Stochastic Modeling, Analysis and Investigation of IoT-Generated Internet Scanning Activities

Article *in* IEEE Networking Letters · May 2020

DOI: 10.1109/LNET.2020.2998045

Stochastic Modeling, Analysis and Investigation of IoT-Generated Internet Scanning Activities

Morteza Safaei Pour*, Sadegh Torabi†, Elias Bou-Harb*, Chadi Assi†, and Mourad Debbabi†
*The Cyber Center For Security and Analytics. University of Texas at San Antonio, San Antonio, United States
†Security Research Centre, Concordia Institute for Information Systems Engineering. Concordia University, Montreal, Canada

***Abstract*—Analyzing the characteristics of scanning activities generated by compromised Internet-of-Things (IoT) devices is instrumental for early detection of IoT malware propagation. In this letter, we leverage about 3 TB of empirical passive network measurements to investigate IoT-generated scanning activities. Specifically, we exploit stochastic processes to model low-rate scans by incorporating the effect of random sampling and jitter on the observed packet Inter-Arrival Times (IAT). We verify the derived formulations using simulated results and empirically explore scans targeting common services (Telnet and HTTP) to demonstrate the effectiveness of our approach towards modeling low-rate scans while generating practical cyber threat intelligence.**

***Index Terms*—IoT botnets, IoT infections, Stochastic analysis, Internet measurements, Darknet.**

I. Introduction

The rise of Internet-of-Things (IoT) malware, which heavily rely on Internet-scale scanning activities to exploit vulnerable devices [1]–[3], highlight the pervasive nature of such scanning activities on the Internet. More specifically, the empirical analysis of scanning activities generated by IoT infections, typically operating within well-orchestrated botnets, highlight different low-rate scanning events, induced by diverse implementations of the employed IoT scanning modules. For instance, the `Mirai` IoT botnet sends batches of low-rate scanning packets followed by an idle state while waiting for SYN-ACK replies, which trigger further operations. Another prevalent low-rate scanning technique is implemented by dividing the traffic into equal sized packet bursts, which are separated by constant idle time intervals (e.g., $T = 1sec$) [4].

Despite the fact that scanning modules can be implement with different settings to control scanning rates and frequencies, low-rate scanners produce distinguishable characteristics that are observed in the distribution of their Inter-Arrival Times (IAT). Moreover, while low-rate scans can be performed by any host, motivated by the prevalence of low-rate stealthy scans generated by compromised devices within well-coordinated botnets [2], [3], in this letter, we draw upon significant empirical data extracted from a large-scale network telescope [5] to infer and characterize low-rate scanning activities generated by exploited devices. While packet IAT has been previously used as an effective feature for characterizing network scans [6], [7], we aim at providing a better understanding of the scanning activities through empirical analysis and probabilistic modeling of the perceived IAT, which would pave the way for exploring IoT-centric open research problems and much needed diverse applications, including IoT device fingerprinting, malware attribution and campaign detection.

To achieve such objectives, we leverage a `/8` network telescope (darknet) to infer scanning traffic generated by compromised devices. We then obtain device information/labels by performing instantaneous scanning and banner analysis of such devices to identify various information such as device type (e.g., IoT/non-IoT) and known malware signatures (e.g., Mirai). Furthermore, we perform empirical analysis of the scanning activities by measuring the IAT Probability Density Functions (PDF) for all devices through implementing a series of dimension reduction techniques, while clustering correlated devices into meaningful groups. Indeed, the obtained results demonstrate the effectiveness of our approach towards classifying IoT and non-IoT devices based on the distribution of their IAT, while showing that devices infected by the same IoT malware family are likely to be correlated due to their similar scanning behaviors. Finally, while we introduce novel stochastic processes for modeling low-rate scanning activities based on observed packet IAT, we provide empirical evidence to support the accuracy of the theoretical model in estimating the behaviors of different groups of correlated devices that perform low-rate stealthy scanning activities.

Along this line of thoughts, we frame the contributions of this letter as follows:

- Executing empirical and probabilistic analysis of scanning activities based on their packet IAT as perceived from a network telescope. Our objective is to explore methods for enhancing IoT security, device fingerprinting, and botnet inference.
- Employing stochastic processes for modeling scanning packets' IAT, while considering network-specific factors such as random packet sampling, path delay, and jitter. The proposed model is validated and shown effective and accurate in modeling different employed scanning modules for various groups of correlated IoT devices.
- Empirically investigating IoT-generated scanning events targeting two prominent services (i.e., Telnet and HTTP) and demonstrating the effectiveness of the proposed approach in differentiating between IoT and non-IoT devices based on their modeled IAT, while correlating devices based on distinguishable scanning characteristics. This can be used to uncover common infections, infer orchestrated campaigns, and provide digital evidence related to IoT malware.

II. Proposed Model

A. Stochastic modeling of stealthy IoT scanning activities

In this letter, we use stochastic modeling to formulate the probability density function of IAT for randomly sampled packets from a given source towards a vantage point on the Internet. The proposed model is founded on three main hypotheses and assumptions: (1) The scanners/infected IoT devices generate stationary behaviors which allow us to model their longitudinal activities; (2) Scanners send scan packets following a burst-idle model; and (3) We are only able to observe small sample of darknet-received packets which are randomly selected.

1) Modeling of the scanning source: In general, we assume scanners send batch of n packets and go dormant/idle for a deterministic or a random period. This dormant period can be due to imposing rate limiting, time required to process response packets, or performing other tasks. We model the scan traffic process as a modulated stochastic point process defined as $X(t) = \Sigma_i B(t-T_i)$, where $B(t) = \Sigma_{j=0}^{n-1}\delta(t-j\Delta)$ is a batch of packets sent out every Δ and $\delta(.)$ is the Dirac Delta impulse. For simplicity, in case of $\Delta \ll E\{T\}$, which is a correct assumption for low-rate scans, we substitute the batch function with $B(t) = n\delta(t)$. The batch arrival times T_i can be deterministic $T_i = \delta(t - iT)$ or can be any renewal process with an arbitrary distribution $T_i - T_{i-1} \sim f(x)$. In general, while the PDF of IAT can form any arbitrary distribution, our proposed model aims at capturing the most prevalent practices of scanners in their stationary phase.

2) Effect of random sampling: It is important to note that a relatively small portion of the overall scanning activities generated by compromised devices is captured at the network telescope. Therefore, the observed PDF of packet IAT is different from the actual distribution before sampling. Inline with that, when random packet sampling is applied, the traffic is only observed at the time of arrival of the selected packets, where each packet is either kept with probability ρ or discarded with probability $1 - \rho$.

Theorem 1. *The inter-arrival PDF of scan packets sent in batches of n packets with inter-batch distribution $T_i - T_{i-1} \sim f(x)$ after sampling is defined as:*

$$g(t) = (1 - \frac{q}{n\rho})\delta(t) + \frac{q}{n\rho}\sum_{i=1}^{\infty} qp^{i-1} f_i(t) \tag{1}$$

where $p = (1-\rho)^n$, $q = 1 - p$ and $f_k(t)$ is the k-fold convolution of PDF $f(t)$ and defined recursively as:

$$f_1(t) = f(t), \quad f_k(t) = \int_0^t f_{k-1}(t-\tau) f(\tau) d\tau \tag{2}$$

Proof. To prove Theorem 1, we mix the discrete process for cases with zero or one packet from each batch with a continuous distribution function $f(x)$. Given the latest received packet from a sample s at time $t_{s_{last}} = 0$, we want to determine the probability of observing subsequent scanning packet in exactly $t_{s_{next}} timeunits$. The probability to receive at least one more packet from the current batch is $1 - \frac{q}{n\rho}$. Further, the probability to receive the next packet from batch i rather than prior batches is $\frac{q}{n\rho} qp^i$. However, the arrival time probability density function of the i^{th} batch should be taken into consideration, which is equal to the k-fold convolution of $f(t)$ [8].

$$\begin{aligned} g(t) &= P(t_{s_{next}} = t | t_{s_{last}} = 0) \\ &= \sum_{i=0}^{\infty} P(t_{s_{next}} = t | t_{s_i} = 0, s_{next} \in B_i, s_{next} \notin B_j \; \forall j < i) \\ &\quad \times P(s_{next} \in B_i, s_{next} \notin B_j \; \forall j < i) \\ &= (1 - \frac{q}{n\rho})\delta(t) + \frac{q}{n\rho}\sum_{i=1}^{\infty} qp^{i-1} f_i(t) \end{aligned} \tag{3}$$

□

Corollary 1.1. *Laplace transform of the probability density function $g(t)$ has closed form:*

$$g(t) \xrightarrow{S\ Transform} G(s) = (1 - \frac{q}{n\rho})\delta(s) + \frac{q^2}{n\rho}\frac{F(s)}{1 - pF(s)} \tag{4}$$

Proof. Since we know that $\delta(t) \xrightarrow{S\ Transform} \delta(s)$ and $f_k(t) \xrightarrow{S\ Transform} F^k(s)$, we have $g(t) \xrightarrow{S\ Transform} G(s)$

$$\begin{aligned} &= (1 - \frac{q}{n\rho})\delta(s) + \frac{q}{n\rho}\sum_{i=1}^{\infty} qp^{i-1} F^i(s) \\ &= (1 - \frac{q}{n\rho})\delta(s) + \frac{q^2}{n\rho}\frac{F(s)}{1 - pF(s)} \end{aligned} \tag{5}$$

□

Given the transform $f(t)$, numerical inversion algorithms can be applied to calculate $g(t)$ for any desired t by inverting the transform $G(s)$ in Eq. (4) [9], [10]. Similarly, we can calculate the source distribution $f(t)$ by rewriting Eq. (4) to find $F(s)$ based on $G(s)$ and numerically inverting it. Subsequently, we investigate the final distribution $g(t)$ in two common cases; exact inter-arrivals $f(t) = \delta(t - T)$ and exponential distribution $f(t) = \lambda e^{-\lambda t} u(t)$

Proposition 1.1. *In case of a precise batch inter-arrival $f(t) \sim \delta(t - T)$:*

$$g(t) = (1 - \frac{q}{n\rho})\delta(t) + \frac{q^2}{n\rho}\sum_{i=1}^{\infty} p^{i-1}\delta(t - iT) \tag{6}$$

Proof. Leveraging Theorem 1 and the fact that $f_i(t) = \delta(t - iT)$, the final result is straightforward. □

Therefore, when we observe equal distant peaks in IAT PDF with peak values reduced by a p factor, we can estimate the number of packets in each batch and the timing between batches T.

Proposition 1.2. *In case of batches of packets, which are sent out with Poisson distribution (inter-arrival time of batches are following exponential distribution $f(t) = \lambda e^{-\lambda t} u(t)$), we observe exponential shape distribution with rate $q\lambda$:*

$$g(t) = (1 - \frac{q}{n\rho})\delta(t) + \frac{q}{n\rho}(q\lambda)e^{-q\lambda t} u(t) \tag{7}$$

Proof. Knowing that $f(t) = \lambda e^{-\lambda t} u(t)$, its Laplace transform is $F(s) = \frac{\lambda}{s+\lambda}$ and using Corollary 1.1:

$$\begin{aligned} G(s) &= (1 - \frac{q}{n\rho})\delta(s) + \frac{q^2}{n\rho} \frac{\frac{\lambda}{s+\lambda}}{1 - p\frac{\lambda}{s+\lambda}} \\ &= (1 - \frac{q}{n\rho})\delta(s) + \frac{q}{n\rho} \frac{q\lambda}{s + q\lambda} \qquad (8) \\ \xrightarrow{S\ Inverse} g(t) &= (1 - \frac{q}{n\rho})\delta(t) + \frac{q}{n\rho}(q\lambda)e^{-q\lambda t}u(t) \end{aligned}$$

□

3) Network jitter: Data traversing over a communication network experience varying delays. Modeling the transit delay requires to take into account several factors such as propagation delay, queuing delay, and congestion in the network. Since we aim at modeling the packet inter-arrival distribution targeting a network telescope, we only consider the effects of network jitter, which is defined as the transit delay of successive packets between the two measure points. Delay variation or jitter is an inherent feature of packet switched communication networks due to various bottlenecks. We assume that we are observing packets in a monitoring point, where all packets traverse a similar path. Assuming a constant jitter τ, the observed inter-arrival distribution is $g(t - \tau)$. Therefore, calculating the final inter-arrival distribution $h(t)$ after taking into account the sampling and the network jitter distribution $j(t)$ would be as follows:

$$h(t) = \int_{\tau=-\infty}^{\infty} j(\tau)g(t - \tau) = j(t) * g(t) \qquad (9)$$

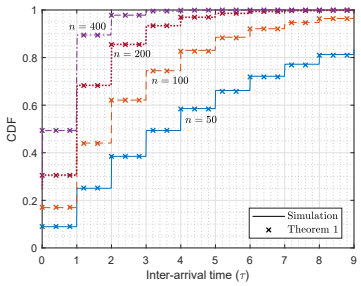
However, the actual distribution function $h(t)$ is slightly distorted due to the fact that often it is challenging to infer the exact order of the scanning packets and therefore, the observed IAT seem to be zero for values less than zero. This slight distortion is however ignored as it possesses negligible impact on the derived outcomes. Despite the fact that different models have been used to describe network jitter [11]–[13], the most common model which is often used in typical conditions (under no strong congestion, and in wide-area networks) is the Laplace distribution.

$$j(t|\alpha, \beta) = \frac{1}{2\beta} e^{-\frac{|t-\alpha|}{\beta}} \qquad (10)$$

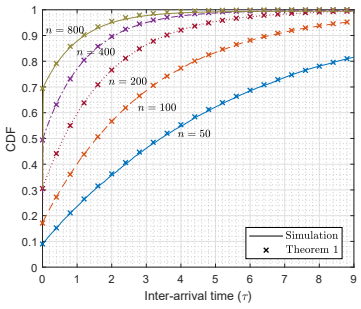
Proposition 1.3. *In case of precise batch inter-arrival* $f(t) \sim \delta(t-T)$*, the final observed inter-arrival distribution is* $h(t) = (1 - \frac{q}{n\rho})j(t) + \frac{q^2}{n\rho} \sum_{i=1}^{\infty} p^{i-1} j(t - iT)$

B. Validation

We perform a series of experiments to simulate packets' IAT at the darknet and compare results to the probability density of observed packet IAT derived using Theorem 1. As illustrated in Figures 1a and 1b, we choose different batch sizes (n packets) for the two modeled batch inter-arrival distributions: $\delta(t - T)$ (Figure 1a), and exponential (Figure 1b). Further, we choose $\rho = \frac{1}{256}$ for sampling probability at the darknet, and $\Delta = 50\mu s$. The analysis indicates a close to perfect accuracy when comparing theoretical and simulation results, which corroborates the validity and soundness of the derived theoretical relations.



(a) (with $T = 1$)



(b) (with $\lambda = 1$)

Fig. 1: Validating the accuracy of the relation in Theorem 1 against simulated batch IATs for (a) distribution $\delta(t - T)$ (Proposition 1.1), and (b) exponential distribution (Proposition 1.2). We select $\rho = \frac{1}{256}$ and $\Delta = 50\mu s$ for all tests.

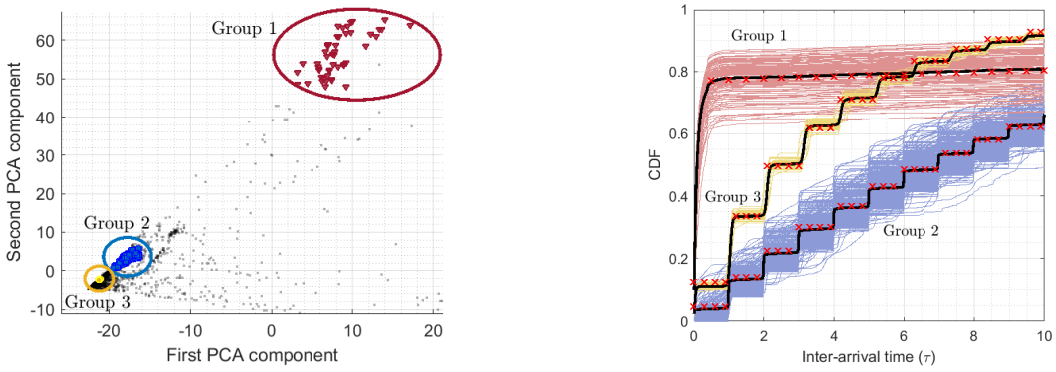
III. Experimental Results

A. Data Collection

We leverage passive data collected at a large-scale network telescope (darknet) maintained by the Center for Applied Internet Data Analysis (CAIDA) [5]. The darknet provides a thorough view of Internet scanning activities by capturing one-way traffic at a large number of routable, yet unused IP addresses (about 16M). CAIDA's network telescope represents a large destination IP address block, where it captures packets from a given source following similar paths with equal delays.

To this end, we utilized the algorithms developed in [2] to analyze about 3.6 TB of darknet data (Oct-08-2019), identify scanning traffic generated by compromised hosts (112,851), and infer device labels (IoT/non-IoT). In addition, we employed the Dvoretzky–Kiefer–Wolfowitz (DKW) inequality [14] to estimate the minimum sample size $s \geq (\frac{1}{2\epsilon^2}) \ln(\frac{2}{\alpha})$, which represents the amount of scanning packets required by each source for achieving acceptable accuracy in the estimated empirical probability functions. Our analysis resulted in selecting the error $\epsilon = 0.02$ and confidence level $\alpha = 0.1$ (i.e., 90% confidence), with a minimum sample size $s > 3,744$ packets.

In general, we identified 112,851 scanners that generated more than 4,000 packets, among which, about 82.4% were IoT. Further, we leveraged Mirai's traffic signatures [3] to identify IoT scanners with Mirai infections (47.8% of all).



(a) (HDBSCAN clustering) (b) IAT CDF

Fig. 2: Visualizing scans targeting port 23 with (a) correlated groups of devices targeting ports 23 (HDBSCAN), and (b) the distribution of IAT CDF for all groups. The solid black lines represent the mean (center) of all CDFs in each group. The fitted CDFs from Propositions 1.1 and 1.2 are marked with asterisks (x).

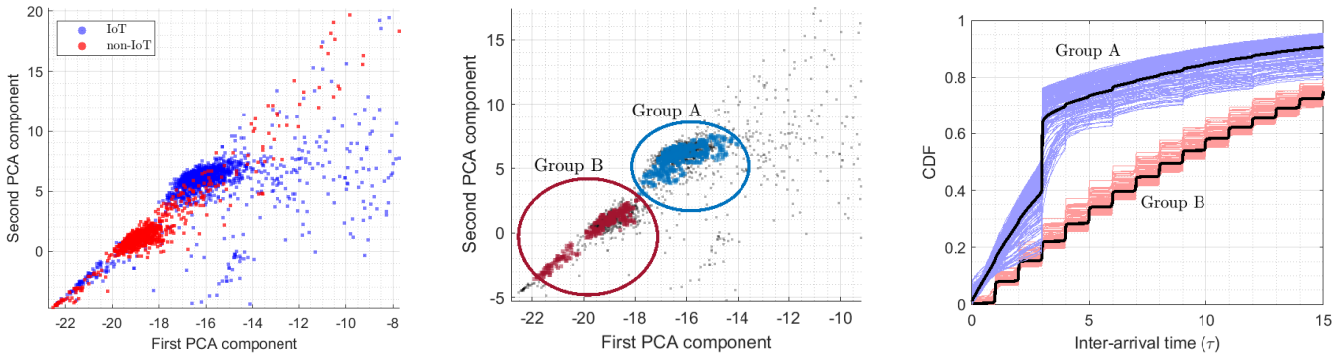
B. Empirical Analysis of Packet Inter-Arrival Times

To compare the IAT Probability Density Functions (PDF) and detect different classes of scans, we adopt the ℓ Wasserstein distance measure, which is an extension of the Euclidean distance metric for comparing distributional-valued data. Subsequently, we leveraged a tailored technique rooted in Principle Component Analysis (PCA) to analyze the probability Density functions [15]. This method reduces the dimensions of the data by measuring differences in a number of characteristics such as position, scale, and shape of their observed distributions. Following this approach, we transform the obtained IAT distributions to a 5-Dimensional space using the `HistDAWass` R package [16]. We used the data2hist function with manual break points, with 0.01 intervals to convert vectors of arrival times to histograms. Finally, we perform subsequent HDBSCAN clustering [17] with minimum number of neighbor points=3, min cluster size=100, and outliers threshold=0.4, to explore further correlations. Note that results are illustrated using 2 main PCA components in 2-D plots. In what follows, we demonstrate the applicability and added-value of the proposed stochastic model through use cases of scanning activities that target two prominent destination ports/services representing Telnet and HTTP.

Telnet port 23: Telnet ports (e.g., 23/2323) have been heavily targeted by compromised IoT devices in recent years. We investigate traffic generated by 7,957 devices that targeted Telnet port 23 by transforming their packet IAT to a 5D space. Furthermore, we perform subsequent clustering using HDBSCAN to identify correlated devices, as presented in Figure 2a. The results highlight three main groups of correlated devices, representing mainly IoT devices. Moreover, it is interesting to see that the majority of devices clustered in group #3 are labeled with Mirai signatures [2]. Furthermore, the vast majority of devices within every group follow almost the same distributions of packet IAT, as illustrated in Figure 2b. Moreover, the scanning signatures for each group represented by the mean value of their given distributions, confirm the theoretic derivations for packet IAT, as presented in Section II-A. We also fit these distributions to the theoretical models (Figure 2b) and show that packet IAT can indeed be a distinguishable feature when comparing scanning activities generated by compromised devices, especially those performing low-rate scans.

Our analysis resulted in identifying an inter-batch arrival distribution $f(t) = 72\delta(t-1.05)$ for devices within group #3, which translates to sending 72 packets every 1.05 seconds. Similarly, devices within group #2 were sending 25 packets every 1 second ($f(t) = 25\delta(t-1)$). Furthermore, each group of scanners might exhibit different overall target size (e.g., number of IP addresses). Therefore, we had to slightly adjust the ρ values to account for these differences while fitting the distributions. Group #1 on the other hand exhibits slightly different IAT distribution, with devices sending batches of 1,100 packets following $f(t) = 0.016e^{-0.016t}$ with $\lambda = 0.016$. This very low rate approximately equals to $\frac{1}{60}$, which means that these scanners send about 1,100 packets before going idle for some random time (average of 60 seconds), possibly due to processing the response packets, before sending the next batch of packets.

In addition to packet IAT, we obtain the scanner rates from the darknet, with groups #1 to #3 having rates equal to 0.0782, 0.0983, and 0.2860, respectively. It is important to note that the observed rates do not provide sufficient details for comparing behavioral characteristics among different scanners. Nevertheless, considering that we observe similar distributions of packet IAT in each identified cluster (Figure 2b), we may have a chance to accurately fingerprint these groups based on their IAT distributions, which in turn, characterize the implementation of their underlying scanning modules and parameters. Indeed, this highlights the importance of packet IAT analysis, which can be used along with other packet/flow features for further clustering/classification of scanning activities generated by different malware variants. In addition, our analysis shed light on an important characteristics of the majority of the explored scanners, which are found to generate low-rate scanning behaviors with some kind of rate limiting techniques, resulting in sending scanning packets in



(a) (2D visualization) (b) (HDBSCAN clustering) (c) IAT CDF

Fig. 3: Visualizing scans targeting ports 80/8080 with (a) two-dimensional IAT distribution, (b) correlated groups of devices targeting ports 80/8080 (HDBSCAN), and (c) CDF of the IAT.

batches separated by specific idle times that can be leveraged to characterize and distinguish between them.

HTTP ports 80 and 8080: We investigate IAT PDFs related to 2,743 compromised devices that targeted HTTP ports 80/8080, which are also among the most targeted ports when analyzing scanning campaigns [18]. Interestingly, our analysis highlights two distinguishable dense areas in Figure 3a, which illustrate high similarities in the scanning packets IAT distributions of the inferred IoT and non-IoT devices, respectively. We also follow the same approach used for analyzing port 23 by performing HDBSCAN clustering on the obtained IAT distributions, and identify two distinctive groups of scanners with correlated IAT probability distributions, with group A to contain mainly IoT devices and group B to correspond to non-IoT (Figure 3b). In addition, devices in group A send a larger number of scans every 3 seconds, while devices in group B send relatively fewer packets per second. Further, while it is clearly observed that the distribution of IATs within group A do not follow any of the proposed models in this letter (Figure 3c), we can still estimate the corresponding IAT distributions $f(t)$ by leveraging the numerical approaches explained following Corollary 1.1. Given this distinguishable behavior, it is clear that the analysis of packet IAT can in fact help in meaningfully classifying IoT and non-IoT devices based on characteristics of their scanning activities.

IV. Conclusion

In this letter, we propose and evaluate novel stochastic processes to model scanning activities generated by compromised IoT devices based on their packet Inter-Arrival Times (IAT), as perceived by a large-scale network telescope. Moreover, we perform empirical analysis using over 3 TB of recent data and evaluate our approach through two use cases of scanning activities targeting Telnet and HTTP services. The obtained results demonstrate the effectiveness of both our empirical approach and probabilistic models towards distinguishing IoT and non-IoT devices based on the distributions of their packet IAT, while correlating groups of IoT devices that are likely to be operating within well-defined orchestrated botnets.

References

[1] P.-A. Vervier and Y. Shen, "Before Toasters Rise Up: A View into the Emerging IoT Threat Landscape," in *Int. Symp. on Research in Attacks, Intrusions, and Defenses*. Springer, 2018, pp. 556–576.

[2] M. S. Pour *et al.*, "On Data-driven Curation, Learning, and Analysis for Inferring Evolving Internet-of-Things (IoT) Botnets in the Wild," *Computers & Security*, p. 101707, 2019.

[3] M. Antonakakis *et al.*, "Understanding the Mirai Botnet," in *26th USENIX Security Symp.*, Vancouver, BC, 2017, pp. 1093–1110.

[4] D. Leonard *et al.*, "Stochastic Analysis of Horizontal IP Scanning," in *2012 Proceedings IEEE INFOCOM*. IEEE, 2012, pp. 2077–2085.

[5] "The CAIDA UCSD Real-time Network Telescope Data," UCSD - Center for Applied Internet Data Analysis. Retrieved from https://www.impactcybertrust.org/, 2018.

[6] A. K. Marnerides and A. U. Mauthe, "Analysis and Characterisation of Botnet Scan Traffic," in *the Int. Conf. on Computing, Networking and Communications (ICNC)*, 2016, pp. 1–7.

[7] Z. Li *et al.*, "Automating Analysis of Large-scale Botnet Probing Events," in *Proc. of the 4th Int. Symp. on Information, Comput., and Commun. Security*, ser. ASIACCS 09, 2009, pp. 11–22.

[8] R. Gorenflo and F. Mainardi, "The mittag-leffler function in the thinning theory for renewal processes," *Theory of Probability and Mathematical Statistics*, vol. 98, pp. 105–113, 2019.

[9] J. Abate and W. Whitt, "A unified framework for numerically inverting laplace transforms," *INFORMS Journal on Computing*, vol. 18, no. 4, pp. 408–421, 2006.

[10] J. Abate, G. L. Choudhury, and W. Whitt, "An introduction to numerical transform inversion and its application to probability models," in *Computational probability*. Springer, 2000, pp. 257–323.

[11] E. J. Daniel, C. M. White, and K. A. Teague, "An interarrival delay jitter model using multistructure network delay characteristics for packet networks," in *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, vol. 2. IEEE, 2003, pp. 1738–1742.

[12] L. Rizo-Dominguez *et al.*, "Jitter in ip networks: a cauchy approach," *IEEE Communications Letters*, vol. 14, no. 2, pp. 190–192, 2010.

[13] Q. Li and D. L. Mills, "Jitter-based delay-boundary prediction of wide-area networks," *IEEE/ACM Transactions on Networking (TON)*, vol. 9, no. 5, pp. 578–590, 2001.

[14] A. Dvoretzky *et al.*, "Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator," *The Annals of Mathematical Statistics*, vol. 27, no. 3, pp. 642–669, 1956.

[15] R. Verde, A. Irpino, and A. Balzanella, "Dimension Reduction Techniques for Distributional Symbolic Data," *IEEE Transactions on Cybernetics*, vol. 46, no. 2, pp. 344–355, Feb 2016.

[16] A. Irpino and M. A. Irpino, "Package 'histdawass'," 2019.

[17] L. McInnes, J. Healy, and S. Astels, "hdbscan: Hierarchical density based clustering." *J. Open Source Software*, vol. 2, no. 11, p. 205, 2017.

[18] S. Torabi *et al.*, "Inferring, Characterizing, and Investigating Internet-Scale Malicious IoT Device Activities: A Network Telescope Perspective," in *Proc. of the 48th Annual IEEE/IFIP Int. Conf. on Dependable Systems and Networks (DSN)*, June 2018, pp. 562–573.