A Probabilistic Model to Preprocess Darknet Data for Cyber Threat Intelligence Generation

Elias Bou-Harb
Department of Computer & Electrical Engineering and Computer Science
Florida Atlantic University
ebouharb@fau.edu

***Abstract*—Internet traffic destined to routable yet unallocated IP addresses is commonly referred to as telescope or darknet data. Such unsolicited traffic is frequently, abundantly and effectively exploited to generate various cyber threat intelligence related, but not limited to, scanning activities, distributed denial of service attacks and malware identification. However, such data typically contains a significant amount of misconfiguration traffic caused by network/routing or hardware/software faults. The latter not only immensely affects the purity of darknet data, which hinders the accuracy of inference algorithms that operate on such data, but also wastes valuable storage resources. This paper proposes a probabilistic model to preprocess darknet data in order to prepare it for effective use. The aim is to fingerprint darknet misconfiguration traffic and subsequently filter it out. The model is advantageous as it does not rely on arbitrary cut-off thresholds, provide separate likelihood models to distinguish between misconfiguration and other darknet traffic, and is independent from the nature of the source of the traffic. To the best of our knowledge, the proposed model renders a first attempt ever to formally tackle the problem of preprocessing darknet traffic. Through empirical evaluations using real darknet traffic and by comparing the proposed model against the baseline and a heuristic approach, we demonstrate the accuracy and effectiveness of the model.**

Keywords*—*Darknet misconfiguration, Darknet data, Big data preprocessing, Passive monitoring, Cyber intelligence

I. Introduction

In our current era where cyber misdemeanors continue to threaten Internet and enterprise security, there is an ever increasing need for cyber threat intelligence. Recently, such misdemeanors were rendered by novel and devastating distributed amplified denial of service attacks [1], the rise of sophisticated malware targeting cyber-physical systems [2], large-scale probing events [3] and targeted exploitations [4]. Although there exists a number of approaches that allow the inference, characterization and attribution of such malicious activities, including, active monitoring of network flows and honeypot deployment, nevertheless, passive monitoring of the Internet remains a highly exploitable and effective approach to achieve these imperative cyber security tasks. Compared to the other mentioned methods, passive monitoring of Internet traffic is typically easier to deploy and manage, covers a significantly larger Internet Protocol (IP) space, and is naturally non-intrusive, which is particularly beneficial from both, privacy as well as performance perspectives.

In order to employ passive monitoring and analysis of Internet traffic, a darknet is usually deployed. A darknet (also commonly referred to as a network telescope) [5] could be defined by a set of routable and allocated yet unused IP addresses. It indeed represents a partial view of the entire Internet address space. From a design perceptive, a darknet is transparent and indistinguishable compared with the rest of the Internet space. From a deployment perspective, it is rendered by network sensors that are implemented and dispersed on numerous strategic points throughout the Internet. Such sensors are often distributed and are typically hosted by various global entities, including Internet Service Providers (ISPs), academic and research facilities, and backbone networks. The aim of a darknet is to provide a lens on Internet-wide malicious traffic; since darknet IP addresses are unused, any traffic targeting them represents a view of anomalous unsolicited traffic. Darknet analysis has demonstrated its effectiveness in generating prompt cyber threat intelligence as well supporting numerous Internet measurement studies [4, 6]. Further, for the past 4 years, we have been working on exploiting such data for numerous goals, including, but not limited to, malware fingerprinting [7] and malicious event characterization [3].

Although darknet data mostly contains malicious packets originating from probes, backscattered packets from victims of distributed denial of service attacks and malware propagation attempts, among others, it might also include what is dubbed as misconfiguration traffic. The latter non-malicious packets might be caused by network/routing or hardware/software faults that were erroneously directed towards a darknet. Such traffic can also be an artifact of an improper configuration when deploying a darknet.
In one of our earlier studies [8], we have estimated that such misconfiguration traffic constituted a momentous 30% of the entire analyzed dataset. Further, on a daily basis, we spot sporadic misconfiguration traffic while analyzing darknet data.

Indeed, misconfiguration traffic “pollute” darknet data as such traffic can not be exploited for cyber threat intelligence. Further, misconfiguration traffic makes it harder for cyber threat intelligence algorithms to operate correctly on darknet data, which often yields to numerous undesirable false positives and false negatives. Another drawback of the existence of misconfiguration traffic within darknet data, is that it wastes valuable storage resources. The latter is especially imperative if one has to index and analyze darknet data for a long period of time (i.e., years for instance), which is common in darknet measurement studies [4].

Motivated by the above, this paper tackles a very particular issue related to darknet data rendered by a preprocessing model that aims at fingerprinting misconfiguration traffic. Specifically, we frame the contributions of this work in the following two core threads:

978-1-4799-6664-6/16/$31.00 ©2016 IEEE

- Proposing a formal probabilistic model that aims at preprocessing darknet data to prepare it for effective use. The model is advantageous as it does not rely on arbitrary cut-off thresholds, provide different likelihood models to distinguish between misconfiguration and other darknet traffic, and is independent from the nature of the source of the traffic. Further, the proposed model neatly captures the natural behavior of misconfiguration traffic as it targets the darknet. To the best of our knowledge, the presented work presents a first attempt ever to systematically fingerprint and thus filter-out darknet misconfiguration traffic.

- Evaluating the proposed model using real darknet data and comparing the outcome against a heuristic approach from the literature.

The paper is organized as follows. In the next section, we present the proposed model and elaborate on its inner details. In Section III, we empirically evaluate the probabilistic model, compare and contrast its findings against other literature approaches, and comment on its performance. In Section IV, we discuss some limitations of the proposed model and offer some improvements for future work. Finally, Section V discusses some related work while Section VI summarizes the contributions of this paper.

II. Proposed Model

In this section, we elaborate on the proposed probabilistic model that is particularly tailored towards the goal of preprocessing darknet data by fingerprinting and thus filtering out misconfiguration traffic.

In a nutshell, the model formulates and computes two metrics that aim at capturing the natural and the characteristical behavior of misconfiguration flows as they target the darknet IP space. The model initially estimates the "rareness of access"; the degree to which access to a given darknet IP address is unusual. The model further considers the "scope of access"; the number of distinct darknet IP addresses that a given remote source has accessed. Subsequently, the joint probability is formulated, computed and compared. If the probability of the source generating a misconfiguration is higher than that of the source being malicious (i.e., scanning or backscattered), then the source is deemed as one that is generating misconfiguration traffic, subsequently flagged, and its corresponding generated darknet flows are filtered out. The above two metrics are elaborated next.

Let $D = \{d_1, d_2, d_3, \cdots\}$ represent the set of darknet IP addresses and D_i a subset of those accessed by source s_i. First, the model captures how unusual the accessed destinations are. The idea behind this metric stems from the fact that misconfigured sources[1] access destinations that have been accessed by few other sources [9]. Thus, the model estimates the distribution of a darknet IP d_i being accessed by such a source as

$$P_{misc}(d_i) = \frac{n_s(d_i)}{\displaystyle\sum_{\forall d_j \in D} n_s(d_j)} \tag{1}$$

where $n_s(d_i)$ is the number of sources that have accessed d_i. In contrary, a malicious darknet source[2] will access a destination at random. Typically, defining a suitable probability distribution to model the randomness of a malicious source targeting a specific darknet destination is quite tedious; often a simplistic assumption is applied to solve this issue. In this context, a very recent work by Durumeric et al. [10] have demonstrated that darknet sources will probe their targets following a Gaussian distribution. By adopting that assumption, one can model the probability of a darknet destination accessed by a malicious source as

$$P_{mal}(d_i) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \tag{2}$$

where σ is the standard deviation, μ is the mean, σ^2 is the variance and x is the location of the darknet destination following the distribution. Recall that equations (1) and (2) allows the model to initially capture how unusual the accessed destinations are. However further, the model considers how many darknet destinations have been accessed by a given source. The latter will be subsequently described.

Given a set of D_i, darknet destinations accessed by a specific source s_i, the model eventually aims at measuring two probability distributions, namely, $P_{misc}(D_i)$ and $P_{mal}(D_i)$. The former being the probability that D_i has been generated by a misconfigured source while the latter is the probability that D_i has been generated by a malicious darknet source.

Let $D_1 = \{d_{i1}, d_{i2}, d_{i3}\}$ be those darknet addresses accessed by s_1. The model captures the probability $P(D_1)$ of the source generating $\{d_{i1}, d_{i2}, d_{i3}\}$ as the probability of s_1 accessing this specific combination of destinations knowing that it targeted three destinations multiplied by the probability of s_1 accessing any three destinations. The latter could be generalized and formalized as

$$P(D_i) = P(D_i = \{d_{i1}, d_{i2}, \cdots, d_{in}\} \mid |D_i| = n) \times P(|D_i| = n) \tag{3}$$

For both, a misconfigured and a malicious source, the first term of equation (3) could be modeled as

$$P_{misc}(D_i = \{d_{i1}, d_{i2}, \cdots\} \mid |D_i|) = \frac{1}{K} \prod_{\forall d_j \in D_i} P_{misc}(d_i) \tag{4}$$

[1] Those sources that are generating misconfiguration traffic towards the darknet IP space.

[2] In a given set of packets received on the darknet IP space, malicious darknet sources represent malicious Internet hosts while darknet addresses represent the destinations in those packets.

$$P_{mal}(D_i = \{d_{i1}, d_{i2}, \cdots\} \ \mid \ |D_i|) = \frac{1}{K} \prod_{\forall d_j \in D_i} P_{mal}(d_i) \tag{5}$$

where K, a normalization constant which is solely employed to allow the probabilities to sum to 1, could be defined as

$$K = \frac{|D|!}{n!(|D| - n)!} \times \frac{1}{|D|^n} \tag{6}$$

Please note that K is a typical normalization constant that is often employed in Bayesian probability[3] [11]. Further, n represents all the sources in the data set, while $|D|$, as previously mentioned, represents the darknet IP space.

The likelihood that a source will target a certain number of darknet destinations (i.e., the second term of equation (3)) depends on whether the source is malicious or misconfigured. Characteristically, misconfigured sources access one or few destinations while malicious sources access a larger pool of destinations. We have modeled such distributions as

$$P_{misc}(|D_i|) = \frac{1}{(e-1)|D_i|!} \tag{7}$$

$$P_{mal}(|D_i|) = \frac{1}{|D|} \tag{8}$$

where the term $(e - 1)$ in equation (7) guarantees that the distribution will sum to 1. It is noteworthy to mention that equation (7) ensures that the probability will significantly decrease as the number of targeted destinations increases. In contrast, equation (8) captures a malicious darknet source accessing a random number of darknet addresses.

By combining the above equations, we can model the probability of a source being a misconfigured or malicious, given a set of darknet destination addresses, as

$$P_{misc}(D_i) = \frac{1}{K(e-1)|D_i|!} \prod_{\forall d_j \in D_i} P_{misc}(d_i) \tag{9}$$

$$P_{mal}(D_i) = \frac{1}{K|D|} \prod_{\forall d_j \in D_i} P_{mal}(d_i) \tag{10}$$

It is imperative to note that equations (9) and (10) provide two distinct likelihood models to distinguish between misconfiguration and other malicious darknet traffic. This permits the simplified and systematic post-processing of the latter two types of darknet traffic. Moreover, as the model generalizes and formalizes the concepts of misconfiguration and other malicious darknet traffic, the proposed model does not make any assumptions related to the nature of the sources of those types of traffic. For example, the model is agnostic to whether the malicious traffic is generated by a worm or a probing tool, or whether the misconfiguration is caused by a malfunctioning Internet router or an invalid connection request.

Algorithm 1: Inferring misconfiguration flows using the probabilistic model

```
   Data: Darknet Flows, $DarkFlows$
   Result: Flag, $MiscFlag$, indicating that the $DarkFlow$
           is originating from a misconfigured source
 1 for $DarkFlows$ do
 2     $MiscFlag \leftarrow 0$
 3     i $\leftarrow$ $DarkFlows$.getUniqueSources()
 4     Amalgamate $DarkFlows_i$ originating from a
       specific source $s_i$
 5     Update $s_i(D_i)$
 6     Compute $P_{misc}(D_i)$, $P_{mal}(D_i)$
 7     if $P_{misc}(D_i) > P_{mal}(D_i)$ then
 8         $MiscFlag \leftarrow 1$
 9     end
10 end
```

To effectively employ the proposed darknet preprocessing model, we present Algorithm 1, which provides a simplistic yet effective mechanism to infer misconfigured sources by employing the model. It is worthy to note that step 6 of the algorithm (i.e., the computation of $P_{misc}(D_i)$ and $P_{mal}(D_i)$) is easily accomplished in practice by computing the negative log-likelihoods,

$$\begin{aligned} L_{misc}(D_i) &= -lnP_{misc}(D_i) \\ L_{mal}(D_i) &= -lnP_{mal}(D_i) \end{aligned} \tag{11}$$

Thus, Algorithm 1 deems a source and its corresponding flows as misconfiguration traffic if $L_{mal}(D_i) - L_{misc}(D_i) > 0$.

III. Empirical Evaluation

In this section, we aim at empirically evaluating the proposed model to validate its accuracy, compare and contrast its advantages, as well as investigating its performance. For this task, we employ a real darknet dataset from the Cooperative Association for Internet Data Analysis (CAIDA)'s educational kit [12].

The aim of this first experiment is to compare the outcome of the proposed model against the baseline; classifying misconfiguration traffic as any traffic that is not scanning or backscattered. The outcome of the latter on CAIDA's dataset is summarized in Table I.

Scanning Traffic	Backscattered Traffic	Misconfiguration
65.1%	8.2%	26.7%

TABLE I: Packets Distribution - Nature of Traffic

To execute the proposed model on this dataset, we aggregate the connections into sessions using an approach similar

[3] http://www.cs.ubc.ca/~murphyk/Bayes/bayesrule.html

to the first step algorithm by Kannan et al. [13]. We consider all those connections within T_{aggreg} of each other as part of the same session for a given pair of hosts. We used the same proposed threshold, T_{aggreg} = 100 seconds, and found that this seems to correctly group the majority of connections between any given pair of hosts.

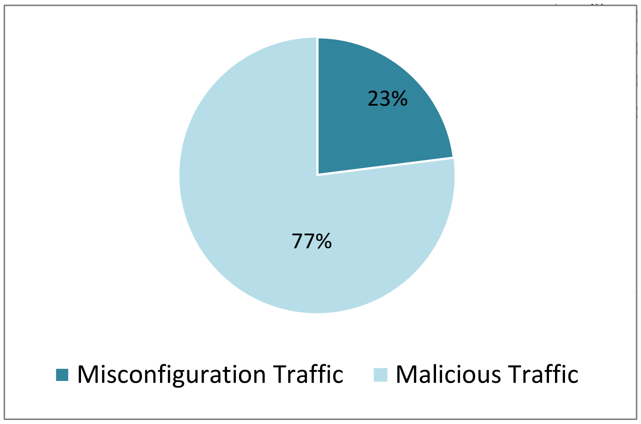

Fig. 1: The distribution of darknet sessions

Figure 1 depicts the outcome of the execution of the proposed model on the extracted sessions. By comparing Table I and Figure 1, we can notice that the proposed model fingerprinted a lower percentage of misconfiguration traffic than our previous study. A semi-automated verification (i.e., using scripts and manual investigation) validated that all the sessions that the model inferred as misconfiguration traffic are true misconfiguration packets, where almost 50% of them are malformed packets while the rest are packets that targeted the darknet IP space only once. We further investigated the 3% darknet sessions that the baseline experiment has inferred as misconfiguration traffic and noticed that they are indeed false positives related to UDP amplification attacks [14]. Thus, from this experiment, we can safely claim that the proposed model was accurate in distinguishing between darknet misconfiguration traffic and other malicious darknet traffic, compared to the baseline. We next compare the proposed model against a heuristic approach from the literature.

A. Comparison against literature

Although darknet preprocessing by fingerprinting misconfiguration traffic has not been excessively dealt with in the literature, arguably, the most prominent work in this topic would be the work by Li et al. [15]. In their work, the authors propose the following heuristic method to infer misconfiguration traffic. First, the researchers filter out scanning traffic caused by botnets or worms based on whether they have malicious payloads or whether they scan a predefined threshold defined by at least 10 IP addresses. Second, they assume and claim that most misconfiguration traffic will be caused by Peer-to-peer (P2P) sources. Thus, based on the latter, the author detect P2P connections based on whether their payloads match known P2P protocol signatures. Third, they aggregate inferred P2P connections into sessions within a 6-hour interval and deem the latter as misconfiguration traffic.

In contrast to our proposed model, we should note the following. First, the work by Li et al. enforces a hard (unexplained) threshold related to the number of scanned IP addresses, which may or may not hold true in all cases of misconfiguration traffic. Second, Li et al. make an assumption on the nature of the source of the misconfiguration traffic. Thus, even if their method is successful, it will only capture P2P misconfiguration. Third, in their work, a second threshold is employed related to when a session is deemed as misconfiguration. There is exists no scientific evidence of how this value was obtained. Fourth, the work by Li et al. is more tailored towards honeypot data, which might not be directly applicable to darknet data. Indeed, the authors' assumption related to P2P payloads can not hold true in the darknet context, as the darknet does not capture any payload due to its passive one-way nature. Nevertheless, we modified the heuristic method by Li et al. by removing the payload assumption and compared it with our proposed model. To achieve this, we utilized one day of CAIDA's darknet dataset.

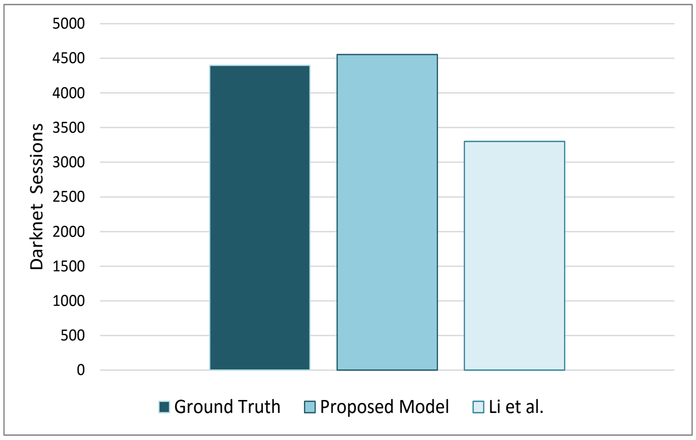

Fig. 2: Misconfiguration Fingerprinting: Proposed model Vs. Literature

Figure 2 illustrates the output of the execution of the proposed model as well as the work by Li et al. [15] on the extracted darknet sessions. Please note that the ground truth was obtained semi-automatically; initially using scripts and subsequently confirmed by stringent and prolonged manual inspection.

Based on the ground truth and experimenting with this specific dataset, we infer around 16% of the darknet sessions as misconfiguration traffic. Comparing this with our proposed model and the literature work, we can pinpoint that our proposed model always inferred the ground truth as well as other sessions, while the literature work missed some of the ground truth on some days and inferred more sessions on other days. Thus, we can state that the proposed model suffered from a 0% false negative rate, while suffering from some false positives. Section IV discusses why the latter occurred. In contrast, the work by Li et al. suffered from a 14% false negative rate, while suffering from a 4% false positive rate. It is evident that the proposed approach significantly outperforms the other literature work in terms of both, false negatives and false positives, in the context of fingerprinting misconfiguration traffic.

Processor	Instruction Set	# of Cores	# of Threads	Processor Frequency	Memory Size & Type
Intel Core i7-2600	64-bit	4	8	3.4 GHz	16 GB, DDR3 1066/1333

TABLE II: The hardware specification of the machine used to benchmark the prototype

B. Performance Evaluation

As darknet data is often exploited in various Internet and cyber security measurement studies, the performance of a preprocessing module should definitely be taken into consideration. A prototype of the proposed model is currently implemented in `Java` and heavily relies on the `jNetPcap`[4] library. As an initial performance benchmark, we were solely interested in inferring the execution time of the prototype; the time from which a darknet dataset is fed into the prototype, till the time the prototype flags the misconfiguration, filters-out such traffic and generates a new "clean" dataset. We executed the experiment on a single machine, where its hardware specification is summarized in Table II. For the experiment, we also employed CAIDA's dataset encompassing a one-hour period. The output disclosed that in order to achieve the intended task, the prototype approximately required 25 minutes to completely process the 1 hr sample. For our current tasks in hand that do not require large measurement studies and given the high accuracy that is offered by the proposed model, we believe that such result is acceptable. However, it might not be tolerable if one has to perform large measurements of darknet data as this preprocessing execution time would rapidly augment. For this reason, we believe that the performance could be significantly improved by (1) dropping the `Java` implementation and rather relying on a `C` implementation that exploits the notorious `libpcap`[5] library, (2) employing multithreading and parallel programming paradigms and (3) replacing the typical hard disk drives with solid state drives (SSDs). Although the latter point might not be cost-effective[6], it would undeniably improve the performance, as we noticed from our executed experiment that a significant performance bottleneck was caused by writing and reading to disk.

IV. Model Limitations

In this section, we discuss two limitations of the current proposed darknet preprocessing model and suggest few possible improvements.

We believe that a core point that is still not dealt with is related to how the model assesses whether the difference of the two probability estimates, namely, $L_{mal}(D_i)\ and\ L_{misc}(D_i)$ of equation (11), is large enough to safely choose one model over the other. To clarify, consider Table III that samples 10 misconfigured sources retrieved from the second experiment of Section III-A.

Source	$L_{mal} - L_{misc}$
1	99
2	132
3	113
4	14
5	146
6	106
7	39
8	97
9	2
10	133

TABLE III: A sample of 10 misconfigured sources

One can notice that although the above 10 sources were all fingerprinted as generating misconfiguration traffic, the difference between their probability estimates varied significantly. Indeed, this variation caused the majority of the false positives that were obtained in the comparative experiment of Section III-A. In this context, we see an opportunity to further improve the model from this perspective by designing and implementing a mechanism that enforces some confidence levels. We envision that the mechanism would aid the model in assessing which difference of the two probability estimates is significant, and thus which misconfiguration sessions are more plausible to be true positives.

Another point that we note that is still not very well-defined is related to the probability distribution of the randomness of a malicious source targeting a specific darknet destination, or namely, equation (2). Indeed, in this work, the assumption related to such probability distribution was borrowed from another literature work, which seems to be realistic and convincing. However, we believe that it would be beneficial and more precise to verify the soundness of that assumption. One way to achieve this is to initially infer malicious traffic targeting the darknet IP space, subsequently perform model fitting and consequently select the best fit model given a certain performance metric, such as, for instance, the Bayesian Information Criterion[7].

The aforementioned two points and their corresponding possible remedies are left for future work.

V. Related Work

In this section, we briefly review some related work in the context of (1) darknet and cyber threat intelligence, and (2) data preprocessing and anomaly detection.

A plethora of conducted research exploited darknet data to generate cyber threat intelligence. On the topic of denial of service attacks, Moore et al. [16] were among the first to leverage darknet data to infer and characterize such attacks. The authors introduced the notion of backscattered packets to permit the analysis and inference of such types of attacks. In a more recent era, Rossow [14] and Fachkha et al. [1] tackled the problem of inferring UDP amplification attacks by analyzing darknet data. Specifically, the latter work designed and implemented algorithms that leverage flow-based features as well as packet headers to fingerprint such attacks. Further, on the topic of probing activities, Dainotti et al. [17] correlated

[4] http://jnetpcap.com/?q=jnetpcap-1.4

[5] http://www.tcpdump.org/

[6] Although well within the financial capabilities of Computer Emergency Response Teams (CERTs) for instance, which highly exploit darknet data.

[7] http://stanfordphd.com/BIC.html

independent scanning activities as perceived by the darknet IP space, to investigate and report on a large-scale, malware-orchestrated, stealthy probing botnet, which scanned the entire IP space in few days.

On the other hand, a limited number of research works assessed the impact of data preprocessing on anomaly detection. A survey on the latter was executed by Davis et al. [18]. Some of the authors' findings include *(i)* the significant impact that data preprocessing possesses on the accuracy and capability of anomaly-based detection approaches and *(ii)* the need for more automated and accurate preprocessing models. On the topic of darknet preprocessing, Yoder [19] estimated that darknet misconfiguration traffic is expected to be one of the largest growing causes of darknet data, which necessitates the design and evaluation of suitable and tailored preprocessing approaches.

VI. Conclusion

Given the fact that numerous Internet-scale entities heavily rely on darknet data to generate cyber threat intelligence, this paper presented, evaluated and discussed a preprocessing model that aimed at cleansing such data. By successfully achieving this, darknet data could be more accurately exploited for the intended tasks, while operators save on valuable storage resources. In contrast with literature works, the proposed approach is more formal, does not depend on any hard thresholds, does not make any assumptions on the nature of the misconfiguration traffic, and was shown to be more accurate. We hope that this work would trigger future discussions related to data quality in the anomaly detection research community.

Acknowledgment

The author is grateful for the Natural Sciences and Engineering Research Council of Canada (NSERC) and Florida Atlantic University for supporting this work.

References

[1] Claude Fachkha, Elias Bou-Harb, and Mourad Debbabi. Inferring distributed reflection denial of service attacks from darknet. *Computer Communications*, 62:59–71, 2015.

[2] Symantec: Western Energy Companies Under Sabotage Threat. http://tinyurl.com/nsyksht.

[3] Elias Bou-Harb, Nour-Eddine Lakhdari, Hamad Binsalleeh, and Mourad Debbabi. Multidimensional investigation of source port 0 probing. *Digital Investigation*, 11:S114–S123, 2014.

[4] Zakir Durumeric et al. The matter of heartbleed. In *Proceedings of the 2014 Conference on Internet Measurement Conference*, pages 475–488. ACM, 2014.

[5] David Moore, Colleen Shannon, Geoffrey M Voelker, and Stefan Savage. *Network telescopes: Technical report*. Department of Computer Science and Engineering, University of California, San Diego, 2004.

[6] Elias Bou-Harb, Mourad Debbabi, and Chadi Assi. Behavioral analytics for inferring large-scale orchestrated probing events. In *Computer Communications Workshops (INFOCOM WKSHPS), 2014 IEEE Conference on*, pages 506–511. IEEE, 2014.

[7] Elias Bou-Harb, Claude Fachkha, Mourad Debbabi, and Chadi Assi. Inferring internet-scale infections by correlating malware and probing activities. In *Communications (ICC), 2014 IEEE International Conference on*, pages 640–646. IEEE, 2014.

[8] Claude Fachkha, Elias Bou-Harb, Amine Boukhtouta, Son Dinh, Farkhund Iqbal, and Mourad Debbabi. Investigating the dark cyberspace: Profiling, threat-based analysis and correlation. In *Risk and Security of Internet and Systems (CRiSIS), 2012 7th International Conference on*, pages 1–8. IEEE, 2012.

[9] Matthew Ford, Jonathan Stevens, and John Ronan. Initial results from an ipv6 darknet13. In *Internet Surveillance and Protection, 2006. ICISP'06. International Conference on*, pages 13–13. IEEE, 2006.

[10] Zakir Durumeric, Michael Bailey, and J Alex Halderman. An internet-wide view of internet-wide scanning. In *USENIX Security Symposium*, 2014.

[11] Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian data analysis*, volume 2. Taylor & Francis, 2014.

[12] T. Zseby, A. King, M. Fomenkov, and k. claffy. Analysis of Unidirectional IP Traffic to Darkspace with an Educational Data Kit, Feb 2014.

[13] Jayanthkumar Kannan, Jaeyeon Jung, Vern Paxson, and Can Emre Koksal. Semi-automated discovery of application session structure. In *Proceedings of the 6th ACM SIGCOMM IMC*, pages 119–132. ACM, 2006.

[14] Christian Rossow. Amplification hell: Revisiting network protocols for ddos abuse. In *Symposium on Network and Distributed System Security (NDSS)*, 2014.

[15] Zhichun Li, Anup Goyal, Yan Chen, and Aleksandar Kuzmanovic. Measurement and diagnosis of address misconfigured p2p traffic. In *INFOCOM, 2010 Proceedings IEEE*, pages 1–9. IEEE, 2010.

[16] David Moore, Colleen Shannon, Douglas J Brown, Geoffrey M Voelker, and Stefan Savage. Inferring internet denial-of-service activity. *ACM Transactions on Computer Systems (TOCS)*, 24(2):115–139, 2006.

[17] A. Dainotti, A. King, K. Claffy, F. Papale, and A. Pescap. Analysis of a "/0" Stealth Scan from a Botnet. *IEEE/ACM Transactions on Networking*, 23(2):341–354, Apr 2015.

[18] Jonathan J Davis and Andrew J Clark. Data preprocessing for anomaly based network intrusion detection: A review. *Computers & Security*, 30(6):353–375, 2011.

[19] James Yoder. Misconfigured addresses created by p2p clients.