ThreatPredict: From Global Social and Technical Big Data to Cyber Threat Forecast

Jérôme François, Inria Nancy Grand Est
Ghita Mezzour, International University of Rabat
Kathleen M. Carley, Carnegie Mellon University
Abdelkader Lahmadi, Université de Lorraine
Mounir Ghogho, International University of Rabat

In a nutshell, ThreatPredict, a project funded by NATO Science for Peace and Security Programme and started in December 2017, aims at characterizing the relationships between security events and social and geographical related data and, using this knowledge, to finally predict future cybersecurity threats and attacks that will occur. We especially aim to improve the research community's understanding of cyber security as a socio-technical problem by analysing and describing large datasets from multiple sources. To realize this objective, the project contribution is three-fold:

(1) Collection, storage and clustering of both technical and social data within a shared and safe repository

(2) Correlation of societal and technical data (security related) to highlight their inter-dependency

(3) Prediction of security threats

The first contribution consists of aggregating the different identified sources of data:

- Security data: darknet (Internet black-hole hosted at Inria), honeypots emulating various services and capturing attacks, CVE and CAPEC databases, OWASP Web hacking incidents, Symantec's Worldwide Intelligence Network Environment,
- Social data: GDELT, a large database classifying events reported in news, and other sources indexing major events (e.g. AllSportDB.com and eventful.com) and social media data (mainly twitter trends)

These data sources will be aggregated within a big data framework for storing and analyzing purposes. Some of these data sources could be shared, under an NDA, with the community depending on the ownership. If not possible to share original data, aggregated data would be considered. In addition, this requires data to be pre-processed including enrichment of data using reverse DNS and geolocation for technical data, clustering and labelling of data instances. Indeed, many sources are purely raw data while we need to extract knowledge. For instance, honeypots and darknets (in the meaning of Internet telescopes) surely contain evidences of attacks but our own experience shows that limitation of these approaches is their lack of being able to precisely identify and characterize attacks.

The second contribution will research on correlation between technical and social data, especially if there exist relationships between the occurrence of a security event and its social context. While some studies have demonstrated such a relation, the project will assess quantitatively and qualitatively those dependencies. Therefore, we plan to evaluate the accuracy of the correlation depending on the granularity of the information (geography, type of attacks, etc.). This will be helpful then to fine-tune prediction models accordingly in the third contribution. As we will rely on technical data, well-formatted and structured, and social data, mainly text-based, text-mining and topic modelling approach will be leveraged to extract vector-based metrics, which fits better correlation approaches.

The third and final contribution of the proposed project is a cybersecurity forecast. Indeed, the goal of this project is to model and forecast the evolution of cyber threats regarding technical criteria. Those models will predict and characterize future threats on both technical and societal aspects (including geographical attributes). A time-series analysis of all collected data will be performed to understand major trends from a purely descriptive point of view. Such a result will support the building of predictive models by reducing the dimension of the problem. This last step will leverage machine learning techniques to create predictive models applied on our datasets. Validation will be performed to assess the veracity of these models depending on the scenarios and granularity of the features of predicted attacks.