Theoretic Derivations of Scan Detection Operating on Darknet Traffic

Article *in* Computer Communications · August 2019

DOI: 10.1016/j.comcom.2019.08.014

2 authors:

Morteza Safaei Pour
University of Texas at San Antonio
6 PUBLICATIONS **16** CITATIONS
SEE PROFILE

Elias Bou-Harb
University of Texas at San Antonio
73 PUBLICATIONS **555** CITATIONS
SEE PROFILE

Theoretic Derivations of Scan Detection Operating on Darknet Traffic

Morteza Safaei Pour*, Elias Bou-Harb

Cyber Threat Intelligence Laboratory, College of Engineering & Computer Science, Florida Atlantic University, Florida, USA

Abstract

Cyber space continues to be threatened by various debilitating attacks. In this context, executing passive measurements by analyzing Internet-scale, one-way darknet traffic has proven to be an effective approach to shed light on Internet-wide maliciousness. While typically such measurements are solely conducted from the empirical perspective on already deployed darknet IP spaces using off-the-shelf Intrusion Detection Systems (IDS), their multidimensional theoretical foundations, relations and implications continue to be obscured.

In this article, we take a first step towards comprehending the relation between attackers' behaviors, the width of the darknet vantage points, the probability of detection and the minimum detection time. We perform stochastic modeling, derivation, validation, inter-correlation and analysis of such parameters to provide numerous insightful inferences, such as the most effective IDS and the most suitable darknet IP space, given various attackers' activities in the presence of detection time/probability constraints. One of the outcomes suggests that the detection strategy employed by the widely-deployed Bro IDS is ideal for inferring slow, stealthy probing activities by leveraging passive measurements. Further, the results do not recommend deploying the strategy utilized by the Snort IDS when the available darknet IP space is relatively small, which is a typical scenario when darknets are operated and employed on organizational networks. In addition, we provide an optimization problem set that identifies a new botnet early infection strategy, which can be leveraged by evolving stealthy bots to circumvent a certain IDS strategy as it operates on the darknet IP space. The implications of this formal derivation are especially factual with the advent of evolving paradigms such as IPv6 deployments, and the proliferation of highly-distributed, orchestrated, large-scale and stealthy probing botnets.

Keywords: Probing activities, Stochastic analysis, Botnet analysis, Darknet traffic, Data analytics, Network Telescope

1. Introduction

Cyber space has radically altered our every day life and impacted a large number of its crucial aspects. This is clearly realized nowadays with the large-scale adoption of the

*Corresponding author. Tel.: +1 561 931 7531

Email address: msafaeipour2017@fau.edu (Morteza Safaei Pour)

Internet-of-Things (IoT) paradigm [1], the modernization of Cyber-Physical Systems (CPS) [2] and the continuous rise and utilization of digital currencies [3], to name a few. Nevertheless, the increasing dependence on cyber space continues to make organizations and Internet-wide services highly vulnerable to targeted threats and exploitations. In an attempt to thwart such malicious attempts, typically, Intrusion Detection Systems (IDS) are often configured, deployed and managed. Complementary, in recent years, security operators and researchers have become increasingly interested in passive monitoring of unused Internet address spaces, which is often known as darknets or network telescopes [4]. A darknet is a collection of routable, allocated yet unused Internet Protocol (IP) addresses. These IP addresses have no interaction with other hosts and only passively gather packets without generating any replies. Since these unused address blocks contain no legitimate hosts, the received packets are characteristically unsolicited and are often the results of Internet-scale probing activities [5], backscattered packets from victims of denial of service attacks [6] or misconfiguration traffic [7].

As noted, one of the most prevalent darknet traffic types is related to probing activities. Such activities are indeed a first step and an enabler of a large number of cyber attacks [8]. The empirical results discussed by Raftopoulos et al. [9] showed that the probability of devices becoming infected by malware increased if they were previously scanned. Autonomously spreading worms [10] employ probing to fingerprint other vulnerable hosts to infect them. Botmasters, orchestrating large-scale botnets [11], adopt probing activities to identify and add more bots to their campaigns [12]. Very recently, the IoT-centric malware Mirai [13] was inferred to be generating a momentous amount of probing activities in an attempt to exploit Internet-facing IoT cameras and video recorders [13]. To this end, promptly detecting such probing activities often aids in preventing actual attacks from occurring or at least contributes in limiting the expansion of botnets. In this context, a darknet has recurrently proven its capability to infer probing activities by analyzing incoming packets to unused IP addresses [14].

There exists a plethora of research contributions which have been conducted on passive detection methods and the practical implementations of darknets [4], yet, to the best of the authors' knowledge, the research effort which endeavors to theoretically derive and analyze **darknet-specific** notions in the context of darknet size, scan detection algorithms operating on such darknet IP spaces, and attackers' behaviors, among various others, have never been attempted before. Indeed, the lack of such formal understanding hinders the optimized deployments and usage of the darknet IP space in a given network subnet. Further, without such formal analysis, one can not determine the best scan detection algorithms to leverage, given a certain attacker's behavior and the available network resources. Additionally, given the proliferation of evolving cyber events such as large-scale, stealthy probing botnets [15], it is required to leverage the formal analysis of the available passive measurement strategies and inference mechanisms, coupled with their implications, in order to select the most suitable approach to employ against these ever-evolving phenomena. Moreover, with the continuous deployment of IPv6, one needs to take into account the implications of passive measurements in such deployment settings, given an operated IDS strategy and certain requirements on detection time and probability. Please note that throughout this paper, when we refer to the open source IDS Snort or Bro, we are indeed specifically referring to their probing detection systems and settings (and not their full blown implementations) and any other probing inference mechanims which closely mimick their behaviors. Further, given that analyzing network information from all sub-networks for detecting bot campaigns is known to be hard [16], darknets typically provide a promising

Internet-scale (macroscopic) approach to infer distributed unsolicited behaviors and activities. In contrast to IoT botnets such as Mirai, Hajime [17] and brickerbot [18] which employ fast, stateless scanning modules to scan networks with the highest possible rate that can be relatively easily detected by IDS operating on the darknet [13], recently discovered IoT botnets similar to IoTroop [19] exploit a non-aggressive vulnerability scan and propagation methodology which makes it relatively stealthy, thus enabling it to pass under the radar of monitoring tools [20]. Another example, would be the Sality botnet which was a generic botnet that attempted to propagate in a low-rate fashion to circumvent deployed IDS on the darknet IP space [21].

Having discussed the aforementioned information and pointed out a few research gaps, we frame the contributions of this article as follows:

- Formalizing the operations of three, widely-deployed detection mechanisms (typically embedded in open source IDS) by focusing on their probing detection *modus operandi* when operated on the darknet IP space.

- For each of the formalized detection approaches, we perform stochastic modeling, derivation and validation of their detection probabilities, their minimum detection time and the minimum number of required darknet IP addresses to achieve a certain detection promptness and accuracy when conducting passive measurements.

- Shedding the light on the impact of detection time, given a certain probing rate and a particular width of the darknet vantage points, and comparing the effectiveness of the investigated detection strategies deployed on darknet traffic in various scenarios. This provides insightful results such as the particular detection strategy employed by the Bro IDS being more architecturally effective in the inference of stealth low-rate scanning cases. Another outcome demonstrates why the detection strategy utilized by the Snort IDS on a relatively small darknet is not recommended.

- Proposing an early infection methodology based on the derived relations of minimum detection time and darknet vantage width by framing it as an optimization problem set where stealthy botnets can exploit. We concur that such an expansion/propagation methodology, although may be generic to typical botnets, could prove advantageous to evolving IoT bots, enabling them to circumvent the radar of measurement techniques, such as those utilizing darknet monitoring.

The road-map of this paper is as follows. In the next section, we review the literature on various topics such as probing events, darknet as a means of probing detection, and stochastic analysis of scanning behavior. In Section 3, we formally define the considered detection strategies and other required preliminaries. To this end, we also present the stochastic derivation, validation and analysis of the defined detection methods in the context of detection probability and time. In Section 4, we execute, compare and contrast several experimentation by leveraging the proposed formalization scheme. In Section 5, we highlight the existence of a botnet (early infection stage) strategy that can be leveraged by evolving bots to avoid detection methods operated on passive measurements. Subsequently, in Section 6, we discuss the implications of some of the results on today's cyber security and Internet measurement challenges. We demonstrate the limitations of the proposed approaches and techniques in Section 7. Finally, in Section 8, we summarize the contributions of this work and pinpoint several topics that aim at paving the way for future work.

2. Related Work

Considering that the contributions of this article are at the intersection of scan detection, passive measurements, botnet analysis and worm propagation, in this section, we review the literature in the context of such inter-related topics.

2.1. Scan Detection

Since probing activities is an important topic in cyber security and Internet measurements, it has been the focus of attention in many contributions. In [5], the authors provided an extensive survey in which its categorize the scanning topic based on their natures, strategies, and approaches. Bhuyan et al. [22] present a taxonomy and survey related to cyber scanning and some of the existing detection mechanisms. Threshold Random Walk (TRW), proposed for scan detection by Jung et al. in [23], is one of the well-known methods which was implemented as part of the Bro [24] IDS. Sridharan et al. in [25] investigated the effectiveness of existing methods such as Snort and TRW, and proposes a new method dubbed as Time-based Access Pattern Sequential hypothesis testing (TAPS). Irwin et al. [26] empirically compared visualized scans with alert outputs of scan detection algorithms employed by Snort and Bro on darknet traffic. Leonard et al. [27] performed stochastic derivation of a number of relations in order to propose an optimal stealth distribution scanning activity based on the probability of detection. The authors undertook the attackers' perspective (and not the measurement point of view) in order to significantly minimize the probability of detection. Further, in [28], the authors analyzed data from a large darknet composed of 5.5 million addresses to detect and study Internet-wide probing activities. Additionally, Fukuda et al. [29] explored the effectiveness of darknets for detecting large-scale IPv6 scan activities. The intuition behind their proposed approach is based on the idea that firewalls, upon detecting any probing activity, will look up reverse DNS names of probes' source IP addresses. As such, the authors demonstrated that DNS backscatter can play an important role in IPv6 Internet measurements.

In contrast to such research contributions, which have been mainly dedicated to the empirical inference and analysis of scan activities, we rather focus on the stochastic analysis of scan activities from the perspective of detection systems as deployed on darknets to comprehensively discover their strengths and weaknesses under various situations and scenarios.

2.2. Passive measurements

The darknet (network telescope) topic have attracted significant initial attention from the research community through its usage in Distributed Denial of Service (DDoS) attack detection [6] and analysis of the propagation of worms [30]. Throughout the past, researchers have shifted their focus to monitoring large-scale cyber events using darknets [31] and towards the role of passive measurements in the study of amplification attacks [32, 33]. Further, a number of research works have been dedicated to studying the impact of reducing the number of utilized darknet IP addresses (i.e., the width of darknet vantage points). For instance, in [34], the authors introduced the concept of the sparse darknet, a network subnet that is sparsely populated with darknet addresses as a way to study the impact of this reduction on its effectiveness. Alternatively, other literature approaches analyzed effective sensor placement strategies such as distributed darknet IP address placement [35], considering placing such IP addresses near live hosts or analyzing the impact of special patterns of localization [36]. Moore et al. [37] studied the relationship between the size of network telescopes and the detection ability of different

network events, along with their precision based on stochastic and probabilistic relations. In [38], the authors collected data from a /20 darknet to show that IP scanning activities can be modeled accurately with mixtures of Poisson distributions. They validated their proposal using well-known scanning methodologies (by using activities generated by the scanning tool Zmap [39] and the Shodan service [40]) and verified that each scanning process has its own signature in terms of the number of mixed distributions and their corresponding parameters. Additionally, in [41], the authors investigated performance metrics such as detection rate, false alarm rate, computational cost and ease of use of two network traffic analysis tools, namely, Corsaro [42] and Cisco ASA 5515-X. The authors also investigated and reported the findings of the application of two machine learning classifiers (i.e., C4.5 Decision Tree, AdaBoost.M1) on diverse darknet datasets. Further, Benson et al. [43] opportunistically leveraged Internet Background Radiation (IBR) or unsolicited traffic sent to darknets, to investigate its effectiveness in generating macroscopic Internet measurements. To this end, the authors considered three application case studies related to identification of open resolvers, determining uptime, and characterizing path changes. Indeed, their goal was to highlight the strengths and limitations of employing IBR as a unique data source for Internet-wide measurements.

In contrast, we present a first attempt ever which exploits darknet-specific parameters and variables to formally comprehend the multidimensional relations between darknet vantage points, various probign detection strategies operating on such darknet IP spaces, the rate of the probing activities and the detection time/probability.

2.3. Botnet Analysis

Li et al. [44] presented analytical schemes as applied on honeynet data to understand the significance of large-scale botnet probing and explored the prevalence of different types of scanning activities. Moreover, they designed mathematical schemes to extrapolate the global properties of scanning events such as total population and target scope, from the limited, local view of a honeynet. Further, in [45], the authors discovered and reported on a large-scale, stealthy orchestrated scanning event by executing darknet analysis. The authors noted that the inferred bots aimed at achieving comprehensive coverage, though enduring higher costs in terms of task completion time. Such unsolicited activity was thoroughly investigated in [21], where a stealthy horizontal scan of the entire IPv4 address space conducted by the Sality botnet was investigated. In a closely related work, Raftopoulos et al. [9] correlated different datasets such as darknet and Snort alerts, with the aim to analyze the same Sality botnet scan. Based on their analysis, only 4% of all the probing flows were shown to have triggered a scan-related intrusion detection signature, which indeed demonstrated the stealthiness of such a large-scale botnet activity. Garcia et al. [46] provided a comprehensive survey about network-based botnet detection methods. The authors presented a new classification, taxonomy, and comparison of network-based botnet detection mechanism, along with extensive highlights in this area. Ban et al. [47] employed an abrupt change-detection algorithm on their darknet data for detecting botnet probing campaigns. Similarly, Bou-Harb et al. [16] proposed a probing botnet detection engine and empirically validated it by uncovering large-scale, previously undocumented stealthy botnets by solely leveraging passive measurements.

In contrast to such contributions, this papers extends network telescope research by initially deriving a stochastic passive measurement scheme to permit the analysis of two commonly employed probing detection strategies. Subsequently, we propose a novel, coordinated propagation

methodology (i.e., early infection strategy), which can be leveraged by ever-evolving botnets, including IoT-centric bots, to avoid a widely deployed detection method as it operates on darknet measurements.

2.4. Worm Propagation

Wang et al. [48] provided a comprehensive survey of worm propagation methods. The authors analyzed various models and highlighted each model's significance. The authors divided worms into two main categories: scan-based and topology-based. A scan-based worm probes the entire cyber space, or a sub-space, and infects vulnerable hosts to propagate itself. Code Red II [49], Witty [50], and Conficker [51] all fall under this scan-based class. In addition, recently emerged IoT botnets such as Mirai [13], Hajime [17], brickerbot [18] and IoTroop [19] use techniques and characteristics similar to scan-based worms. However, a topology-based worm locates new host targets by using information in the victim's machine. Examples are email worms [52, 53] and social network [54] worms such as Koobface [55]. Scan-based techniques can be further classified into random scanning (which consists of methods such as uniform scanning, hit-list scanning [56], routable scanning [57]) and localized scanning. Among all the proposed methodologies to increase the spreading speed and effectiveness, there exist several methods which concentrate on the early scanning stage with the aim of increasing the number of initially infected hosts prior to starting the large-scale probing stage. Hit-list scanning is an example that works by infecting all vulnerable hosts on the hit-list before initiating the scanning events. Clearly, there exists many challenges to this process of building such a hit-list, including accuracy and timeliness. Further, such process might alarm the deployed detection system about an incoming worm attack, especially when they are targeting the entire Internet space.

In this research, we present a novel orchestrated and stealthy scan-based early infection scheme. In this stage, scanners probe the entire network IP address space once (in an orchestrated fashion) and immediately infect identified vulnerable hosts without being detected by detection systems (as elaborated an analyzed in Section 3). This aids such an attacker to spread more efficiently in their early stage to increase the number of infected hosts. Subsequently, when their population is sufficient, they can employ any orchestrated stealthy probing technique, such as those noted in [27], or execute any other misdemeanors.

3. Formal Modeling and Stochastic Analysis

The purpose of this section is to formalize and define various kinds of probing detection method with the aim of finding relations between different parameters, such as minimum number of required darknet IP addresses, minimum detection time and the probability of detection for different scanning rates. Given a subnet S, consisting of $|S|$ IP addresses, we notate the set of darknet IP addresses, distributed within S and utilized in the detection process, as DIP. In this paper, following the natural behavior of large-scale probing events [15], we consider that the attacker intends to scan all IP addresses in S.

Indeed, there exists various scanning patterns such as sequential and uniform probing, which are typically employed for scanning Internet networks. In this context, we note the average scanning rate r and the average inter-probe delay $\frac{1}{r}$.

It is noteworthy to mention that as long as we assume that the darknet IP addresses are distributed uniformly in the intended network (or the attacker probes uniformly which is a common

practice [16]), the final relations and outcomes will continue to hold true. Thus, we consider the uniformity of the scans, which indicates that, on average, every $\frac{1}{r}$, the scanner would send a packet to an IP address in the analyzed network.

Performing the stochastic modeling of the detection methods would be the first required step for comprehending and analyzing the detection probability and the relations between various passive measurement parameters. For the sake of this work, we focus on three different detection strategies based on well-known, highly-deployed IDS.

Definition 1: $\rho(\tau)$ is the probability of detecting a probing activity X in less than τ time units from the start of the scan.

$$\rho(\tau) = \int_0^{\tau} Pr(alarm(t) = TRUE)dt \tag{1}$$

Definition 2: For a probing activity X with an average scanning rate r over a subnet S and given a certain Detection System (DS), the minimum detection time τ^{ϵ}_{min} is the minimum required time for DS to detect the scan with probability more than $1 - \epsilon$.

$$\tau^{\epsilon}_{min} = inf\{t \geq 0 : \rho(t) \geq 1 - \epsilon\} \tag{2}$$

$1 - \epsilon$ represents how confident the detection system should be to raise the alarm. A higher ϵ refers to a more relax condition on detection and consequently a larger false alarm rate.

3.1. FH Detection System

The first considered detection method is the First-Hit (FH) algorithm, which raises an alarm on the detection of the very first darknet packet. Indeed, this represents the simplest detection system that we analyze here to specify some bounds on the parameters. After the first hit, this DS raises the alarm. This method intuitively uses the lowest amount of memory and processing requirements for detection. While this approach might be effective, it undoubtedly could lead to a high false positive rate; it might identify received darknet packets, caused by backscattered activities or misconfiguration, as probing activities. Thus, we consider this technique and its detection time/probability as a reference model rather than a DS that can actually be operated in practice.

Recall that there exists $|DIP|$ darknet IP addresses in the subnet S. Thus, the probability of one of these darknet IP addresses being hit by probing packets is $q = \frac{|DIP|}{|S|}$. Therefore, the effective rate λ, the scanning rate that would actually be sensed by the darknet, would be $\lambda = qr$. Now, given an average scanning rate r, we can write $\rho(\tau)$ as in

$$\rho(\tau) = 1 - e^{-\lambda\tau} \tag{3}$$

Based on (2) and some mathematical operations, we can easily derive τ^{ϵ}_{min} from equation (3), as in

$$\tau^{\epsilon}_{min} = \frac{\log(\epsilon)}{-\lambda} \tag{4}$$

Further, we can infer the minimum required darknet IP addresses for specific τ^{ϵ}_{min} and ϵ, as follows.

$$\min |DIP| = \frac{|S| \ln(\epsilon)}{-r\tau^{\epsilon}_{min}} \tag{5}$$

3.2. *Detection System* I *(DS*I*)*

The second detection method is a window-based detection technique that is based on the widely-deployed, open source Snort [58] IDS. We refer to this detection system as DSI and subsequently describe its operations. Consider a counter $C_i(t) = 0$ for each observed source IP address i. After its reset (at time t), it starts counting received packets in a time window $[t, t + \Delta_{DSI}]$. During this time window, if the counter hits the threshold α_{DSI}, DSI raises an alarm, otherwise, the counter and the time window will be re-initiated. This algorithm is clearly more complex than the FH algorithm because it requires a timer to check the window's timeout, and thus memory is required for storing $C_i(t)$ for all packets arriving from different source IP addresses i. The operation of DSI is summarized in Algorithm 1.

Algorithm 1: DSI detection algorithm

```
C_i(0) = 0;
alarm = FALSE;
t_reset = 0;
while do
    if t ≤ t_reset + Δ_DSI then
        if A packet from source i is received then
            C_i(t) = C_i(t) + 1;
            if C_i(t) ≥ α_DSI then
                alarm = TRUE;
            end
        end
    else
        t_reset = t_reset + Δ_DSI;
        C_i(t) = 0;
    end
end
```

The DSI is defined with the parameter pair $(\Delta_{DSI}, \alpha_{DSI})$. Letting $\tau = p\Delta_{DSI} + \nu$ where $0 \leq \nu \leq \Delta_{DSI}$, we can then compute $\rho(\tau)$, as follows.

$$\rho(\tau) = 1 - W_0^p W_1 \tag{6}$$

where

$$W_0 = e^{-\lambda\Delta} \sum_{k=0}^{\alpha-1} \frac{(\lambda\Delta)^k}{k!} = \frac{\Gamma(\alpha, \lambda\Delta)}{(\alpha-1)!} = Q(\alpha, \lambda\Delta) \tag{7}$$

and

$$W_1 = e^{-\lambda\nu} \sum_{k=0}^{\alpha-1} \frac{(\lambda\nu)^k}{k!} = \frac{\Gamma(\alpha, \lambda\nu)}{(\alpha-1)!} = Q(\alpha, \lambda\nu) \tag{8}$$

where $\Gamma(\alpha, x)$ is the upper incomplete gamma function and $Q(\alpha, x)$ is the regularized upper incomplete gamma function. We use the probability of events for a Poisson distribution, which is a typical distribution observed for malicious packets targeting the darknet IP space [59], to derive W_0 and W_1. W_0 can be interpreted as the probability of $alarm = False$ at the end of the

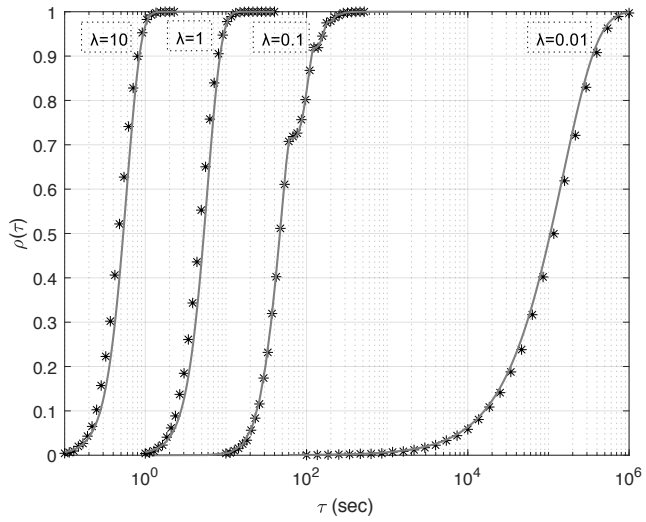

Figure 1: Validating the accuracy of the relation derived in (6) against simulation results (marked with asterisks) for DSI.

time window Δ, therefore, $W_0^p W_1$ is the probability of alarm not being raised from the beginning until time τ. The default values of $(\Delta_{DSI}, \alpha_{DSI})$ for the Snort IDS are $(60, 5)$.

We validate the accuracy of the formulation in Figure 1, which shows the derived relation in (6) against executed simulation results for DSI. For different values of λ, we can note a near-perfect accuracy, which corroborates the soundness of the derived relation.

For DSI, based on (6) and Definition 2, we have $\rho(\tau) = 1 - W_0^p W_1 = 1 - \epsilon$. Therefore, knowing that $W_0 \leq W_1 \leq 1$ leads to $1 - W_0^{p+1} \geq 1 - W_0^p W_1 \geq 1 - W_0^p$. For this purpose, we can write $(p+1)\ln(Q(\alpha, \lambda\Delta)) \leq \ln(\epsilon) \leq p \ln(Q(\alpha, \lambda\Delta)) \implies \lfloor \frac{\tau}{\Delta} \rfloor \leq \frac{\ln(\epsilon)}{\ln(Q(\alpha,\lambda\Delta))} \leq \lfloor \frac{\tau}{\Delta} \rfloor + 1$. Thus,

$$\tau^{\epsilon}_{min}(\lambda, \epsilon) \approx \frac{\Delta \ln(\epsilon)}{\ln(Q(\alpha, \lambda\Delta))}. \tag{9}$$

Equivalently, we can derive $\lambda = \frac{1}{\Delta} Q^{-1}(\alpha, \epsilon^{\frac{\Delta}{\tau_{min}}})$, where $Q^{-1}(\alpha, z)$ is the inverse regularized gamma function. If $Q(\alpha, x) = z$, then the inverse regularized gamma function $Q^{-1}(\alpha, z)$ is equal to x. Applying this results and recalling that $\lambda = \frac{|DIP|}{|S|} r$, we will obtain

$$\min |DIP| \approx \frac{|S|}{r} \frac{Q^{-1}(\alpha, \epsilon^{\frac{\Delta}{\tau_{min}}})}{\Delta} \tag{10}$$

3.3. Detection System II *(DS*II*)*

The third inference method is also a window-based DS, based on TRW [23] and related to the well-known Paxon's Bro IDS [60]. We use DSII to refer to this detection method. In such an IDS, for each source IP address i, a counter $C_i(t)$ is created. After receiving a packet from source host i at time t, the technique will wait $\Delta_{DS\,\mathrm{II}}$ time units to receive another packet. In case a packet hit the detection system during $[t, t + \Delta_{DS\,\mathrm{II}}]$, the IDS will increment $C_i(t)$; otherwise, it

will reset $C_i(t)$. Algorithm 2 summarizes the modus operandi embedded within DSII.

Algorithm 2: DSII detection algorithm

```
C_i(0) = 0;
alarm = FALSE;
t_reset = 0;
while do
    if t ≤ t_reset + Δ_DSII then
        if A packet from source i is received then
            C_i(t) = C_i(t) + 1;
            t_reset = t;
            if C_i(t) ≥ α_DSII then
                alarm = TRUE;
            end
        end
    else
        t_reset = t;
        C_i(t) = 0;
    end
end
```

$\rho(\tau)$ for DSII can be calculated based on (11), where $p_\alpha(t)$ is the probability of $Pr(Alarm(t) = True)$ for DSII with threshold α. To this end, we compute the Probability Distribution Function (PDF) of DSII with parameter α, recursively, based on the PDF of DSII with threshold $\alpha - 1$. Consequently, we can derive the CDF, which in fact refers to $\rho(\tau)$.

$$p_\alpha(t) = \begin{cases} \frac{1}{A}\int_{x=0}^{\Delta_{DSII}} p_{\alpha-1}(x)\lambda e^{-\lambda(t-x)}dx, & \text{if } t \geq \Delta_{DSII} \\ \frac{1}{A}\int_{x=0}^{t} p_{\alpha-1}(x)\lambda e^{-\lambda(t-x)}dx, & \text{if } t < \Delta_{DSII} \end{cases} \tag{11}$$

where $A = (1 - e^{-\lambda\Delta_{DSII}})$. Equation (11) can be shown with the convolution operator, as in (12). We employ Laplace Transform for calculating these recursive convolutions.

$$p_\alpha(t) = p_{\alpha-1}(t) * \frac{1}{A}g(t) = p_{\alpha-1}(t) * \frac{1}{A}\lambda e^{-\lambda t}(u(t) - u(t-\Delta)) \tag{12}$$

Therefore, $p_1(t) = \lambda e^{-\lambda t}u(t) \xrightarrow{S\ Transform} P_1(s) = \frac{\lambda}{s+\lambda}$ and $g(t) = \lambda e^{-\lambda t}(u(t) - u(t-\Delta)) \xrightarrow{S\ Transform} G(s) = \frac{\lambda}{s+\lambda}(1 - e^{-\Delta(s+\lambda)})$. We know that in S-Transform, we have the relation $f(t) * g(t) \leftrightarrow F(s)G(s)$. Thus, we can rewrite (12), as in (13).

$$\begin{aligned} P_\alpha(s) &= \frac{1}{A}P_{\alpha-1}(s)G(s) = \frac{1}{A^{\alpha-1}}P_1(s)G^{\alpha-1}(s) \\ &= \frac{1}{A^{\alpha-1}}(\frac{\lambda}{s+\lambda})^\alpha(1 - e^{-\Delta(s+\lambda)})^{\alpha-1} \\ &= \frac{1}{A^{\alpha-1}}(\frac{\lambda}{s+\lambda})^\alpha(\sum_{k=0}^{\alpha-1}(-1)^k\binom{\alpha-1}{k}e^{-k\Delta(s+\lambda)}) \end{aligned} \tag{13}$$

Inverse Laplace Transform of (13) can be calculated, and the result of (14) would be the PDF of the detection at time t. Now, we transfer the equation to the time domain, as in:

$$p_\alpha(t) = \frac{\lambda^\alpha e^{-\lambda t}}{A^{\alpha-1}(\alpha-1)!} \sum_{k=0}^{\alpha-1} (-1)^k \binom{\alpha-1}{k} (t-k\Delta)^{\alpha-1} u(t-k\Delta) \tag{14}$$

If we define the integral of the first term of (14) as in $X_0^\tau(t) = \frac{1}{A^{\alpha-1}} \int_0^\tau \frac{\lambda^\alpha}{(\alpha-1)!} t^\alpha e^{-\lambda t} u(t) dt$, then

$$\begin{aligned} \rho(\tau) = \int_0^\tau p_\alpha(t)dt &= \sum_{k=0}^{\alpha-1} (-1)^k \binom{\alpha-1}{k} e^{-k\lambda\Delta} X_0^{\tau-k\Delta}(t) \\ &= X_0^\tau(t) - e^{-\lambda\Delta} \binom{\alpha-1}{1} X_0^{\tau-\Delta}(t) + \dots \end{aligned} \tag{15}$$

Therefore, based on (15) for $\tau \le \Delta$, only the first term is nonzero, and for $\Delta < \tau \le 2\Delta$, only the first and second terms are nonzero and so forth. Further, because the Δ values are usually large (the default values of $(\Delta_{DS\,\mathrm{II}}, \alpha_{DS\,\mathrm{II}})$ for Bro are $(600, 20)$), the coefficient $e^{-k\lambda\Delta}$ for $k \ge 1$ is very small (for $\lambda = 0.1$ and $\Delta = 600$, $e^{-60} \approx 8.75\mathrm{e}{-27}$). Thus, we can solely consider the first term in our formulation. Additionally, the value of $\frac{1}{A^{\alpha-1}}$ is approximately equal to 1 for $\lambda \ge 0.01$ and $\Delta = 600$. After some mathematical manipulations, we can derive the probability of detection for DSII as in (16).

$$\begin{aligned} \rho(\tau) = X_0^\tau(t) &= \int_0^\tau \frac{\lambda^\alpha}{(\alpha-1)!} t^{\alpha-1} e^{-\lambda t} u(t) dt \\ &= -\sum_{j=0}^{\alpha-1} \left[\frac{(\lambda t)^j e^{-\lambda t}}{j!}\right]\Big|_{t=0}^{\tau} \\ &= 1 - e^{-\lambda\tau}\left(1 + \lambda\tau + \frac{(\lambda\tau)^2}{2!} + \dots + \frac{(\lambda\tau)^{\alpha-1}}{(\alpha-1)!}\right) \\ &= 1 - \frac{\Gamma(\alpha, \lambda\tau)}{(\alpha-1)!} = 1 - Q(\alpha, \lambda\tau) \end{aligned} \tag{16}$$

Numerical results reveal that for $\lambda \ge 0.001$, the exact formulation in (15) and the approximation in (16) have similar values, demonstrating high accuracy. Additionally, Figure 2 clearly depicts that the derived equation in (16) is quite accurate in comparison with generated simulation results.

For DSII, based on (16) and Definition 2, we can derive $\rho(\tau) = 1 - Q(\alpha, \lambda\tau) = 1 - \epsilon \implies Q(\alpha, \lambda\tau) = \epsilon$, and therefore,

$$\lambda\tau = Q^{-1}(\alpha, \epsilon) \tag{17}$$

$$\tau_{min}^\epsilon(\lambda) \approx \frac{Q^{-1}(\alpha, \epsilon)}{\lambda} \tag{18}$$

$$\min |DIP| \approx \frac{|S|}{r} \frac{Q^{-1}(\alpha, \epsilon)}{\tau_{min}} \tag{19}$$

In addition, (17) and (18) imply that for DSII with specific threshold α and a chosen ϵ, the $\lambda\tau_{min}^\epsilon(\lambda)$ is always constant.

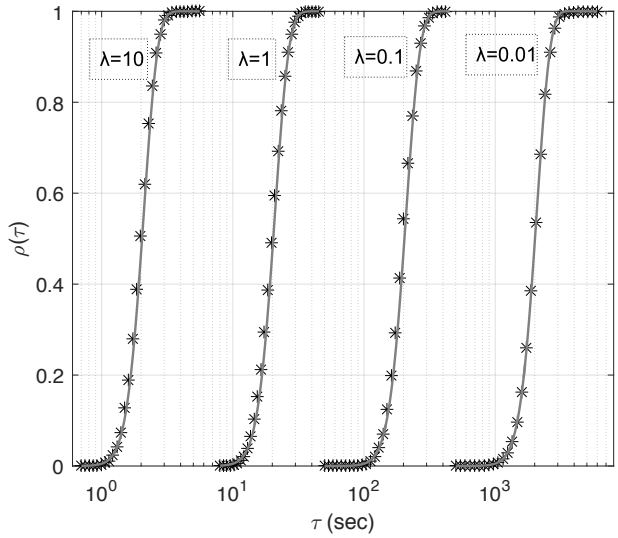

Figure 2: Validating the accuracy of the relation derived in (16) against simulation results (marked with asterisks) for DSII.

4. Experimentation and Results

In this section, we execute several experimentations to comprehend (i) the impact of the probing rate on the minimum detection time related to various employed IDS operated on the darknet IP space and (ii) the implication of the width of the darknet IP space on detection time. Further, motivated by real deployments of darknets, we analyze two case studies to shed light on the implications of the discussed detection systems in contrast with their detection promptness when operated on those specific darknet IP spaces.

Figure 3 shows various values of effective rate λ in contrast with the minimum detection time for DSI and DSII. Recall that $\lambda = qr = \frac{|DIP|}{|S|}r$, and therefore, λ is clearly dependent on the scanning rate r and the ratio of number of darknet IP addresses to the subnet size $|S|$. It is revealed from Figure 3 that for $\lambda > 0.1$, DSI outperforms DSII with respect to the minimum detection time, and for $\lambda \leq 0.1$, DSII outperforms DSI. From such results, one can extract that for stealthy, low-rate probing events, the detection strategy employed by the Bro IDS is more suited to perform the detection when operated on the darknet IP space. Please note that the result for the FH detection technique is solely depicted to show the lower bound for the minimum detection time; a DS can not reach a lower detection time than the minimum detection time of the FH algorithm for a specific λ.

We proceed by illustrating Figure 4, which shows the minimum required portion of deployed darknet IP addresses in the intended subnet in order to achieve a specific minimum detection time. We notate $/x$, which refers to the number of darknet IP addresses; $\frac{1}{2^x}$ of all the subnet address space. Therefore, $x = \log_2(\frac{|S|}{|DIP|})$, with a larger value for x indicating a lower portion of allocated darknet IP addresses. We compare DSI and DSII with their default parameters for Snort and Bro, respectively, given a fixed scanning rate $r = 100$. Figure 4 demonstrates that for $r = 100$ and $\tau_{min} < 300sec$, DSI requires less darknet IP addresses (thus reducing cost and management/monitoring resources) in comparison with DSII. Therefore, by employing the Snort

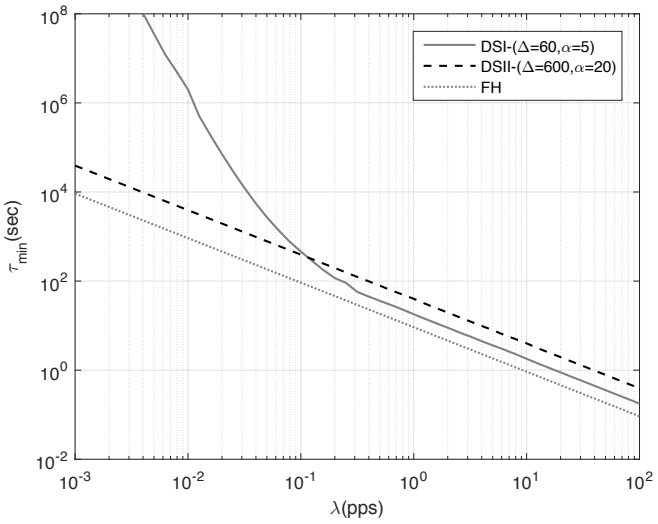

Figure 3: Minimum detection time τ_{min} versus effective rate λ for $\epsilon = 0.0001$

IDS, one can achieve the same minimum detection time by utilizing less darknet IP addresses. On the other hand, for $\tau_{min} \geq 300sec$, the minimum required darknet IP addresses is far lower for DSII. For instance, when $\tau_{min} = 10000$, the required darknet IP space would be a /11 for DSI and about /15 for DSII. This indicates that $2^{32-11} = 2^{21}$ darknet IP addresses are required to detect a large-scale probing activity targeting the entire IPv4 address space with a probability of more than 0.9999 in 10000 seconds if one employs the Snort IDS, and only $2^{32-15} = 2^{17}$ darknet IP addresses would be needed if one employs the Bro IDS to achieve the same objective.

We now consider two various darknet deployments representing two practical darknet setups that are currently deployed "in the wild". One refers to a /8 darknet, which resembles a large network telescope that is operated by the Center for Applied Internet Data Analysis (CAIDA)[1], while the other represents a /13 darknet operated by Farsight Security Information Exchange (SIE)[2]. On one hand, Figure 5 shows that for the /13 darknet, the execution of DSII on passive measurements leads to a lower minimum detection time in comparison with DSI (for a practical range of probing rate $1 \leq r \leq 1000$). Therefore, for a /13 darknet, the Bro IDS seems to be a more effective detection system, resulting in a lower detection time. On the other hand, Figure 6 shows comparative results for the /8 darknet. We can deduce that for this darknet setup, DSI (employed by the Snort IDS) appears to be a more suitable choice for detection, given an average probing rate $30 \leq r \leq 10000$. In the sequel, we propose an early infection methodology based on the derived relations, where stealthy botnets can leverage.

[1] http://www.caida.org/data/passive/telescope-near-real-time_dataset.xml
[2] https://archive.farsightsecurity.com/SIE_Channel_14/

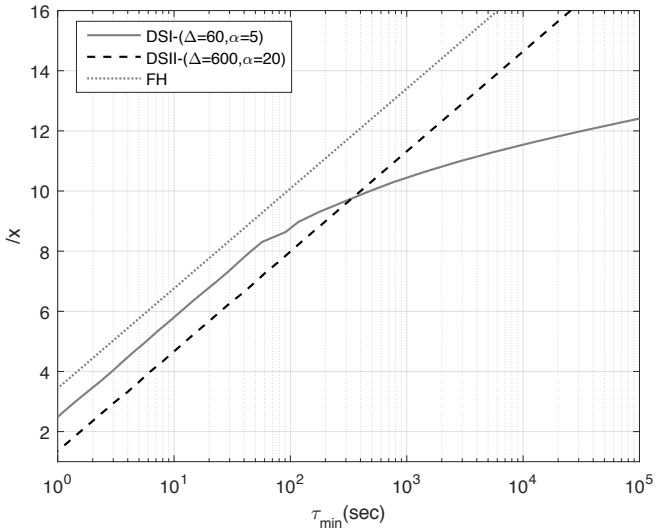

Figure 4: Portion of darknet IP addresses deployed within a certain subnet versus minimum detection time (τ_{min}) for scanning rate $r = 100$; $\epsilon = 0.0001$

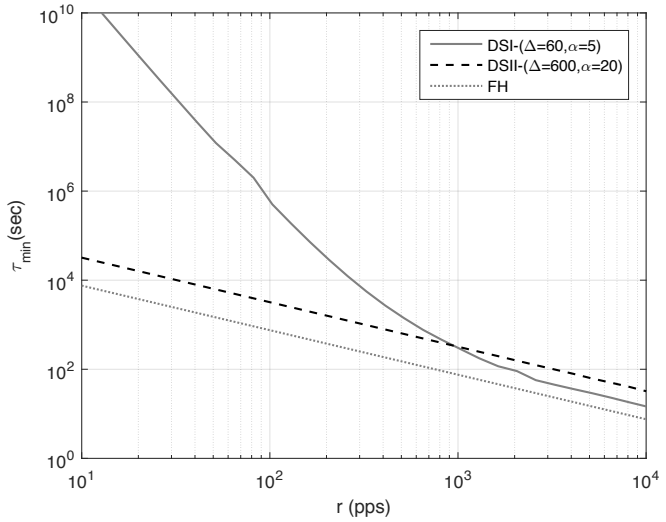

Figure 5: A /13 Network Telescope; $\epsilon = 0.0001$

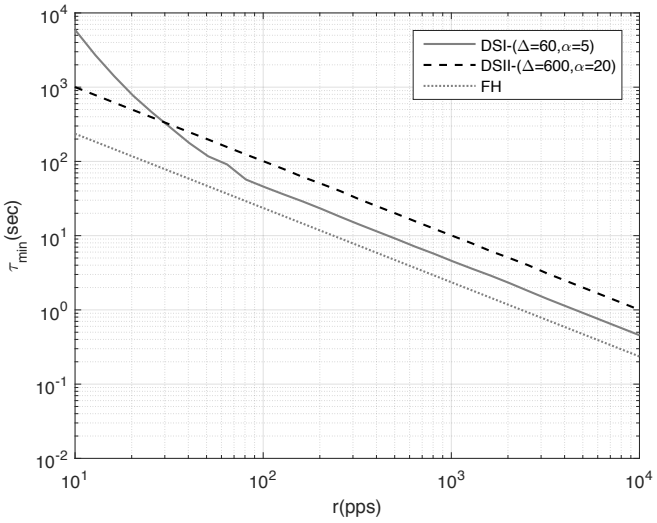

Figure 6: A /8 Network Telescope; $\epsilon = 0.0001$

5. Stealthy Botnet Early Infection Methodology

Distributing a probing activity among several bots that individually scan with intensely low rates tremendously increases the minimum detection time, and thus avoids passive measurement detection schemes. To this end, obtaining stealthiness using a large-scale probing botnet is relatively straightforward and typically involves scanning with very low rates in conjunction with using a divide-and-conquer technique or employing more sophisticated scanning methodologies [27]. To put things in perspective, to obtain optimal stealthiness, the rates for each bot in the context of the proposed stealth-optimal scanner in [27] for Snort and Bro are respectively 4 $packet/min$ and 1.9 $packet/min$. Thus, these techniques require a large number of bots, and therefore, they would not be effective at the beginning of the botnet life-cycle when there is only a handful of bots. Consequently, the botnet has no choice but to scan the Internet intensely (which will trigger detection methods operated on network telescopes) or use methods (at the early infection stage) to increase its initial size, before initializing the actual large-scale stealthy scan stage.

Therefore, the stealthy botnet goal herein is to (1) without raising any warning alarm of scan detection systems operated on the dark IP space, reach the intended population and (2) minimize the duration of this early infection stage to be less than several days. Thus, we focus here on the early infection stage of scan-based botnets; the stage where the botmaster, with a low number of bots, starts to scan the entire network IP address space (each IP address probed once in this stage) and immediately infects (in a negligible time compared to scanning time) any vulnerable hosts to join its scanning campaign. Subsequently, when the number of infected hosts reaches a level which permits the botnet to perform stealthy scans, it can resume the typical stealthy botnet probing (to infer other vulnerable hosts that will join the campaign) or execute its intended malicious activities such as DDoS, phishing, etc.

While we deem that this approach related to the early infection stage is generic to all scan-

based botnets, we believe that evolving IoT botnets might particularly take advantage of this scheme; indeed, there are different monitoring capabilities and resources in a deployed network, such as large distributed honeypots and darknets, as well as aggregated alarms from firewalls and IDS systems, which security experts exploit to gather information about Internet-wide botnet populations and evolutions to alarm the community as early as possible. Therefore, stealthy botnets typically intend to hide their activities from these sources. Firewalls and IDS systems are often not very effective when dealing with stealthy botnets, which can avoid alarms by methods such as random scanning. In this context, IoT devices being highly heterogeneous are quite hard to mimic using honeypots. Additionally, IoT devices typically have limited resources and consequently lack a sophisticated security measures. Therefore, it would be rational for a stealthy IoT botnet to focus more on methods that hide its track from darknets. Moreover, IoT devices are numerous in number, where a simple vulnerability in a specific model can lead to a large number of infected devices. For instance, the Mirai botnet compromised more than 1 million IoT devices just by brute forcing default credentials. Another example is the IoTroop [19] botnet which compromised a massive number of IoT devices by leveraging a small set of known attack vectors and vulnerabilities, rather than only compromising devices that use default credentials.

To this end, we propose a stealthy botnet probing methodology for the early infection stage with the aim of alerting about the possibility of leveraging this technique by future botnets. In this methodology, instead of randomly probing with the highest possible rate, we formulate the scanning bots in a divide-and-conquer manner as an optimization problem to find the best rate for each bot at each step to minimize the total early infection stage duration. We start with the most general state of an (IoT) botnet;

$$
\begin{aligned}
\text{minimize} \quad & T = \sum_{i=1}^{n} \tau_i = \tau_1 + \tau_2 + \cdots + \tau_n \\
\text{subject to} \quad & \sum_{i=1}^{n} (\tau_i \sum_{j=1}^{n} r_j) = |S|, \\
& \sum_{i}^{n} \tau_i \leq \tau^{\epsilon}_{min}(\frac{|DIP|}{|S|} r_i),\ i \in \{1, \ldots, n\}, \\
& \tau_i \sum_{j=1}^{i} r_j \geq \zeta,\ i \in \{1, \ldots, n\} \\
& \tau_i \geq 0,\ i \in \{1, \ldots, n\} \\
& r_i \geq 0,\ i \in \{1, \ldots, n\}
\end{aligned}
\tag{20}
$$

Here, T is the total scanning time to scan all addresses in S, and $\tau^{\epsilon}_{min}(\frac{|DIP|}{|S|} r_i) = \tau^{\epsilon}_{min}(\lambda_i)$ is the darknet minimum scan detection time with probability $1 - \epsilon$, if the scanner scanned with rate r_i. $\zeta = \frac{|S|}{\text{expected \#vulnerable hosts}}$ is the expected value of IP addresses that should be scanned to find a vulnerable host. This value can be estimated based on published empirical security research [61] or estimated on the fly after scanning some portion of S. In this formulation, we assume that the botnet is initiated with a bot scanning rate r_1, and after τ_1 time units, it will find a new vulnerable IoT device, which turns it into a new bot. Subsequently, the first bot continues scanning with rate r_1, and the new bot starts scanning with rate r_2; therefore, both of the bots will scan with overall

rate $r_1 + r_2$ for τ_2 to find a new vulnerable host and so forth untill they scan all the addresses in S and infect all vulnerable bots. Particularly, in the i^{th} step, i infected hosts will scan with overall rate $\sum_{j=1}^{i} r_j$ for τ_i time units to find the next vulnerable host.

In this problem set, n is equal to the total number of expected vulnerable hosts. If we assume that every bot starts scanning with a specific constant rate from the time it has joined the scanning campaign until T, the total scanning time of a bot which joined later is less than the bots that have joined the botnet earlier on. In other words, the total scan duration of the i^{th} bot is $T - \sum_1^{i-1} \tau_j$, and for the $(i+1)^{\text{th}}$ bot, it is $T - \sum_1^{i} \tau_j = T - \sum_1^{i-1} \tau_j - \tau_i$. Further, it is clear that the minimum detection time function $\tau^{\epsilon}_{min}(\lambda)$ will be descending with respect to λ. This means that the $(i+1)^{\text{th}}$ bot can adopt a higher probing rate r_{i+1} than that of the i^{th} bot r_i with the constraint of limiting its total scan duration $T - \sum_1^{i} \tau_j$ to less than the corresponding minimum detection time $\tau^{\epsilon}_{min}(\lambda_i)$. Thus, the main idea here is that the new bots will scan with higher rates without being detected.

However, parameter n, the expected number of vulnerable hosts, are usually very high which render the optimization problem quite complex. Hence, we simplify it by dividing all the processes into $m = \log_2 n$ steps. In this new simplified scenario, we assume that the botnet starts with a single bot and scans the network with rate r_1 for τ_1 to find two vulnerable hosts to infect. Additionally, it continues scanning with rate r_1, and two new bots will join the scanning campaign with rate r_2 for τ_2. These three bots will scan with overall rate $r_1 + 2r_2$ to find four new vulnerable hosts. Likewise, at step i, $\sum_{k=1}^{i} 2^k$ bots will scan with overall rate $\sum_{k=1}^{i} 2^k r_k$ to find 2^{i+1} vulnerable hosts and continue until they scan all the addresses in S. Thus, we obtain the following optimization problem, where $m = \log_2 n$:

$$
\begin{aligned}
\text{minimize} \quad & T = \sum_{i=1}^{m} \tau_i = \tau_1 + \tau_2 + \cdots + \tau_m \\
\text{subject to} \quad & \sum_{i=1}^{m} \left(2^{i-1} r_i (\sum_{j=i}^{m} \tau_j) \right) = |S|, \\
& \sum_{i}^{m} \tau_i \leq \tau^{\epsilon}_{min}(\frac{|DIP|}{|S|} r_i), \ i \in \{1, \ldots, m\}, \\
& (\sum_{j=1}^{i} 2^{j-1} r_j) \tau_i \geq 2^i \zeta, \ i \in \{1, \ldots, m\} \\
& \tau_i \geq 0, \ i \in \{1, \ldots, n\} \\
& r_i \geq 0, \ i \in \{1, \ldots, n\}
\end{aligned}
\tag{21}
$$

5.1. Case Study

In this section, given the lack of viable literature methodologies which we can use to directly perform comparisons against the proposed early botnet propagation approach, we analyze and assess the effectiveness of the proposed methodology for different scenarios under various employed detection systems, to find the probing rates of the bots and their duration at each step, to obtain a stealthy early infection stage. For practical reasons, we consider the following assumptions: (1) the botnet starts with only one bot, (2) scanning and spreading is done over the entire IPv4 IP space ($|S| = 2^{32}$), and (3) the expected number of vulnerable IoT devices is 2^{13}

Table 1: Rates and the duration at each step of the stealthy early infection stage with the assumption that Snort is employed on the /13 darknet

Steps	1	2	3	4	5	6	7	8	9	10	11	12	13	
Rates (pps)	193.3	206.5	220.5	230.4	237.6	243.9	251.2	260.4	272.6	288.3	308.5	334.0	366.1	
Duration (min)	613.4	475.1	371.0	314.6	281.0	255.0	229.3	201.3	171.2	140.7	111.7	85.7	63.8	3314

[61]; thus, $m = 13$ and $\zeta = \frac{|S|}{\text{expected \#vulnerable IoTs}} = \frac{2^{32}}{2^{13}} = 2^{19}$.

For implementation purposes, we employed MATLAB optimization toolbox [62] to find the set of rates that satisfy all the conditions of the problem set (21). However, because of the high number of nonlinear constraints, converging to the optimal solution is not simple. Nonetheless, we uncovered a set of rates that, while they are not exactly optimal, satisfy all the conditions. In the case where Snort's probing detection method is deployed on a /13 darknet IP space, the rates (with $\epsilon = 0.0001$), the duration at each step and the total early infection stage duration are summarized in Table 1. Based on these values, we can note that new bots joining the botnet in each step adopt higher rates, from 193.3 pps (first step) to 366.1 pps (last step), while the duration of each step decreases from 36809 seconds (first step) to 3830 seconds (last step). The total duration of the early infection stage will be 198863 sec = 55.24 $hours$ which demonstrates that the rates and total duration time are viable in practice.

However, for the case when Bro's probing detection method is employed on the same darknet IP space, there is no solution to satisfy the problem set. The maximum probed cyber space in a stealthy manner (i.e., without raising an IDS scan alert) is equal to $r_i * \tau^{\epsilon}_{min}(\lambda_i)$, which is the rate multiplied by the minimum detection time corresponding to that rate. Regarding Eq. (17) and Figure 3 (on page 13), we can derive that for DSII with specific threshold α and a chosen ϵ, the $\lambda\tau^{\epsilon}_{min}(\lambda)$ is always constant (it is a linear line in the $\log(\tau^{\epsilon}_{min}(\lambda))$-$\log(\lambda)$ plane). This means that the maximum probed cyber space without raising a scan alarm for DSII is always constant. For instance, in this case, this value is independent of the rate r and is $r\tau^{\epsilon}_{min}(\lambda) = 2^{13}\lambda\tau^{\epsilon}_{min}(\lambda) = 2^{13}Q^{-1}(\alpha, \epsilon) = 336127 \approx 2^{18}$ which is less than $2\zeta = 2^{20}$; this means that there is no solution that can meet the noted constraints. On the other hand, DSI shows a completely different behavior; reducing λ (equivalently r) leads to an increase in the maximum probed cyber space $\lambda\tau^{\epsilon}_{min}(\lambda)$ without raising the scan alarm for DSI. Therefore, a stealthy scanner can decrease its rate in exchange for probing larger portions of the targeted IP space.

In contrast, we consider the same assumptions, but when the size of the darknet is a /8. The results are reported in Table 2. Similarly to the /13 use case, there is no solution for the case when Bro is deployed. For the case when Snort is operated, the rates are from 2.4 pps to 3.8 pps, and the total duration of the early infection stage hugely surges towards the impractical value of 3225.5 $hours$ = 134.4 $days$.

From the aforementioned analysis, we can infer that DSI (i.e., Snort) is more susceptible to stealthy botnet propagation schemes, similar to the proposed methodology; as opposed to DSII (i.e., Bro). We observed that in both cases with the /8 and /13 darknet IP spaces, by employing the Bro IDS, we can mitigate the proposed propagation methodology. Further, in the case when the Snort IDS is employed, raising the darknet vantage width from /13 to /8 tremendously increases

Table 2: Rates and the duration at each step of the stealthy early infection stage with the assumption that Snort is employed on the /8 darknet

Steps	1	2	3	4	5	6	7	8	9	10	11	12	13	
Rates (pps)	2.421	2.506	2.627	2.712	2.765	2.803	2.844	2.900	2.981	3.102	3.275	3.522	3.868	
Duration (hours)	501.1	428.1	346.1	300.3	275.3	258.8	242.7	222.6	196.7	165.2	130.1	94.8	63.3	3225.5

the total duration of the (stealthy) infection stage from 55.24 *hours* to 134.4 *days*.

6. Discussion

The outcomes of the proposed darknet formalization scheme can be discussed in the context of three topics. First, with the continuous transition from IPv4 to IPv6, the IP address space has intensively increased from 2^{32} to 2^{128}. This larger cyber space indeed requires much more efforts and resources to be monitored, measured and assessed. The darknet IP space, being one of the main sources of Internet measurements for cyber threat intelligence, should also be adapted. Although the exhaustive and large-scale scanning of the entire IPv6 address space is still infeasible (due to its extremely large address space), there have been some contributions to find sub-spaces of IPv6 to probe, with the help of passive sources and specific target generation algorithms [63]. Indeed, the derived darknet formalized relations is helpful in selecting the best scan detection algorithm and efficient darknet size to deploy on the entire IPv6s space or any sub-space S of IPv6 based on the attacker's scope. In the following, we discuss the effect of scan detection algorithms on the minimum required daknet size within the context of IPv6 (as the target scope of the attackers).

As deduced from Section 4, in case of $r = 100\ pps$, $\epsilon = 0.0001$ and $\tau^{\epsilon}_{min} = 10000\ sec$, the minimum required portion of darknet IP addresses for DSI is /15 and for DSII is /11. Recall that this indicates that at least $2^{32-11} = 2^{21}$ darknet IP addresses will be required to detect a probing activity targeting the entire IPv4 address space using the Snort IDS and $2^{32-15} = 2^{17}$ using the Bro IDS, respectively. In contrast, when dealing with IPv6, these numbers are orders of magnitude larger and the implications are even more imperative; for the Snort IDS, one requires $2^{128-11} = 2^{117}$ darknet IP addresses and for the Bro IDS, $2^{128-15} = 2^{113}$ darknet IP addresses are needed, to infer a complete scan of the IPv6 address space. Thus, for IPv4, the difference in terms of required darknet IP addresses related to various IDS types is 15×2^{17} while for IPv6, it is a momentous 15×2^{113}. One can hence note that the choice of the probing detection method deployed on passive measurements can severely affect (and amplify) the cost of the resources as well as the darknet management efforts.

Second, we ought to consider highly-distributed scans, similar to the large-scale event reported in [15]. With distributed scans, the probing activity is divided among a large number of bots and as a result, the effective scanning rate that is sensed by the darknet is divided by the number of bots participating in the probing campaign. This phenomena can significantly reduce the effective rate λ. Hence, as observed in Figure 3, as λ continues to decrease, the gap between DSI and DSII increases vastly, pinpointing the importance of selecting a suitable detection methodology for combating such ever-evolving events. Nevertheless, one has to note that as seen in Figure 3, related to the minimum detection time, none of the investigated detection system is ideal for inferring such large-scale, orchestrated and distributed probing events, paving

the way for more tailored detection methodologies to be researched, designed and implemented in the near future.

Last but not least, considering the proposed technique for the botnet early infection stage, botmasters can take advantage of DSI weaknesses to circumvent detection. In DSII, the maximum probed cyber space without raising an alarm, $\lambda\tau^{\epsilon}_{min}(\lambda)$, independent of the scanning rate, is always constant. On the contrary, for DSI, the maximum probed cyber space without raising a scan alarm $\lambda\tau^{\epsilon}_{min}(\lambda)$ will increase as the scanning rate continues to decrease. For DSI (i.e., the Snort IDS) deployed on the darknet, this feature provides an alarming weakness that can be adopted by stealthy botnets (including ever-evolving IoT bots) to practically spread without leaving a trace in the logs of deployed detection systems.

7. Considerations and Limitations

In this section, we discuss several noteworthy points related to the proposed work. Initially, we have to note that Snort and Bro, which are both widely used on various vantage points in Internet and organizational networks, are primarily designed to operate on two-way traffic and thus are not optimized to specifically be deployed on darknet vantage points. However, we have to mention that their corresponding scan detection algorithms are widely considered as the primary steps for inferring unsolicited activities from darknet traffic. We hope that the presented formal schemes and outcomes of this paper are enlightening and could potentially be employed as building blocks for further investigation related to deployed scan detection algorithms and stealthy activities, in the context of darknet deployments and usages. Additionally, in this work, we assumed a constant probing rate for a scanner (i.e., for each bot) throughout the duration of its scan to avoid unnecessary complexity, which indeed might not be a completely valid assumption in the real world. For instance, in a more general form, for a multi-rate scanner, it would be challenging to determine the probability of detection and minimum detection time in the case when probing is executed with rate r_1 for τ_1, then when the rate is modified to r_2 for τ_2 and so forth. This interesting task and its implications on the formalized scheme is left for future work. Moreover, in this work, we solely considered a horizontal scan while in practice, the scanner might probe using strobe or vertical scanning. Nevertheless, given the proliferation of horizontal scans on the Internet (at least from the darknet perspective), which is confirmed by our recent empirical observations, we deem that exploring horizontal scans is a significant first step, especially in the context of formal approaches and the analysis of darknet-centric notions. Furthermore, in Section 5, providing an estimation of the expected number of vulnerable hosts (or IoT devices) might not always be feasible. However, apart from the empirical research which can assist with this task, there exists a number of fast scanning tools like Masscan [64] and Zmap [39] that can provide supplementary information to provide a lower bound on the expected number of vulnerable hosts with some level of accuracy. We also assumed that the duration of the infection process (such as the time for downloading and executing malware on a vulnerable host) is negligible. Further, the total early infection stage duration T is strongly dependent on the darknet vantage width and the number of vulnerable hosts, which might limit the practicality of the proposed early infection stage. Finally, the optimization problem set of Eq. (21), given its high number of nonlinear constraints, is hard to converge to its global optimum; therefore, we had to apply some simplification assumptions for the case study to obtain the sub-optimal values. Indeed, future work will explore auxiliary mathematical techniques in an attempt to address this limitation.

8. Concluding Remarks

Motivated by the fact that passive measurements by way of exploiting darknet IP spaces are significantly effective in generating various cyber threat intelligence in addition to the lack of formal modeling of darknet parameters, this article is among the first to present a formal perspective in such contexts. Several detection systems based on highly-employed methods were formalized and a number of derivations were computed and validated to shed light on the relations between detection probability/time, scanners' rates and the size of darknet. Some of the outcomes suggested the practical usage of the Bro IDS for inferring low-rate probing, its effective application in smaller darknet IP spaces given a setup that somehow tolerates a delay in detection, and its cost-reduction characteristics when implemented in IPv6 darknet deployment settings. Another outcome confirmed that the Bro IDS, by employing DSII, is resilient against stealthy botnet spreading, while the detection strategy employed by Snort IDS is susceptible to such methodology. Broadly, the outcomes pinpointed the lack of effective passive detection methodologies that are capable of inferring large-scale, distributed probes in a timely and practical manner.
As for future work, apart from addressing a number of current limitations as discussed in Section 7, we are also conducting various experimentations to find the optimal pair (Δ, α) with respect to the required scan detection precision in addition to using real darknet data to better situate the formalization scheme. Further, we are in the process of formally investigating the impact of contemporary IoT attacks in the context of passive measurements.

Acknowledgments

The authors would like to express their sincere gratitude in advance to the anonymous reviewers and editors for their constructive feedback. This work was supported by grants from the U.S. National Science Foundation (NSF) (Office of Advanced Cyberinfrastructure (OAC) #1755179 and OAC #1829698).

References

[1] A. Al-Fuqaha, A. Khreishah, M. Guizani, A. Rayes, M. Mohammadi, Toward better horizontal integration among iot services, IEEE Communications Magazine 53 (9) (2015) 72–79.
[2] E. Bou-Harb, W. Lucia, N. Forti, S. Weerakkody, N. Ghani, B. Sinopoli, Cyber meets control: A novel federated approach for resilient cps leveraging real cyber threat intelligence, IEEE Communications Magazine 55 (5) (2017) 198–204.
[3] T. Moore, The promise and perils of digital currencies, International Journal of Critical Infrastructure Protection 6 (3) (2013) 147–149.
[4] C. Fachkha, M. Debbabi, Darknet as a source of cyber intelligence: Survey, taxonomy, and characterization, IEEE Communications Surveys & Tutorials 18 (2) (2016) 1197–1227.
[5] M. D. E. Bou-Harb, C. Assi, Cyber scanning: a comprehensive survey, IEEE Communications Surveys & Tutorials 16 (3) (2014) 1496–1519.
[6] D. Moore, C. Shannon, D. J. Brown, G. M. Voelker, S. Savage, Inferring internet denial-of-service activity, ACM Transactions on Computer Systems (TOCS) 24 (2) (2006) 115–139.
[7] C. Fachkha, E. Bou-Harb, A. Keliris, N. Memon, M. Ahamad, Internet-scale probing of cps: Inference, characterization and orchestration analysis, in: Proceedings of NDSS, Vol. 17, 2017.
[8] V. Bartos, M. Zadnik, An analysis of correlations of intrusion alerts in an nren, in: Computer Aided Modeling and Design of Communication Links and Networks (CAMAD), 2014 IEEE 19th International Workshop on, IEEE, 2014, pp. 305–309.
[9] E. Raftopoulos, E. Glatz, X. Dimitropoulos, A. Dainotti, How dangerous is internet scanning, TMA (2015) 158–172.

[10] B. K. Mishra, S. K. Srivastava, B. K. Mishra, A quarantine model on the spreading behavior of worms in wireless sensor network, Transaction on IoT and Cloud Computing 2 (1) (2014) 1–12.
[11] M. Feily, A. Shahrestani, S. Ramadass, A survey of botnet and botnet detection, in: Emerging Security Information, Systems and Technologies, 2009. SECURWARE'09. Third International Conference on, IEEE, 2009, pp. 268–273.
[12] C. Fachkha, E. Bou-Harb, M. Debbabi, On the inference and prediction of ddos campaigns, Wireless Communications and Mobile Computing 15 (6) (2015) 1066–1078.
[13] M. Antonakakis, et al., Understanding the mirai botnet, Usenix Security Symposium.
[14] E. Bou-Harb, M. Debbabi, C. Assi, On fingerprinting probing activities, computers & security 43 (2014) 35–48.
[15] A. Dainotti, A. King, K. Claffy, F. Papale, A. Pescap, Analysis of a "/0" Stealth Scan from a Botnet, IEEE/ACM Transactions on Networking 23 (2) (2015) 341–354.
[16] E. Bou-Harb, C. Assi, M. Debbabi, Csc-detector: A system to infer large-scale probing campaigns, IEEE Transactions on Dependable and Secure Computing.
[17] W. Grange, Hajime worm battles mirai for control of the internet of things, Symantec Blog, April.
[18] C. Kolias, G. Kambourakis, A. Stavrou, J. Voas, Ddos in the iot: Mirai and other botnets, Computer 50 (7) (2017) 80–84.
[19] Iotroop botnet: The full investigation, https://research.checkpoint.com/iotroop-botnet-full-investigation/, online; accessed 11 July 2019.
[20] N.-K. Nguyen, Iotroop, a new pandemic affecting connected objects, https://www.digital.security/en/blog/iotroop-new-pandemic-affecting-connected-objects, online; accessed 11 July 2019.
[21] A. Dainotti, A. King, F. Papale, A. Pescape, et al., Analysis of a/0 stealth scan from a botnet, in: Proceedings of the 2012 ACM conference on Internet measurement conference, ACM, 2012, pp. 1–14.
[22] M. H. Bhuyan, D. Bhattacharyya, J. K. Kalita, Surveying port scans and their detection methodologies, The Computer Journal 54 (10) (2011) 1565–1581.
[23] J. Jung, V. Paxson, A. W. Berger, H. Balakrishnan, Fast portscan detection using sequential hypothesis testing, in: Security and Privacy, 2004. Proceedings. 2004 IEEE Symposium on, IEEE, 2004, pp. 211–225.
[24] I. Bro, Homepage: http://www. bro-ids. org (2017).
[25] A. Sridharan, T. Ye, S. Bhattacharyya, Connectionless port scan detection on the backbone, in: Performance, Computing, and Communications Conference, 2006. IPCCC 2006. 25th IEEE International, IEEE, 2006, pp. 10–pp.
[26] B. Irwin, J.-P. van Riel, Using inetvis to evaluate snort and bro scan detection on a network telescope, in: VizSEC 2007, Springer, 2008, pp. 255–273.
[27] X. W. D. Leonard, Z. Yao, D. Loguinov, Stochastic analysis of horizontal ip scanning, in: INFOCOM, 2012 Proceedings IEEE, IEEE, 2012, pp. 2077–2085.
[28] Z. Durumeric, M. Bailey, J. A. Halderman, An internet-wide view of internet-wide scanning., in: USENIX Security Symposium, 2014, pp. 65–78.
[29] K. Fukuda, J. Heidemann, Who knocks at the ipv6 door?: Detecting ipv6 scanning, in: Proceedings of the Internet Measurement Conference 2018, ACM, 2018, pp. 231–237.
[30] M. Bailey, E. Cooke, F. Jahanian, D. Watson, The blaster worm: Then and now, IEEE Security & privacy 3 (4) (2005) 26–31.
[31] A. Dainotti, R. Amman, E. Aben, K. C. Claffy, Extracting benefit from harm: using malware pollution to analyze the impact of political and geophysical events on the internet, ACM SIGCOMM Computer Communication Review 42 (1) (2012) 31–39.
[32] C. Fachkha, E. Bou-Harb, M. Debbabi, Fingerprinting internet dns amplification ddos activities, in: New Technologies, Mobility and Security (NTMS), 2014 6th International Conference on, IEEE, 2014, pp. 1–5.
[33] C. Rossow, Amplification hell: Revisiting network protocols for ddos abuse., in: NDSS, 2014.
[34] W. Harrop, G. Armitage, Defining and evaluating greynets (sparse darknets), in: Local Computer Networks, 2005. 30th Anniversary. The IEEE Conference on, IEEE, 2005, pp. 344–350.
[35] Z. M. D. W. F. J. E. Cooke, M. Bailey, D. McPherson, Toward understanding distributed blackhole placement, in: Proceedings of the 2004 ACM workshop on Rapid malcode, ACM, 2004, pp. 54–64.
[36] J. Göbel, P. Trinius, Towards optimal sensor placement strategies for early warning systems., in: Sicherheit, 2010, pp. 191–204.
[37] D. Moore, C. Shannon, G. M. Voelker, S. Savage, Network telescopes: Technical report, Department of Computer Science and Engineering, University of California, San Diego, 2004.
[38] G. De Santis, A. Lahmadi, J. Francois, O. Festor, Modeling of ip scanning activities with hidden markov models: Darknet case study, in: New Technologies, Mobility and Security (NTMS), 2016 8th IFIP International Conference on, IEEE, 2016, pp. 1–5.
[39] Z. Durumeric, E. Wustrow, J. A. Halderman, Zmap: Fast internet-wide scanning and its security applications., in: USENIX Security Symposium, Vol. 8, 2013, pp. 47–53.
[40] J. Matherly, Shodan search engine, Available at [Online]: https://www. shodan. io.

[41] E. Balkanli, A. N. Zincir-Heywood, Highlights on analyzing one-way traffic using different tools, in: 2015 IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA), IEEE, 2015, pp. 1–8.
[42] CAIDA, Corsaro, http://www.caida.org/tools/measurement/corsaro/, online; accessed 11 July 2019.
[43] K. Benson, A. Dainotti, k. claffy, A. C. Snoeren, M. Kallitsis, Leveraging internet background radiation for opportunistic network analysis, in: Proceedings of the 2015 Internet Measurement Conference, IMC '15, ACM, New York, NY, USA, 2015, pp. 423–436. doi:10.1145/2815675.2815702.
URL http://doi.acm.org/10.1145/2815675.2815702
[44] Z. Li, A. Goyal, Y. Chen, V. Paxson, Towards situational awareness of large-scale botnet probing events, IEEE Transactions on Information Forensics and Security 6 (1) (2011) 175–188.
[45] A. Dainotti, A. King, K. Claffy, Analysis of internet-wide probing using darknets, in: Proceedings of the 2012 ACM Workshop on Building analysis datasets and gathering experience returns for security, ACM, 2012, pp. 13–14.
[46] S. García, A. Zunino, M. Campo, Survey on network-based botnet detection methods, Security and Communication Networks 7 (5) (2014) 878–903.
[47] T. Ban, L. Zhu, J. Shimamura, S. Pang, D. Inoue, K. Nakao, Detection of botnet activities through the lens of a large-scale darknet, in: International Conference on Neural Information Processing, Springer, 2017, pp. 442–451.
[48] Y. Wang, S. Wen, Y. Xiang, W. Zhou, Modeling the propagation of worms in networks: A survey, IEEE Communications Surveys & Tutorials 16 (2) (2014) 942–960.
[49] D. Moore, C. Shannon, et al., Code-red: a case study on the spread and victims of an internet worm, in: Proceedings of the 2nd ACM SIGCOMM Workshop on Internet measurment, ACM, 2002, pp. 273–284.
[50] C. Shannon, D. Moore, The spread of the witty worm, IEEE Security & Privacy 2 (4) (2004) 46–50.
[51] G. Lawton, On the trail of the conficker worm, Computer 42 (6).
[52] D. M. Kienzle, M. C. Elder, Recent worms: a survey and trends, in: Proceedings of the 2003 ACM workshop on Rapid malcode, ACM, 2003, pp. 1–10.
[53] Email-worm:vbs/loveletter, https://www.f-secure.com/v-descs/love.shtml, online; accessed 12 January 2019.
[54] W. Fan, K. Yeung, Online social networksparadise of computer viruses, Physica A: Statistical Mechanics and its Applications 390 (2) (2011) 189–197.
[55] w32.koobface, https://www.symantec.com/security_response/writeup.jsp%3Fdocid%3D2008-080315-0217-99, online; accessed 12 January 2019.
[56] S. Staniford, V. Paxson, N. Weaver, et al., How to own the internet in your spare time., in: USENIX Security Symposium, Vol. 2, 2002, pp. 14–15.
[57] C. C. Zou, D. Towsley, W. Gong, S. Cai, Routing worm: A fast, selective attack worm based on ip address information, in: Proceedings of the 19th Workshop on Principles of Advanced and Distributed Simulation, IEEE Computer Society, 2005, pp. 199–206.
[58] M. Roesch, et al., Snort: Lightweight intrusion detection for networks., in: Lisa, Vol. 99, 1999, pp. 229–238.
[59] E. Bou-Harb, M. Debbabi, C. Assi, A novel cyber security capability: Inferring internet-scale infections by correlating malware and probing activities, Computer Networks 94 (2016) 327–343.
[60] V. Paxson, Bro: a system for detecting network intruders in real-time, Computer networks 31 (23) (1999) 2435–2463.
[61] M. Galluscio, N. Neshenko, E. Bou-Harb, Y. Huang, N. Ghani, J. Crichigno, G. Kaddoum, A first empirical look on internet-scale exploitations of iot devices, Proceedings of 28th International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC) (2017).
[62] Matlab optimization toolbox, https://www.mathworks.com/products/optimization.html.
[63] A. Murdock, F. Li, P. Bramsen, Z. Durumeric, V. Paxson, Target generation for internet-wide ipv6 scanning, in: Proceedings of the 2017 Internet Measurement Conference, ACM, 2017, pp. 242–253.
[64] Masscan, https://github.com/robertdavidgraham/masscan, online; accessed 23 January 2019.