A hybrid approach to allow unsupervised clustering of both connections and hosts in networked data

Azqa Nadeem[1,*], Mark Patrick Roeling[1,2,*], and Sicco Verwer[1]

[1] *Delft University of Technology, The Netherlands*
[2] *University of Oxford, United Kingdom*
[*] *Authors contributed equally. Corresponding author m.p.roeling@tudelft.nl*

Abstract

The data-volumes generated by inter-connected devices present a challenge for network analysis. Since the number of connections can far exceed the number of hosts, analysis often requires data reduction. To this end, a plethora of network clustering algorithms have been proposed, but studies usually model either hosts or connections. We propose a combination of connection-clustering and host-classification by using two popular state-of-the-art unsupervised clustering algorithms, Hierarchical Density-based Clustering (HDBScan) and Stochastic BlockModels (SBM). We used data from the Stratosphere IPS project (CTU-Malware-Capture-Botnet-91) that includes malicious and benign hosts.

Network traffic from different hosts was split into unidirectional connections ($a \rightarrow b$). Each connection was represented by 4 sequential features to ensure the capture of temporal information (packet size, inter-arrival time, source and destination ports). An optimized packets-threshold was used to ensure effective behavioural modeling, after which the filtered connections were clustered with HDBScan. The resulting class labels (for every connection) were fed into a SBM, using the posterior probability as a covariate to accommodate imperfect class assignment. The SBM was also fitted on the raw features to investigate the merit of using HDBScan output as SBM-input.

Model fitting commended a 5-class solution for connection-clustering and 4-class solution for host-clustering. Applying the SBM directly to features resulted in a 2-class solution, indicating that connection clustering increased SBM sensitivity. Misclassification of the nodes was also lower in the combined method. Further investigation of class assignment revealed clusters of behaviourally similar connections (see Figure 1). Conjointly, 88.5% of nodes were labelled correctly, albeit that classification was less accurate for benign hosts.

Conclusively, this study provides a rationale for using clustered connections as input for host-classification in this context, by sequential modeling of two powerful clustering methods, without the need for labelled data.

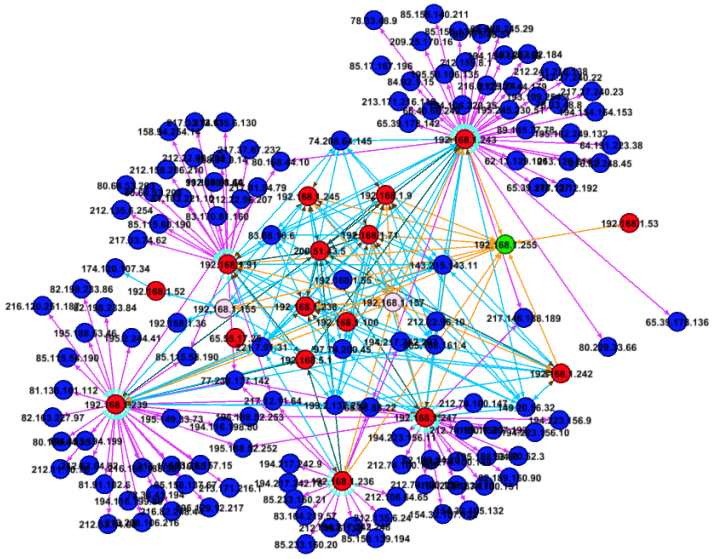

Figure 1: Graph of the network with connections and hosts classes coloured.