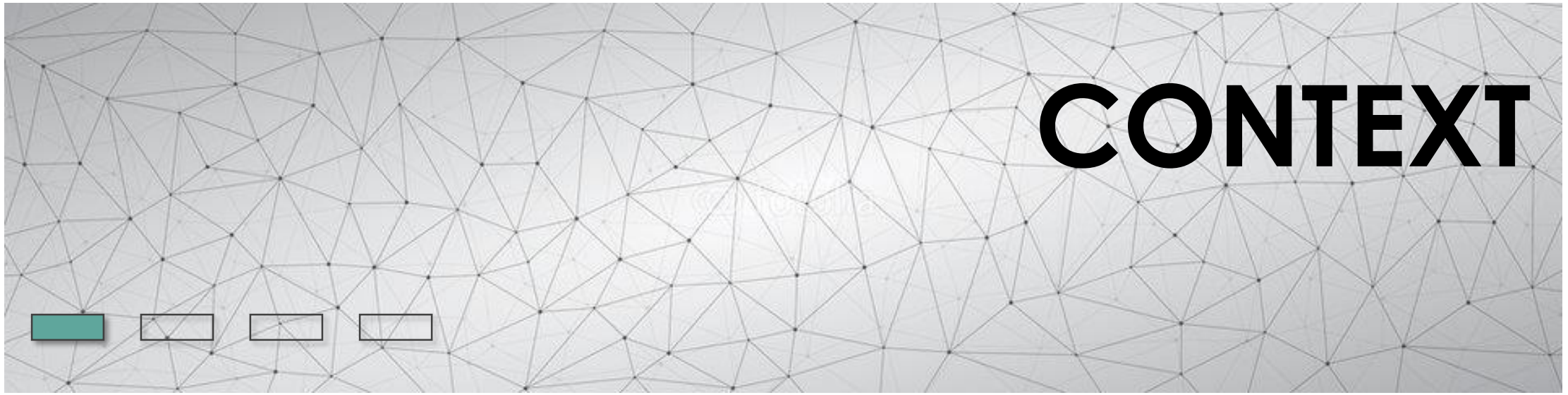
DATA ANALYSIS ALGORITHM FOR MALWARE DETECTION IN NETWORK TRAFFIC INSPECTION

Andrea Morichetta
Politecnico di Torino

Part 1

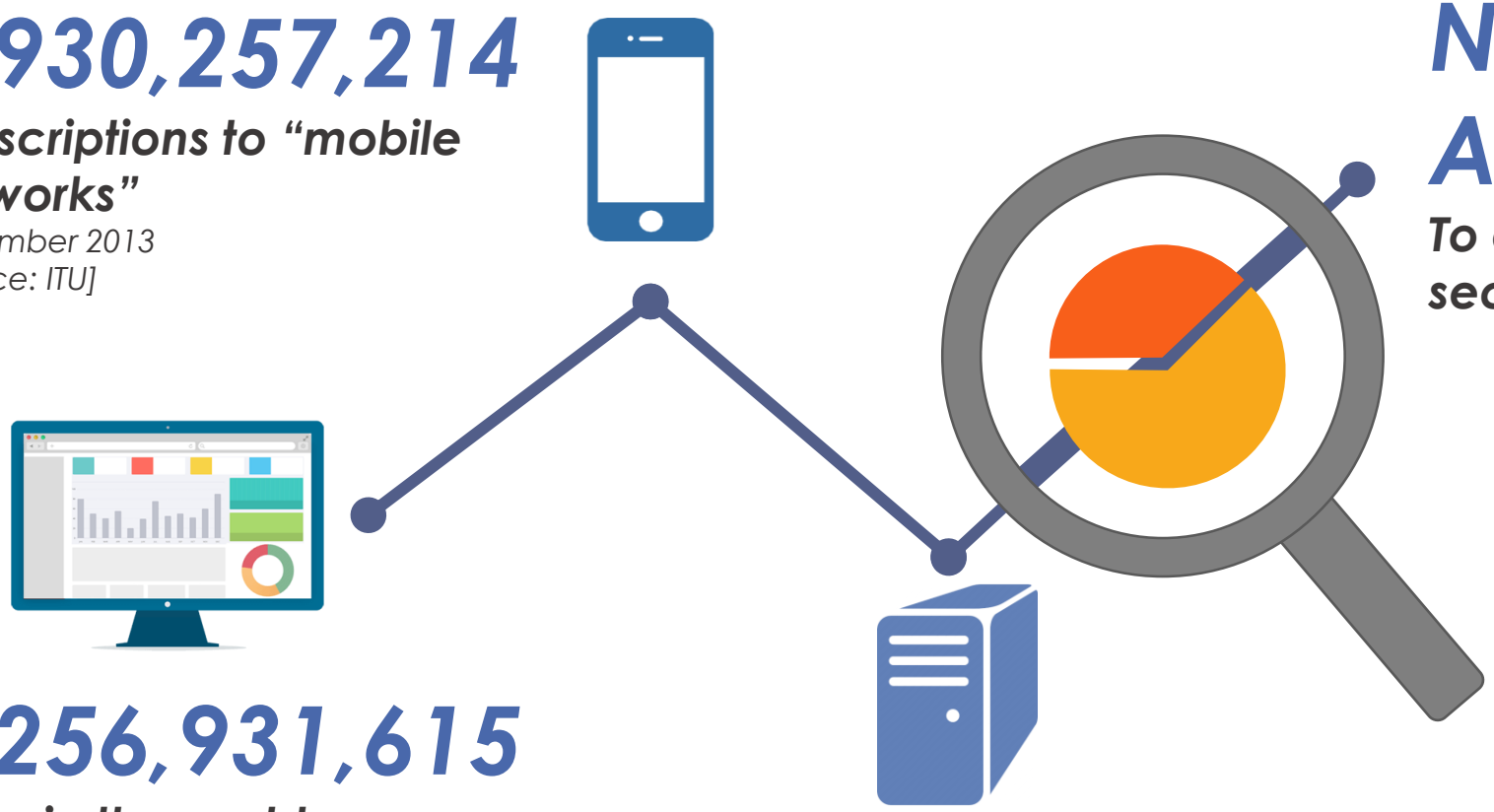


SCENARIO

1,930,257,214

Subscriptions to "mobile networks"

December 2013
[Source: ITU]



3,256,931,615

Users in the world

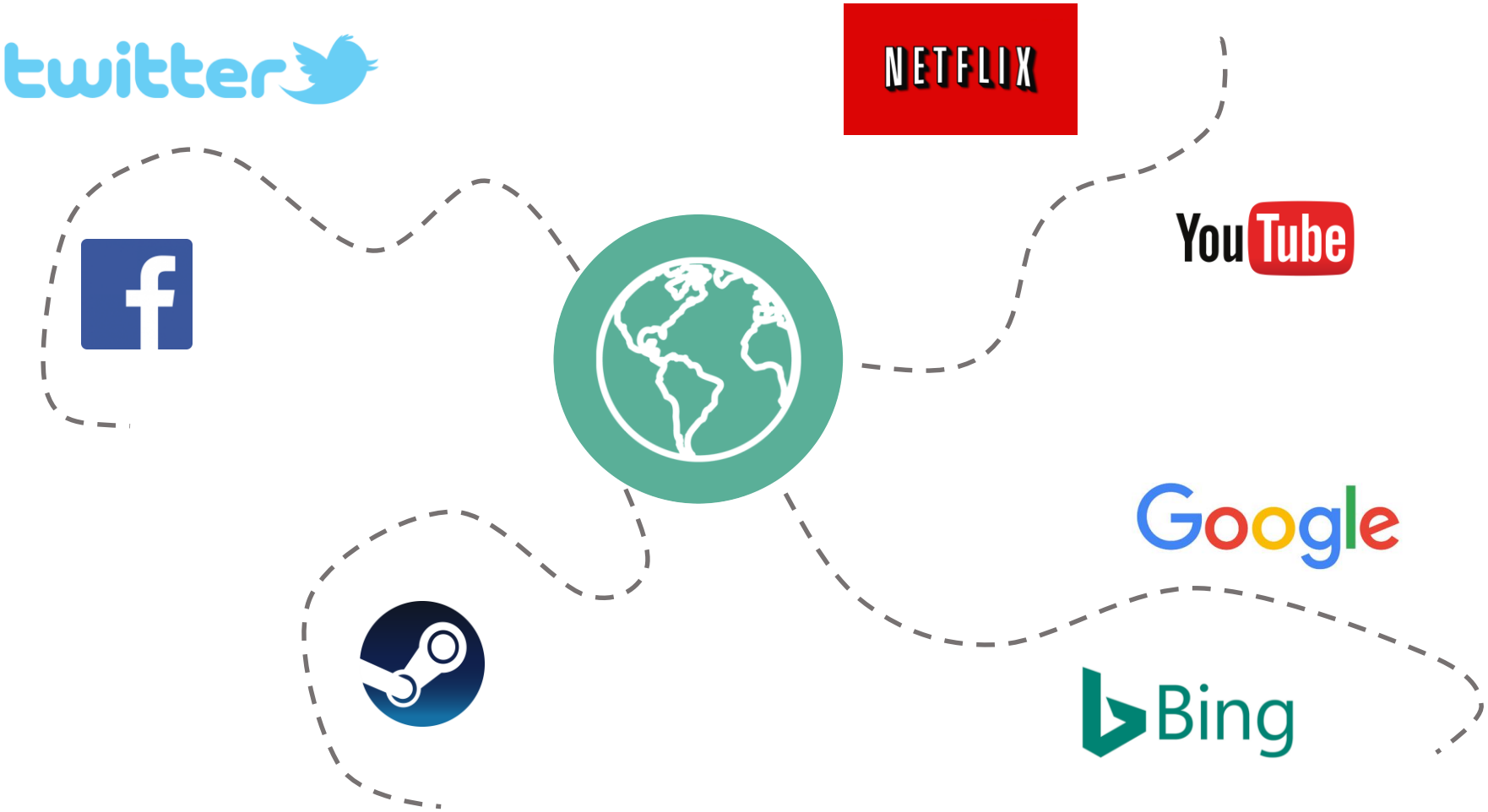
December 2nd 2015
[http://www.internetlivestats.com]

Network Analysis

To obtain quality and security

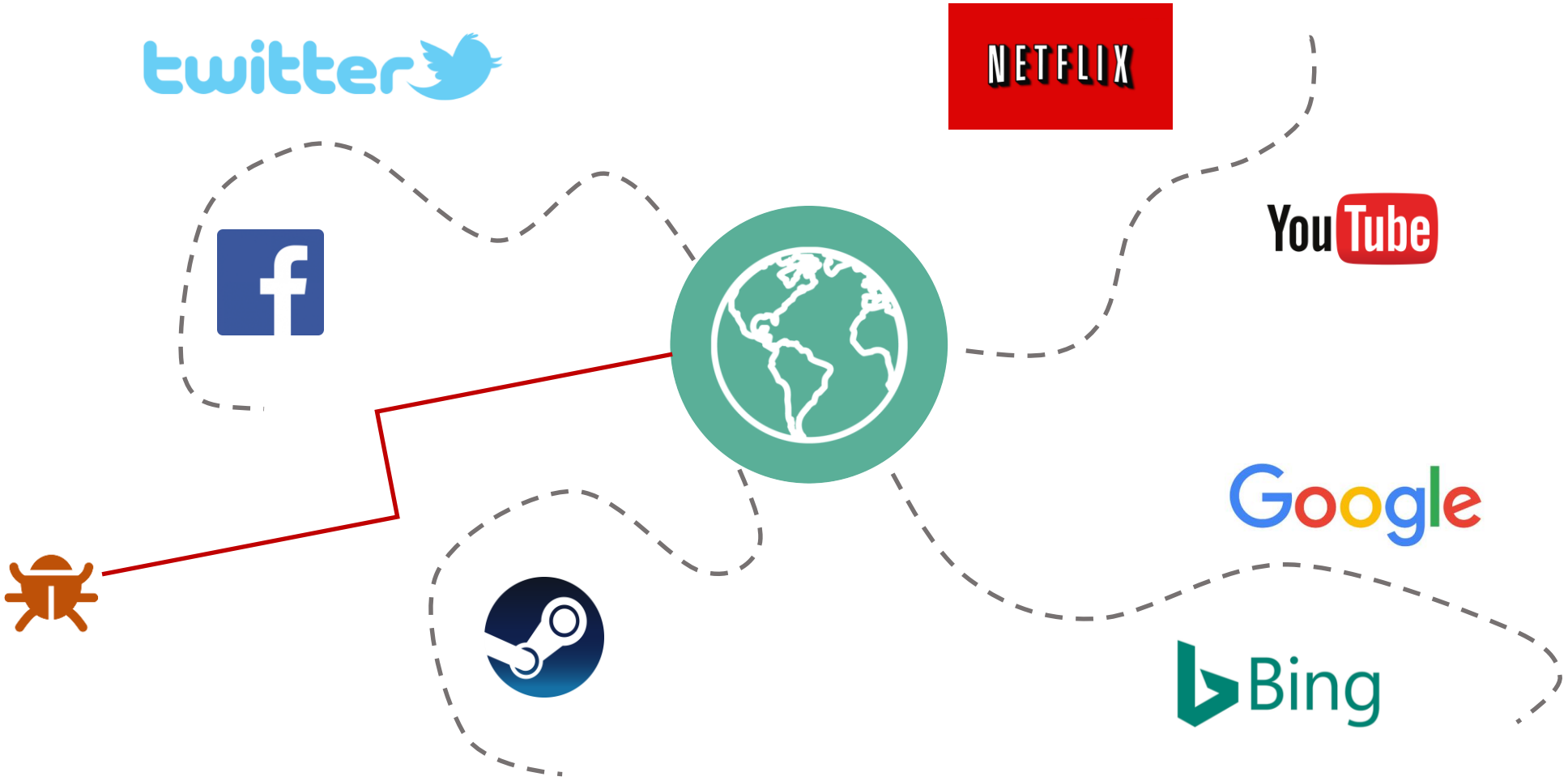
THE WEB TODAY

Internet increasing complexity.



THE WEB TODAY

Internet increasing complexity.



PURPOSE

How to find anomalous behaviors in the huge traffic flow?

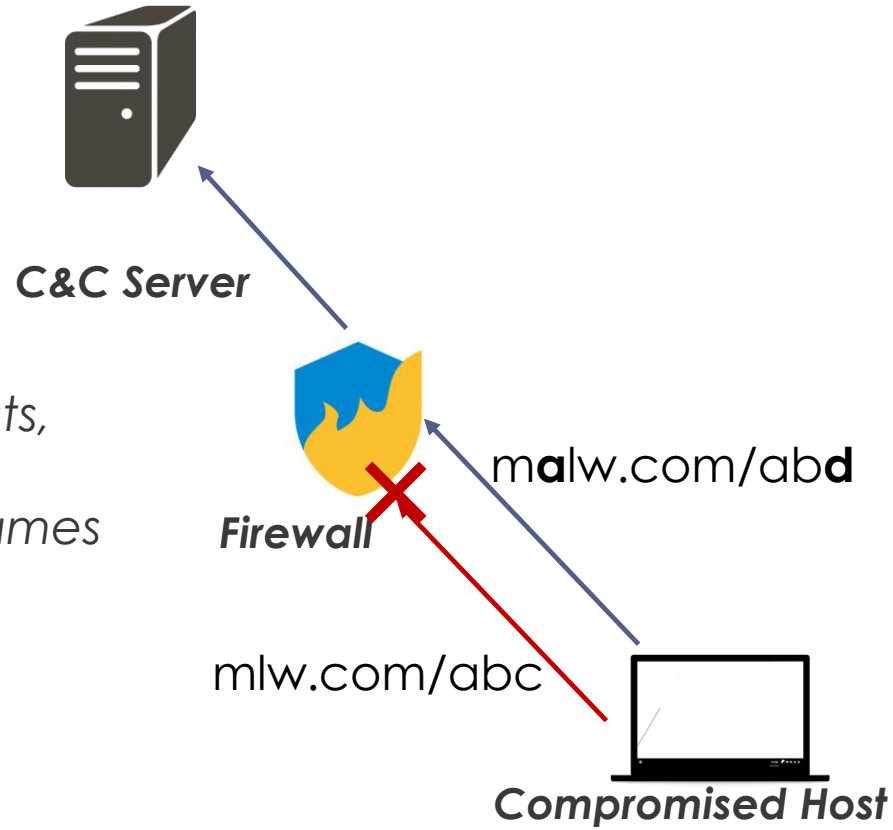
Timestamp	Hostname	Path
1339130937	www.emilbanca.it	/emilbanca/img2009/angolo_blu_dx.gif
1339130938	fpdownload.adobe.com	/pub/swz/crossdomain.xml
1339130941	8lnxpg8vwyuzhbol.com	/rkF4Tx3x8N4YR8C5dj0xLjkmaWQ9NmQzNDZkY2FkYzU4Yzk0ODBkZDliODNkNjYxYzIzMmNjZTZhZDY4ZCZhaWQ9MzA0Mjgmc2lkPTQmb3M9NS4xIDAwMDAgU1AwLjAma3c9WlcxcGJHSmhibU5oRFFwbGJXbHNZbUZ1WTJFTkNtVnlibVZ6ZEc4cmMzUnliM3A2YVN0bVpYSnlZWEpoRFFvPSZ1cmw9YUhSMGNEb3ZMM2QzZHk1cGJtSmhibXN1YVhRdlpuVnVZM1JwYjI0dmJHOW5hVzR2YVc1a1pYZ3Vhbk53UDJ4aGJtYzlhWFFtWVdKcFBUQTNNRGN5Sm1OemN6MHdOekEzTWc9PSZyZWY9ZDNkM0xtVnRhV3hpWVc1allTNXBkQzl3YjNKMFlXd3ZjR0ZuWlQ5ZmNHRm5aV2xrUFRJMk9ERXNNU1pmWkdGa1BYQnZjblJoYkNaZmMyTm9aVzFoUFZCUFVsUkJUQT0935A
1339130945	83.133.121.147	/c/kaw0hOOD6x5Jpso2440a89f7bdeb9da9f4b5af9160e66aa908c
1339130946	delivery.jemacpv.com	/network/c/adclick.php
1339130946	delivery.jemacpv.com	/network/c/adclick.php
1339130947	www.peaktube.com	/video_play
1339130948	www.peaktube.com	/redirect.php
1339130949	cdn1.static.videobash.com	/css/ie8-style-new.css
1339130949	www.videobash.com	/video_play
1339130949	cdn1.static.videobash.com	/css/style_new.css
1339130954	www.emilbanca.it	/emilbanca/img2009/labanca_con.jpg
1339130965	www.emilbanca.it	/emilbanca/img2009/labanca_con.jpg
1339130980	img3.iol.it	/s/sport/med/balotelli-quotma-quale-derby-di-mercato-quot.jpg
1339130980	rta.criteo.com	/dis/rtt.js
1339130980	img1.iol.it	/img107/share/pubblicita/07/76/2012/4/nome.jpg
1339130980	img3.iol.it	/img107/coldx/appl/01/1047/2011/3/3.gif
1339130980	img1.iol.it	/img107/coldx/appl/03/3043/2012/3/03_110x107.jpg
1339130980	img3.iol.it	/s/lavoro/116/autogrill-licenzia-in-massa.jpg
1339130980	www.libero.it	/

PURPOSE

HTTP traffic analysis to track anomalous and potentially malicious behaviors.

Malware Zero-day

They elude static controls, based on blacklists, changing URLs' paths and hostnames

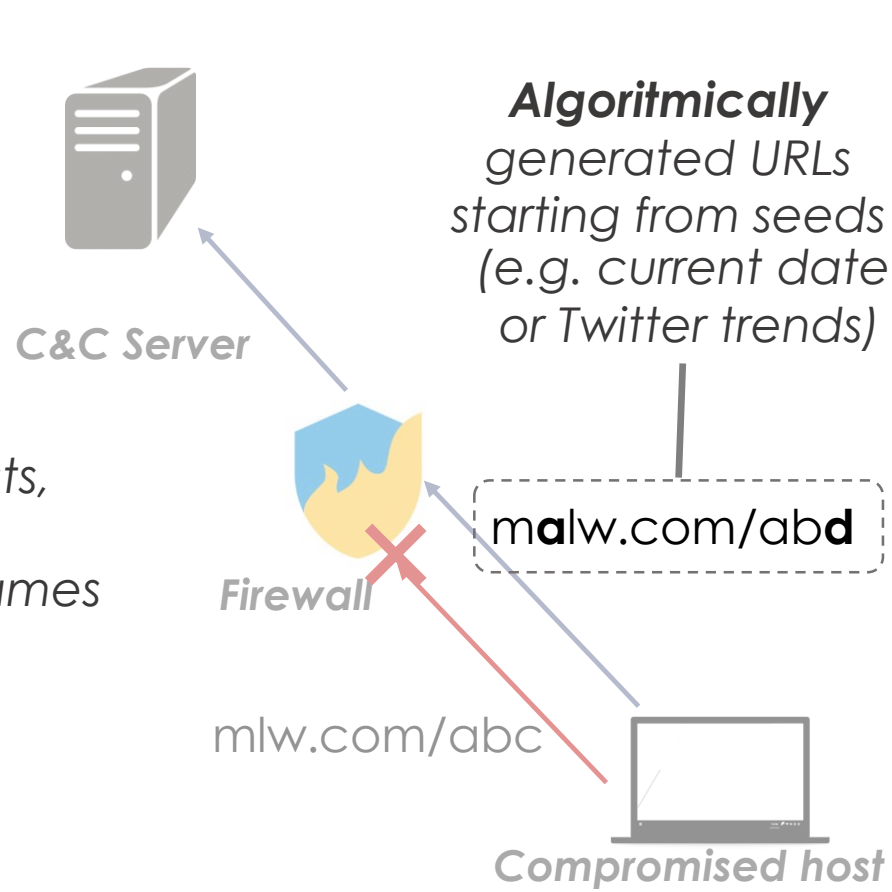


PURPOSE

HTTP traffic analysis to track anomalous and potentially malicious behaviors.

Malware Zero-day

They elude static controls, based on blacklists, changing URLs' paths and hostnames

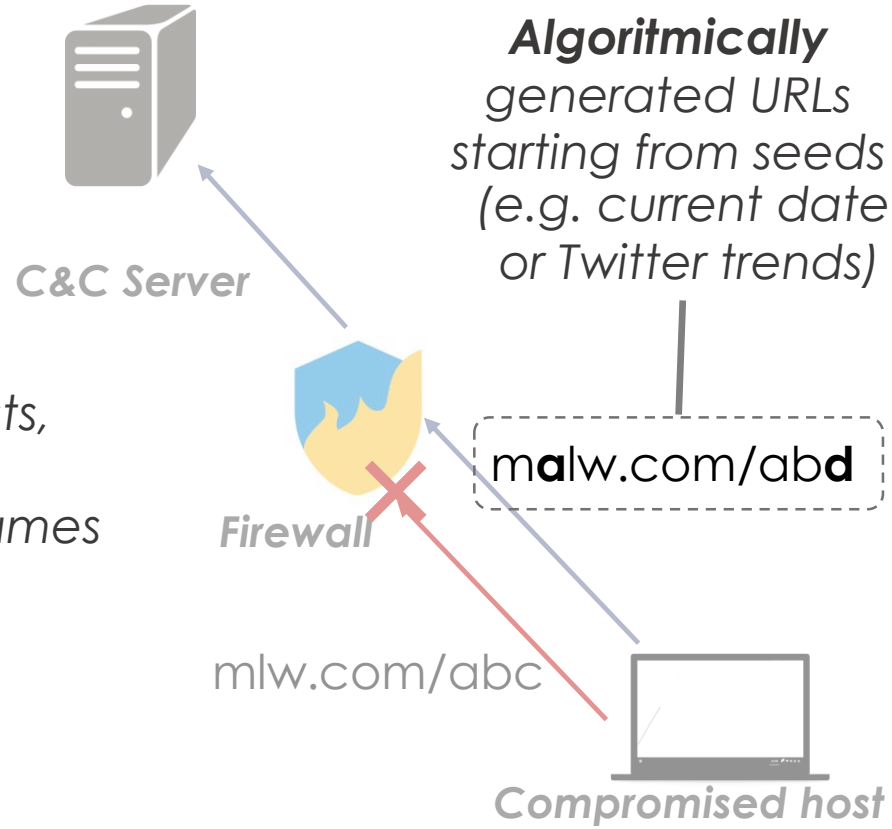


PURPOSE

HTTP traffic analysis to track anomalous and potentially malicious behaviors.

Malware Zero-day

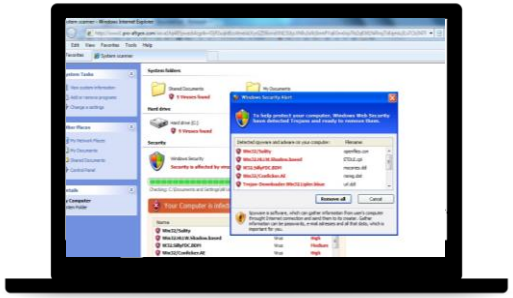
They elude static controls, based on blacklists, changing URLs' paths and hostnames



HTTP traffic analysis

Detect and group algoritmically generated URLs.
Control and monitoring on possible, not-checked, malicious behaviors.

TIDSERV



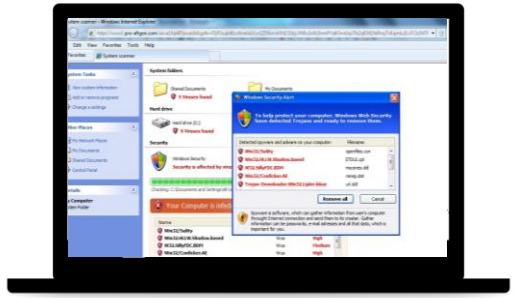
Trojan Rootkit

Profit-making purpose

It spreads with users complicity

URLs characterized by pseudo-randomness

TIDSERV



Trojan Rootkit

Profit-making purpose

It spreads with users complicity

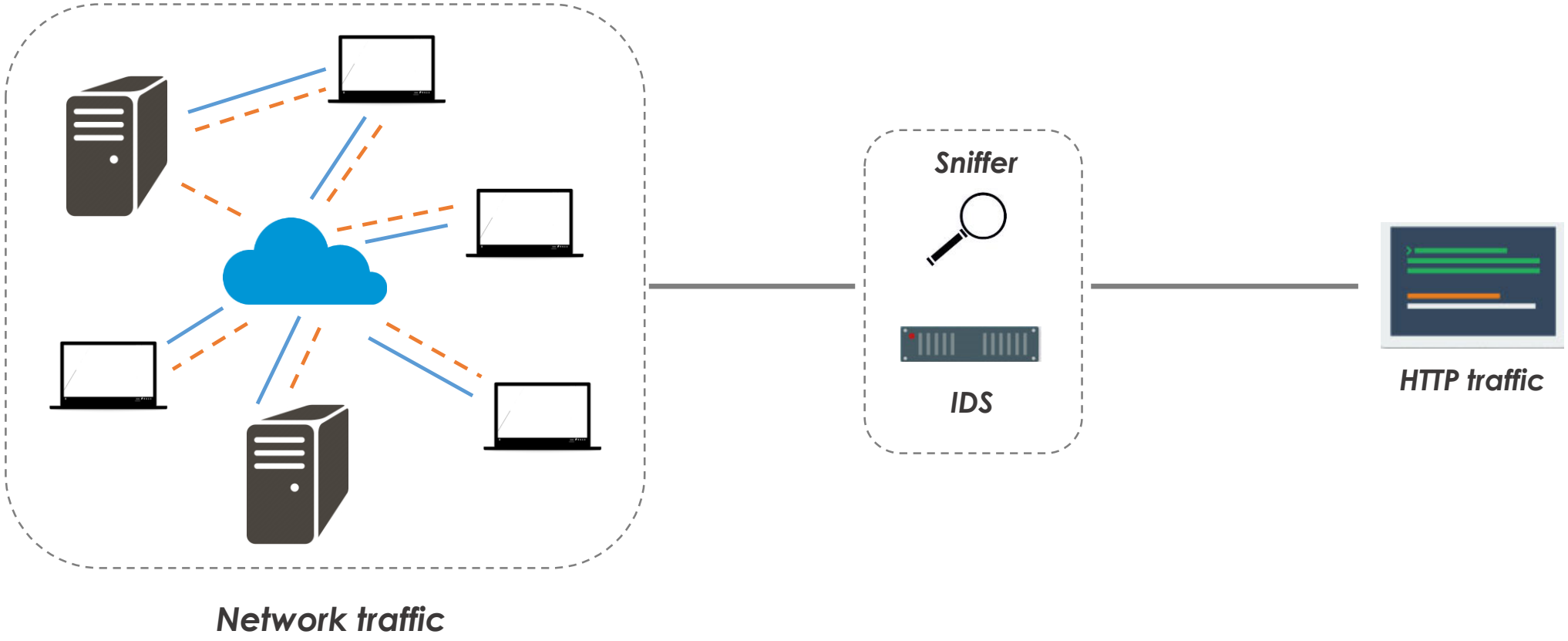
URLs characterized by pseudo-randomness

swltcho81.com/NZf4A07d7r7yE1C1**dmVyPTQuMCZiaWQ9YjZjYWVhNjE0NjhhMmQ4ZTc0OGQ3ZTEzMTI yMDZiMDQ4NWY2MjJhYSZhaWQ9NDAxOTcmc2lkPTAmcmQ9MCZlbmc9d3d3Lmdvb2dsZS5pdCZxP**X VpbmZlIG5ZGVzaw==38c

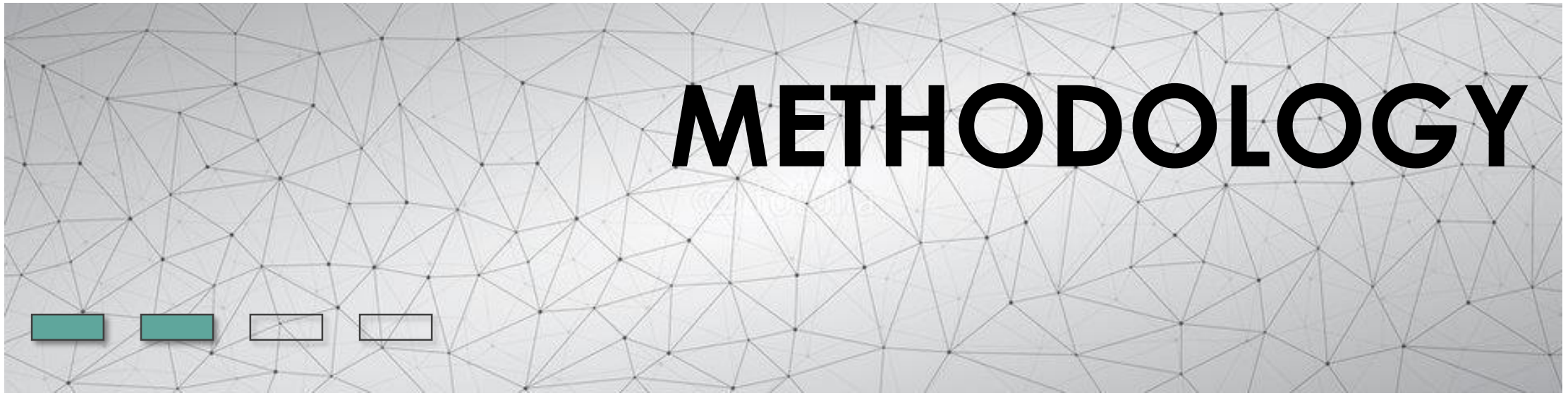
rammyjuke.com/kaI1wWRd8Y5yfbU9**dmVyPTQuMCZiaWQ9YjZjYWVhNjE0NjhhMmQ4ZTc0OGQ3ZTEzM TIyMDZiMDQ4NWY2MjJhYSZhaWQ9NDAxOTcmc2lkPTAmcmQ9MCZlbmc9d3d3Lmdvb2dsZS5pdCZxP** WZvcnVtIGFybWF0YSBkZWxsZSB0ZW5lYnJl37g

SCENARIO

Traffic collected from a network with more than 20000 Hosts connected.



Part 2



Comparison between elements with no good understanding a priori.

Edit Distance

Class of distance functions in which, given two strings *s* and *t*, distance is the cost of **best** sequence of ***edit operations*** that convert *s* to *t*.

LEVENSHTEIN DISTANCE

Simple *Levenstein distance:* assigns a unit cost to all edit operations

JARO DISTANCE

The *Jaro* algorithm is a measure that evaluates the number and order of features in common

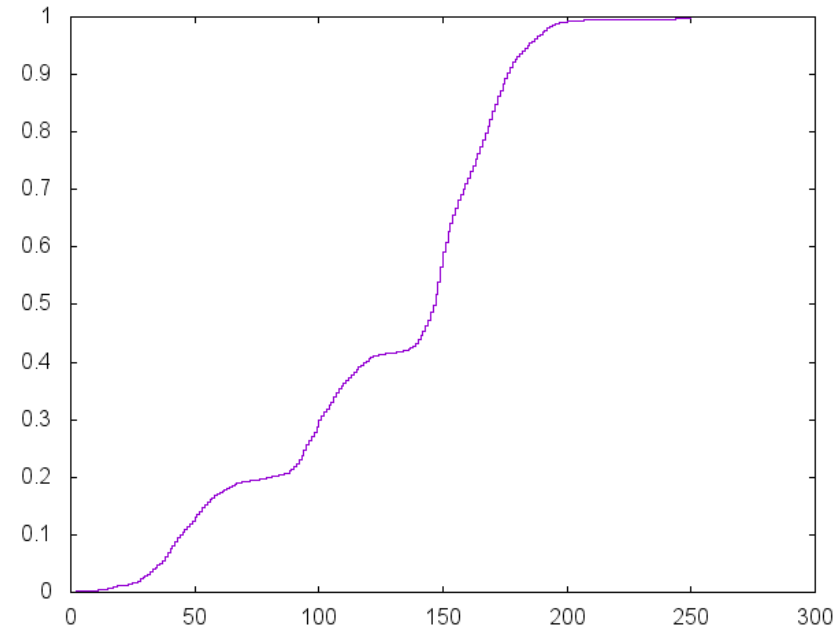
URL DISTANCE

Levenshtein modified: unitary weight for adding and removing edit operations, double weight for replacements

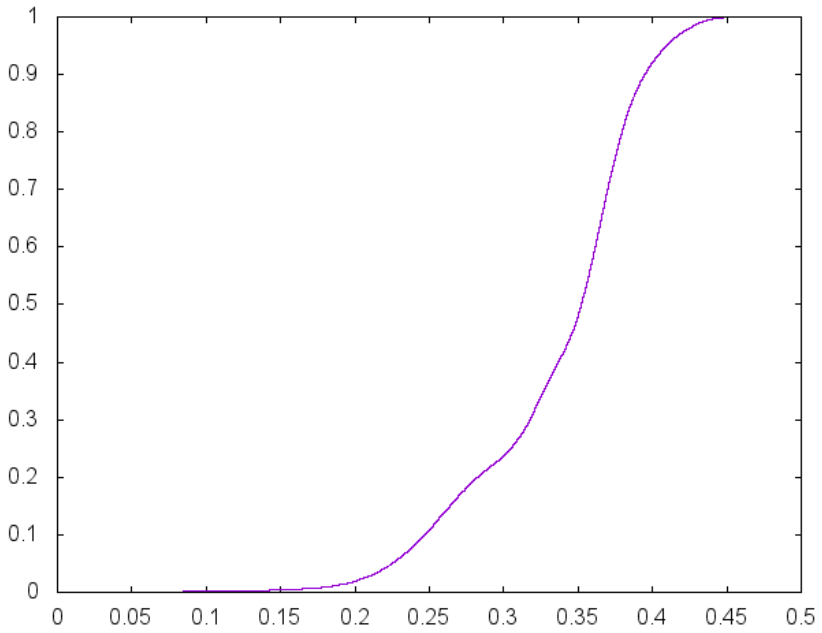
DISTANCE EVALUATION

Comparing distance measures behavior with TidServ elements.

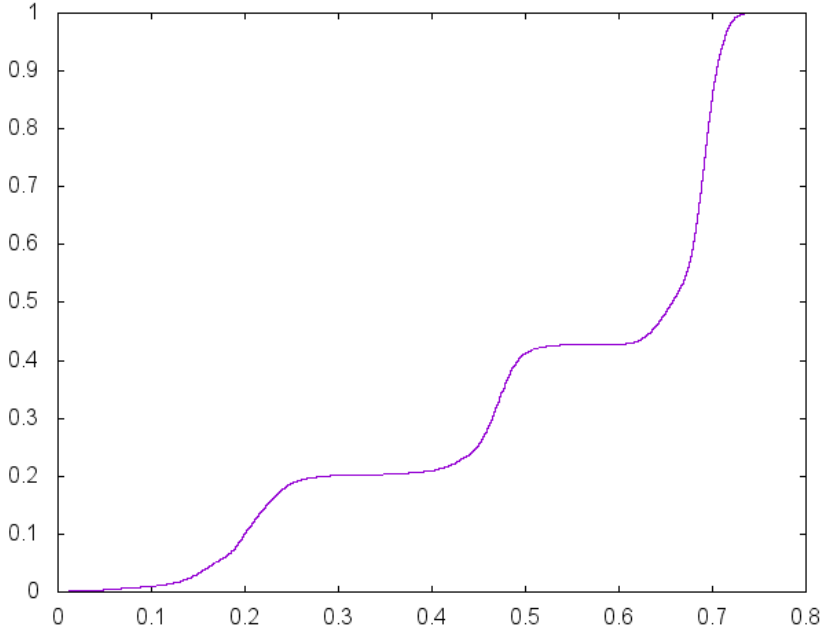
LEVENSHTEIN DISTANCE



JARO DISTANCE

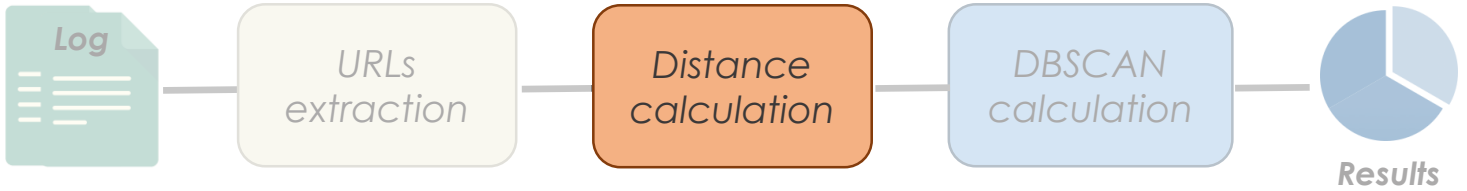


URL DISTANCE



CDF functions

URL DISTANCE



Measure to calculate strings similarity.

Based on Levensthein distance, it provides a similarity value normalized between 0 and 1

Levenshtein distance: min. number of character operations necessary to match the two strings

Levenshtein modified: unitary weight for adding and removing, double weight for replacements

FORMULA

$$D_{URL}(string_1, string_2) = 1 - \frac{|string_1| + |string_2| - Levenshtein\ distance\ mod(string_1, string_2)}{|string_1| + |string_2|}$$

EXAMPLE

$url_1 =$ '$google.com$' $(10\ caratteri)$;
$url_2 =$ '$1goggle.com$' $(11\ caratteri)$;

$Levenshtein\ distance\ mod(url_1, url_2) =$
$1\ add(weight{:}\ 1) + 1\ replacement(weight{:}\ 2) = 3;$

$$D_{URL}(url_1, url_2) = 1 - \frac{10+11-3}{10+11} = 0.143$$

DBSCAN

Clustering algorithm used for grouping URLs together.

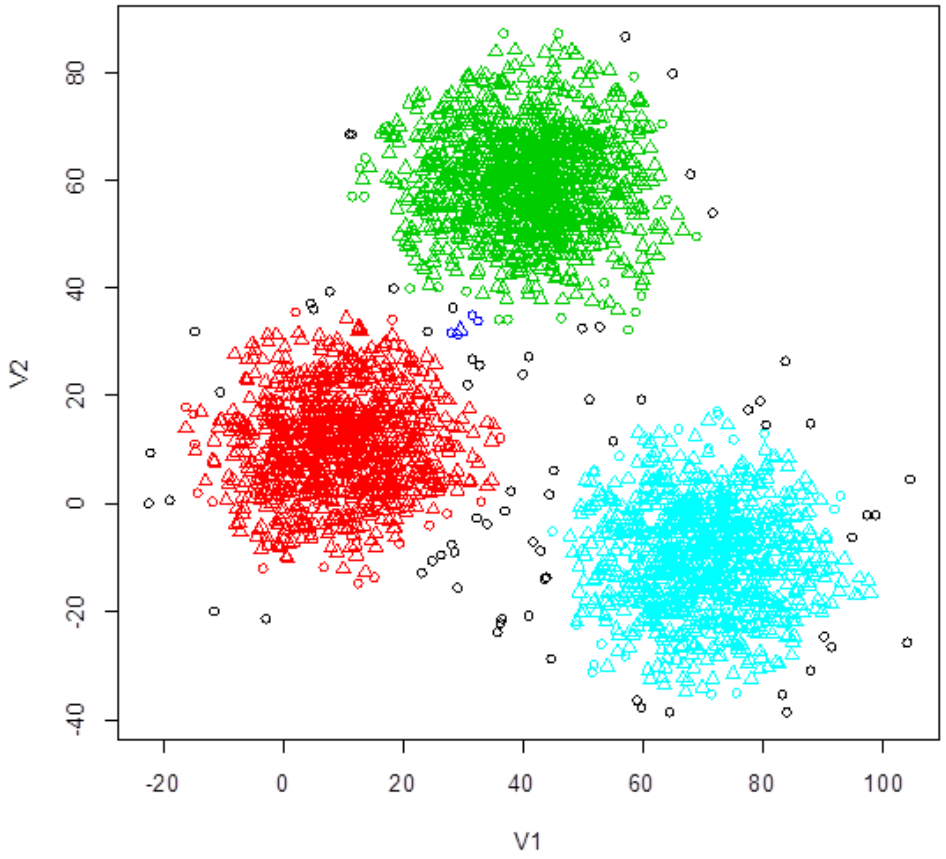
Based on the idea of density, intended as the number of points in a specific area; compared to other algorithms families it provides partial solutions.

Features

- It includes the presence of **noise**: prevents **non-coherent** elements to be added to the cluster.
- **Must not** define the number of clusters **a priori**
- **Must not** define centroids
- **Do not** mandatory require points in **Euclidean** space
- Can handle **different shaped clusters** and not only globular ones

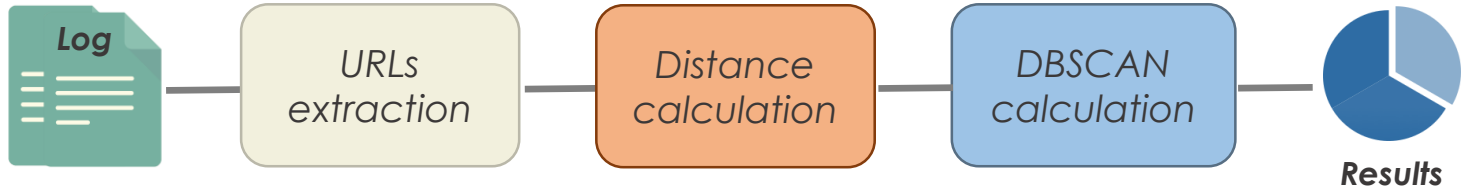
Parameters

- ***Epsilon***, radius of the considered area
- ***Min points***, minimum number of points inside the area

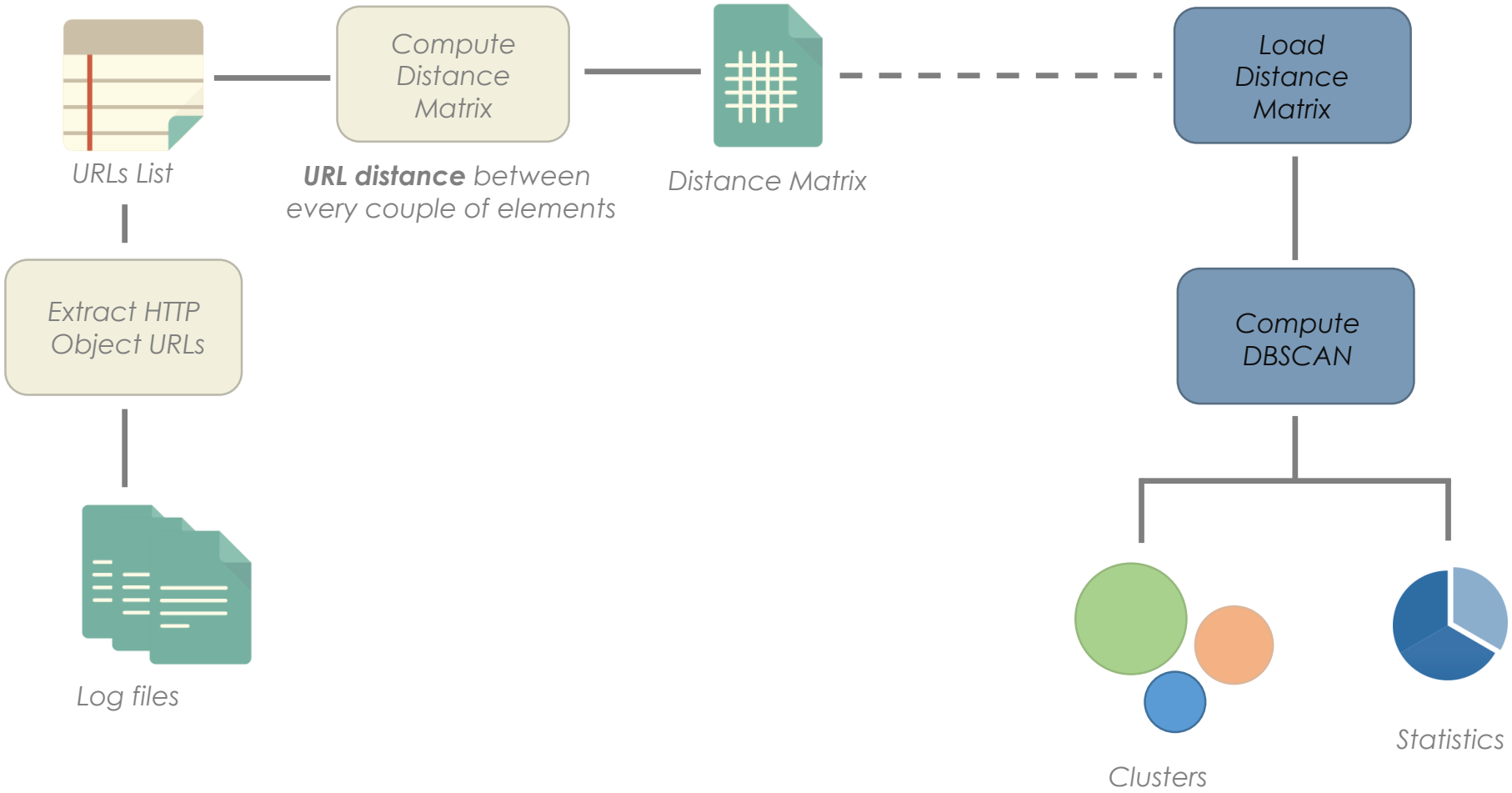


Example of clustering with DBSCAN

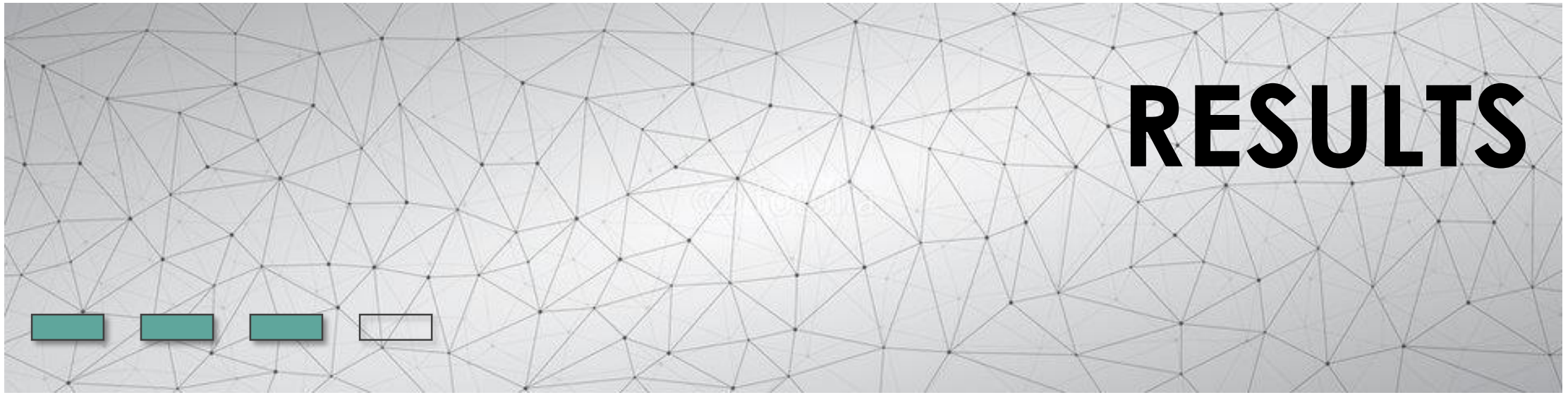
SCHEMA



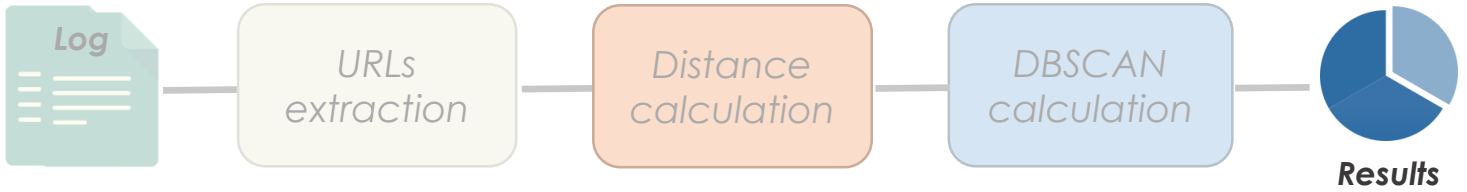
Final schema. Developed in Python.



Part 3



ANALYSIS



About

I focused most of the analysis observing the behavior and the performances of DBSCAN on 14 Hosts infected by TidServ and other 20 randomly selected Hosts.

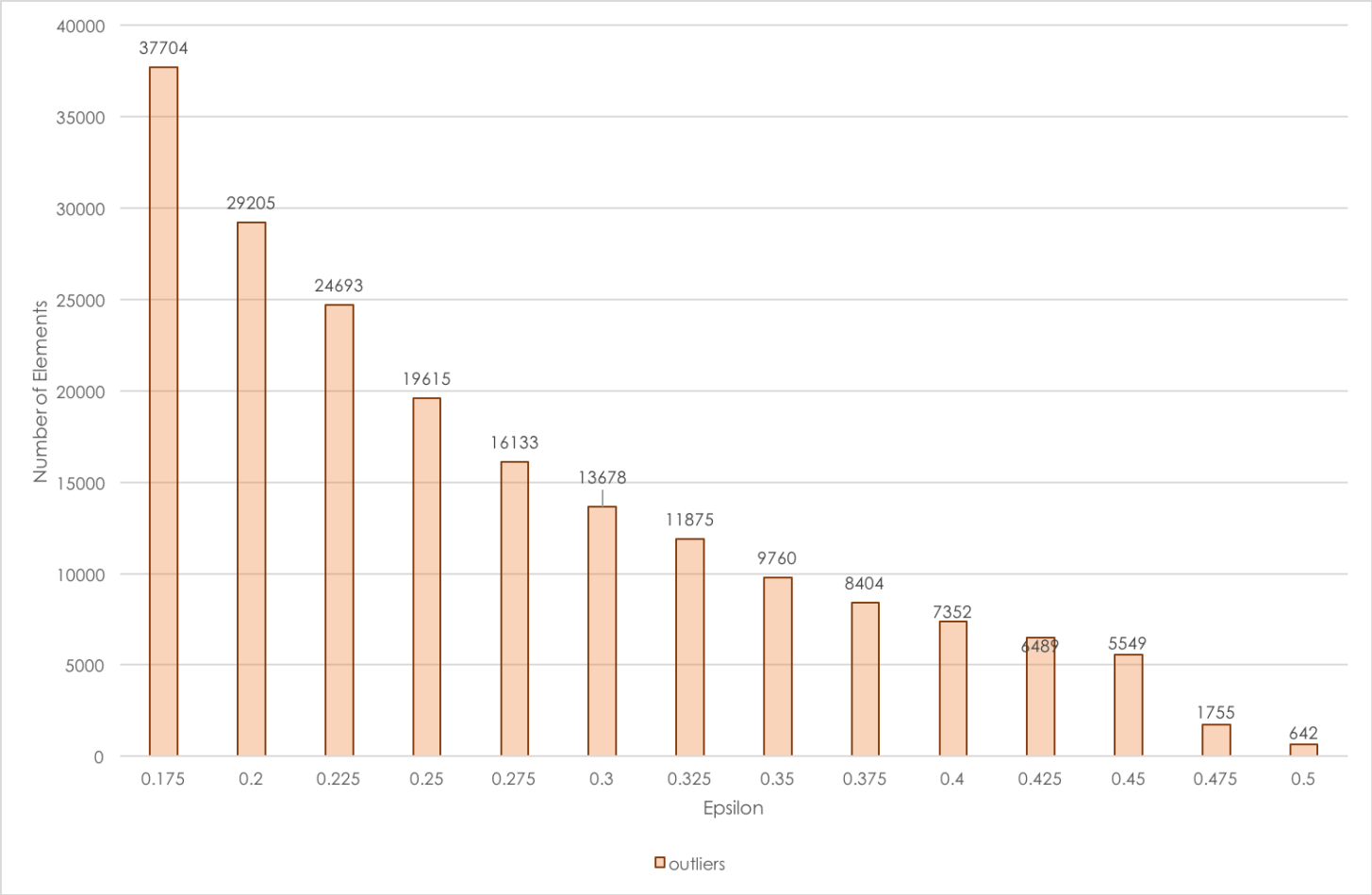
I performed many tests setting *MinPoints* value to 4 (the best value according to previous analysis) and varying *Epsilon*.

DISTINCT URL ELEMENTS

TidServ	228
Other malware	33
Benign	78160
Total	78421

CLUSTERING

Results for 34 Hosts infected by TidServ.



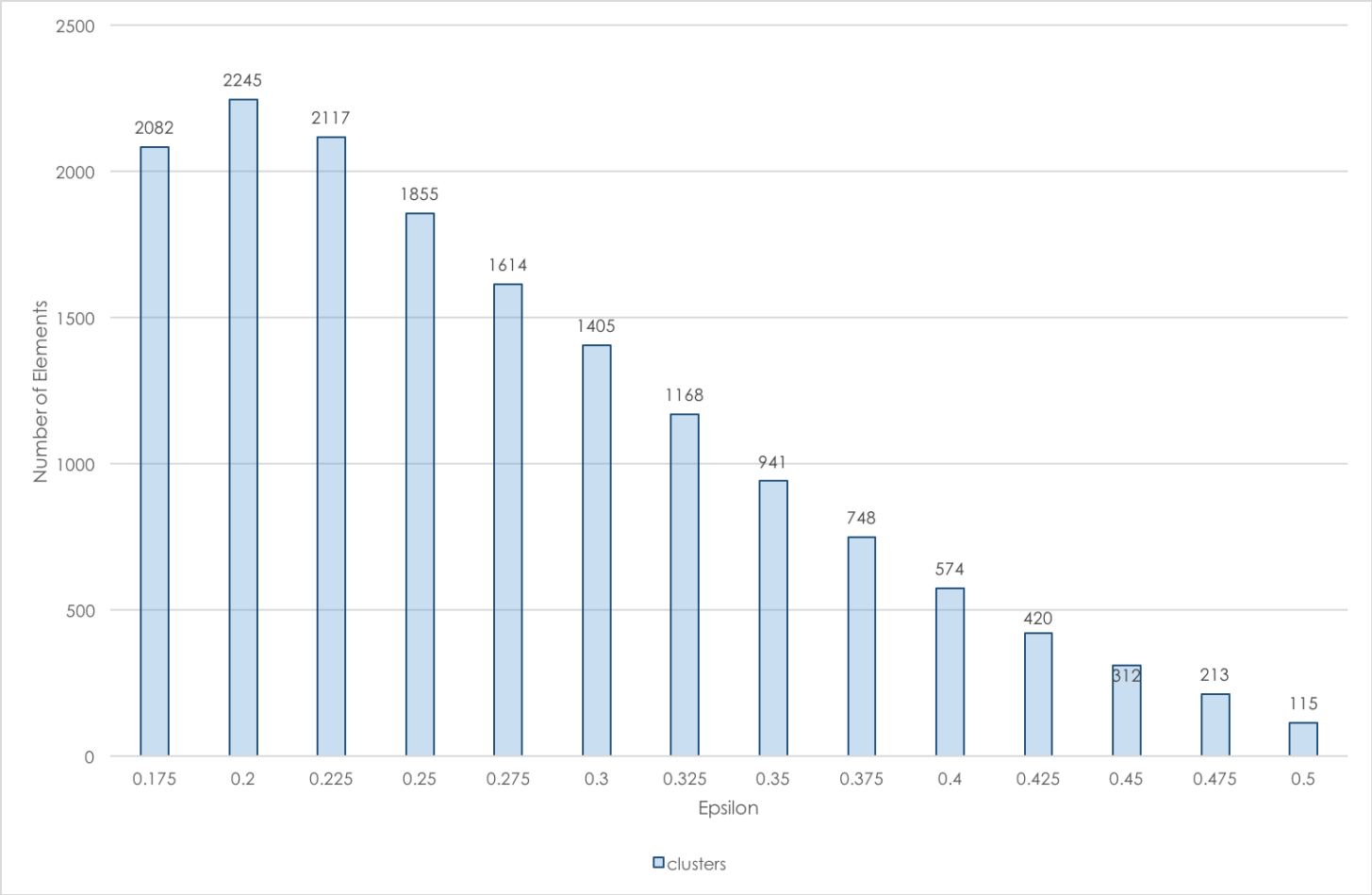
Performance

NUMBER OF OUTLIERS

Decrease in the number of outliers, for *Epsilon* growing.

CLUSTERING

Results for 34 Hosts infected by TidServ.



Performance

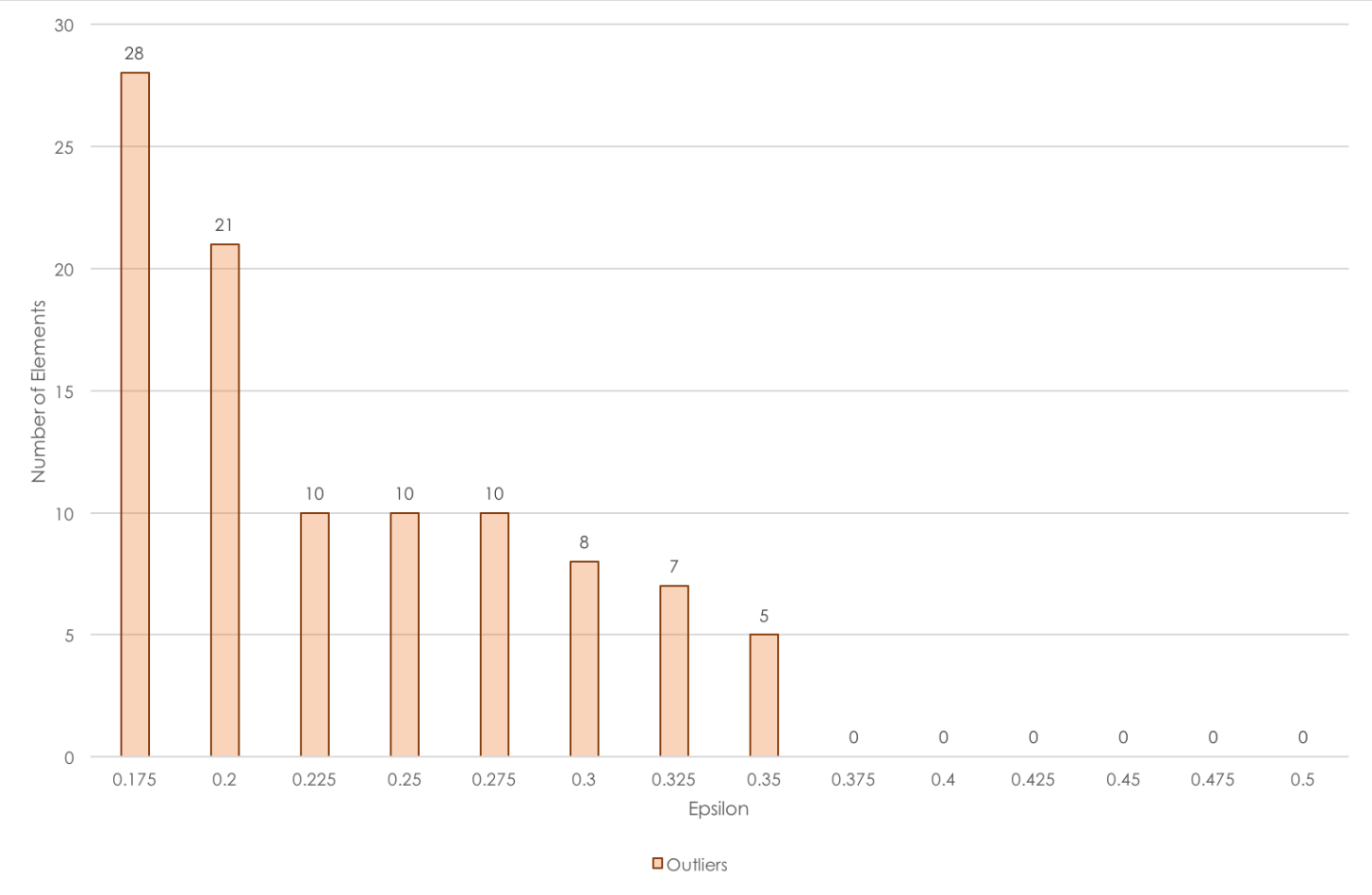
NUMBER OF CLUSTERS

Decrease of the number of clusters starting from *Epsilon* = 0.25, with a halving in the transition from 0.3 to 0.4.

Increase in the number of clusters for *Epsilon* = 0.2 and 0.225, due to the fact that many elements previously considered noise constitute new clusters.

CLUSTERING

Results for 34 Hosts infected by TidServ.



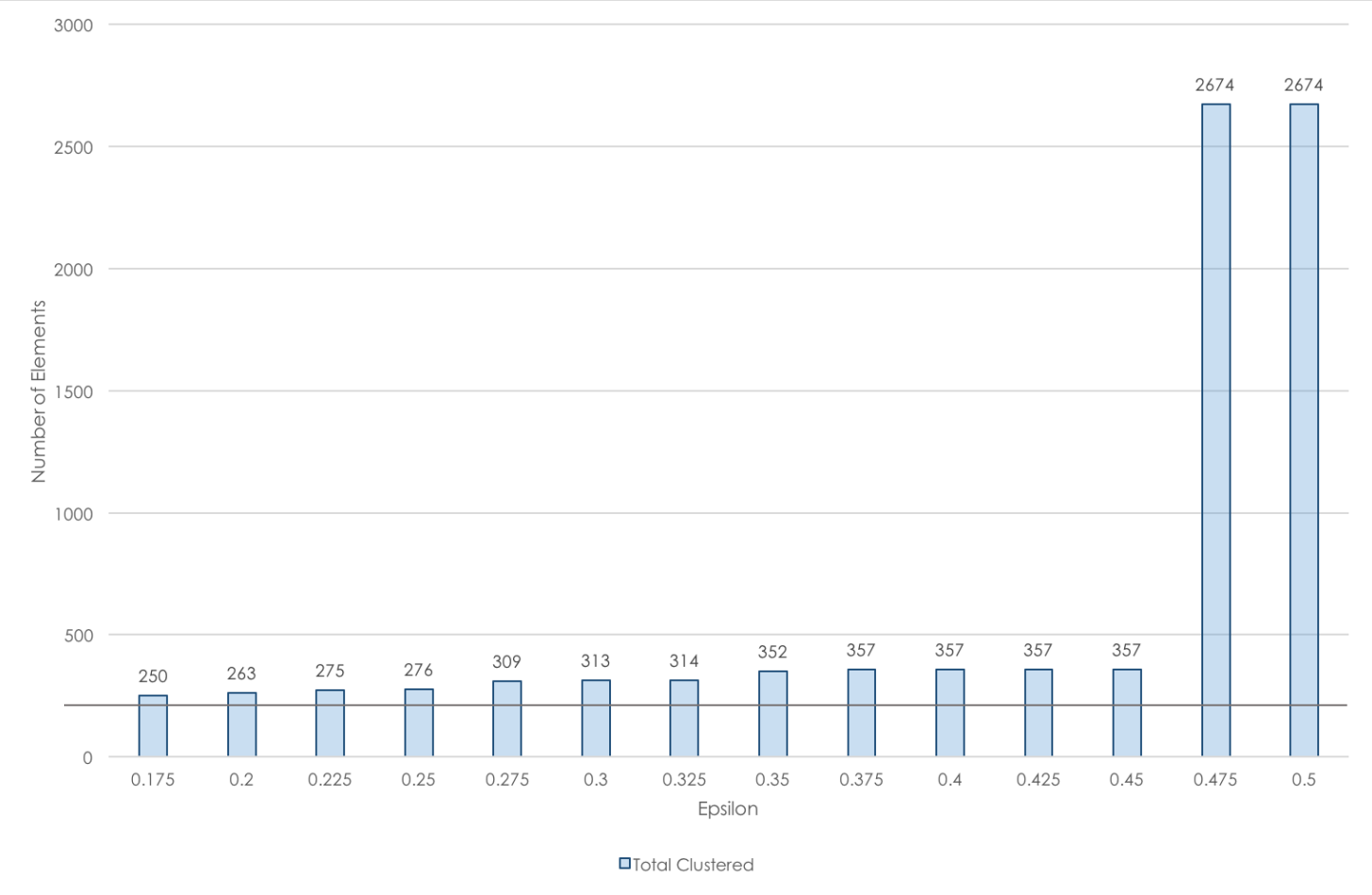
Performance

CLUSTERING RESULTS FOR TIDSERV - OUTLIERS

Decrease in the number of outliers, until reaching 0 for *Epsilon* = 0.4.

CLUSTERING

Results for 34 Hosts infected by TidServ.



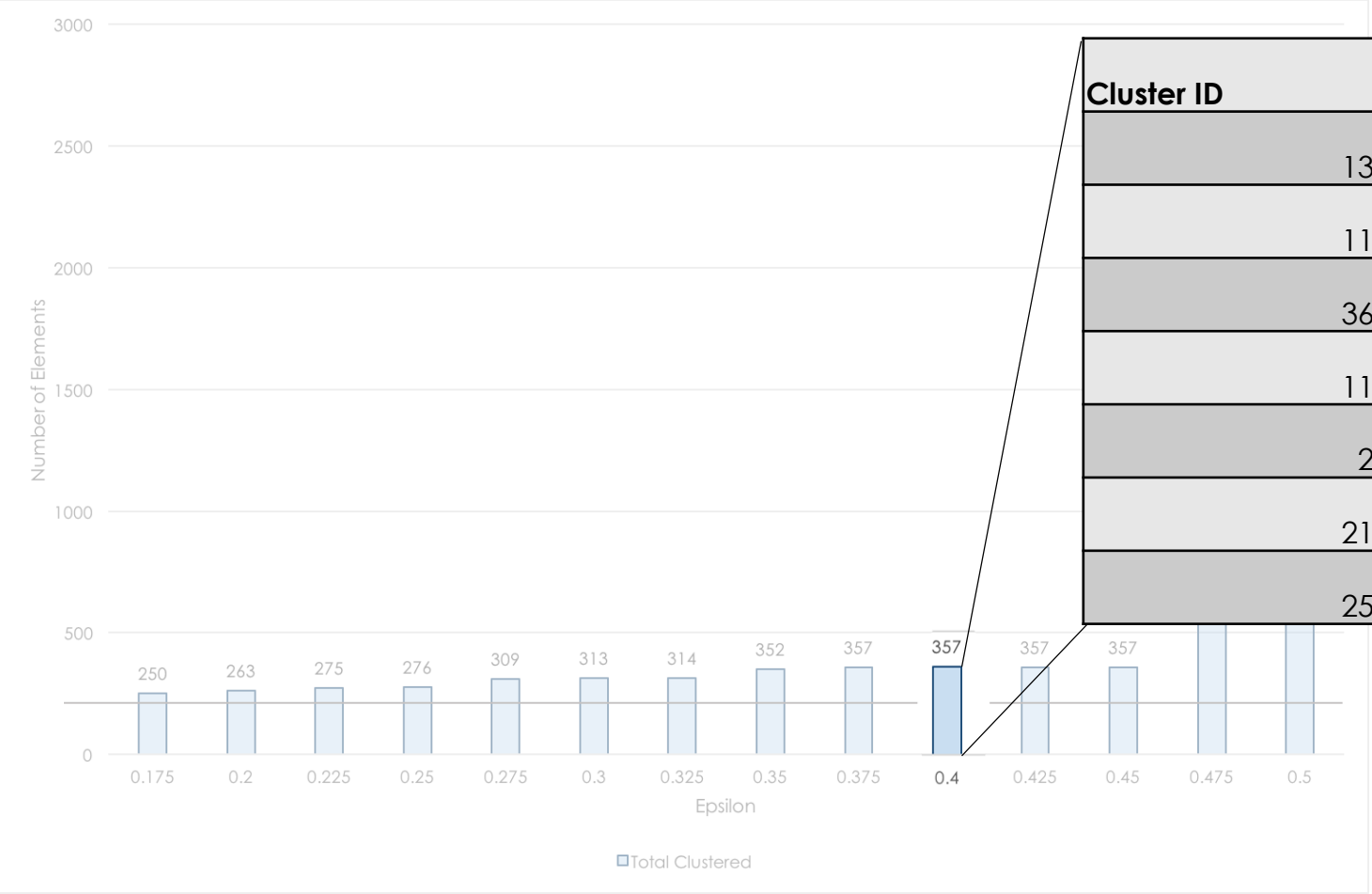
Performance

CLUSTERING RESULTS FOR TIDSERV

Constant and coherent growing of the number of known elements included and ability to aggregate additional not-reported elements.

CLUSTERING

Results for 34 Hosts infected by TidServ.



Cluster ID	TidServ - IDS Count	All elements Count
133	5	5
110	18	32
369	5	6
117	75	79
22	118	192
215	6	6
252	1	37

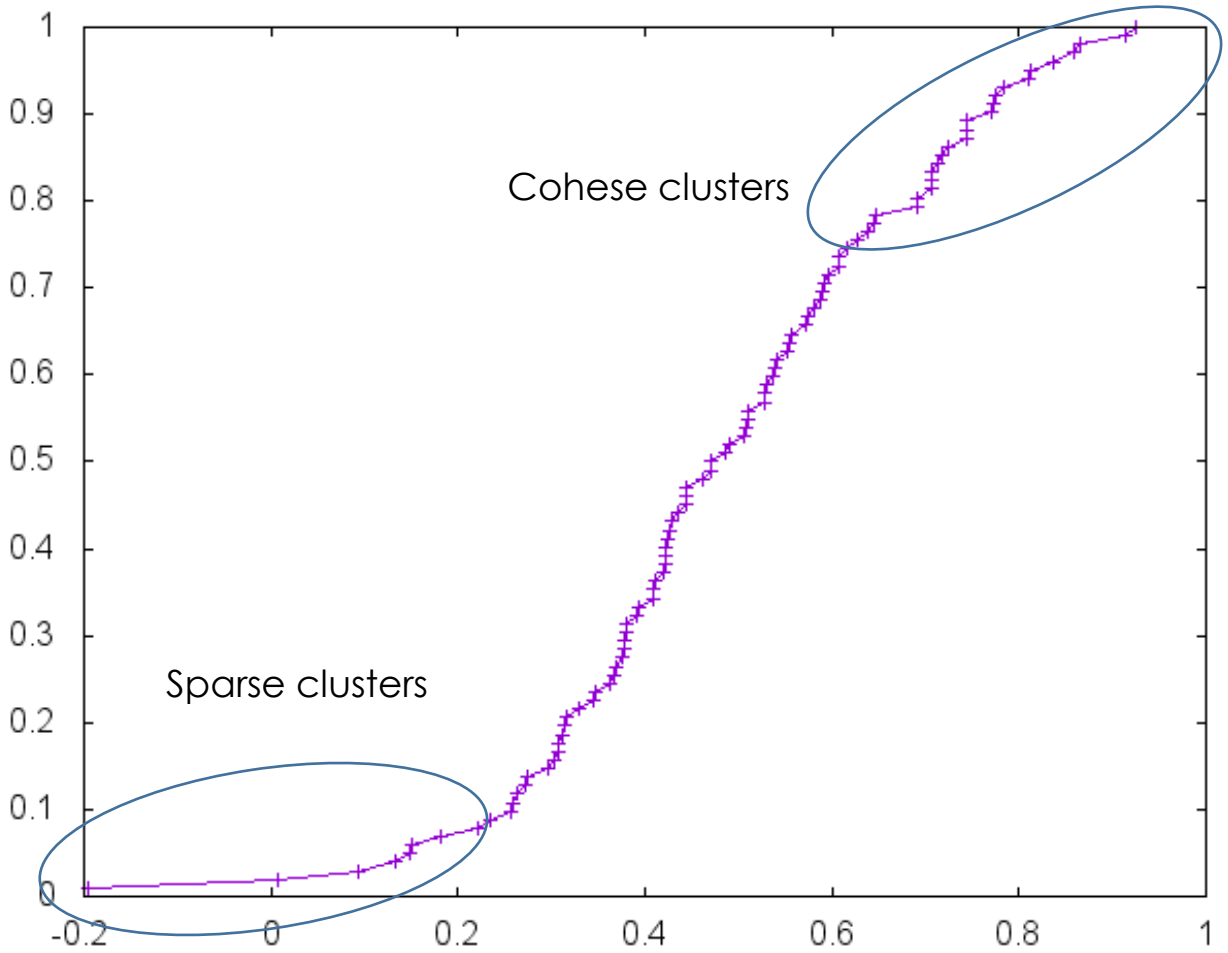
TIDSERV ANALYSIS

Cluster 252 – Compare Elements 17/20

- gnu4oke0r.com/4VY00y9P7Z5xiPs9dmVyPTQuM**CZiaWQ9NWJjNWFiMjE1YjRmN2I4ZjM3OTRmODNkZjhmNWY0ZjFmODZkYjE1YyZhaWQ9MzAwMDEmc2lkPTA**mcmQ9M**CZlbmc9d3d3Lmdvb2dsZS5pdCZxPWxvdWlzIGNydWlzZXM=16h
- lkckclckIii1i.com/TAR3vUsX844qz1c5Y2xrPTIuN**CZiaWQ9NWJjNWFiMjE1YjRmN2I4ZjM3OTRmODNkZjhmNWY0ZjFmODZkYjE1YyZhaWQ9MzAwMDEmc2lkPTAmcmQ9M**A==27g
- lkckclckI1i1i.com/TAR3vUsX844qz1c5Y2xrPTIuN**CZiaWQ9NWJjNWFiMjE1YjRmN2I4ZjM3OTRmODNkZjhmNWY0ZjFmODZkYjE1YyZhaWQ9MzAwMDEmc2lkPTAm**cmQ9M**A==27g
- lkckclckIii1i.com/ZvP1nw3P6z6XLSs7Y2xrPTIuN**CZiaWQ9NWJjNWFiMjE1YjRmN2I4ZjM3OTRmODNkZjhmNWY0ZjFmODZkYjE1YyZhaWQ9MzAwMDEmc2lkPTAmcmQ9M**A==26g
- lkckclckI1i1i.com/ZvP1nw3P6z6XLSs7Y2xrPTIuN**CZiaWQ9NWJjNWFiMjE1YjRmN2I4ZjM3OTRmODNkZjhmNWY0ZjFmODZkYjE1YyZhaWQ9MzAwMDEmc2lkPTAmcmQ9M**A==26g
- lkckclckIii1i.com/yVv4l79D5E7yT8u9Y2xrPTIuN**CZiaWQ9NWJjNWFiMjE1YjRmN2I4ZjM3OTRmODNkZjhmNWY0ZjFmODZkYjE1YyZhaWQ9MzAwMDEmc2lkPTAmcmQ9M**A==18x
- lkckclckI1i1i.com/yVv4l79D5E7yT8u9Y2xrPTIuN**CZiaWQ9NWJjNWFiMjE1YjRmN2I4ZjM3OTRmODNkZjhmNWY0ZjFmODZkYjE1YyZhaWQ9MzAwMDEmc2lkPTAmcmQ9M**A==18x
- lkckclckIii1i.com/3Zh2DpoP583XBvc2Y2xrPTIuN**CZiaWQ9NWJjNWFiMjE1YjRmN2I4ZjM3OTRmODNkZjhmNWY0ZjFmODZkYjE1YyZhaWQ9MzAwMDEmc2lkPTAm**cmQ9M**A==05Z
- lkckclckI1i1i.com/3Zh2DpoP583XBvc2Y2xrPTIuN**CZiaWQ9NWJjNWFiMjE1YjRmN2I4ZjM3OTRmODNkZjhmNWY0ZjFmODZkYjE1YyZhaWQ9MzAwMDEmc2lkPTAm**cmQ9M**A==05Z
- lkckclckIii1i.com/ZaW4pfQP6P4Q7EO9Y2xrPTIuN**CZiaWQ9NWJjNWFiMjE1YjRmN2I4ZjM3OTRmODNkZjhmNWY0ZjFmODZkYjE1YyZhaWQ9MzAwMDEmc2lkPTA**mcmQ9M**A==06c
- lkckclckI1i1i.com/ZaW4pfQP6P4Q7EO9Y2xrPTIuN**CZiaWQ9NWJjNWFiMjE1YjRmN2I4ZjM3OTRmODNkZjhmNWY0ZjFmODZkYjE1YyZhaWQ9MzAwMDEmc2lkPTA**mcmQ9M**A==06c
- lkckclckIii1i.com/SVn4kZCE8Y6MEes8Y2xrPTIuN**CZiaWQ9NWJjNWFiMjE1YjRmN2I4ZjM3OTRmODNkZjhmNWY0ZjFmODZkYjE1YyZhaWQ9MzAwMDEmc2lkPTAm**cmQ9M**A==38A
- lkckclckI1i1i.com/SVn4kZCE8Y6MEes8Y2xrPTIuN**CZiaWQ9NWJjNWFiMjE1YjRmN2I4ZjM3OTRmODNkZjhmNWY0ZjFmODZkYjE1YyZhaWQ9MzAwMDEmc2lkPTAm**cmQ9M**A==38A

SILHOUETTE

Silhouette values distribution for some representative clustering results.



Performance

CALCULATIONS

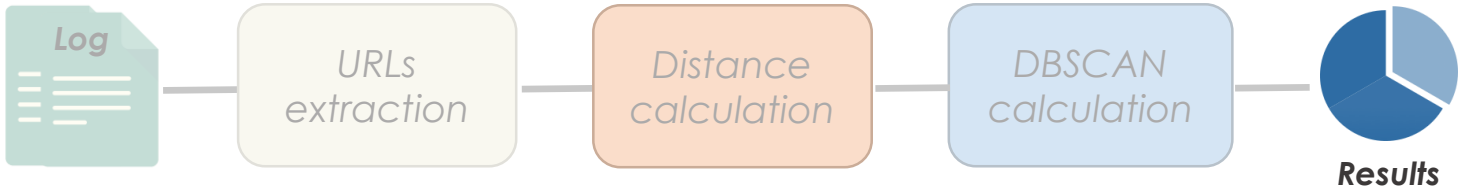
They were taken into analysis clusters with more than 20 elements

It was found that almost all clusters have an average value of silhouette greater than 0

Tidserv's clusters are not those with greater degree of cohesion; their silhouette values range between 0.7 and 0.4

Generally clusters with silhouette > 0 are associated to items created algorithmically. In particular, this behavior is highlighted for silhouette values greater than 0.7

SILHOUETTE



Examples of groupings (Eps = 0.4, MinPts = 4)

skygo_streaming-i.akamaihd.net (551 elements)

Silh = 0.92

ad.doubleclick.net (99 elements)

Silh = 0.91

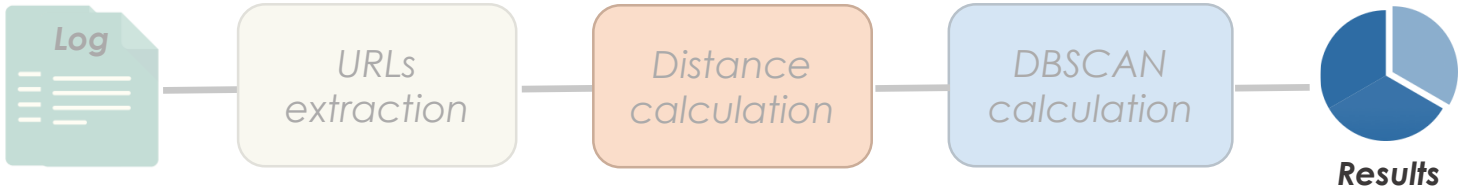
mfdclk001.org (27 elements)

Silh = 0.81

cookex.amp.yahoo.com (61 elements)

Silh = 0.87

SILHOUETTE



Examples of groupings (Eps = 0.4, MinPts = 4)

skygo_streaming-i.akamaihd.net (551 elements)

Silh = 0.92

mfdclk001.org (27 elements)

Silh = 0.81

Video Streaming

(H1) skygo_streaming-i.akamaihd.net
(P1) /hls/live/201
(P2) //20

substrings

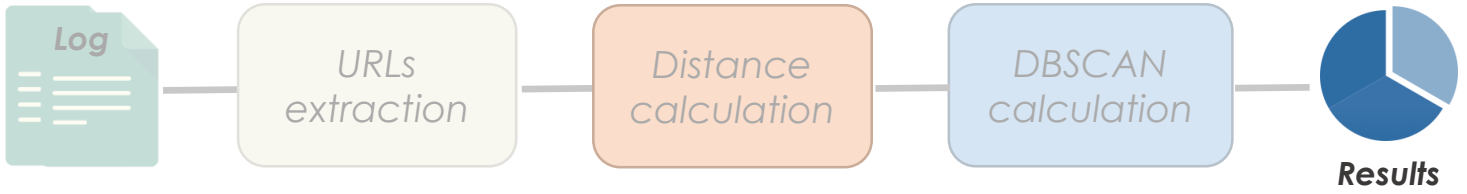
- **(H1) / (P1)** 287 **(P2)**929/latest.ptr
- **(H1) / (P1)** 287 **(P2)** 929/2100/000F9A04.ts
- **(H1) / (P1)** 286 **(P2)** 920/latest.ptr
- **(H1) / (P1)** 286 **(P2)** 920/ 2100/000FA476.ts
- **(H1) / (P1)** 304 **(P2)** 362/latest.ptr
- **(H1) / (P1)** 304 **(P2)** 362/1600/000F9CFF.ts
- **(H1) / (P1)** 304 **(P2)** 362/1600/000F9D01.ts
- **(H1) / (P1)** 287 **(P2)** 929/2100/000F9A06.ts
- **(H1) / (P1)** 287 **(P2)** 929/2100/000F9A08.ts
- **(H1) / (P1)** 287 **(P2)** 929/2100/000F9A0A.ts
- **(H1) / (P1)** 287 **(P2)** 929/2100/000F9A0C.ts
- **(H1) / (P1)** 287 **(P2)** 929/2100/000F9A0E.ts
- **(H1) / (P1)** 287 **(P2)** 929/2100/000F9A10.ts
- **(H1) / (P1)** 287 **(P2)** 929/2100/000F9A12.ts
- **(H1) / (P1)** 287 **(P2)** 929/2100/000F9A14.ts
- **(H1) / (P1)** 287 **(P2)** 929/400/000F9A3D.ts
- **(H1) / (P1)** 287 **(P2)** 929/1100/000F9A2A.ts
- **(H1) / (P1)** 287 **(P2)** 929/400/000F9A3F.ts
- **(H1) / (P1)** 287 **(P2)** 929/1600/000F9A39.ts
- **(H1) / (P1)** 287 **(P2)** 929/600/000F9A59.ts
- **(H1) / (P1)** 287 **(P2)** 929/600/000F9A5B.ts

[…]

= 0.91

n = 0.87

SILHOUETTE



Examples of groupings (Eps = 0.4, MinPts = 4)

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=

substrings

- **(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)**
- **(H1) / (P4)(P5)** 300x25 **(P2)** 242694111/Right **(P3)** 300x25 **(P1)**
- **(H1) / (P4)** .2 **(P5)** 728x9 **(P2)** 867018368/Top **(P3)** 728x9 **(P1)**
- **(H1) / (P4)(P5)** 300x25 **(P2)** 327982228/Right **(P3)** 300x25 **(P1)**
- **(H1) / (P4)** .2 **(P5)** 728x9 **(P2)** 2104424489/Top **(P3)** 728x9 **(P1)**
- **(H1) / (P4)** .2 **(P5)** 728x9 **(P2)** 2099868064/Top **(P3)** 728x9 **(P1)**
- **(H1) / (P4)(P5)** 300x25 **(P2)** 2000194920/Right **(P3)** 300x25 **(P1)**
- **(H1) / (P4)** .2 **(P5)** 728x9 **(P2)** 387141438/Top **(P3)** 728x9 **(P1)**

[…]

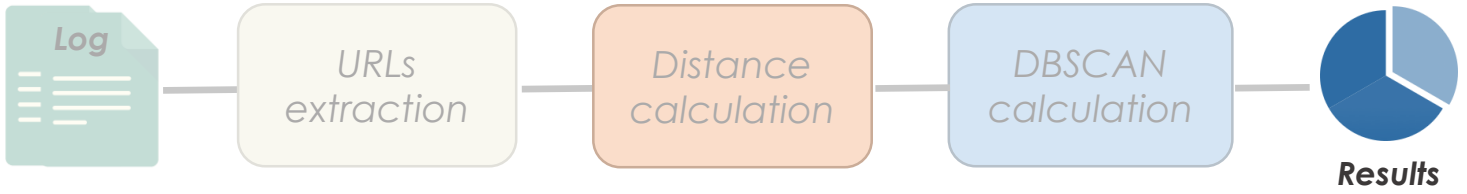
ad.doubleclick.net
(99 elements)

Silh = 0.91

cookex.amp.yahoo.com
(61 elements)

Silh = 0.87

SILHOUETTE



Examples of groupings (Eps = 0.4, MinPts = 4)

Malware

(H1) cookex.amp.yahoo.com
(P1) /*http%3A//ad.yieldmanager.com/imp
(P2) /v2/cexposer/SIG=1

} substrings

- **(H1) / (P2)** 2gbvl6c7 **(P1)**
- **(H1) / (P2)** 3rofdl53 **(P1)**
- **(H1) / (P2)** 429i9hfr **(P1)**
- **(H1) / (P2)** 4vnv3q8e **(P1)**
- **(H1) / (P2)** 3llrt2ba **(P1)**
- **(H1) / (P2)** 4l7rqd9q **(P1)**
- **(H1) / (P2)** 46il2hv3 **(P1)**
- **(H1) / (P2)** 5e9dlgqo **(P1)**
- **(H1) / (P2)** 43rlndbi **(P1)**
- **(H1) / (P2)** 42kkkcmp **(P1)**
- **(H1) / (P2)** 670ekdpa **(P1)**
- **(H1) / (P2)** 429fgkl3 **(P1)**
- **(H1) / (P2)** 3tt2g651 **(P1)**
- **(H1) / (P2)** 7d0ll6co **(P1)**
- **(H1) / (P2)** 7c2fdctr **(P1)**
- **(H1) / (P2)** b5d0bpkl **(P1)**
- **(H1) / (P2)** 366n865c **(P1)**
- **(H1) / (P2)** 34eqqqe3 **(P1)**
- **(H1) / (P2)** 49qh7jna **(P1)**
- **(H1) / (P2)** 48sftpdt **(P1)**
- **(H1) / (P2)** 4kdh9i43 **(P1)**
- **(H1) / (P2)** 4jallfpf **(P1)**
- **(H1) / (P2)** 4h4i0g74 **(P1)**
- **(H1) / (P2)** 361vg7cb **(P1)**
- **(H1) / (P2)** 7hur6jlt **(P1)**
- **(H1) / (P2)** 7g6nn14s **(P1)**
- **(H1) / (P2)** m3ikrg5f **(P1)**
- **(H1) / (P2)** b5kebu4s **(P1)**
- **(H1) / (P2)** aneoasrb **(P1)**
- **(H1) / (P2)** 3pncvt94 **(P1)**

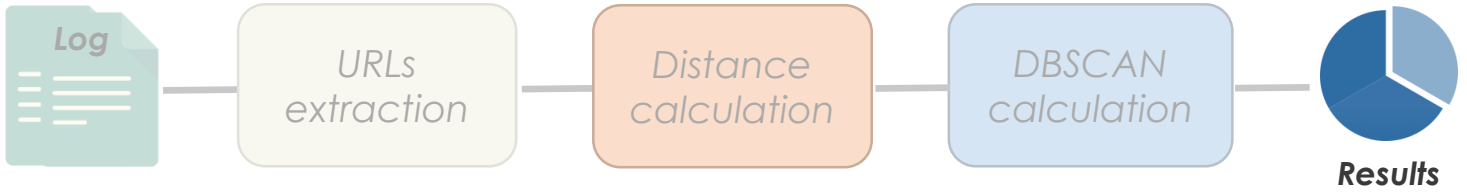
ad.doubleclick.net
(99 elements)

Silh = 0.87

cookex.amp.yahoo.com
(61 elements)

Silh = 0.87

SILHOUETTE



Examples of groupings (Eps = 0.4, MinPts = 4)

skygo_streaming-i.akamaihd.net
(551 elements)

Silh = 0.92

mfdclk001.org
(27 elements)

Silh = 0.81

Malware

(H1) mfdclk001.org
(P1)
Y2xrPTEuMjEmYmlkPTUwMGFhNzVjLWY1ZTUtNDhhOC05ZjlkLWY2ODQ3NGYzOGQwZCZhaWQ9MTAwMTAmc2lkPTAmcmQ9MTcuMTEuMjAxMQ==

substrings

- **(H1)** /UVw07aeI7p4qVcS6 **(P1)** 26c
- **(H1)** /wZI1ELDd7N5quws3 **(P1)** 26A
- **(H1)** /SZY35xzx6x4q5Ks1 **(P1)** 26x
- **(H1)** /7V10LUdI7Z5mxcS2 **(P1)** 25A
- **(H1)** /LZK2BDxe5k4mQiO2 **(P1)** 17g
- **(H1)** /WZb3fvgI643q33U7 **(P1)** 05c
- **(H1)** /AAc2B3xD6b6YLhs5 **(P1)** 17x
- **(H1)** /Ca90Id3P725XYAc3 **(P1)** 16g
- **(H1)** /IVy0fWZp7g5qgFU5 **(P1)** 16k
- **(H1)** /oAO2pxHx7x5MsUo4 **(P1)** 26h
- **(H1)** /wkD3ca5x785XbDC4 **(P1)** 16x
- **(H1)** /qat3k6Mp775q5rS4 **(P1)** 06h
- **(H1)** /uai0hjnp6M3yGSS4 **(P1)** 15c
- **(H1)** /cVd1TQGe5j4mwEC0 **(P1)** 07c
- **(H1)** /5kP3xEkx7x5yKOO0 **(P1)** 06h
- **(H1)** /wac2AMsx6b5yUkU2 **(P1)** 05h
- **(H1)** /oZb3ix1l5h4Yn6S4 **(P1)** 17x
- **(H1)** /4kP1tYcD5w6XzUo6 **(P1)** 25g
- **(H1)** /XAB1G2qP6J5JPjU8 **(P1)** 25c
- **(H1)** /1Zk2UQ3P634jzHc1 **(P1)** 15c
- **(H1)** /YZi17W8L5G4YeBU1 **(P1)** 26k
- **(H1)** /WAy1u5ux6f5Qiqs2 **(P1)** 16h
- **(H1)** /gKr0IsTE7g5YnES0 **(P1)** 05x
- **(H1)** /gzi0odrd7e6jyxo4 **(P1)** 27A
- **(H1)** /ca70FnQE785me4S4 **(P1)** 05g
- **(H1)** /xVD02q2D6w6MVoC1 **(P1)** 15A
- **(H1)** /fAM37YIL714QLuc5 **(P1)** 07k