Modeling of IP scanning activities with Hidden Markov Models: Darknet case study

Giulia De Santis*, Abdelkader Lahmadi†, Jérôme François*, Olivier Festor*†
*INRIA - Nancy Grand-Est, France   †LORIA, University of Lorraine, France

Email: {giulia.de-santis, abdelkader.lahmadi, jerome.francois, olivier.festor}@inria.fr

Abstract—**We propose a methodology based on Hidden Markov Models (HMMs) to model scanning activities monitored by a darknet. The HMMs of scanning activities are built on the basis of the number of scanned IP addresses within a time window and fitted using mixtures of Poisson distributions. Our methodology is applied on real data traces collected from a darknet and generated by two large scale scanners, ZMap and Shodan. We demonstrated that the built models are able to characterize their scanning activities.**

Index Terms—**Network scanning, ZMap, Shodan, Poisson distribution models, Hidden Markov Models, HMMs**

1. Introduction

Scanning of IP addresses is widely used by attackers to exploit vulnerabilities as well as during the reconnaissance phase of Advanced Persistent Threats [1]. Even if about 50% of cyber attacks follow some scanning activities [2], [3], only little progress has been made to improve the security of networks [3] using IP traffic collected by a darknet space. All traffic arriving to such IP dark space is undesired since it has no active hosts, and it contains many patterns of large scale scans, DDoS (Distributed Denial-of-Service) attacks and misconfigurations at the scale of Internet [4]. In this paper we aim to model the scanning activities of two large scale scanners, which are ZMap [5] and Shodan[1], with Hidden Markov Models. We obtained their scanning traces from a /20 darknet. We modeled for each of them the number of scanned IP addresses in given time windows. Since an over-dispersion is present, i.e. a single Poisson distribution is not enough to describe the observations, mixtures of Poisson distributions have been used for each scanning activity. Since it is unknown from which subset the observation comes, Hidden Markov Models (HMMs) are then needed. We found that each unique scanning activity has its own model and it is distinguishable from the others, because of the difference of their inferred parameters and probabilities. This work finds also applicability to security issues of modern technologies, like automating systems working in Smart Houses that have connections to the Internet [6].

This paper is organized as follows. Section 2 reviews scanning techniques and their existing models. Section 3 provides details of our methodology for building HMM models of a scanning activity. Section 4 is dedicated to the models of ZMap and Shodan activities. Section 5 concludes the work and draws future work.

2. Related work

Large scale, or horizontal, scanning tools, like ZMap [5], Shodan [7] and Masscan [8], are used by attackers to discover vulnerable hosts [9]. An Internet-wide scan consists in a permutation and split of IP addresses into groups assigned to a probing source IP address and the schedule of the scan [10]. ZMap [5] scans the IPv4 address space while permuting its elements using a multiplicative group. The more the coverage of the scan increases, the more it is likely that the scan is made with ZMap [9]. Shodan [7] is a computer search engine able to identify Internet-facing devices and acts as black-box from the user perspective. The impact of faster tools to scan the IP space is described in [9], which fingerprints Masscan and ZMap and also analyzes who performs large scans, their targets and what software is used. The authors do not provide a model of the faced scanning techniques. In [3], authors performed fingerprinting of probing activities by assessing if IP addresses are permuted, if the scan is random or follows a pattern and if it belongs to certain clusters of scanning techniques, but don't provide models of probing activities. Our work may be useful to assess what is the currently faced scanning activity. Mazel et al. [11] aim to uncover coordinated scans from single-source scans previously detected.

Closer to our work is [2], that uses the discrete Fourier transform and the Kalman filter to manage missing values when observing a scan from a darknet. It aims at inferring probing campaigns, whereas our work aims at modeling scanning techniques, and the obtained models may be used to infer if and what probing campaign is underway. Sperotto et al. [12] instead use Hidden Markov Models to model malicious traffic on SSH. Their work focuses on malicious activities and attacks on port 22, whereas ours is more focusing on scanning activities, not being limited to a single port. Their proposed model emulates the behavior of an attacker, whereas we aimed to create a descriptive model to assess if and what scanning technique is running.

[1] https://www.shodan.io/

978-1-5090-2914-3/16/$31.00 ©2016 IEEE

3. Methodology

We focused on the modeling of traffic collected from a darknet and generated by two large-scale scanners, ZMap [5] and Shodan [7]. For each packet we have considered its timestamp, source and destination IP addresses and ports. We identified the scanning traffic through the domain names of source IP addresses, since ZMap is mainly used by the University of Michigan and the sources used by Shodan belong to its domain name.

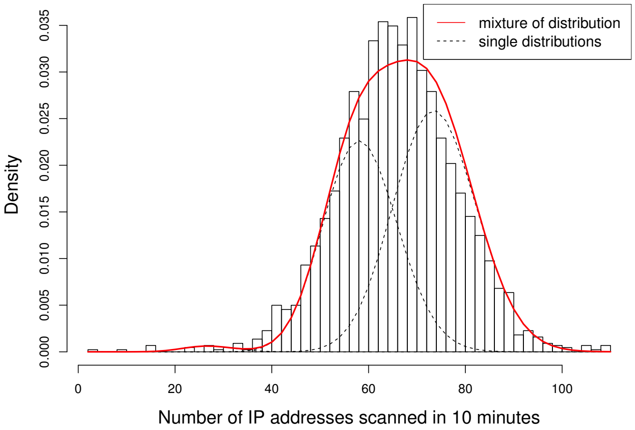


Figure 1. Fitting with Poisson distributions of the number of IP addresses scanned within 10 minutes by ZMap on port 23.

We considered, for each of them, their scanning intensity characterized by the number of IP addresses scanned in 10 minutes. Figure 1 shows the Probability Density Function (PDF) of the scanning intensity of ZMap while scanning port 23. Assumed that the searched distribution is Poisson, because the considered variable is discrete and the arrival of probe packets is coherent with this distribution [3], [13], a single Poisson distribution is not sufficient to fit the data as shown in Figure 1. This is due to the over-dispersion of the observations, that have a mean equal to 66.85 and a variance of 147.06. Thus, we relied on mixture of Poisson distribution models to fit the number of IP addresses scanned in a time window for each observed scanning activity. As shown in Figure 1, the number of scanned IP addresses in 10 minutes fits well with the mixture of Poisson distributions represented by the red line.

Our proposed approach consists in creating, from sets of observations represented as the count of IP addresses scanned in 10 minutes, mixture distribution models and Hidden Markov Models (HMMs) that describe the considered scanning activity. Our modelling method relies on two steps. The first step consists in selecting a mixture of m Poisson distributions for each scanning trace, while considering its Cumulative Distribution Function (CDF) of the mixture and the Bayesian Information Criterion (BIC) which is defined as $BIC = k\ln(n) - 2\ln(L)$ where k is the number of estimated parameters, L the maximum value of the likelihood function and n the sample size. The outputs are the vectors $\boldsymbol{\delta} = (\delta_1, \delta_2, \dots, \delta_m)$ containing the probabilities that the observation comes from the i^{th} distribution (where $i \leq m$) and $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \dots, \lambda_m)$ containing the parameters of the Poisson distributions. The next step consists in creating the HMM, whose m states are the m distinct distributions provided by the previous step. To select the HMM that best fits the considered scanning activity, we used the AIC (Akaike Information Criterion, $AIC = 2k - 2\ln(L)$) and BIC indicators to measure the relative quality of statistical models that keep into account the complexity of the model.

3.1. Mixture Distribution Models

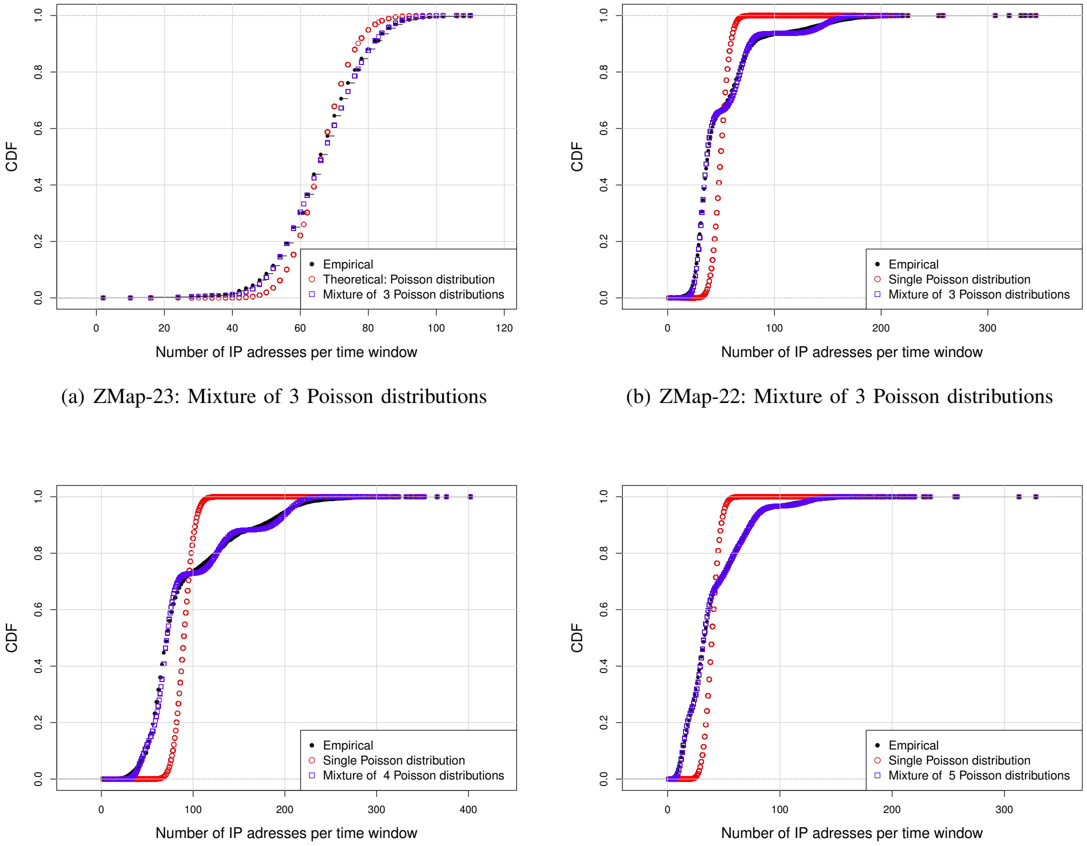
Independent mixtures of Poisson distribution models are used when the population consists of unobserved groups, each of them with its own Poisson distribution, and the selection of one of the groups is independent from the previous selection. They consist of m Poisson distributions $p_1, p_2, \dots, p_m$ and a *mixing distribution* $\delta = (\delta_1, \delta_2, \dots, \delta_m)$ which selects one of these components. The selection of the distribution is established by a random variable, say C, performing the mixing: once its value is known, an observation is drawn from the group i. The crucial point is that the value of C is unknown, i.e. the distribution $p_i, i = 1, 2 \dots, m$ being active when the observation was done. Let X denote the discrete variable which counts IP addresses scanned in a given time window, whose distribution is a mixture of multiple Poisson distributions. Its probability function is thus given by $p(x) = \sum_{i=1}^{m} \delta_i p_i(x)$.

3.2. Hidden Markov Models

Mixture distribution models do not consider probabilities to move between distributions. This information is provided by HMMs, that consist of an unobserved parameter process $\{C_t : t = 1, 2, \dots, m\}$ satisfying the Markov Property: $\Pr(C_t | C_{t-1}, \dots, C_1) = \Pr(C_t | C_{t-1})$, and a state-dependent process $\{X_t : t = 1, 2, \dots\}$ such that, when C_t is known, the distribution of X_t depends only on the current state of C_t: $\Pr(X_t | \boldsymbol{X}^{(t-1)}, \boldsymbol{C}^{(t)}) = \Pr(X_t | C_t)$. C_t establishes the Poisson distribution of the random process X_t. We used HMMs because they are able to handle partial knowledge, which here is the absence of knowledge of what distribution each observed count of IP addresses belongs to.

4. Experimental results

We considered five datasets for ZMap and one for Shodan as shown in Table 1. *ZMap-22*, *ZMap-23*, *ZMap-80* and *ZMap-443* are respectively ZMap scans on port 22 (SSH), 23 (Telnet), 80 (HTTP) and 443 (HTTPS) while *ZMap-All* contains the scanning of ZMap over multiple ports. The dataset named *Shodan* contains all the scanning traffic from the Shodan scanner since, after manual investigation, it is a large mix between horizontal and vertical scanning.



(a) ZMap-23: Mixture of 3 Poisson distributions

(b) ZMap-22: Mixture of 3 Poisson distributions

(c) ZMap-443: Mixture of 4 Poisson distributions

(d) ZMap-80: Mixture of 5 Poisson distributions

Figure 2. Cumulative Distribution Functions for *ZMap-22*, *ZMap-23*, *ZMap-80*, *ZMap-443*.

TABLE 1. DETAILS OF SCANNING DATASETS FOR ZMAP AND SHODAN.

Dataset	ZMap-22	ZMap-23	ZMap-80	ZMap-443	ZMap-All	Shodan
# Sources	178	80	246	222	253	13
# Destinations	4096	4096	4096	4096	4096	4096
# Ports	1	1	1	1	28	244
Duration (days)	532.00	119.16	541.73	216.80	533.69	12
# packets	487056	147340	1078053	1536667	7992496	708160

4.1. Poisson mixture distribution models of scanning activities

4.1.1. ZMap over a single port, 10 minutes. A single Poisson Distribution does not well describe the dataset *ZMap-23* as shown in Figure 2(a). Multiple Poisson distributions are required as previously discussed for Figure 1.

Appropriate models have been obtained with a mixture of 3 and 6 Poisson distributions respectively. We fitted *ZMap-23* with a mixture of 3 Poisson distributions, with parameters and probabilities $\boldsymbol{\lambda} = (27.16, 58.55, 73.92)$ and $\boldsymbol{\delta} = (0.008, 0.434, 0.558)$ respectively, for the following reasons. First, the CDF plot shown in Figure 2(a) for the mixture of 3 Poisson distributions fits well the observations. Furthermore, the difference between the obtained BIC values as shown in Figure 3 is really low, and in the mixture with 6 Poisson distributions three of them are really rare to appear.

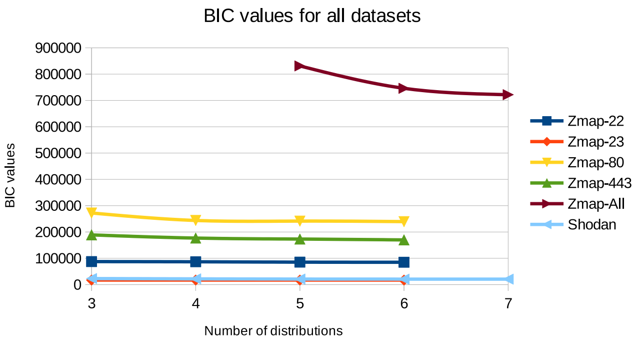


Figure 3. BIC values of the mixture of Poisson distribution models.

Also to describe the dataset *ZMap-22* a single Poisson distribution is not enough (see Figure 2(b)). A mixture model with 3 Poisson distributions gives good results. Vectors $\boldsymbol{\lambda}$ and $\boldsymbol{\delta}$ are $\boldsymbol{\lambda} = (33.28, 68.39, 147.65)$ and $\boldsymbol{\delta} = (0.661, 0.276, 0.063)$. We compared this model with the mixture of 4 Poisson distributions. The first two parameters are really close, as their probabilities. The third parameter of the mixture of 3 distributions is split into two parameters in the mixture of 4 distributions, whose probabilities sum to the probability of the third parameter of the mixture with 3 distributions. The improvements provided by the presence of one more parameter are thus marginal (see Figure 3). Even if ports 23 and 22 are both reserved for remote connections, their scanning procedure behaves differently when handling them (Figures 2(a), 2(b)). This is confirmed also by the different vectors $\boldsymbol{\lambda}$ and $\boldsymbol{\delta}$, not comparable: the parameters of the first distribution of *ZMap-23* and *ZMap-22* may be considered close (27.16 and 33.28 respectively), but the probabilities of the corresponding Poisson distributions are highly different (0.008 against 0.661).

For the dataset *ZMap-443* a mixture of 4 Poisson distributions is needed, with $\boldsymbol{\lambda} = (44.78, 69.69, 127.37, 201.83)$ and $\boldsymbol{\delta} = (0.145, 0.583, 0.155, 0.117)$. A mixture of 5 Poisson distributions would better fit the dataset, but the improvements are again too marginal as shown in Figure 3. It is not reasonable to merge the models for *ZMap-23* or *ZMap-22* with the one obtained for *ZMap-443*. Indeed, even if it is possible to model *ZMap-23* and *ZMap-22* also with 4 Poisson distributions, their parameters will not be comparable with the ones for *ZMap-443*.

ZMap-80 is instead well described by a mixture of 5 Poisson distributions, with $\boldsymbol{\lambda} = (14.73, 31.75, 54.75, 73.79, 124.76)$ and $\boldsymbol{\delta} = (0.242, 0.442, 0.1415, 0.1415, 0.033)$. The model for *ZMap-80* is not merged with the ones for *ZMap-23* and *ZMap-22* for the same reasons showed for *ZMap-443*, and neither with the one for *ZMap-443* because of the important difference between their parameters.

4.1.2. ZMap-All and Shodan. *ZMap-All* is modeled with a mixture of 6 Poisson distributions (Figure 4(a)), with $\boldsymbol{\lambda} = (32.00, 72.83, 122.19, 183.49, 276.42, 502.56)$ and $\boldsymbol{\delta} = (0.121, 0.216, 0.345, 0.220, 0.083, 0.015)$.

Shodan can be fairly good fitted with a mixture of 4 Poisson distributions, but we preferred the model with 6 Poisson (Figures 4(b), 3), with $\boldsymbol{\lambda} = (226.82, 274.87, 360.97, 432.85, 496.08, 573.15)$ and $\boldsymbol{\delta} = (0.163, 0.157, 0.057, 0.182, 0.248, 0.193)$. Even if both *ZMap-All* and *Shodan* are modeled with mixtures of 6 Poisson distributions, their parameters are highly different as also highlighted in in Figure 4. Their scanning processes are thus very different.

4.2. Hidden Markov Models

Here HMMs are built where each Poisson distribution of the mixture model is a state of the unobserved Markov chain and the elements of the vector $\boldsymbol{\delta}$ are the initial probabilities of each state.

4.2.1. ZMap over a single port. Both the HMMs describing the datasets *ZMap-23* and *Zmap-22* respectively have 3 states. The initial one of the HMM for *ZMap-23* is State 1, (initial state probabilities are in $v = (1, 0, 0)$), and transition probabilities are in Table 2.

TABLE 2. TRANSITION MATRIX OF THE HMM OF ZMAP-23.

	to State 1	**to State 2**	**to State 3**
from State 1	0.253	0.421	0.326
from State 2	0.004	0.394	0.602
from State 3	0.003	0.417	0.579

For *ZMap-22* the initial state is the second ($v = (0, 1, 0)$), and transition probabilities between states are in Table 3. Probabilities to "escape" from a state are really low. This doesn't happen for *ZMap-23*. This difference is added to the ones mentioned about mixture models, and confirms the need to keep the HMMs for the two datasets separated.

TABLE 3. TRANSITION MATRIX OF THE HMM OF ZMAP-22.

	to State 1	**to State 2**	**to State 3**
from State 1	0.967	0.000	0.033
from State 2	0.000	0.994	0.006
from State 3	0.007	0.013	0.980

The HMM for *ZMap-443* has 4 states. The initial state is the fourth ($v = (0, 0, 0, 1)$), and transition probabilities are in Table 4.

TABLE 4. TRANSITION MATRIX OF THE HMM OF ZMAP-443.

	to State 1	**to State 2**	**to State 3**	**State 4**
from State 1	0.885	0.010	0.028	0.078
from State 2	0.020	0.680	0.013	0.287
from State 3	0.066	0.000	0.925	0.010
from State 4	0.015	0.045	0.005	0.935

ZMap-80 is fitted with a 5-state Hidden Markov Models: the initial state is the third ($v = (0, 0, 1, 0, 0)$, whereas the transition probabilities are shown in Table 5.

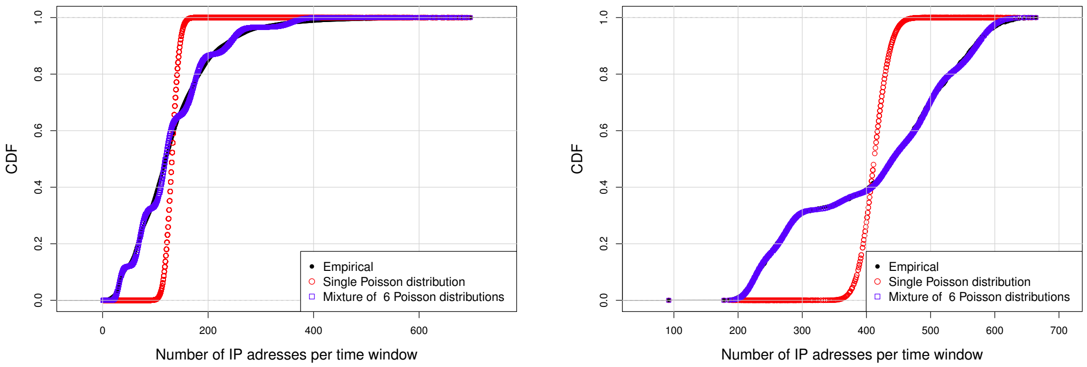
TABLE 5. TRANSITION MATRIX OF THE HMM OF ZMAP-80.

	to S.1	**to S.2**	**to S.3**	**to S.4**	**to S.5**
from S.1	0.962	0.000	0.002	0.031	0.005
from S.2	0.000	0.988	0.004	0.008	0.000
from S.3	0.003	0.003	0.982	0.012	0.000
from S.4	0.020	0.003	0.006	0.970	0.000
from S.5	0.030	0.000	0.000	0.000	0.970

4.2.2. ZMap over multiple ports and Shodan. *ZMap-All* is described by a 6-state HMM, with the first state as initial ($v = (1, 0, 0, 0, 0, 0)$) and the transition matrix in Table 6. Also the HMM describing *Shodan* has 6 states, but the initial one is the fifth ($v = (0, 0, 0, 0, 1, 0)$) and the transition matrix is in Table 7. In both cases, transition probabilities between states are low. This feature is common to all modeled datasets, with the exception of *ZMap-23*: its different behavior is clear also from its CDF (Figure 2(a)).

5. Conclusion and future work

This work presented a method, based on mixtures of Poisson distributions and HMMs, to model IP scanning



(a) ZMap-All: Mixture of 6 Poisson distributions

(b) Shodan: Mixture of 6 Poisson distributions

Figure 4. Cumulative Distribution Functions for *ZMap-All* and *Shodan*

TABLE 6. TRANSITION MATRIX OF THE HMM OF ZMAP-ALL

	to S.1	**to S.2**	**to S.3**	**to S.4**	**to S.5**	**to S.6**
from S.1	0.905	0.000	0.002	0.008	0.015	0.070
from S.2	0.001	0.877	0.112	0.009	0.001	0.000
from S.3	0.001	0.019	0.868	0.104	0.007	0.001
from S.4	0.001	0.001	0.037	0.864	0.093	0.004
from S.5	0.003	0.000	0.002	0.052	0.886	0.057
from S.6	0.048	0.001	0.002	0.009	0.083	0.857

TABLE 7. TRANSITION MATRIX OF THE HMM OF SHODAN.

	to S.1	**to S.2**	**to S.3**	**to S.4**	**to S.5**	**to S.6**
from S.1	0.909	0.000	0.000	0.000	0.000	0.091
from S.2	0.000	0.758	0.000	0.068	0.162	0.012
from S.3	0.000	0.000	0.905	0.000	0.095	0.000
from S.4	0.000	0.140	0.000	0.820	0.040	0.000
from S.5	0.000	0.157	0.089	0.000	0.753	0.001
from S.6	0.066	0.007	0.000	0.000	0.005	0.922

activities from datasets collected by a darknet. The models were built on the number of scanned IP addresses within a time window. We showed that there is no a single generic model to fit the used scanning traces of two well known scanners, ZMap and Shodan. Indeed, even if all the datasets can be modeled with the same number of Poisson distributions, their parameters and probabilities would be too different to be compared. The same holds for the initial state and transition probabilities of the corresponding HMMs. Therefore, each scanning process has its own signature.

Future work will consist in using the same approach for both discrete and continuous observations, such as time and distance between two successive scans.

Acknowledgement

This work was partially funded by HuMa, a project funded by Bpifrance and Region Lorraine under the FUI 19 framework. It is also supported by the High Security Lab hosted at Inria Nancy Grand Est (http://lhs.inria.fr).

References

[1] P. Chen, L. Desmet, and C. Huygens, "A study on advanced persistent threats," in *Communications and Multimedia Security*, ser. Lecture Notes in Computer Science, B. De Decker and A. Zúquete, Eds. Springer Berlin Heidelberg, 2014, vol. 8735, pp. 63–72. [Online]. Available: http://dx.doi.org/10.1007/978-3-662-44885-4_5

[2] E. Bou-Harb, M. Debbabi, and C. Assi, "A time series approach for inferring orchestrated probing campaigns by analyzing darknet traffic," in *2015 10th International Conference on Availability, Reliability and Security (ARES)*. IEEE, 2015, pp. 180–185.

[3] ——, "On fingerprinting probing activities," *Computers & Security*, vol. 43, pp. 35–48, 2014.

[4] C. Fachkha and M. Debbabi, "Darknet as a source of cyber intelligence: Survey, taxonomy, and characterization," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 2, pp. 1197–1227, 2016.

[5] Z. Durumeric, E. Wustrow, and J. A. Halderman, "Zmap: Fast internet-wide scanning and its security applications." in *Usenix Security*, vol. 2013, 2013.

[6] E. Isa and N. Sklavos, "Smart home automation: Gsm security system design & implementation," in *3rd Conference on Electronics and Telecommunications (PACET15)*, 2015.

[7] R. Bodenheim, J. Butts, S. Dunlap, and B. Mullins, "Evaluation of the ability of the shodan search engine to identify internet-facing industrial control devices," *International Journal of Critical Infrastructure Protection*, vol. 7, no. 2, pp. 114–123, 2014.

[8] R. D. Graham, "Masscan: Mass ip port scanner," *URL: https://github.com/robertdavidgraham/masscan*, 2014.

[9] Z. Durumeric, M. Bailey, and J. A. Halderman, "An internet-wide view of internet-wide scanning," in *23rd USENIX Security Symposium (USENIX Security 14)*, 2014, pp. 65–78.

[10] D. Leonard, Z. Yao, X. Wang, and D. Loguinov, "Stochastic analysis of horizontal ip scanning," in *INFOCOM, 2012 Proceedings IEEE*. IEEE, 2012, pp. 2077–2085.

[11] J. Mazel, R. Fontugne, and K. Fukuda, "Identifying coordination of network scans using probed address structure," in *Traffic Monitoring and Analysis - 8th International Workshop, TMA 2016, Louvain la Neuve, Belgium, April 7-8, 2016*.

[12] A. Sperotto, R. Sadre, P.-T. de Boer, and A. Pras, "Hidden markov model modeling of ssh brute-force attacks," in *International Workshop on Distributed Systems: Operations and Management*. Springer, 2009, pp. 164–176.

[13] Z. Li, A. Goyal, and Y. Chen, "Honeynet-based botnet scan traffic analysis," in *Botnet Detection*. Springer, 2008, pp. 25–44.