Botake: Detecting BOTs based on Weighting Algorithm and Entropy Behavior in DNS Traffic

Azar Hosseini[1†] and Arezoo Hosseini[2††],

[†]School of Electrical Engineering, Iran University of Science and Technology, Tehran, Iran
[2††]Pardis Nasibe-Shahid Sherafat, Farhangian University,Tehranm, Iran

Summary
Bots abuse legitimate protocols privileges for unauthorized purposes. Due to the large-scale of these subversive intentions, paying attention to the expansion of botnet and control channels has a major proportion of recent researches. Attackers use this type of network to carry out widespread attacks and access to confidential information as well as access to sensitive organizational systems within countries. A vastly expanded potential of Bots is ability of using a variety of protocols, different patterns of behavior in communication and variety of social networks for the exchange of information. In this paper, we focus on exploring the neighboring domains on the DNS traffic and identifying patterns by feature extraction, behavioral entropy calculation, and a system weighting algorithm to detect the active network in the DNS as an essential substrate of the Internet.

Key words:
Botnet, DNS, Weighting Algorithm, Entropy

1. Introduction

Nowadays, investigation of threats and secure data transfer emphasize on the importance of the potential threats from botnets. Because according to the reports of the well-known security agencies such as Symantec, a speedup in Botnets growth has been 84% in recent years. A particular approach to addressing the threats of this type of network is consideration to smart devices and smart TVs. In the new generation, botnets have a large number of control centers and communication channels. For example, the "dridex" bot, which infected thousands of banking systems in 2015, continued its a key role in malicious activities in 2016. As far as reports it was considered one of the most important banking issues. The destruction rate is estimated at around $ 40 million [1].

From another perspective, bots can be considered in terms of its application. One of a usage of botnets is the exploitation of the victim systems in the denial of service attacks. Often, these kinds of attacks occur in less than 30 minutes with sending multiple gigabytes of incomplete or forged communication to disrupt server servicing. These attacks are substantial while encountering banking exchanges, electronic payment systems, news sites and political sites in crucial times such as the election of the

countries. In summary, the application of botnets can be divided into nine sections: 1. Phishing and hijacking the financial information; 2. Denial of service attacks; 3. Identity theft; 4. Spam; 5. Advertising tools; 6. Spoofing the click; 7. Spyware, 8. Spreading new malware and 9. Spamming network traffic like man in the middle attacks.

Figure 1 shows the process of botnet investigation during its life time. The structure of this diagram is as follows: time division between new and old generations, the type of structure of the command and control centers, the different kinds of commands between C&Cs, the various types of bots, protocols and the data used in the botnets, communication techniques and ultimately the collection and compare the features of the botnets.

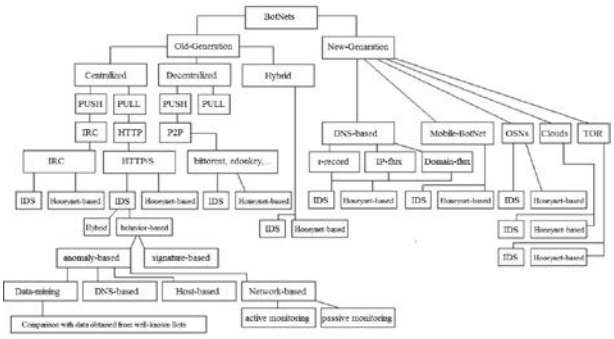

Fig. 1 Botnets Review Process

The time division between two generations of the old generation (2004-2010) and the new generation (2010-2016) helps to make the subject matter of the bots and their targets of attacks more evident. The basis of this kind of differentiation implies to their communication and their features. Network traffic, protocols and vulnerabilities of protocols and systems are all clear reasons for choosing time separation. The next point that has been taken into consideration in the old generation of bots is the structure of C&Cs. In the old generation, due to communication

Manuscript received November 5, 2017
Manuscript revised November 20, 2017

constraints and widespread attacks, much of the research has focused on the structure of C&Cs. In the review of the old botnets, the mechanism of operation of C&Cs is divided into two groups: PUSH and PULL. The first is the "PUSH" style, where commands are pushed or sent to bots. IRC-based C&C is an example of the push style. The second is the "PULL" style, where commands are pulled or downloaded by bots. HTTP-based C&C is an example of the pull style [2]. These operations refer to direct contacts with the Bot-Master, which are only carried out by the command and control centers.

The next step is to investigate the bots from the point of view of protocol and transit traffic. The purpose of the protocol and transitional traffic is the same as the use of bots in various attacks. Finally, at the last level of the pyramid, we will identify and study different behaviors in different ways. This level of review is more similar to the methods used to detect bots in intrusion detection systems. In these types of systems, the behaviors between the C&Cs, the bots, and the bot master are examined. At this stage, the features extracted from known bots and their comparison with current contaminated traffic can be helpful. Todays, analytical systems use data processing methods, DNS traffic, host's features that deal with the protocols used by the C&Cs and the active and passive network behavior to identify unknown bots. The purpose of active and passive behavior of the network, on the one hand, is the devices that are involved in generating, conducting and amplifying the signal, such as modems, but on the other hand is online and offline monitoring. Numerous studies have demonstrated that the analysis of suspicious traffic not only needs prompt scrutiny of the current traffic features, but also requires pre-stored information about some protocols such as DNS and other IP-registered information acting as Whois and malware detection sites like Virustotal.

The main approach of this paper is to provide a system to detect botnets with relying on behavioral analysis of domains. We have tried to find the relationship between domains and duplicate IPs, or vice versa by using weighting algorithms. The advantage of this algorithm is the short running time and the utility of the calculation. The impact of this algorithm has been investigated on fast-flux communications and the results were more effective than data processing on extractive features of network, host and C&C. The various sections of this article are as follows: The second part relates to DNS network attacks, the third part presents the bot detection system along with the diagram of the various parts of it, the fourth part examines the neighboring domains behavior and the proposed algorithm, the fifth part exposes the algorithm results in consonance with big-data and the sixth section concludes this paper.

2. DNS Network Attacks

Different kinds of attacks in botnets can be divided into twelve separate categories: 1. phishing and the exploiting confidential information; 2. DoS / DDoS; 3. Fake click attacks to disrupt the function of advertising; Spamming and consuming sacrificial resources, 5- Identity theft, 6 - Exploiting information leakage, 7 – ransomwares to create fears and impose malwares on users, 8 - Eavesdropping and stealing communication and robbing other bots through this way , 9. Installing the keyloggers to swipe usernames, passwords and IDs, 10- Distributing malwares to victim systems, 11- attacks against Internet Relay Chat (IRC) networks, 12-manipulate online programs such as online games and inject malicious codes. Attacks often have interdependent characteristics, because combining features can provide more complex and unidentified attacks. As reported by Symantec we can examine the role of various protocols, such as ICMP, DNS and etc. These reports state that attacks based on DNS protocol make up 29.4% of all attacks associated with this type of traffic [1]. Figure 2 shows the twelve branches of DNS-based attacks.

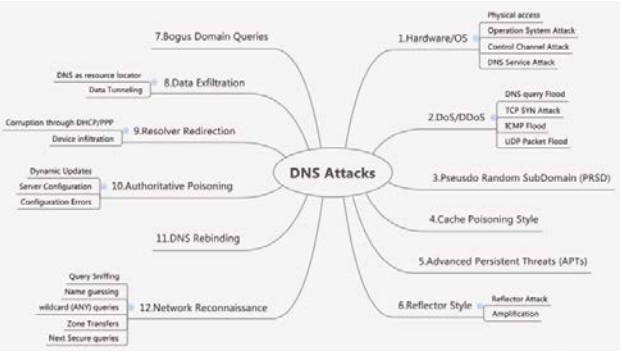

Fig. 2 DNS Attacks

The first type is the physical access to the victim's system with two aspects of hardware and software. Vulnerabilities within the server operating system and applications may be exploited by attackers in order to severely hamper or crash the server [3]. The second type is one of the most important types of attacks to use extensively the DNS protocol for denial-of-service attacks. In this type, the attacker substantially increases the volume of queries in an effort to inundate the server and reduce performance or consume resources. The third type is related to generate similar subdomains to the victim's server domain. The effects of this attack are mostly on recursive servers, because this kind of server is obliged to manage the response of sent packets, among which the valid packages of trusted users may be removed. In fourth-instance attack, the attacker penetrates into the memory cache of the DNS

server and the recursive server then redirects client requests to himself.

The target of fifth attack is the local information of requests that can be exploited by an attacker to infect with malware. The sixth type alludes to the attacker who attempts to endanger victim access to the network by sending a large number of requests to several DNS servers with a variety of DNS types and changing the source IP with the victim's IP. The seventh attack involves fake domains and creating a huge flood of NX Domain responses. The eighth type of DNS attack allows attackers to use the compression method to send their data through the DNS protocol, because the firewalls pass the DNS packets without checking. Attackers in the 9th, access to the client and change the resolver settings to their favor. Tenth type of DNS attack focuses on infecting or hijacking local files. Another way to redirect users to the attacker's path is feasible with the 11th attack. In this attack, the attacker provides a cheap server with two addresses. The first is the attacker's address and the second is the address of his desired destination. After adjustments mentioned, the first response to the user request is the attacker's site, and the second-time response to his request, is the attacker's intention. The last type of DNS attacks is to penetrate network devices and receive packets of requests and replies. In this type of attack, the attackers are guessing user's IDs and passwords by collecting information.

3. Bot Detection System

Figure 3 shows the modular identification of the unknown C&C system. The inputs of this system are DNS and HTTP/S logs to find suspicious behavior of unknown bots. Other entries include the blacklist of infected domains, the whitelist published by Alexa [4], domain information[1], intrusion detection tools, crawlers, malware databases such as Virustotal and certificates issued for reputable domains and sites. The purpose of the list of infected domains is the same domains as the phishing detection system result, DGA, Honeypot and DNS open-resolvers identification.

Online tools in the Analyzer Block are as follows:

Registrants' information in Whois_ where records domain information such as date, buyer's name, expiration date and server names, malware information in Virustotal_ it maintains domains/IPs and their history of malicious activities as well as the hash of their infected files, Sslyze_ A tool for checking the validity of certificates issued for domains and ultimately online crawlers to find related links to domains/subdomains.

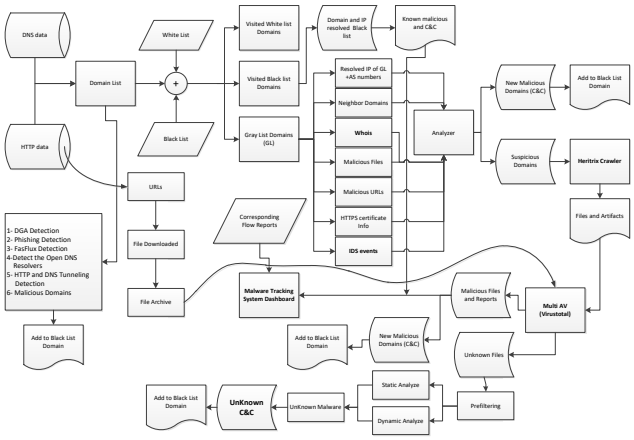

Fig. 3 Botake can detect botnets

Botake system considers identifying the unknown bots by seeking suspicious manner in DNS and HTTP/S logs. After matching these two types of log and passing through the blacklist filter, the number of domains and corresponding IPs being limited. A gray list composed of clean and infected domains as an outcome of filtering will be brought to the analysis phase. In this phase resolved IPs and AS numbers as well as Whois information, SSL certificates, related events of the IPs corresponding to these domains which are derived from the intrusion detection system are examined. Weighting algorithm is a decisive part of analyzer to explore the behavior of neighboring domains and IPs in communications. Two reports are attained from analyzer: 1.IPs and infected domains declared to be severely contaminated; 2.Suspicious domains and IPs that their recent files should be downloaded and analyzed or their links to the other infected links should be found by strong crawlers. The address of recent files must be extracted from HTTP/S traffic.

Static and dynamic analysis blocks in Figure 3 are used to execute malware detection commands online and offline in real and experimental environments.

4. Behavior of neighboring domains and proposed algorithm

The central and sensitive part of Botake is the analyzer block that needs solid information. One of this section's requires to make decisions is information about neighboring domains. Some researchers [5] believe that bots and their neighbors in a network use a different

[1] Whois

template to introduce themselves by domain name or IP value. Consequently, they emphasize on extracting the properties of domain names and IPs. Others [6] try to combine features extracted from domains/IPs with features related to the flow of domains and suspicious IPs for clustering. Extending features of this method takes a long time to run. Therefore, we propose the limited elicitation of features then using weighing algorithms as a unit of computation of behavior and Snort data as a database Passive data.

4.1 Eliciting Features

As discussed in the previous sections, the scope of traits in URLs, IPs and well-known behavior, as well as the comparison of suspicious behavior with white traffic is very wide. Table 1 lists the most important features examined in recent papers [7].

Table 1: Collection of the most important features of the Bot

Multi AVs (e.g. virustotal)	***Sslyze***
Creation Date	Certificate Creation Date
Expiration Date	Certificate Expiration Date
Resolve ip addresses into domain names	***DNS Features***
The number of sub-domains of each domain	*Based on FQDN*
Assessment of contamination factor	Number of sections in DNS requests
known infected urls in relation to each domain	Average length of sections in DNS requests
Netflow Features	The Ratio of numerical characters
Average Bytes per sent Packet	Length of SLD
Average Bytes per Received Packet	*Based on Request*
Average Sent Bytes per Second	Query Type
Average Received Bytes per Second	Query Ratio
Flow rate per hour	*Based on Response*
Package rate per Flow	Total Number of Response
Average Packet per received Flow	Total Number of NXDomains
Average Packet per sent Flow	Total Number of No-Error type
Different ports contacted	Average Number of IPs Resolved
Different IPs contacted	Average TTL
The percentage of SYN, SYN-ACK, ACK and ACK-PUSH packets in the TCP protocol	Standard deviation of TTL
The percentage of SYN, SYN-ACK, ACK and ACK-PUSH packets in the UDP protocol	*Based on*
	geographic location
--	Number of ASs resolved for IPs

Reviewing all the features quoted increases the complexity of the calculations. When Botake is encountered the bulky data for example, 1,500,000 requested domains over a 24-hour it will exploit weighting algorithms as a catalyst for detecting bot as well as restricted features of DNS traffic. These features are shown in Table 2.

Table 2: List of features in the Botake system

	Botake's Features
1	The number of DNS servers which the domain is associated with (req / res)
2	Query Type
3	Number of queries for each domain
4	Number of responses for each domain
5	Number of NXDomain for each domain
6	Number of resolved IPs for each domain
7	TTL Value
8	The number of sub-domains returned for each domain
9	Number of queries related to subdomains
10	The number of countries and ASs that are related to IPs
11	Entropy Names of Domains
12	behavioral entropy of each domain
13	The number of files downloaded from resolved IPs
14	the contaminated factor of each IP
15	Using the snort traffic features

The 12th feature of Table 2 highlights the main aim of this paper and can cause radical consequences to rely on our feature selection in clustering. For computing this feature, the number of IPs corresponding to each domain plus the communication volume associated with these IPs are stored as an input of the weighting algorithm in the database. The contaminated factor in fourteenth feature sheds light on negative reports of approximately 63 antivirus and malware detector. This factor (see Eq. 1) which acquired by Virustotal's reports reminds us the suspicious activities of one IP.

$$Contaminated\_Factor = \frac{\sum Negative\_Reports / Total\_Reports}{Total\_FilesDownloaded} \quad (1)$$

The fifteenth feature can be considered as a kind of black/white features for clustering or reference features for classification to detect unknown bots or C&Cs. Actually, an error message between two IPs in the intrusion detection system opens new way to exploit correlation methods to detect bots in clustering [8].

4.2 Investigating behavioral entropy between domains

4.2.1 Integration of Data and computing Entropy

In domain behavior analysis, the IP network and DNS traffic must be merged. This integration provides appropriate threshold for measuring features collected in Table 2.

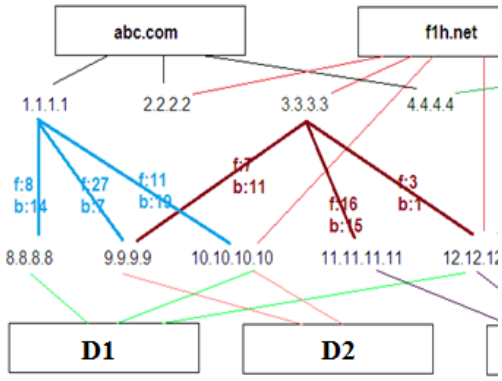

Fig. 4 Flow Network between Domains and IPs

Figure 4 illustrates the exchanges between source and destination IPs with their corresponding domain as well as other IPs that play the role of relay for packet transfer or completion of the attack process.

Calculating the entropy of these exchanges from the perspective of the IPs and their corresponding domains based on the packet volume is shown in Equations 2 to 4. The standard size of packets is assumed between 8 and 65535 bytes.

$$En(abc.com) = \{En(IP1) + En(IP2) + En(IP4)\} \quad (2)$$

$$En(IP1) = \{En(f) + En(b)\} \quad (3)$$

$$\begin{aligned} En(b) &= \{En(1,14) + En(1,7) + En(1,19)\} = \\ &\{-[\,(-14/65527)\log(14/65527) + \\ &(-7/65527)\log(7/65527) + \\ &(-19/65527)\log(19/65527)\,]\} \end{aligned} \quad (4)$$

Eq. 2 computes the entropy of the domain “abc.com”, whose relationship with the IPs of 1.1.1.1 and 2.2.2.2 and 4.4.4.4 is shown in Figure 4. En(IP1), the entropy of IP1, sums up the entropy of the packets (f[1]) and (b[2]). The entropy of backward in En(IP1) considered to the 14,7 and 19 received bytes respectively. The entropy of sent bytes as forward packets is also calculated as backward entropy. To complete the behavioral analysis of domains, we need consider to the entropy of the time and volume of communication between IPs and domains.

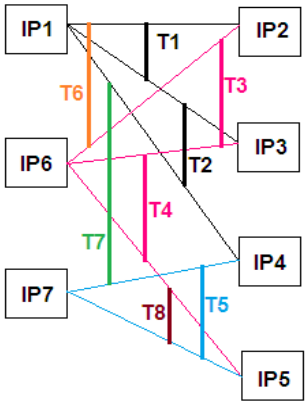

Fig. 5 Time gap between IPs connected

Probing abusive manner in traffic proves that cycle of exchanging between bots follows the rules of their network. Figure 5 shows the entropy of IPs connections for exploring suspicious pattern. These communication times are the same as the intervals between sending packets. Equations 5 and 6 describe how to calculate behavioral entropy of time gap between IPs connections. In these equations, the time interval T1 is assumed 10 minutes, and T2 is considered to be 1 hour during 24 hours.

$$En(IP1) = \{En(T1) + En(T2)\} \quad (5)$$

$$\begin{aligned} En(IP1) = \{&-[\,(-10\,\text{min}/24\,\text{h})\log(10\,\text{min}/24\,\text{h}) + \\ &(-1\,\text{h}/24\,\text{h})\log(1\,\text{h}/24\,\text{h})\,]\} \end{aligned} \quad (6)$$

Last part of computing is about exposing entropy of domains connection. This entropy leads us to distinguish C&Cs particularly DGAs from normal domains. Because the most detected patterns were between arbitrary domains like DGAs and short-life command and control centers.

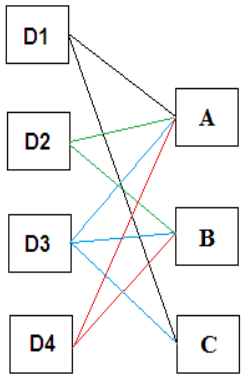

Fig. 6 Front-End exchange between Domains

[1] forward
[2] backward

If we assume that the relationship between domains D1 to D4 with domains A, B, and C is shown in Figure 6, the entropy of domains behavior in terms of time intervals will be calculated according to Eq. 7.

$$En(A) = En(D1) + En(D2) + En(D3) + En(D4) \quad (7)$$

In Eq. 7, the entropy of the domain (A) is equal to the sum of the entropies obtained for domains D1 to D4, because the final connection of these domains ends to domain (A). The method for calculating the entropy of each domain is given in Eq. 2.

4.2.2 Weighing Matrix

In this section, by providing the matrix of domains and intermediate IPs, as shown in Figure 4, we tried to weigh the domains to discover the neighboring domains in the botnets. This idea is based on the fact that in a botnet a unique pattern for communication often has been used [9]. The sequel of this matrix brings the set of domains with their intermediate connections, which will be considered in equations 2 to 7 for discovering bots' pattern.

$$\begin{bmatrix} & \overbrace{IP1 \quad IP2 \quad \cdots \quad IPm}^{i} \\ j\begin{cases} D1 \\ D2 \\ \vdots \\ Dn \end{cases} & \ddots \end{bmatrix} \quad (8)$$

$$W_{Dk} = \sum_{i=1}^{m}\left(\sum_{j=1}^{n} M_{ji}\right) \times M_{ki} \quad (9)$$

In matrix (8), it is assumed that the left-hand column contains "n" domains studied in the network, and the highest row representations "m" IPs which related to the "n" domains.

Figure 7 shows the result of identifying the suspicious domains, along with the IPs in a connected network. Bot masters always try to hide their control channels by helping two patterns of Domain-Flux and IP-Flux. The merit of this algorithm and equations 2 to 7 is being versatile to identify domains based-on DGA algorithm, Discover temporary C&Cs, random IPs and short/long-lived domains and IPs.

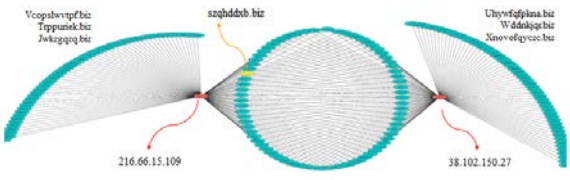

Fig. 7 Malicious communications of neighboring domains

As shown in Figure 7, uhywfqfpkna.biz, wddnkjqr.biz and xnovofqyczc.biz are the substantial result of our analysis in Botake. Virustotal reported them as CONFICKER bots with three Name Servers: NS.CONFICKER-SINKHOLE.COM, NS.CONFICKER-SINKHOLE.NET and NS.CONFICKER-SINKHOLE.ORG during 2015-2016. Details of attained Domains and IPs can be found in Table 3.

Table 3: acquired domains with their details

Domains	***Related IP***	***Weight***	***Entropy***	***Bot***
Vcops1wvtpf.biz	216.66.15.109	[4]	0. 56	□
Trppuriek.biz	216.66.15.109	[4]	0. 56	□
Jwkzgqzq.biz	216.66.15.109	[4]	0. 56	□
Szqhddxb.biz	216.66.15.109 and 38.102.150.27	[8]	0.23	□
Uhywfqfpkna.biz	38.102.150.27	[4]	0. 56	■
Wddnkjqr.biz	38.102.150.27	[4]	0. 56	■
Xnovofqyczc.biz	38.102.150.27	[4]	0. 56	■

Actually the domain "Szqhddxb.biz" with the lowest entropy and highest weight (cluster [8]) is the turning point of calculation because it brings us to two important IPs and bot domains. The IPs, 216.66.15.109 and 38.102.150.27 are both contaminated and they have been used in malicious connections. The relation between these two IPs has been recorded in 2015-08-29 with following URLs:

http://216.66.15.109/search?q=7&aq=7
http://38.102.150.27/search?q=7&aq=7

These IPs are detected by three anti-malwares out of the 63.

5. Conclusions

Research into APTs and botnets has been progressing for almost a decade. In this paper, we tried to discover various types of bot attacks based on DNS by Botake detection system. The results indicate that 24-hour data analysis with a volume of about 80GB requires 15 hours of processing. This processing has been run on a system with CPU, core i5 and 8GB of RAM. The main parameters in processing were the fewer features which reduce clustering levels, weighting algorithm to filter more active domains and comprehensive entropy calculations. The behavior of sending identified volume, regular connections, specific duration for exchanging, a certain number of IP in each cycle of connection, injection of random domains into the association cycle and finally, exchange malformed data can be recognized by Botake.

Acknowledgments

I would like to express my sincere gratitude to Mrs. Mansoureh Ghannadi for her patience, motivation, enthusiasm, and immense knowledge.

References

[1] Symantec, "Internet Security Threat Report", Vol. 21, APRIL 2016.
[2] Guofei Gu , "BotSniffer: Detecting Botnet Command and Control Channels in Network Traffic", Proceedings of the 15th Annual Network and Distributed System Security Symposium, 2008.
[3] Michael Dooley and Timothy Rooney, "DNS SECURITY MANAGEMENT", Published by John Wiley & Sons, Inc., Hoboken, New Jersey, ISBN: 978-1-119-32827-8, 2017.
[4] Top Domains Alexa : https://github.com/csirtgadgets/CIF-Data-Providers/wiki/Alexa-Top-Sites.
[5] Stefano Schiavoni and et al., "Phoenix: DGA-based Botnet Tracking and Intelligence", International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment, pp. 192-211, 2014.
[6] Ting-Fang Yen and et al., "Beehive: Large-Scale Log Analysis for Detecting Suspicious Activity in Enterprise Networks", ACSAC, ACM, 2013.
[7] Guofei Gu and et al., "BotMiner: Clustering Analysis of Network Traffic for Protocol- and Structure-Independent Botnet Detection", SS'08 Proceedings of the 17th conference on Security symposium, pp.139-154, 2008.
[8] Guofei Gu and et al., "BotHunter: Detecting Malware Infection Through IDS-Driven Dialog Correlation", 16th USENIX Security Symposium, pp. 167–182, 2007.
[9] Elias Bou-Harb and et al., "Big Data Behavioral Analytics Meet Graph Theory: On Effective Botnet Takedowns", IEEE Network, 2017.

Azar Hosseini Master of Science (M.Sc.) in Secure Communication Eng., Iran University of Science and Technology, 2013. Bachelor of Science (B.Sc.) in Electrical Eng., Dr. Shariaty Technical College, Iran, 2008. Fields of Interest: Machine Learning, Data Mining, Information Security, Internet of Things (IoT), Cognitive Radio, Communication Networks Mobile Communications, Wireless Sensor Networks. Email: st.azar.hosseini@gmail.com
Web: http://azar-hosseini.com/

[2††]Arezoo Hosseini
Doctor of Philosophy (PhD) in Pure Mathematics at Topological Groups, University of Guilan, Iran, 2012. Master of Science (M.Sc.) in Pure Mathematics at Topological Groups, 2008. University of Guilan, Iran. Bachelor of Science (B.Sc.) in Mathematics, Iran University of Shahid Rajaee , 2006.
Fields of Interest: Topological Groups and Dynamical system, cohomological Groups. Email: a.hosseini@cfu.ac.ir.
[2††]Corresponding Author