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

**A ML-based 2-stage URL Clustering Framework**

**Overview**

- **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

**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

**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

**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

**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.

**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