Machine Learning and Big Data Approaches for Automatic Web Traffic Monitoring
Summary

Andrea Morichetta

December 19, 2019

The analysis and monitoring of network traffic are of fundamental importance for numerous activities related to its management. These activities are involved in quality control of the service, support the planning and updates of the network based on traffic load, and contribute to the development of advanced security systems and the identification of malicious attacks. Therefore, approaches oriented to the processing of data represented by traffic traces are appropriate to understand the conditions and behavior of the network.

Particularly critical is the analysis of the Hypertext Transfer Protocol (HTTP) traffic. The last years have seen the proliferation of applications and services that rely on HTTP. The complexity of the Web is increased and, consequently, its analysis. What is more, cyber-criminals in the years have deployed more sophisticated and stealthy ways to generate and spread their malicious content through HTTP traffic. In this direction, many researchers and companies are focusing on data analysis and machine learning techniques. Many solutions have been developed, but often pinpointing just particular problems.

The thesis, therefore, proposes to provide a generic methodology for monitoring HTTP traffic; in particular, it aims to identify new services, anomalies, and suspicious traffic, looking at URLs. The Uniform Resource Locator (URL) is a unique address for a particular web resource, such as an image, a video or a HyperText Markup Language (HTML) file and is characterized by two components: the hostname, or the name of the server that owns the object and the object path or the name (and path) of the object. We have seen that, in general, malicious URLs tend to have characteristics that make them visually different from benign ones. For example, malicious organizations tend to use artificially generated hostnames composed of random strings that do not contain common or easily memorable names, as is the practice for legitimate time organizations, to avoid detection by black-lists. In general, the URLs contain essential information about the related services, creating in that way structures that are identifiable.

In detail, the work of this thesis develops the idea mentioned above, addressing the following research questions. The first one is (i) how to automatically reduce the amount of traffic, creating meaningful groups, (ii) how to let the grouping technique being adaptive for different kinds of data, i.e., URLs, without a constant need to manually tune the parameters, (iii) how to scale up to a big data problem, (iv) how to check the occurrence of new traffic and how to build a history of the previous collected information.

This thesis presents a self-tuning clustering solution, as grouping technique, called Iterative DBSCAN. It consists of iteratively run DBSCAN, a popular clustering algorithm, each time using a different value of the input parameters, in order to extract clusters that, after an evaluation, result to be well-shaped, according to quality metrics. To group URLs, I considered them as strings, sequences of characters using metrics that allow measuring the difference between two sequences.

Clustering execution is, however, a computationally demanding task, especially with complex distance functions. This thesis aims at untieing the distance computation, from the algorithm execution, to overcome this performance bot-

tleneck. This approach, together with the gains of distributed platforms like Apache Spark or MapReduce, guarantees a faster execution of the algorithms, together with more flexibility in the choice of the clustering method.

In order to analyze the evolution of the traffic over time, this thesis presents the implementation of a self-learning methodology, where the system grows its knowledge, which is used in turn to automatically associate traffic to previously observed services, and identify new traffic generated by possibly suspicious applications. The whole system takes the name of LENTA - Longitudinal Exploration for Network Traffic Analysis.

The developed methodologies are essential tools for the network administrator, be it a corporate network or a provider. The operator will be able to generate clusters starting from the URLs contacted by the employees of the company (or ISP customers) and, starting from this aggregate view, identify the activities related to malicious behavior. Following this analysis, the administrator can apply filters on these unwanted contents within the network. These approaches ensure greater security against malicious attacks, for the network itself and for the hosts that make it up, without affecting the quality of the user's navigation. Furthermore, the proposed methodologies let the analysts easily observe changes over time in network traffic, identify new services, and unexpected activities. The work is applied over HTTP and HTTPS data. The former case makes use of passive traces, while the latter is the outcome of data collected using a proxy installed on users' devices. In particular, this second scenario requires specific care concerning privacy and shows the potential of the proposed techniques in an enterprise context.