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

Jérôme François, Frédéric Beck, Ghita Mezzour, Kathleen M Carley, Abdelkader Lahmadi, Mounir Ghogho, Abdellah Houmz, Hicham Hammouchi, Mehdi Zakroum, Narjisse Nejjari, et al.

THREATPREDICT: FROM GLOBAL SOCIAL AND TECHNICAL BIG DATA TO CYBER THREAT FORECAST

Jérôme François (jerome.francois@inria.fr) [1]

Frederic Beck (frederic.beck@inria.fr) [1]

Ghita Mezzour (ghita.mezzour@uir.ac.ma) [2]

Kathleen M. Carley (kathleen.carley@cs.cmu.edu) [3]

Abdelkader Lahmadi (abdelkader.lahmadi@loria.fr) [1]

Mounir Ghogho (mounir.ghogho@uir.ac.ma) [2]

Abdellah Houmz (abdellah.houmz@uir.ac.ma) [2]

Hicham Hammouchi (hicham.hammouchi@uir.ac.ma) [2]

Mehdi Zakroum (mehdi.zakroum@uir.ac.ma) [2]

Narjisse Nejjari (narjisse.nejjari@uir.ac.ma) [2]

Othmane Cherqi (othmane.cherqi@uir.ac.ma) [2]

[1] *Inria - LORIA - University of Lorraine, France*

[2] *International University of Rabat, Morocco*

[3] *Carnegie Mellon University , USA*

Abstract.

Predicting the next threats that may occurs in the Internet is a multifaceted problem as the predictions must be enough precise and given as most as possible in advance to be exploited efficiently, for example to setup defensive measures. The ThreatPredict project aims at building predictive models by integrating exogenous sources of data using machine learning algorithms. This paper reports the most notable results using technical data from security sensors or contextual information about darkweb cyber-criminal markets and data breaches.

Key words: threat, prediction, darknet, darkweb, data breaches

1. Introduction

Predicting future cyber-threats with reasonable lead-time and accuracy can give security practitioners sufficient time to prepare for upcoming major attacks. For example, these practitioners can increase network provisioning to deal with an upcoming major Denial-of-Service (DoS) attack or purchase security insurance. These preventive measures can stop future attacks or at least reduce their impact.

However, most prior work focused on detecting attacks instead of predicting them. A notable exception (Zhan et al., 2015) predicts future cyber rates by taking into account statistical temporal properties of cyber attacks. Unfortunately, it only achieves reasonable accuracy for a limited prediction time horizon (a few hours). (Liu et al., 2015) forecasts security incidents on organizations by using security features about the organization such as network mismanagement symptoms. Unfortunately, the authors were unable to predict geo-politically motivated attacks.

A promising direction to improve these prediction models is to take into account exogenous global social and technical data such as Twitter sentiments and Darkweb trades and discussions. Social and traditional media data has been successfully used to predict civil unrest (Ramakrishnan et al., 2014), disease outbreaks (Lazer et al., 2014), and the stock market (Bollen et al., 2011). Exogenous technical events such as the discovery of new software vulnerabilities are also known to affect cyber attack rates.

ThreatPredict aims so at developing machine-learning models that predict future cyber threat rates up to a few days in advance. Our models will take into account past cyber attack temporal evolution as well as global social, geo-political, economical, and technical events that are known to affect cyber attack rates.

To reach our ultimate goal of predicting cyber threats days in advance, three steps must be achieved:

- Data collection: the goal is to collect all relevant data and to preprocess them to clean and index them for further correlation;
- Data analysis: based on collected data, the first analysis consists into refine them in order to have a more fine-grained view about activities we can observe. For example, a security sensor can emit a large number of alerts even if they all concern a single attack. This is usually the case for scanning or DDoS attacks that are launched from multiple hosts.
- Threat prediction: once all data is indexed and refined, we can pursue the analysis by mining historical data to predict and characterize the next threats or attacks to occur will be.

This paper focuses on describing relevant exogeneous data sources for cyber-security in Section 2 but also on an initial sets of data-analysis results. First, darknet data will be investigated to show how such raw attack traffic data can be better integrated into machine learning algorithms to support predicting scanning activities of cyber-criminals 3. We will also leverage contextual information such as twitter data to increase our predictive model for attack rates. Finally, Sections 4 and 5 focus on contextual data, darkweb market trends and data breaches respectively.

2. Overview

As briefly mentioned in Introduction, we rely on various sources of data to infer the future threats of attacks. Figure 2 summarizes them. The initial set of data are collected by security sensors in our own infrastructure. They provide very fine-grained information about attacker activities and with a large history over multiple years. We can differentiate the darknet and the honeypots. Honeypots are artificial systems with vulnerabilities the attackers can interact with them. They allow to observe attacker behaviors on different services such as SSH, telnet, web or even industrial systems. However, this paper mostly focuses on darknet data analysis. A darknet or telescope is using a whole subnetwork, which is announced over Internet such that packets sent to the IP addresses are properly routed over. However, in a darknet, this subnetwork does not host any services or devices and so no legitimate traffic is supposed to reach it. The entity hosting the darknet is then silently collecting all incoming packets, i.e. without replying to any of them. Such an infrastructure is mainly characterized by the size of the subnetwork defined by the prefix length. In our case, we host a darknet of a /20 network meaning that we monitor more than 4000 adresses.

Darknet data consists so in network traffic flow, which are monitored in real-time. It is completed by other security technical data that are freely accessible over Internet as highlighted on top of figure 2. They are used for two purposes. First, data collected by the darknet are raw data. So, all traffic is abnormal but classifying and grouping observed attacks is not trivial. We rely on external information such as blacklists or attack and threats reports. The latter contains information about hosts (IP addresses) which have been involved in attacks. These IP addresses are matched against our darknet data and serve as a refinement of data but also to validate our classification algorithms. Indeed, darknet data can be only partially enriched or annotated through public information. We thus develop methods to automatically perform this task (mainly to group traffic flows by attack instances and label them with the attack type). Other technical and public data refer to SCAP (Security

Content Automation Protocol) components such as Common Vulnerabilities and Exposures (CVE) or Common Attack Pattern Enumeration and Classification (CAPEC) provided by MITRE and to scanning data like Rapid7 sonar project[1] or Censys(Durumeric et al., 2015). Scanning data allows to evaluate the exposure of certain devices or services and indirectly their risk of being attacked if there exist past or new threats against them, which can be discovered in SCAP components.

All sources previously described can be qualified as technical. However, the added value of ThreatPredict is to leverage non technical data. Firstly, evidences about past attacks can be found in PRC[2] (Privacy Rights Clearinghouse) data breaches providing evidences of security issues. They can be used to assess the risk of exposure of a given organization to a data breach. this risk can be measured by the configuration of the network, which can be used to forecast future attacks. Finally, to build these predictive models contextual information is considered. For example, analysis of sentiment or trendy topics over Twitter are good indicators to identify potential victims of attacks, e.g. major events usually attracts attackers. Also, Darkweb tradings and discussion could reveal important indicators on hackers interests and future attacks. Moreover, Organizational data such as economic value would give insights about an organization and could be a valuable predictor in our predictive models.

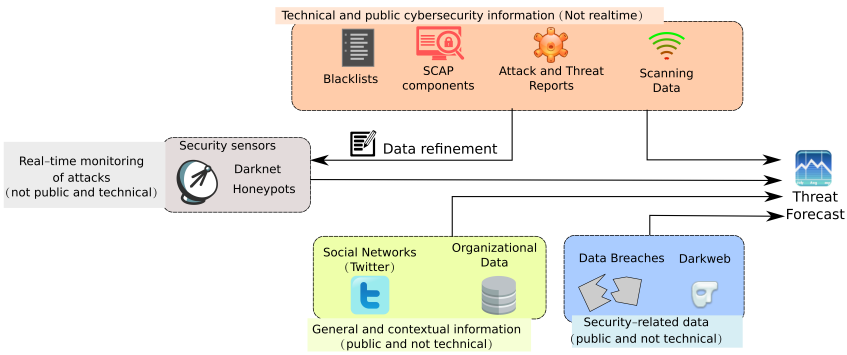

Figure 1. ThreatPredict overview

[1] https://www.rapid7.com/research/project-sonar/
[2] https://www.privacyrights.org/

3. Darknet monitoring and prediction

3.1. PREDICTING TCP SCANS

When performing an attack, the first stage usually consists in identifying the potential targets. Discovering accessible machines and services often relies on IP sweeping and/or scanning TCP and UDP ports. Naive approach testing all ports numbers and all IP addresses of a targeted subnetwork is time-consuming and has a large footprint which can be easily detected. However, smarter attackers would look for specific ports to search for particular services with potential vulnerabilities.

As a result, observing the strategy of performed port scans revealed the nature of the targeted threats and can be leveraged to predict the next sequence of scanned ports. Full results have been reported in (Evrard et al., 2019) and demonstrated that such sequences can be modeled as a graph and mined to derive semantic distances between ports using a shortest path-based distance. Such a distances that is publicly accessible[3] can be exploited in other algorithms, for example to predict what are the next targeted ports by an attacker when we observe his or her first scans. We analyzed around 815 million of darknet packets collected during 9 months to infer the values of distances. Based on them, K-nearest neighbors (with K=3) has been applied to the MAWI Labs dataset(MawiLab, 2018) to deduce future port scans. We consider the period from 2 to 9 September 2015 (except the 5th and 7th of September because of dataset unavailability) with a total of 590,173,645 IP packets with a mean of 98,362,274 packets a day. Our technique has been succesfully applied since 99.98% (True Positive Rate) of scans have been predicted and would have been ideally proactively blocked while the false positive is limited 0.0015%. The interested reader is encouraged to read (Evrard et al., 2019) for all methodological details and results.

3.2. PREDICTION ATTACK RATES

In this section we study different aspects of the traffic captured by the /20 network telescope including the prediction of the traffic rate.

As a first step, we performed an exploratory data analysis of 3 years of traffic. We found that 90% of probing activities are targeting only 550 ports of the port's space. The latter includes remote access services which are the most sought by network probers, followed by database management systems, web services and miscellaneous services as well.

Network probers tend to scan only a limited number of ports. Thus, extracting the network probers exploration patterns reveals their interest in terms

[3] `http://port2dist.lhs.inria.fr/`

of services and also the sets of related ports. We model these behaviors using transition graphs decorated with probabilities of switching from a port to another. Each graph represents the scanning activities of one prober, where a node represents a port targeted by the prober, and the edges represent the probability of transition between ports. The results of the analysis showed strong relations between ports of similar services, as depicted in figure 2, such as the ports used for remote access including the port 23 (TelNet), 22(SSH) and their alternatives 2323 and 2222, and the same is found for database systems, web services etc.

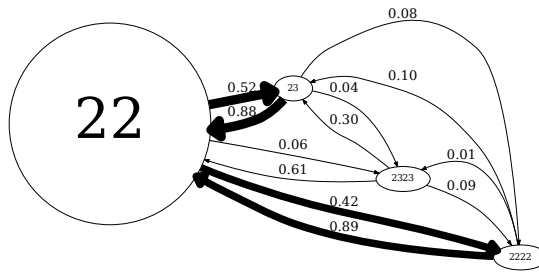

Figure 2. Targeted probing pattern (remote access services)

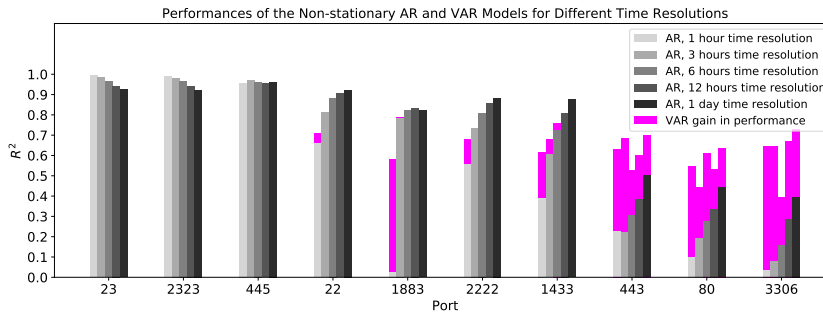

Figure 3. Comparison of performances of the non-stationary *AR* and *VAR* models

Besides the descriptive analysis, observing a significant disparity between the predicted probing rate and the actual value may help detecting an imminent threat. Furthermore, exploiting the similarities found between ports can improve the predictions. Thus, we extracted time series of traffic targeting each port in the darknet, then we assessed the capacity of Non-stationary Autoregressive (AR) and Vector Autoregressive (VAR) models in predicting port probing rates. We applied the two models on the extracted time series of

probing rates for each port. The result in figure 3 represents the R^2 score of the models, applied on different ports' time series with different time resolutions. We conclude that VAR models generally give better results than AR models, as they use data of related ports to predict rates on the target port. We found that the non-stationary AR model produces satisfying results for services exhibiting low short term probing rates variability such as the ports 23, 2323 and 445, with an R^2 of more than 0.90 for time resolutions ranging from one hour to one day. Where non-stationary VAR model consistently produces better results for services exhibiting high probing rate variability. More details could be found in the paper (Zakroum et al., 2018).

3.3. PROBING RATE SEVERITY PREDICTION

Since we analyzed the probing activities and performed probing rates prediction hours in advance, we explore in this section the aspect of predicting probing rate severity by leveraging twitter sentiments as exogenous data (Hammouchi et al., 2019b). We investigate the relationships between probing rates and Twitter sentiments towards France through a statistical analysis. It revealed a significant correlation between the darknet attack rates and negative sentiments which further motivated the use of these sentiments as additional features for prediction. We cluster the probing rates into four homogeneous levels of severity using spectral clustering, and we cast the prediction as a classification problem. Then, we predict future rate levels using several classifiers. Incorporating negative sentiments to our models allows to predict the next day severity with 81% accuracy using AdaBoost classifier. It is equivalent to a 5% improvement compared to only using past probing rates for prediction. This demonstrates the importance of incorporating social signalsin models when predicting future probing rates.

4. Darkweb analysis

Another research aspect to consider is Darkweb since it contains important amounts of information on hackers tradings and discussions. The Darkweb is a subset of the Deep Web, which cannot be accessed using standard web browsers. Notably, tt is used by cyber-criminals to offer all kinds of illegal services and goods.

In this section, we analyze the illegal trade in the darkweb's marketplaces to provide a cyber threat intelligence. The analysed services include the most common and severe malwares, zero-day threats, exploits, DDOS services etc. Full details are available in (Cherqi et al., 2018).

We focused on analyzing four of the most popular and well known marketplaces: Agora, Silkroad 2.0, Alphabay and Nucleus. We retrieved the ad-

vertisements that have been published and extracted all information about the offers. We were able to collect a total of 972.655 items for our analysis, starting from 7 July 2013 until 20 December 2015.

The main information we collected about the products are: Vendors alias, Item's Category, Items description, Price of the article, Sellers location, Destinations vendor ships to, Rating given by consumers to vendors based on a stars rating scale.

Our descriptive analysis highlights the large extent of this phenomenon with markets worth up to millions of USD each year. It is a scale of organized crime, well anchored in the field and whose demand and supply is booming. Also we were able to highlight the existence of well organized cells that monopolize the market. Indeed, 98% of the market is controlled by only 20% of sellers. In addition to this, it is a market that is based mainly on the reputation of sellers and the feedback of customers. Users are generally satisfied by the outcome of their transactions.

Being proactive is the best and most effective defense. By gathering threat intelligence from a largely anonymous and underground network, we could obtain targeted, relevant dark web intelligence, and facilitate the understanding of threats in order to prevent or mitigate the damages. For example, getting alerts if a breached data is found on the Dark Web, allows to take active measures to reduce the legal liability.

5. Data breaches analysis

Cyber attacks are diverse and having a bigger picture on cyber threats requires analyzing different sources of data. We explore in this section another perspective of attacks that is not purely technical but allows to give context about attacks and threats. Indeed, data breaches represent a permanent threat to all types of organizations since they disclose personal and confidential information to the public. In accordance with the new regulations such the EU General Data Protection Regulation (GDPR), and with the objective of predicting these breaches in advance to prevent the damage, we

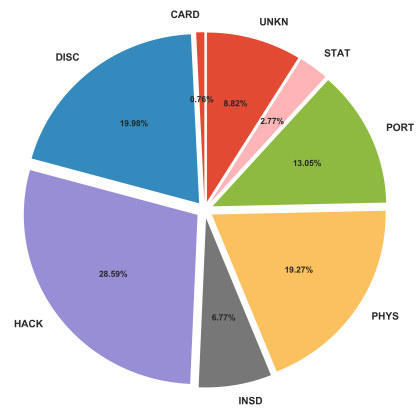

Figure 4. : Methods used in Data Breaches

conducted an exploratory analysis
to study the frequency, trends, and the causes of these leaks. In this work (Hammouchi et al., 2019a), we considered a dataset of data breaches mainly encountered in the USA, consisting of about 9,000 breaches involving organizations in different sectors, collected in the period between 2005 and 2019. These breaches led to the loss of 11.5 Billion individual records, which have a significant financial loss and technical impact. There exist several methods that involves a breach, from insiders (INSD) and unintended disclosure (DISC) to hacking activities (HACK) as illustrated in figure 4. Hacking is the most prevalent method with 29% and also the most devastating in terms of breached records and financial impact. For example, the Yahoo breach as the biggest reported breach was estimated to have a financial loss of **$**444**B**. In addition, we found that the most targeted sectors are medical institutions and online services and platforms, which could be explained by the amount of sensitive and personal data they hold on their users which makes them the top target to the attackers. From 2014, we observed that the frequency of attacks has increased significantly in medical and online sectors with an inerarrival time (time bewtween two consecutive attacks) of 4 and 7.5 days respectively.

As breaches due to hacking are the most interesting to research on, we are currently working on predicting these breaches by analyzing an organization's network and looking at several aspects such as DNS misconfigurations, IP blacklists, and expired certificates in order to assess the risk of exposure to breach attacks.

6. Conclusion

In this paper, we introduced the ThreatPredict project that pursues the goal to provide an accurate forecasting of internet threats day in advance. It resides in co-jointly analyzing various sources of security and contextual data such as darknet, honeypots, darkweb, data breaches... Although the sources of data can should be constantly updated to capture the multiple sides of the security threats, we showed in this paper the veractiy of some of them, in particular darknet, Twitter, darkweb and data breaches. Although the analysis is mostly limited to a single or two sources of data, the next stage of the project will be focused on improving the joint analysis of multiple sources of data.

Acknowledgments: This work has been partially supported by the NATO Science for Peace and Security Programme under grant G5319 Threat Predict: From Global Social and Technical Big Data to Cyber Threat Forecast and by the project SecureIoT, funded from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 779899.

It is also supported by the High Security Lab (https://lhs.loria.fr) providing the French darknet data and we gratefully thank its technical director, Frederic Beck

from Inria Nancy Grand Est. We also thank the NICT providing the Japan darknet data (https://www.nicter.jp).

We acknowledge the support from the National Center of Scientific and Technical Research (CNRST), Rabat, for the grant of an excellence scholarship.

References

Bollen, J., Mao, H., and Zeng, X. (2011) Twitter mood predicts the stock market, *Journal of Computational Science* **2**, 1 – 8.

Cherqi, O., Mezzour, G., Ghogho, M., and El Koutbi, M. (2018) Analysis of Hacking Related Trade in the Darkweb, In *2018 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pp. 79–84.

Durumeric, Z., Adrian, D., Mirian, A., Bailey, M., and Halderman, J. A. (2015) A Search Engine Backed by Internet-Wide Scanning, In *Proceedings of the 22Nd ACM SIGSAC Conference on Computer and Communications Security*.

Evrard, L., François, J., and Colin, J.-N. (2019) Attacker Behavior-Based Metric for Security Monitoring Applied to Darknet Analysis, In *Proceedings of IFIP/IEEE International Symposium on Integrated Network Management*, Washington, DC, USA.

Hammouchi, H., Cherqi, O., Mezzour, G., Ghogho, M., and El Koutbi, M. (2019)a Digging Deeper into Data Breaches: An Exploratory Data Analysis of Hacking Breaches Over Time, *Procedia Computer Science* **151**, 1004–1009.

Hammouchi, H., Mezzour, G., Ghogho, M., and El Koutbi, M. (2019)b Predicting Probing Rate Severity by Leveraging Twitter Sentiments, In *2019 15th International Wireless Communications & Mobile Computing Conference (IWCMC)*, pp. 883–888.

Lazer, D., Kennedy, R., King, G., and Vespignani, A. (2014) The Parable of Google Flu: Traps in Big Data Analysis, *Science* **343**, 1203–1205.

Liu, Y., Sarabi, A., Zhang, J., Naghizadeh, P., Karir, M., Bailey, M., and Liu, M. (2015) Cloudy with a Chance of Breach: Forecasting Cyber Security Incidents, In *24th USENIX Security Symposium (USENIX Security 15)*, Washington, D.C.

MawiLab (2018).

Ramakrishnan, N., Butler, P., Muthiah, S., Self, N., Khandpur, R., Saraf, P., Wang, W., Cadena, J., Vullikanti, A., Korkmaz, G., Kuhlman, C., Marathe, A., Zhao, L., Hua, T., Chen, F., Lu, C. T., Huang, B., Srinivasan, A., Trinh, K., Getoor, L., Katz, G., Doyle, A., Ackermann, C., Zavorin, I., Ford, J., Summers, K., Fayed, Y., Arredondo, J., Gupta, D., and Mares, D. (2014) 'Beating the News' with EMBERS: Forecasting Civil Unrest Using Open Source Indicators, In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Zakroum, M., Houmz, A., Ghogho, M., Mezzour, G., Lahmadi, A., François, J., and El Koutbi, M. (2018) Exploratory Data Analysis of a Network Telescope Traffic and Prediction of Port Probing Rates, In *2018 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pp. 175–180, IEEE.

Zhan, Z., Xu, M., and Xu, S. (2015) Predicting Cyber Attack Rates With Extreme Values, *IEEE Transactions on Information Forensics and Security* **10**, 1666–1677.