Research context and motivation

• These years have seen the proliferation of applications and services that rely on HTTP, thus increasing the complexity of Web and consequently its analysis.
• What’s more, cyber criminals in the years have deployed more complex and stealth 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.

Addressed research questions/problems

• Every day hundreds of thousands of unique URL are generated in user navigation.
• How to automatically reduce this amount of traffic creating meaningful groups?
• How to let the grouping technique being suitable for different kinds of URLs?
• How to scale up to a big data problem?
• How to check the occurrence of new traffic and how to build an history of the previous collected information?

Adopted methodologies

• URL Distance: based on edit distance, i.e., given two strings express the number of edit necessary to let one string equal to the other, but normalized and with different weights for edit operations.
• Self Tuning Iterative DBSCAN: extraction of the best value of the DBSCAN (density-based clustering algorithm) distance algorithm, while processing the clustering. Recursive clustering of bad formed clusters.
• Percentile Sampling: Performed on clusters to ease the comparison between clusters, to reduce computational complexity and keep traffic digest.
• 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.

Novel contributions

• LENTA, a methodology for the fast identification of HTTP-based services based by looking at URLs string similarities.
• Implements a self-learning methodology to automatically associate previously observed services and identify new traffic generated by possibly suspicious applications.
• Reduce the amount of traffic to be manually checked.
• Ease the observation of changes in the network behavior.
• It extracts well-formed URLs clusters which greatly simplify the identification of possibly malicious and undesired traffic.
• Promising results in identifying anomalies in the traffic.

Future work

• Focus on HTTPS traffic to have a complete view on the network activities.
• Extend big data approaches to all the stage of the system, to scale the analysis.
• Application of LENTA over different lexical features, e.g., hostname in DNS queries or user agents in HTTP requests.

List of attended classes

• 02LWHRV – Communication (07/06/2017, 1)
• 01QTEIU – Data mining concepts and algorithms (06/04/2017, 4)
• 01PJMRR – 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)
• 01RQXVR– Pattern recognition and neural networks (05/05/2017, 8)
• 01RISRV – Public speaking (07/06/2017, 1)
• 01ORRVR – 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 (03/09/2018, 6)

Submitted and published works