Capturing DDoS Attack Dynamics behind the Scenes

An Wang[1], Aziz Mohaisen[2], Wentao Chang[1], and Songqing Chen[1]

[1] Department of Computer Science, George Mason University
[2] Verisign Labs

{awang10, wchang7, sqchen}@gmu.edu amohaisen@verisign.com

Abstract. Despite continuous defense efforts, DDoS attacks are still very prevalent on the Internet. In such arms races, attackers are becoming more agile and their strategies are more sophisticated to escape from detection. Effective defenses demand in-depth understanding of such strategies. In this paper, we set to investigate the DDoS landscape from the perspective of the attackers. We focus on the dynamics of the attacking force, aiming to explore the attack strategies, if any. Our study is based on 50,704 different Internet DDoS attacks. Our results indicate that attackers deliberately schedule their controlled bots in a dynamic fashion, and such dynamics can be well captured by statistical distributions.

1 Introduction

Internet Distributed Denial of Service (DDoS) attacks have been a challenge for many years. Today, many DDoS attacks are launched via different botnets. Recent years have witnessed the rapid increase of such DDoS attacks in terms of both the number and the volume. For example, according to a recent attack report for Q2 2014 by Prolexic [19], compared to Q2 a year ago, "the average bandwidth of DDoS attacks was up 72%, and the peak bandwidth increased 241%, while the attack duration was only half as long". As a matter of fact, today botnet DDoS attacks have become a mainstream commodity in the cybercrime ecosystem, where they could be rented or loaned.

To understand the fundamentals of DDoS attacks and defend against them, enormous efforts are continuously made from both academia and industry [3,8,16,21]. Driven by the underlying profits and the lifted bar by the ever-improving defense mechanisms, DDoS attacking strategies are becoming increasingly sophisticated in order to evade various detection systems. Therefore, a timely and in-depth understanding of latest DDoS attack strategies is a key to improve the existing defenses. However, most of our understanding of DDoS attacks is based on the indirect traffic measures and static characterization of DDoS attacks [5,21,6,8,15,14]. For example, our previous study shows that most DDoS attacks today are not widely distributed, but highly regionalized [21]. Most of such characterizations only touch the surface of attackers' strategies, which is far from sufficient for us to design more effective defenses against many attacks.

To win the arms race, we set out to investigate the attacking strategies behind the scenes. For this purpose, we aim to explore the attackers' strategies in deploying the attack force, focusing on the dynamic control of the attack forces in different DDoS attacks. Our study is based on a DDoS dataset collected for a period of 7 continuous months. Our dataset was provided by Team Cymru and is collected using passive and active techniques from multiple anchor points. The data was collected from August 28, 2012 to March 24, 2013, a total of 209 days (about seven months of valid and marked attack logs). In this seven-month period, a total of 50,704 different DDoS attacks were observed—more details are in [4]. Through our analysis, we find several interesting results. We find that a botnet family often uses a limited number of sophisticated patterns in dynamically scheduling bots to participate in various DDoS attacks. This dynamic scheduling is indicated by the shifting patterns of participating bots. Further, the bot shifting pattern in different botnet families can be well captured by statistical distributions, with parameters depending on the corresponding family.

The preliminary findings in this study not only refresh our understanding of today's Internet DDoS attacks, but also offer new insights for security analysts to identify botnet families and help predict how the attacking forces evolve over time during attacks.

The rest of the paper is organized as follows. In Section 2, we describe our dataset including the overall data statistics and the data fields we utilized for our analysis. In Section 3, we study the bot shifting pattern of each botnet family, the basic characteristics of each pattern We discuss related work in Section 4 and conclude with a concise summary of our analyses and their implications in Section 5.

2 Dataset Collection

The dataset is based on Team Cymru's constant monitoring of Internet critical infrastructure to aid intelligence gathering concerning the state of the art of attack posture, using both active and passive measurement techniques.

Even though there might be some potential skews in our dataset, our preliminary studies [21,5,4] suggest that our dataset still preserves the geographical features of botnet families. The unit constantly monitors Internet attacking traffic to aid the mitigation efforts of its customers, using both active and passive measurement techniques. For active measurements and attribution, malware families used in launching the various attacks are reverse engineered, and labeled to a known malware family using best practices. A honeypot is then created to emulate the operation of the reverse-engineered malware sample and to enumerate all bots across the globe participating in the particular botnet. A similar approach has been proposed by Kang et al. [9] As each botnet evolves over time, new generations are marked by their unique hashes.

Traces of traffic associated with various DDoS campaigns are then collected at various anchor points across the globe in cooperation with various ISPs. The traces are then analyzed to attribute and characterize attacks on various targets. The collection of traffic is guided by two general principles: 1) that the source of the traffic is an infected host participating in a DDoS campaign, and 2) the destination of the traffic is a targeted client, as concluded from eavesdropping on C&C of the campaign using a live sample, or where the end-host is a customer of the said DDoS mitigation company.

The analysis is high level in nature to cope with the high volume of ingest traffic at peak attack times—as shown later, on average there were 243 simultaneous verified DDoS attacks launched by the different botnets studied in this work. High level statistics associated with the various botnets and DDoS attacks are recorded every one hour. The

Table 1. Summary of the workload information

Summary of Attackers		Summary of Victims	
description	count	description	count
# of bot_ips	310950	# of target_ip	9026
# of cities	2897	# of cities	616
# of countries	186	# of countries	84
# of organizations	3498	# of organizations	1074
# of asn	3973	# of asn	1260

workload we obtained ranges from August 28, 2012 to March 24, 2013, a total of 209 days (about seven months of valid and marked attack logs). In the log, a DDoS attack is labeled with a unique DDoS identifier, corresponding to an attack by given DDoS malware family on a given target. We cannot reveal the capability of the capturing facility because attackers would learn such information, which is also critical to the business of the data source.

Table 1 sums up some statistics of our dataset, including information from both the attacker and the target sides. Over a period of 28 weeks, 50,704 different DDoS attacks were observed. These attacks are launched by 674 different botnets. These attacks targeted victims located in 84 different countries, 616 cities, involving 1074 organizations, residing in 1260 different autonomous systems. In our analysis, we focus on the botnets involved in DDoS attacks. However, [4] contains more detailed information about botnet family activities and patterns.

The attackers' IP information enables us to study the geolocation distribution of each botnet family. Contrary to the traditional understanding of DDoS attacks, the attacks are not very distributed but rather highly regionalized [21]. Each family has its own geolocation preferences. Among all the families, *Dirjumper* covers the largest number of countries: 164. A comparable coverage is Optima's: 153. Even though these families have very broad country coverages, the average number of bots participating in each attack pertaining to those botnets is small.

3 Attack Dynamics

To seek an in-depth understanding of attackers' strategies, we set to explore attacks from the adversary's perspective. By doing that, we are motivated to find out how their controlled bots are scheduled to participate in attacks. To this end, we use the IP information of the bots captured in our dataset. Our analysis starts off from two different perspectives, namely the bot shift pattern dynamics and the multi-owned bot attacking interval, both of which are related to DDoS attack strategies.

For any DDoS attack, it evolves over time in terms of the attacking force. In our dataset, each entry represents a snapshot of the DDoS attacks captured at that time point. As a result, each DDoS can be represented by a chronological sequence of snapshots. Dynamic characterizations can be captured by analyzing each data record.

For each entry in our dataset, we have the IP information of all the bots participating in that DDoS attack at that moment, of which the country code (cc) could also be obtained from such information (the snapshots are updated hourly). Thus, the dataset contains all the IP information of bots involved. After further organizing the bots based on their country code, each entry in the dataset can be denoted by $< cc_1 : n_1, cc_2 : n_2, \ldots, cc_m : n_m >$ where each $cc_i, i \in [1 \ldots m]$ represents the country code where the bots locate; while for each $n_i, i \in [1 \ldots m]$ denotes the number of bots located in $cc_i, i \in [1 \ldots m]$. Since each of such vector represents a snapshot, so if we line up all the vectors belonging to the same DDoS attack together, we can observe the deployment differences by comparing the number of bots in each country and the number of countries involved. For example, if we have two such records, denoted by vec_1 and vec_2, the change can be denoted by $vec_2 - vec_1 =< cc_1 : \Delta_1, cc_2 : \Delta_2, \ldots, cc_j : \Delta_j >= vec_{\Delta_v}$. Notice that the lengths of vec_1 and vec_2 may not be equal and the length of vec_{Δ_v} will be the same as the longer one. Thus, the difference vector reflects the changes of the bots numbers at the country level for the the given attack, which is defined as *shift* in our analysis.

To further quantify such changes, we use the notion of *shift expectation* to represent each attacking force shift. In another word, each vector described above will be denoted by a single value called *shift expectation*, whose calculation will be elaborated later. In this way, each DDoS attack can be denoted by a vector whose elements are shift expectation, i.e. $< E_{shift_1}, E_{shift_2}, \ldots, E_{shift_m} >$, since each DDoS attack can be denoted by a time series of snapshots. And the length of this vector is determined by both the number of magnitude changes happened in each attack as well as the number of snapshots taken for each attack.

The *shift expectation* is calculated as $\sum_{i=1}^{m} p_i \times \Delta_i$, where Δ_i is obtained from vec_{Δ_v} and p_i denotes the probability estimator of the shift. And p_i is computed as follows. From our dataset, we obtain the geolocation information of all bots involved in the DDoS attacks. For each family, we generate a table that has two columns; the first column contains all the country codes that are covered by this family while the second one has the corresponding number of bots that locate in that country. So each entry in this table is denoted by (cc_i, n_i), for $i \in [1 \ldots l]$. On the other hand, $p_i, i \in [1 \ldots l]$, is calculated as $\frac{n_i}{\sum_{j=1}^{l} n_j}$. With both p_i and Δ_i, the expectation of each shift E_{shift} can be calculated. After converting each DDoS attack into a time series vector, we have all the vectors with various lengths for all the DDoS attacks in our dataset. Our following analyses will be built on top of these vectors.

3.1 Bots shift pattern analysis

First, we use the K-means clustering on all attack vectors of each family. Since the lengths of vectors may vary, we cannot calculate the Euclidean distance between vectors directly. The Dynamic Time Warping (DTW) has been widely used for shape matching and time series classification. Accordingly, we use DTW to calculate the distance and similarity between attack vectors. To reduce the distortion under the influence of attack magnitude, we normalize the vector before we calculate the DTW distance on them.

The results are shown in Figure 1 - Figure 4. In these figures, x-axis represents Shift Times, which is determined by the number of snapshots belonging to each DDoS

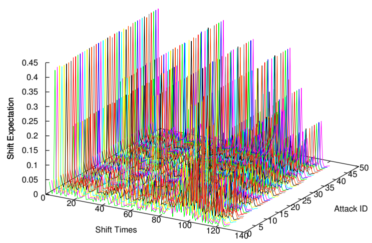

Fig. 1. *Dirtjumper*: attacks with same bot shift pattern

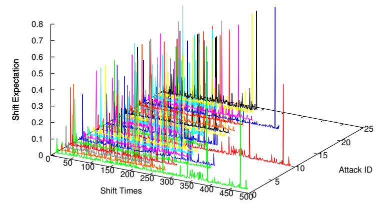

Fig. 2. *Dirtjumper*: attacks with similar bot shift pattern

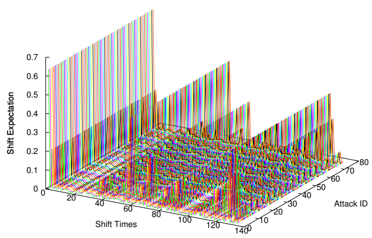

Fig. 3. *Pandora*: attacks with same bot shift pattern

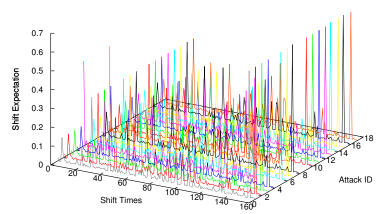

Fig. 4. *Pandora*: attacks with similar bot shift pattern

attack; y-axis represents Attack ID, which is used to differentiate different DDoS attacks; z-axis represents the calculated Shift Expectation values. Figure 1 and Figure 2 illustrate two of the four largest clusters discovered by the K-means algorithm of the *Dirtjumper* family, where $K = 10$. We cluster these vectors into 5, 10, and 20 clusters. It is shown that clustering them into 10 clusters yields better results. So, we present the two largest clusters of the 10 clusters for brevity. The two clusters contain 54 and 24 attacks, respectively. In each figure, the x-axis represents the length of the attack vector, i.e., the shifts happened in a single attack; the y-axis represents the unique DDoS ID; and the z-axis represents the shift expectation of each shift. Note that since *Dirtjumper* has too many DDoS attacks with different lengths of shifts, we first group the attacks by size. In this study, we focus on the analysis of attack vectors that have more than 100 shifts, which include 242 attacks launched by *Dirtjumper*. Our following analyses are based on this subset of our dataset as well.

In these figures, the expectations should be discrete values. To more clearly show the changes, we use lines to connect these dots. Figure 1 shows that in these attacks, bots are being scheduled with the exact same pattern in different attacks, while Figure 2 indicates a similar pattern—although not identically—in different attacks. With further inspection, we find that in Figure 1 there are 46 simultaneous DDoS attacks ongoing to-

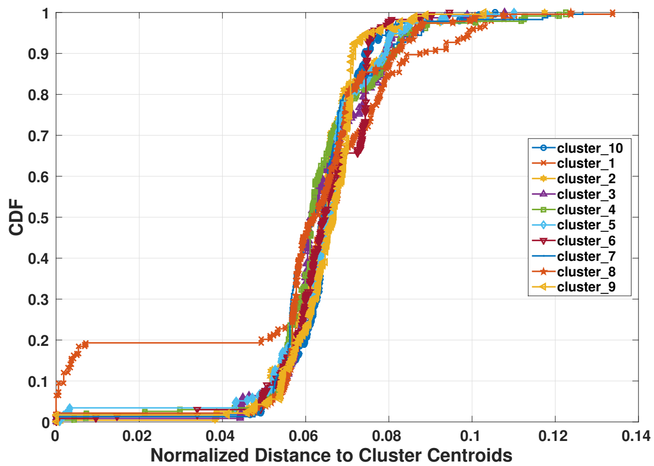

Fig. 5. Vector Distances CDF

wards the same target located in Finland, which is a company providing communication services from basic broadband to high-speed fiber connections.

These results suggest that the attacking forces are not randomly scheduled by the attackers in *Dirtjumper*. Also, simultaneous attacks cannot be arranged by a completely random deployment strategy. There has to be certain strategies behind DDoS attacks launched by each family. To see if such a pattern is specific to *Dirtjumper* or generalizable to others, we examine other families. Figure 3 and Figure 4 illustrate two clusters of another active botnet family *Pandora*. We use the same K-means clustering with 10 clusters for attacks and more than 100 shifts as before. We have similar observations on *Pandora* as on *Dirtjumper*. While other families show similar results, we omit them due to page limit. These findings also suggest that there might be a way to detect DDoS attacks based on these shift behaviors. But this only will be possible if we can precisely model these pattern, which is the aim of our next step.

3.2 Mathematical representation of shift patterns

To further explore the pattern behind these vectors, we first find the centroid vector of each cluster and then calculate the distance between each attack vector in that cluster and the centroid. The centroid vector cannot be calculated simply by averaging all the vectors involved since they are of different length. We define centroid vector as the vector that has the smallest total distance to all other vectors. And all the distances involved are measured by DTW distance. We use *Dirtjumper* as an example since it is the most active family.

The CDF of the distance distribution for *Dirtjumper* is shown in Figure 5. In this figure, each curve represents a cluster. If we observe these curves, the distances seem to follow the normal distribution very well except for cluster-1. To verify the distribution, we further fit the data into multiple distributions, including *tlocationscale distribution*,

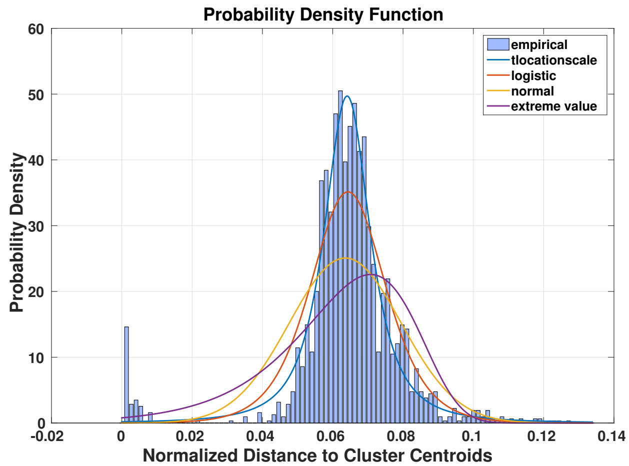

Fig. 6. Distribution Fit

normal distribution, *logistic distribution* and *extreme value distribution*. The fitting results are shown in Figure 6.

Except for the *extreme value distribution*, all other distribution functions are symmetric distributions. Figure 6 shows that the data fit the *tlocationscale distribution* best. *tlocationscale distribution* is the generalized *Student's t-distribution* into location-scale family. Location-scale family is a family of univariate probability distributions parameterized by a location parameter and a non-negative scale parameter. The *tlocationscale distribution* is useful for modeling data distribution with heavier tails than the *normal distribution*, meaning that it is more prone to producing values that fall far from its mean. This makes it useful for understanding the statistical behavior of certain types of ratios of random quantities. In this case, the distribution describes the distances between multiple shift patterns of botnets. It means that if we use the centroids of different clusters as a baseline, we can learn and predict how the bots are going to shift based on this distribution.

To this end, it is likely that attackers are utilizing this feature to arrange and control bots during attacks, especially with a large number of bots. From a defense perspective, such information can be very useful. On one hand, with this information—even though there might be more than one shift pattern per family—we can predict how attacks shift based on the distribution. On the other hand, we can simulate DDoS attacks behaviors, not only based on traffic volume but also by incorporating dynamics behind them.

Similar to Figure 5, we also plotted a CDF for *Pandora*'s clusters, which confirmed the similar behavioral patterns. However, compared to *Dirtjumper*, *Pandora* exhibits a slight deviation in the distribution, perhaps due to the smaller number of attacks in *Pandora* compared to *Dirtjumper*. Results obtained by analyzing other families reveal similar findings.

Besides the pattern clustering graphs, Table 2 summarizes some statistical information about *Pandora* clusters. In this table, *Size* shows the size of each cluster; *Max_Diss*

Table 2. Statistic Information of *Pandora* Cluster

Size	Max_Diss	Avg_Diss	Diameter	Separation	Avg_Exp	Max_Exp	Std
97	3.69	0.18	3.73	3.46	0.03	0.74	0.076
74	0.06	0.02	0.09	3.62	0.03	0.63	0.08
20	4.52	1.46	4.99	2.39	0.025	0.88	0.073
17	3.48	0.41	3.48	2.39	0.028	0.68	0.075
7	4.18	1.06	4.30	3.57	0.03	0.88	0.08

represents the maximum distance between any two vectors in the same cluster; *Diameter* represents the largest dissimilarity between any two pairs of the observations within the same cluster; *Separation* represents the minimal dissimilarity between an observation of the cluster and an observation of another cluster; *Avg_Exp* shows the average shift expectation of each cluster and *Std* is the standard deviation of expectations of each cluster. Statistically speaking, the smaller the *Diameter*, the better the cluster. From this table, we can see that the second cluster is the best, which also conforms with Figure 3. Another observation from this table is that for most clusters, the *Diameter* is larger than *Separation*, meaning that these clusters are not totally isolated. The total isolation means that the patterns might be attack-specific. However, results show the opposite: each cluster still shares some similarities with other clusters. This further indicates that there might be certain dynamic mechanisms behind each family.

4 Related Work

DDoS attacks have been intensively investigated and numerous measurement works have been done to help achieve better understanding of them. In 2006, Mao et al. [16] presented their measurement work of DDoS attacks relying on both direct measurement of flow-level information and more traditional indirect measurements using backscatter analysis. Moore et al. [17] conducted a backscatter analysis for quantitatively estimating DoS activity in the Internet based on a three-week dataset. Due to the growth of network address translation and firewall techniques, much of the Internet was precluded from the study by the traditional network measurement techniques. Thus, in the early days, the work [3] proposed an opportunistic measurement approach that leverages sources of spurious traffic, such as worms and DDoS backscatter, to unveil unseen portion of Internet. In 2010, a more recent study [23] revisited the same topic and characterized the current state of background radiation specifically highlighting those which exhibit significant differences. Our work serves as a revisit to those studies with new insights. Bailey et al. [1] designed and implemented the Internet Motion Sensors (IMS), a globally scoped Internet monitoring system to detect Internet threats, which includes a distributed blackhole network with a lightweight responder and a novel payload signature and caching mechanism. Xu et al. [24] presented a general methodology to build behavior profiles of Internet backbone traffic in terms of communication patterns of end-hosts and services.

In our work, we focus on DDoS dynamics. We use several techniques including K-means clustering and Dynamic Time Warping (DTW). DTW was first introduced in the data mining community in the context of mining time series proposed by Berndt et al. [7]. Several techniques have been introduced to speed up DTW and to reduce the space overhead [12,11]. The K-means clustering methods we use were first proposed by Lloyd et al. [18]. And it remains a very popular method of clustering after many years perhaps due to the simplicity of the algorithm and its effectiveness in practice. These techniques successfully helped us discover the principles of the dynamics behind the scenes. Several other works focused on DDoS dynamics analysis as well, including Arne et al. [22], Armin et al. [2], Mohammad et al. [10], Kührer et al. [13]. In Stringhini et al. [20], a similar approach was proposed to model the country distribution of bots and cluster together botnets.

5 Conclusion

DDoS attacks remain one of the most challenging threats on the Internet, despite numerous efforts to characterize, model, and defend against them. This indicates that increasingly sophisticated strategies are being employed by the DDoS attackers. Successful defenses demand in-depth understanding of their strategies. In this work, we have conducted a preliminary analysis on a large scale DDoS dataset, aiming to understand the dynamics of the DDoS attack strategies behind the scenes. With the help of Dynamic Time Warping and clustering, we have found that attackers are deliberately and dynamically deploying their attack forces in individual or collaborative attacks, indicating the strong bond and organization of different botnet families in various attacks. Furthermore, such dynamics can be well captured by statistical distributions. These results add to the existing literature of DDoS characterization and understanding. More importantly, they lay a promising foundation for us to predict the dynamics during a DDoS attack in the future, which could be utilized to enhance existing defenses.

6 Acknowledgment

This work is partially supported by National Science Foundation (NSF) under grant CNS-1117300. The views and opinions expressed in this paper are the views of the authors, and do not necessarily represent the policy or position of NSF or VeriSign, Inc.

References

1. Bailey, M., Cooke, E., Jahanian, F., Nazario, J., Watson, D., et al.: The internet motion sensor-a distributed blackhole monitoring system. In: NDSS (2005)
2. Büscher, A., Holz, T.: Tracking ddos attacks: Insights into the business of disrupting the web. In: Proceedings of the 5th USENIX Workshop on Large-Scale Exploits and Emergent Threats (LEET), San Jose, CA, USA (2012)
3. Casado, M., Garfinkel, T., Cui, W., Paxson, V., Savage, S.: Opportunistic measurement: Extracting insight from spurious traffic. In: Proc. 4th ACM Workshop on Hot Topics in Networks (Hotnets-IV) (2005)

4. Chang, W., Mohaisen, A., Wang, A., Chen, S.: Measuring botnets in the wild: Some new trends. In: ACM ASIACCS (2015)
5. Chang, W., Wang, A., Mohaisen, A., Chen, S.: Characterizing botnets-as-a-service. In: Proceedings of the ACM SIGCOMM (poster) (2014)
6. Feinstein, L., Schnackenberg, D., Balupari, R., Kindred, D.: Statistical approaches to DDoS attack detection and response. In: DARPA Information Survivability Conference and Exposition (2003)
7. J.Berndt, D., Clifford, J.: Using dynamic time warping to find patterns in time series. In: KDD workshop (1994)
8. Jin, S., Yeung, D.: A covariance analysis model for ddos attack detection. IEEE International Conference on Communications (2004)
9. Kang, B.B., Chan-Tin, E., Lee, C.P., Tyra, J., Kang, H.J., Nunnery, C., Wadler, Z., Sinclair, G., Hopper, N., Dagon, D., et al.: Towards complete node enumeration in a peer-to-peer botnet. In: Proceedings of the 4th International Symposium on Information, Computer, and Communications Security. pp. 23–34. ACM (2009)
10. Karami, M., McCoy, D.: Understanding the emerging threat of ddos-as-a-service. In: LEET (2013)
11. Keogh, E., Ratanamahatana, C.A.: Exact indexing of dynamic time warping. In: Knowledge and Information Systems (2005)
12. Kim, S.W., Park, S., W.Chu, W.: An index-based approach for similarity search supporting time warping in large sequence databases. In: Proceedings of International Conference on Data Engineering (2001)
13. Kührer, M., Hupperich, T., Rossow, C., Holz, T.: Exit from hell? reducing the impact of amplification ddos attacks. In: USENIX Security Symposium (2014)
14. Lee, K., Kim, J., Kwon, K.H., Han, Y., Kim, S.: DDoS attack detection method using cluster analysis. Expert Systems with Applications 34, 1659–1665 (2008)
15. Li, M.: Change trend of averaged hurst parameter of traffic under ddos flood attacks. Computers and Security (2006)
16. Mao, Z.M., Sekar, V., Spatscheck, O., van der Merwe, J., Vasudevan, R.: Analyzing Large DDoS Attacks using Multiple Data Sources. In Proceedings of ACM SIGCOMM Workshop on Large-Scale Attack Defense (2006)
17. Moore, D., Shannon, C., Brown, D.J., Voelker, G.M., Savage, S.: Inferring internet denial-of-service activity. ACM Transactions on Computer Systems (TOCS) 24(2), 115–139 (2006)
18. P.Lloyd, S.: Least squares quantization in pcm. In: IEEE Transactions on Information Theory (1982)
19. Prolexic: Q2 2014: Malicious actors switch tactics to build, deploy and conceal powerful botnets. http://www.prolexic.com/knowledge-center-ddos-attack-report-2014-q2.html (June 2014)
20. Stringhini, G., Holz, T., Stone-Gross, B., Kruegel, C., Vigna, G.: BOTMAGNIFIER: Locating Spambots on the Internet. In: USENIX Security Symposium (2011)
21. Wang, A., Chang, W., Mohaisen, A., Chen, S.: How distributed are today's ddos attacks? In: Proceedings of the ACM CCS (poster) (2014)
22. Welzel, A., Rossow, C., Bos, H.: On measuring the impact of ddos botnets. In: Proceedings of the Seventh European Workshop on System Security. p. 3. ACM (2014)
23. Wustrow, E., Karir, M., Bailey, M., Jahanian, F., Huston, G.: Internet background radiation revisited. In: Proceedings of the 10th ACM SIGCOMM conference on Internet measurement. pp. 62–74. ACM (2010)
24. Xu, K., Zhang, Z.L., Bhattacharyya, S.: Profiling internet backbone traffic: behavior models and applications. In: ACM SIGCOMM Computer Communication Review. vol. 35, pp. 169–180. ACM (2005)