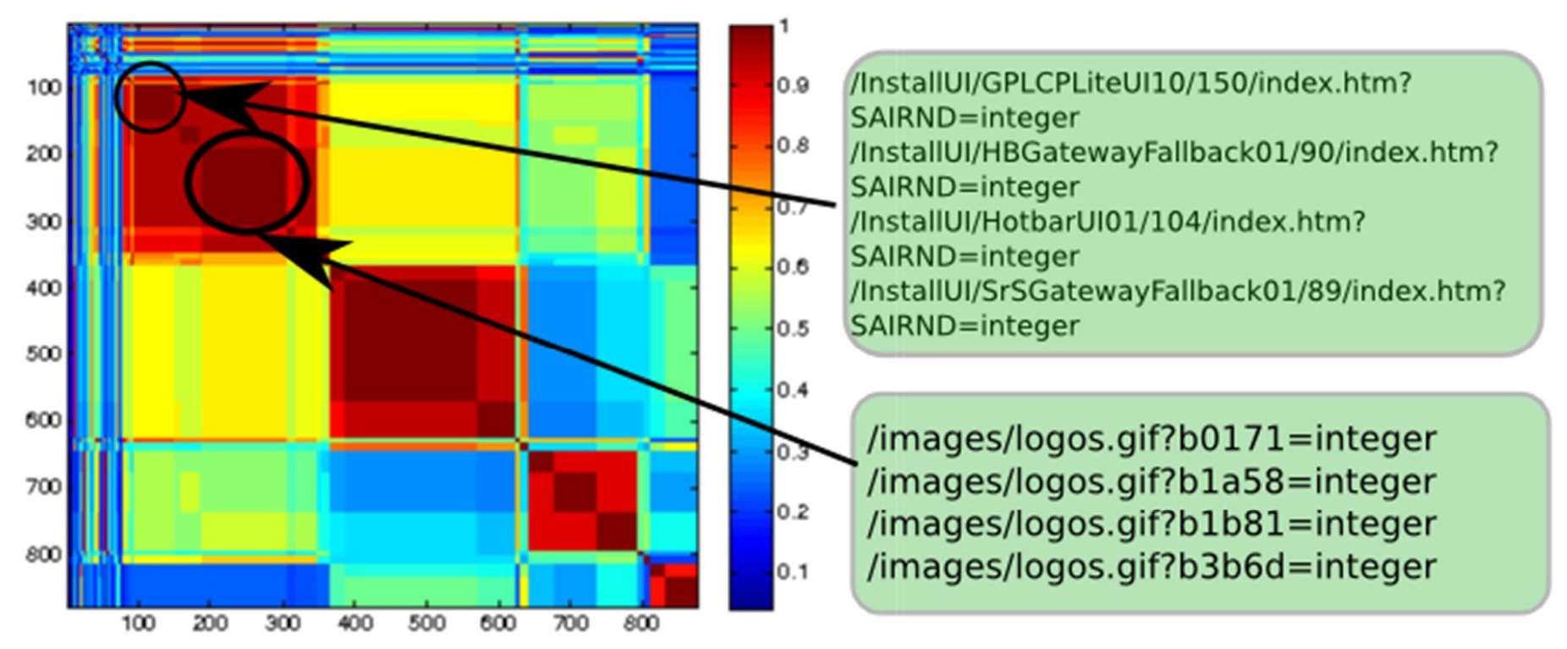
➲ 3D visualization of coarse-grained clusters using projection. One color is associated with each fine-grained cluster

Fine-grained Clustering

- ***Method:*** Unsupervised density-based clustering with DBSCAN
- ***Distance function:*** 1) path distance using the longest common substring algorithm; 2) key/value pair distance based on Jaccard distance on sets of keys associated with a value type



➊ Similarity matrix of a coarse-grained cluster before DBSCAN clustering



➊ Similarity matrix of a coarse-grained cluster after DBSCAN clustering

Quality

- Visualizing the density of a similarity matrix gives on the quality of a clustering algorithm
- Dunn index was also used to assess the quality of coarse-grained clusters

Visualization

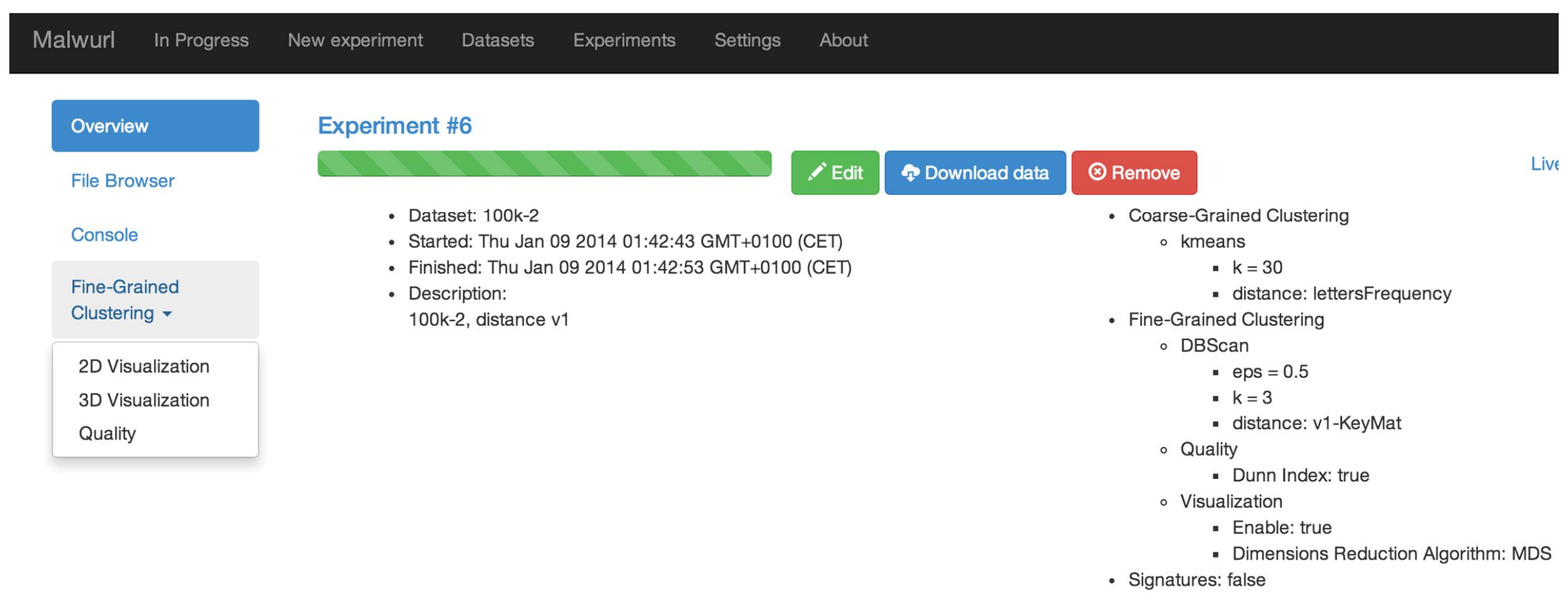
- To confirm that a density-based clustering algorithm fits well with the dataset by visualizing the shapes of clusters
- Using multidimensional scaling on the cluster distance matrix, it is possible to compute the main contributing axes on which will be based the 2D and 3D visualizations

Future Work

- URL signature generation for a family of malware
- Signature-matching-based incremental DBSCAN
- Improve first stage through early typing or Canopy clustering
- Apply typing to paths and possibly keys, try refined typing using length of values

Web Interface

➊ Screenshot of the web interface to launch experiments and analyze results



Experiment #6

- Dataset: 100k-2
- Started: Thu Jan 09 2014 01:42:43 GMT+0100 (CET)
- Finished: Thu Jan 09 2014 01:42:53 GMT+0100 (CET)
- Description:
  100k-2, distance v1

- Coarse-Grained Clustering
  - kmeans
    - k = 30
    - distance: lettersFrequency
- Fine-Grained Clustering
  - DBScan
    - eps = 0.5
    - k = 3
    - distance: v1-KeyMat
  - Quality
    - Dunn Index: true
  - Visualization
    - Enable: true
    - Dimensions Reduction Algorithm: MDS
- Signatures: false

NECOMA Project

January 2014

Contact gregory.blanc@telecom-sudparis.eu

Site web www.necoma-project.eu