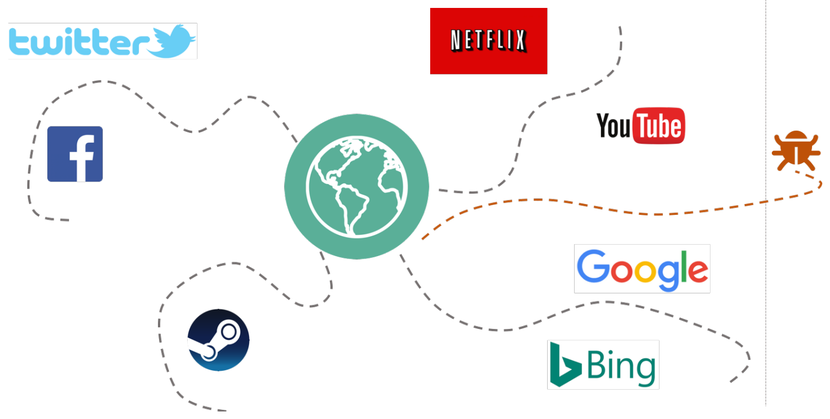
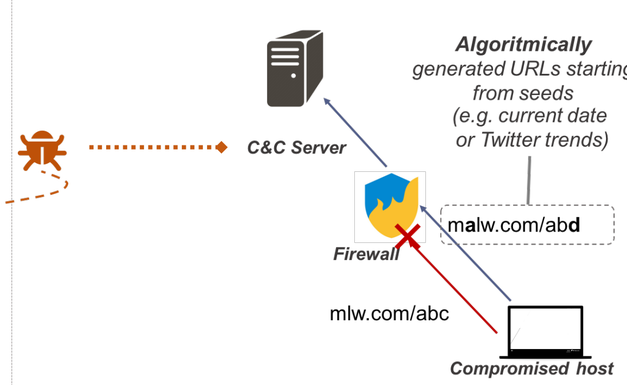
CLUE: CLUSTERING FOR WEB MINING

Proliferation of applications and services that rely on HTTP



Is it possible to automatically mine data to discover potential threats?



GOAL: Develop methodologies and tools to dig into this data and extract useful information

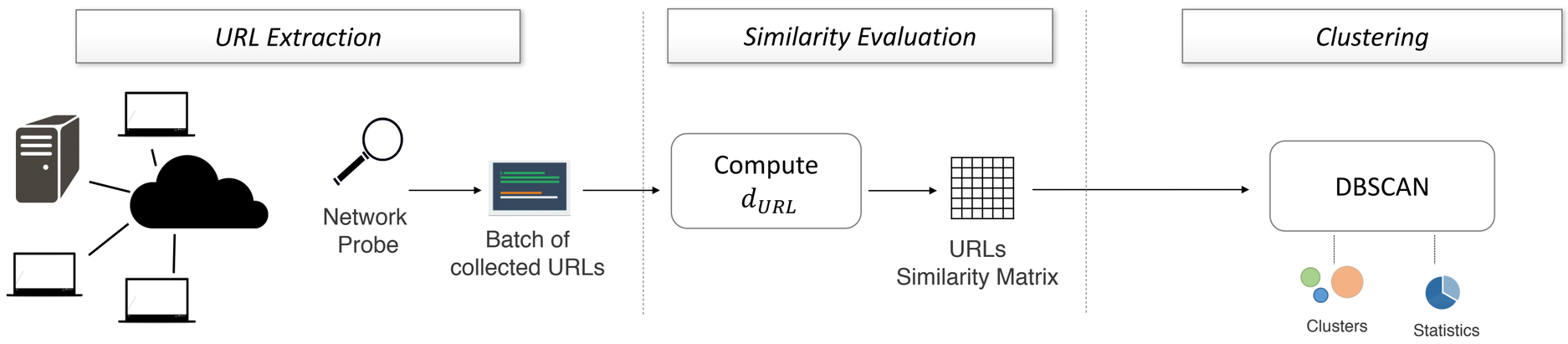
- Looking for similar patterns in traffic
- Clustering similar URLs
- Analyze results

» **Automatic** method that provides aggregated views of URLs
  - **Simplifies** network administrator's tasks

» Use of **passively** monitored network traffic
  - **Transparent** for the user

» Completely **unsupervised** methodology

Workflow



URLs are extracted from a network by means of a passive probe and collected in batch

Similarity between each pair of URLs is computed through d_{URL} metric (Levenshtein distance variant)

DBSCAN clustering algorithm is applied. Additional statistics are provided for the analyst

Final view & Analysis

Silhouette analysis
an unsupervised methodology to find how well each object lies within its cluster

↓

*Can **ease** the **choice***
those clusters whose points are very similar inside the cluster, and very different from points in the rest of clusters are likely to be interesting

S(C)	Main hostname (unique number)	Elements	Activity
0.92	skygo_streaming-i.akamaihd.net (1)	551	Streaming
0.91	ad.doubleclick.net (1)	99	Advertising
0.87	**cookex.amp.yahoo.com (1)**	**61**	**Malware**
0.85	static.simply.com (1)	25	File Hosting
0.81	d24w6bsrhbeh9d.cloudfront.net (1)	63	File Hosting
0.81	**mfdclk001.org (1)**	**27**	**Malware**
0.78	adserver.webads.it (1)	35	Advertising
0.77	**.com (3)**	**37**	**Malware**
0.75	pixel.quantserve.com (1)	57	Advertising
0.72	watson.microsoft.com (1)	29	Windows Debug

Clusters sorted by silhouette coefficient

Advertising
(H1) ad.doubleclick.net
(P1)
/0_AcquisitionRtr_Apr12_AmericanExpress.html/5854707559307a644238674141515767
(P2) 0;click0=http://oase00821.247realmedia.com/5c/msn.it/Female/L-13/
(P3) /GroupM-IT/AmericanExpress_Acquisition_Apr12_Rtr/
(P4) /adj/N4199.456584.XAXIS.COM1/B6490067
(P5) ;sz=
- (H1) / (P4) (P5) 300x25 (P2) 1403202186/Right (P3) 300x25 (P1)
- (H1) / (P4) .2 (P5) 728x9 (P2) 523922702/Top (P3) 728x9 (P1)
- (H1) / (P4) .2 (P5) 728x9 (P2) 717876294/Top (P3) 728x9 (P1)
- (H1) / (P4) .2 (P5) 728x9 (P2) 309206097/Top (P3) 728x9 (P1)
- (H1) / (P4) .2 (P5) 728x9 (P2) 2064492282/Top (P3) 728x9 (P1)
- (H1) / (P4)(P5) 300x25 (P2) 1934004172/Right (P3)300x25 (P1)

Silh = 0.91

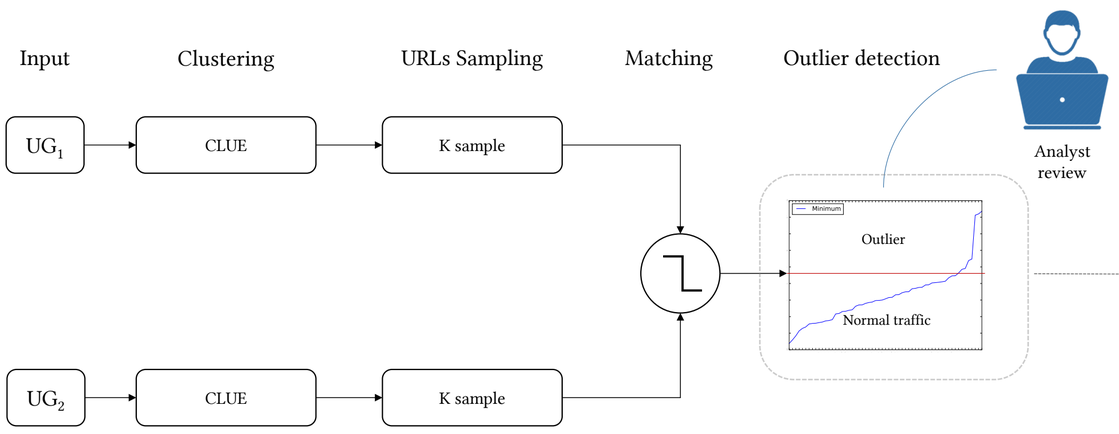
Malware
(H1) mfdclk001.org
(P1)
Y2xrPTEuMjEmYmlkPTUwMGFhNzVjLWY1ZTUtNDhhOC05ZjlkLWY2ODQ3N
GYzOGQwZCZhaWQ9MTAwMTAmc2lkPTAmcmQ9MTcuMTEuMjAxMQ==
- **(H1)** /UVw07ael7p4qVcS6 **(P1)** 26c
- **(H1)** /wZl1ELDd7N5quws3 **(P1)** 26A
- **(H1)** /SZY35xzx6x4q5Ks1 **(P1)** 26x
- **(H1)** /7V10LUdl7Z5mxcS2 **(P1)** 25A
- **(H1)** /LZK2BDxe5k4mQiO2 **(P1)** 17g
- **(H1)** /WZb3fvgl643q33U7 **(P1)** 05c

Silh = 0.81

Current Developments – questions and experiments

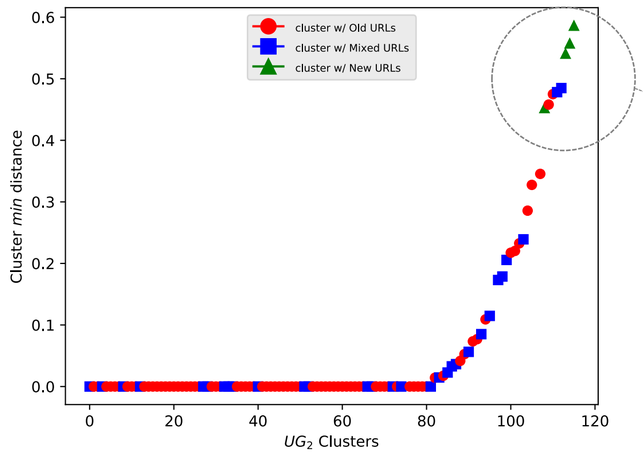
Potential Applications

✦ Use it to compare activities of different groups of hosts (**U**RLs **G**roups)
✦ Detecting the web traffic changes in time for a group of hosts



Experiment

Compare two groups of URLs, the second has the same traffic as the first, plus other injected traffic. Then, look for anomalies.



UG_2 clusters that contain only new traffic are at a greater distance from UG_1 clusters.
In the graph some noise is visible, caused by the URLs sampling.

Morichetta Andrea
PhD Student
andrea.morichetta@polito.it

Mellia Marco
Associate Professor
marco.mellia@polito.it

POLITECNICO DI TORINO
BG DAMA