ScuDo
Scuola di Dottorato – Doctoral School
WHAT YOU ARE, TAKES YOU FAR

XXXII Cycle

Big Data and Machine Learning Methodologies for Network Traffic Monitoring and Cyber Security

Andrea Morichetta

Supervisor: Prof. Marco Mellia

Research context and motivation

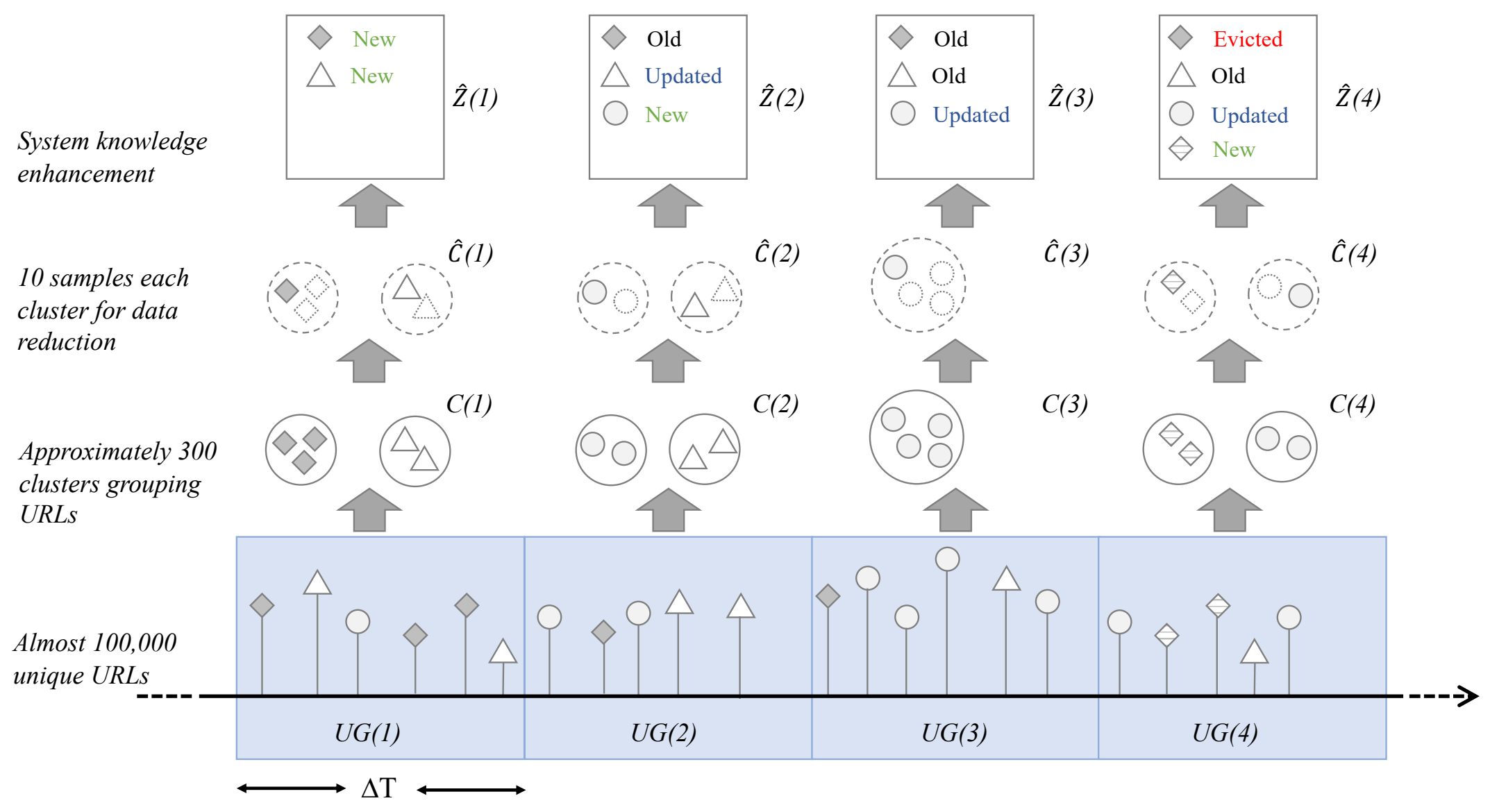
- These years have seen the proliferation of applications and services that rely on HTTP, thus increasing the complexity of the Web and consequently its analysis.
- What is more, cybercriminals in the years have deployed more sophisticated and stealthy ways to generate and spread their malicious contents through HTTP traffic.

- Essential to ease network monitoring, with a logical view of the traffic instead of log processing
- Develop a systematic analysis tool that allows to periodically check changes and novelties in the traffic, in order to detect interesting and possibly suspicious URLs

Adopted methodologies

LENTA: Longitudinal Exploration for Network Traffic Analysis

- **URL Distance**: based on edit distance, i.e., given two strings express the number of edits necessary to let one string equal to the other, but normalized and with different weights for edit operations.
- **Self Tuning Iterative DBSCAN**: Iteratively run DBSCAN, each time using a different value of the parameter ε, accepting only those clusters that, after an evaluation, result to be well-shaped, according to Silhouette quality metric.
- **Percentile Sampling**: Performed on clusters to ease the comparison between clusters, to reduce computational complexity and keep traffic digests
- **System Knowledge Enhancement**: Using URL distance, new clusters are compared to the ones in the System Knowledge and added to it if the distance to the closest old cluster is higher than a threshold α. Old clusters that are not anymore active are evicted.

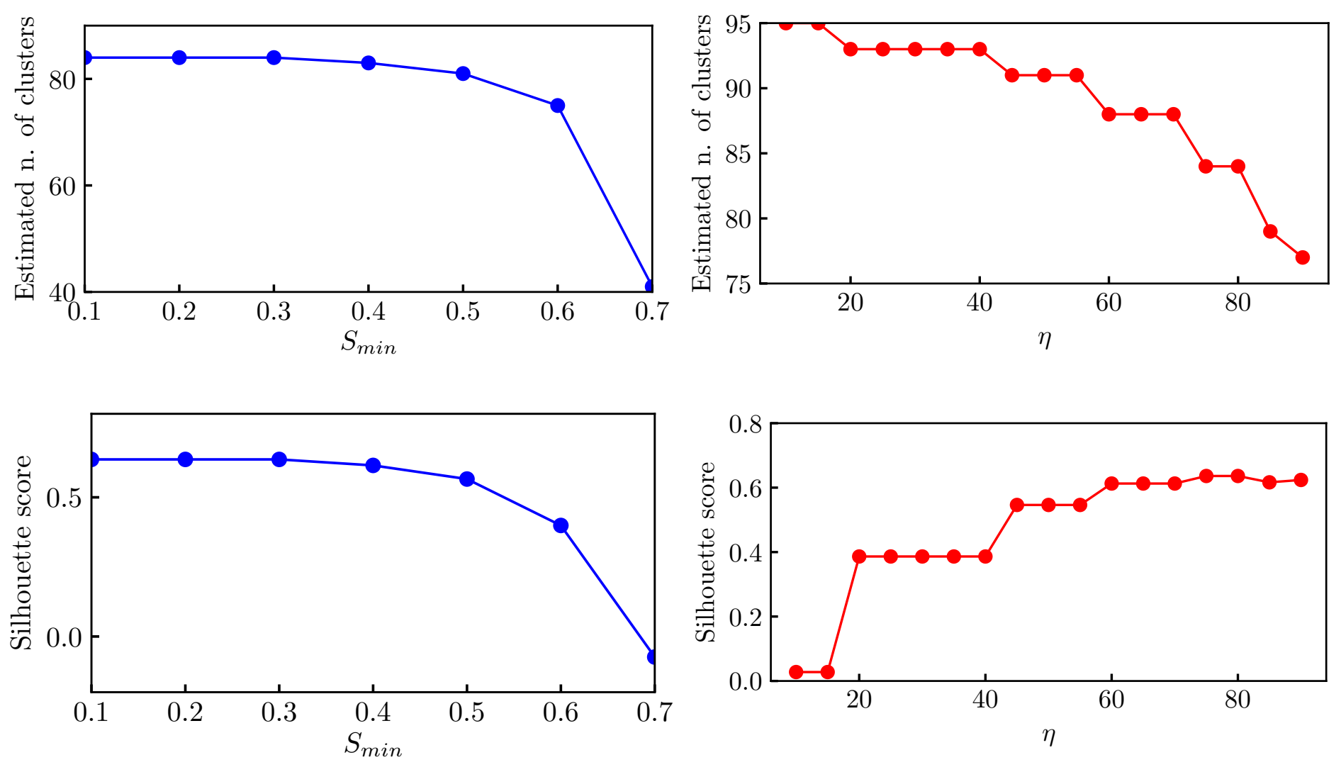


Submitted and published works

- 2019 [Journal Paper] *Morichetta, Andrea*; Mellia, Marco, LENTA: Longitudinal Exploration for Network Traffic Analysis from Passive Data, in: IEEE Transaction on Network and Service Management, 2019
- 2019 [Journal Paper] D'Alconzo, Alessandro; Drago, Idilio; *Morichetta, Andrea*; Mellia, Marco; Casas, Pedro, A Survey on Big Data for Network Traffic Monitoring and Analysis, in: IEEE Transaction on Network and Service Management, 2019
- 2019 [Journal Paper] *Morichetta, Andrea*; Mellia, Marco, Clustering and Evolutionary Approach for Longitudinal Web Traffic Analysis, in: PEVA - Performance Evaluation, 2019
- 2019 [Conference Paper] *Morichetta, Andrea*; Trevisan, Martino; Vassio, Luca, Characterizing Web Pornography Consumption from Passive Measurements, in: Passive and Active Measurement, 2019
- 2018 [Conference Paper] Faroughi, Azadeh; Javidan, Reza; Mellia, Marco; *Morichetta, Andrea*; Soro, Francesca; Trevisan, Martino, Achieving Horizontal Scalability in Density-based Clustering for URLs, in: Proceedings of the 2018 IEEE International Conference on Big Data, Seattle, WA, 2018
- 2018 [Conference Paper] *Morichetta, Andrea*; Mellia, Marco, LENTA: Longitudinal Exploration for Network Traffic Analysis, in: 30th International Teletraffic Congress (ITC 30), Vienna, Austria, 2018
- 2017 [Conference Paper] Ciociola, Alessandro; Cocca, Michele; Giordano, Danilo; Mellia, Marco; *Morichetta, Andrea*; Putina, Andrian; Salutari, Flavia, UMAP: Urban Mobility Analysis Platform to Harvest Car Sharing Data, in: Proceedings of the IEEE Conference on Smart City Innovations, 2017
- 2016 [Conference Paper] *Morichetta, Andrea*; Bocchi, Enrico; Metwalley, Hassan; Mellia, Marco, CLUE: Clustering for Mining Web URLs, in: 28th International Teletraffic Congress (ITC 28), Würzburg, Germany, 2016

Novel contributions

Tuning Parameters for IDBSCAN



IDBSCAN needs, beyond MinPoints, only two parameters:

- η that specifies the percentage of points to clustering, used to find ε, according to the K-dist graph
- S_{min} that sets the minimum value for the silhouette quality metric, in order to accept the cluster in the final result $\mathcal{C}$= $\cup_j \{C_j | S(C_j) > Sm_{in}\}$

Comparing IDBSCAN with other Density-based algorithms

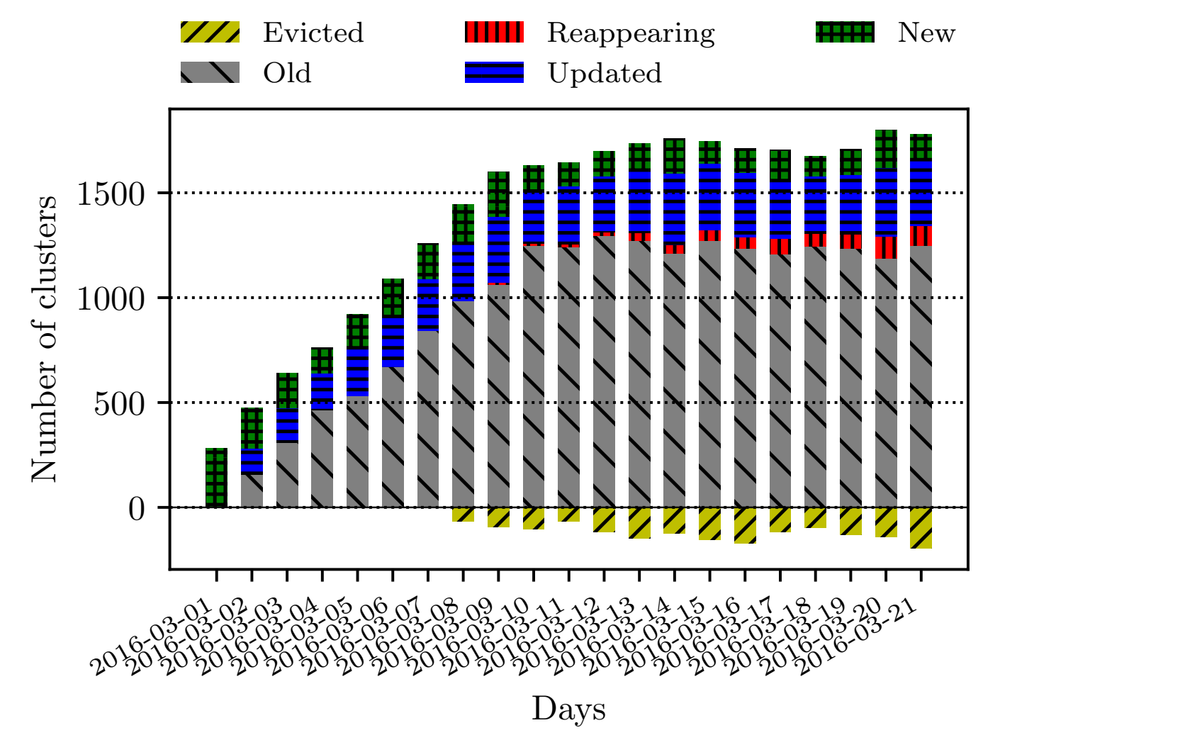
IDBSCAN has been compared to other density based algorithms, using both artificial and real datasets

Algorithm	Percentage clustered $(S(C) \geq S_{min})$	N. clusters	Size largest cluster	$S(C)$ largest cluster	Size smallest cluster	$S(C)$ smallest cluster	Mean cluster size	25% cluster size	50% cluster size	75% cluster size	Computational Time (s)
DBSCAN	45.14	238	15246	-0.15	16	0.41	148.28	5.0	16.5	57.75	113.70
HDBSCAN	53.16	563	4360	0.52	20	-0.17	82.34	28.0	41.0	65.0	218.43
OPTICS	44.65	227	15214	-0.15	2	0.44	205.45	27.0	44.0	89.0	8175.82
CANF	29.67	233	15946	-0.09	2	0.84	148.28	5.0	16.5	57.75	1500.73
IDBSCAN	**55.55**	**283**	**4359**	**0.52**	**12**	**0.41**	**147.87**	**27.0**	**44.0**	**83.0**	**843.63**

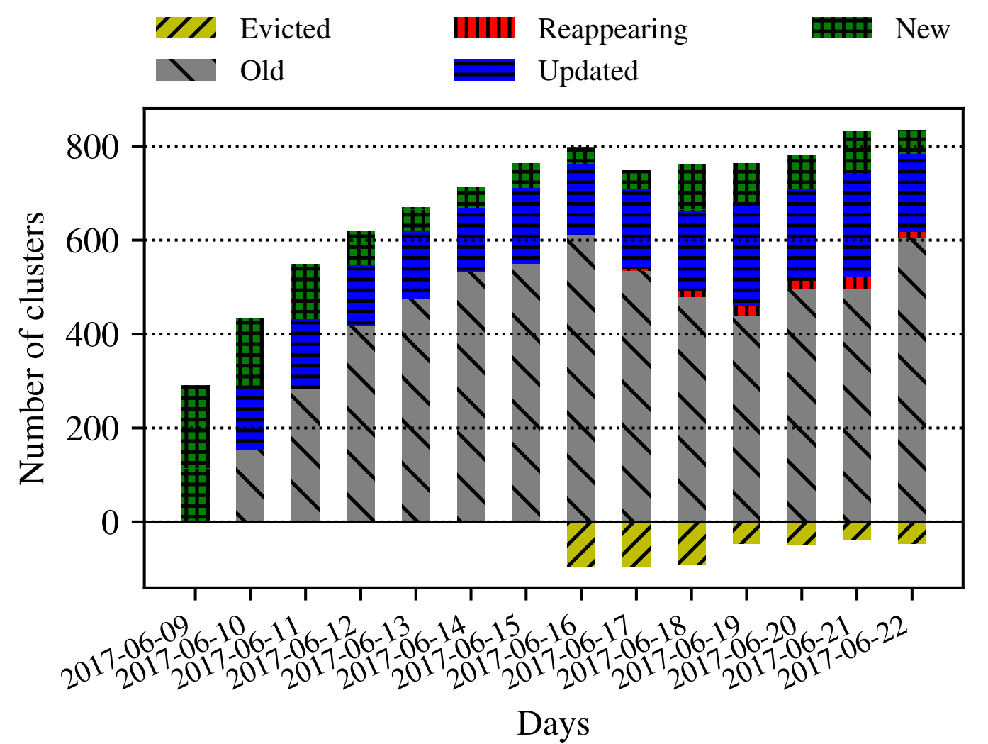
Comparison of different density-based algorithms over one day of HTTP traffic.

Trying LENTA over different datasets

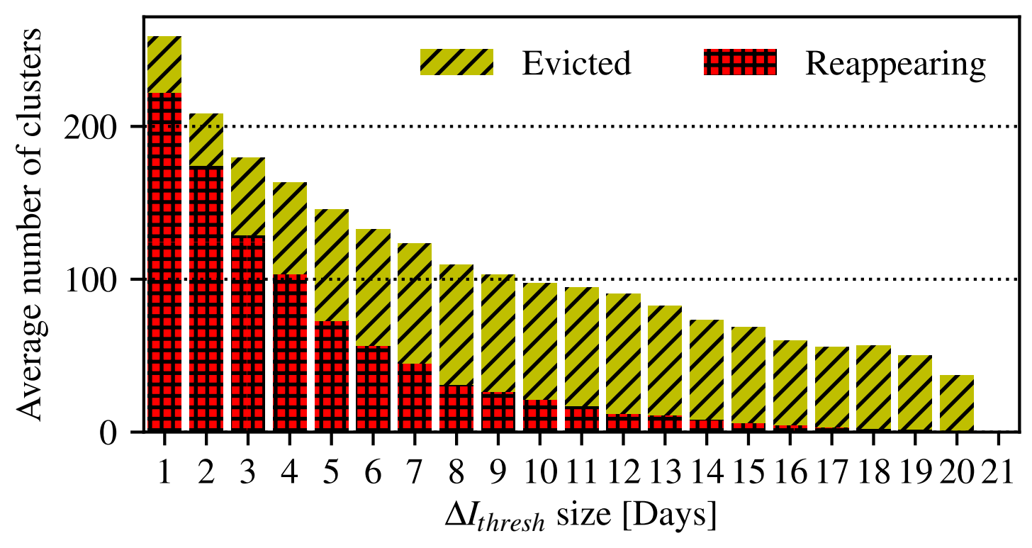
LENTA system has been tested over different datasets, both considering the activity of users as a whole and singularly. Different tests has been performed, checking the capability of the system to adapt over time, and checking the capability of understanding user behaviors.



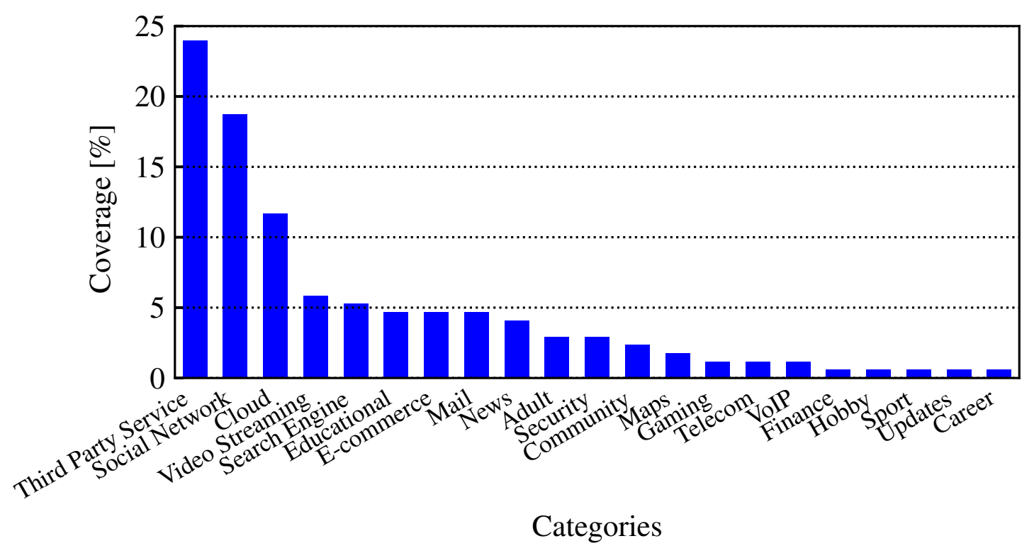
System Knowledge Enhancement for three weeks of HTTP traffic, users aggregated.



System Knowledge Enhancement for two weeks of HTTP+HTTPS traffic, users considered individually.



Number of evicted and reappearing clusters given different time windows as thresholds.



Most popular categories extracted from the clusters visited by at least two users.

Future work

- Explainable AI: explain better the functioning and the decisions provided by machine learning models, such that human experts can understand these.
- Work in the direction of providing better insight into the knowledge discovery process in addition to studying and developing systems which combine this field with unsupervised learning.

List of attended classes

- 02LWHRV – Communication (07/06/2017, 1)
- 01QTEIU – Data mining concepts and algorithms (06/04/2017, 4)
- 01PJMRV – Etica informatica (05/05/2017, 4)
- 01QSAIU – Heuristics and metaheuristics for problem solving (11/05/2017, 4)
- 01RZTRV – Il criterio di responsabilità nella ricerca e nell'innovazione 1 (06/06/2017, 4)
- 01RZURV – Il criterio di responsabilità nella ricerca e nell'innovazione 2 (06/06/2017, 4)
- 01RQXRV– Pattern recognition and neural networks (05/05/2017, 8)
- 01RISRV – Public speaking (07/06/2017, 1)
- 01QORRV - Writing Scientific Papers in English (08/06/2017, 4)
- 02RHORV - The new Internet Society: entering the black-box of digital innovation (25/07/2017, 1)
- 01RELKG – Probabilità applicata e machine learning (09/07/2018, 6)

POLITECNICO DI TORINO

PhD program in
Electrical, Electronics and Communications Engineering