REFLECTION REPORT FOR THE DATA SCIENCE FOR CYBERSECURITY CONFERENCE

I. About the Data science for cyber-security workshop

The data science for cybersecurity workshop has been organised jointly by imperial college London, Los Alamos National Laboratory and the University of Bristol.

While data science is an established field of research, its application to cybersecurity is not. Indeed, "the use of data science for cyber-security applications is a relatively new paradigm. This includes deployment of statistical methodology, machine learning, and Big Data analytics for network modelling, anomaly detection, forensics, risk management, and more".

The conference's aim was to showcase cutting-edge research in statistical cyber-security in academia, business and government, as well as help align these endeavours

The rest of the report will present the main takeaways from the talks which have been most relevant to me

1. Transcend: Detecting Concept Drift in Malware Classification Models

The speaker in this talk focused on how can be build sustainable malware classifiers using data science techniques. Indeed, if we are to apply machine learning techniques to malware classification, then we should start by figuring out how to train such classifiers. While the question might be straightforward for another dataset, malware samples have the particularity that they evolve rapidly, making them thus becomes hard—if not impossible—to generalize learning models to reflect future, previously-unseen behaviours.

This characteristic renders many malware classifiers obsolete very quickly. The speaker presented Transcend, a framework to identify aging classification models in vivo during deployment, much before the machine learning model's performance starts to degrade. This is not a common technique for classical classifiers, where aging models are kept even if poor performance is observed.

In order to achieve this, the speaker uses a statistical comparison of samples seen during deployment with those used to train the model, thereby building metrics for prediction.

The speaker also shared experimental results of his solutions, when run against both android and windows malware. The experiment proved sound results for both binary and multi-class classification scenarios on different datasets and algorithms using proper training, calibration and validation, and testing datasets.

2. Detecting Botnet Activities Based on Abnormal DNS traffic

The speaker focused in this talk on one of the most popular forms of attack on the internet is through botnets. He pointed out that in recent attacks, one can observe that botnets make use of DNS servers and query them through the same interface used by any other legitimate host. This makes It difficult to distinguish between legitimate DNS traffic and illegitimate DNS traffic.

The speaker presented his proposed solution to classify DNS traffic with the goal of detecting botnet traffic. The main assumption of this solution is that botnets appear as a group of hosts periodically. Hence, the classifier proposed in this solution tries to distinguish and classify group behaviour from single host behaviour.

The results of this solution are promising. Indeed the experiment conducted on the NAv6 network

shows that this proposed solution has an average detection rate of about 89%.

3. General remarks

While many of the talks focused on anomaly detection based of machine learning applied to different datasets, one common criticism to these approaches was that an anomaly does not necessarily mean a security vulnerability. Combining the netflow data sets which were extensively used by the speakers ought to be combined with other different sets which can provide context and add a semantic layer to their results. This way, one can close the gap between the detection of an anomaly and the detection of an actual security vulnerability or the presence of malware in the network or the platform.

II. Relevance to the research projects at the University of Oslo

The informatics department at the University of Oslo has an ongoing project which also aims to use data science techniques to solve cybersecurity problems.

One of the outcomes of this conference was that we are considering using the Loa Alamos netflow dataset in this project in order to run comparative studies.

It was also interesting to see that system logs on individual machines are also used by other researchers in order to detect anomalous behaviour within this platform.