2011 Second International Conference on Networking and Distributed Computing

A Survey of Botnet Size Measurement

Shangdong Liu, Jian Gong, Wang Yang, Ahmad Jakalan
Key Laboratory of Computer Network Technology of Jiangsu Province
Southeast University, Nanjing, China
{sdliu, jgong, wyang, ahmad}@njnet.edu.cn

***Abstract*—Botnet size is one of the most important characteristics to evaluate the threat of botnet. Previous studies of botnet size basically focus on details and concreteness instead of nature of problem due to polymorphic, versatile, complex behavior of botnet and partial observation on network. This paper investigates the nature of botnet size, upon which four issues are introduced. The paper summarizes the existing solutions to the four issues and analyzes the challenges to resolve them deeply. Finally, some valuable research works of next step are proposed.**

Keywords- Botnet Size; Botnet Migrating; Botnet Clone; NAT; DHCP

I. INTRODUCTION

Botnet is a network of compromised computers named "Bots" which are all victims of botnet hacker's malware propagation. Bots are controlled by attacker (named by Bot Master or Bot Herder). Botnet is communicated through command and control channel which is also called C&C or C2 channel. C2 channel is one of the most distinguishing characteristics compared with the traditional computer virus or network worms.

Botnets have exerted serious threat against cyber-security. From the first botnet appeared in 2000 on, making advantageous of modular design, versatile functions, intelligent and stealthy propagation, highly controlled behavior, distributed and large scale attacks, botnet has become one of the ideal platforms for malicious activities on internet. Symantec has reported that there were 6,798,338 bots detected in 2009, and 85% of spam is from botnet [1].

For the detection of botnet, obtaining type and size is the basis of leveraging the degree of threat. It can be achieved to determine the type of a botnet by capturing samples of bot or analyzing characteristics of botnet C2 communication. However, as far as the size of botnet is concerned, it is more difficult to calculate, for all existing approaches of observation to networks are localized. Though it has been 10 years since the first malicious bot appeared and lots of findings have been achieved in research community on detection and prevention of botnet, a great many of challenges to measure botnet size still exist, for example, how to eliminate the influence of DDNS、NAT、DHCP and botnet migration or clone ("migration" / "clone" is that bot master asks bots "transfer" / "copies" itself to another channel or C2 controller)? For the purpose of understanding botnet size measurement, promoting further research on effective measurement and helping network security manager to learn the situation of botnet infections, it is significant to summarize research progress of botnet size measurement and clarify nature of botnet size measurement.

Up to the present, the definition of botnet size has been clearly proposed by M. A. Rajab in the meeting of USENIX HotBots 2007. They are botnet Footprint and Live Population [2]. Botnet footprint refers to the overall size of infected population of botnet at any time in its lifetime. Botnet live population is the number of live bots simultaneously present in C2 channel.

Based on the definition, measurement of botnet size should be dissertated by four issues: (1), the measurement of botnet live population, which is a problem of botnet detection in nature. (2), the measurement of botnet footprints. (3), dynamic tracing of botnet size. (4), area issue of botnet size. For: (1), live population can be obtained by network anomaly detections, but footprint contains offline bots which can not be detected by network anomaly detections. So there is different between live population and footprint of botnet. From the perspective of threat evaluation, botnet live population represents attack volume but botnet footprint stands for range of infection. (2), though live population and footprint can be obtained by anomaly detection and statistical inference, when considering dynamic change of size, migration and clone of botnet, situation is entirely different. Meanwhile, with the widely using of DDNS, DHCP, NAT technology, identification of botnet might get different result in different observing location and time. In short, it is different between static size of botnet and dynamic size of botnet. Therefore, it requires a separate study for dynamic tracking of botnet size and obviously it is a worth study for assessment of botnet threat. (3), global size should be measured differ with local size of botnet. The reason is, for the measurement of global size, that the influence of time zone, global estimation model should be considered, but without same requirement for local size. Since global size is helpful to understand botnet completely, it has important significance to study global size of botnet and all related questions can be summarized as "area issue of botnet size". In summary, the four issues of botnet size measurement have different characteristics and ideas to response. Meanwhile difficulties and challenges confronted with are also different. So discussing separately is required.

The rest of the paper is organized as follows: section 2 summarizes the methodologies to measure botnet live population. In section 3, some ideas to measure botnet footprint is discussed. Section 4 shows the problems to dynamically track botnet size. Section 5 describes the area issue of botnet size. Summary and future work is concluded in section 6.

978-0-7695-4427-4/11 $26.00 © 2011 IEEE
DOI 10.1109/ICNDC.2011.15

II. MEASUREMENT OF BOTNET LIVE POPULATION

Almost all methods of network intrusion detection can be used to measure botnet live population directly or indirectly. In terms of detection principle, there are three classes of methods: (1), detection methods based on active/passive DNS detection; (2), detection methods based on botnet C2 features; (3), detection methods based on correlation of multiple bases.

Active DNS detection (e.g. [3][4][5]) indicates that detection of botnet is based on actively utilizing domain name of C2 controller. Take DNS redirection [3] for example, it maps the domain name of botnet C2 server to prepared sinkhole and records all the connections between bots and C2 server. After counting the number of hosts who take connections, the live population of the botnet can be obtained. As for exploiting botnet C2 domain name, another example is botnet infiltration used in [5], in which, a behavior controlled bot program is lead to join a botnet C2 channel and the information fetched from broadcast message of C2 communication can be used to infer the live population. The advantageous of this method are obviously, that result of live population is precise, but possessing bot source code and semantics of C2 communication in advance are great challenges. More important, matured botnets have no longer broadcasted any message with member information. As a result, the method will be invalid. The difficulties and limitations of active DNS detections lie in: (1), the domain name of botnet must be known in advance; (2), some botnets have the ability to probe the DNS redirection; (3), some botnets employ hierarchical management (bot master shepherds bots with multiple C2 servers [6]). Obtaining only subset of all C2 servers would result in inaccurate botnet size.

Passive DNS detection (e.g. [7][8][9]) indicates that detection of botnet is based on special pattern of botnet DNS query which is collected from network passively. The DNS query launched by botnet has 3 characteristics: (1), sending volume is fixed (legal DNS query has randomness); (2), synchronism; (3), Aimed to increase stealthy, botnet frequently adopts DDNS technology but not so for legal hosts. Last but not least, DNS detection methods are available only on botnets with centralized C2 structure.

Typical features of C2 communication include: C2 channel ID of IRC botnet, kinds of IRC botnet commands ([10][11]), URL in Spam from botnet [12], stable pattern of botnet C2 communication [13], abnormal in/out degree [14], abnormal metrics of network flow [15] and so on. One of the examples exploring spam to detect is method by analyzing spam content [16]. The idea is hunting hosts which send a large amount of e-mail in the short term on mail servers as much as possible. These hosts will be judged to suspects. The mails embedded with same key URL will be classified as spam from the same botnet. Counting the spammers from one botnet will obtain the botnet's live population. An example employing features of C2 communication to measure botnet live population is: probabilistic algorithm based on P2P peer scanning [17], which is based on UDP peer-scan event of Conficker-C P2P botnet. The principle is that input UDP scanning volume in the monitored network is examined, and then the live population of Conficker-C botnet is inferred in statistical way. There are 2 problems for the algorithm: (1), how to determine the received UDP packet is Conficker-C's peer-scan packet? (2), how to choose the model to infer? For the first problem, as the destination port of Conficker-C's peer-scan is produced by Src-IP and date and which has been cracked, the scan packets can be accurately identified through UDP destination port. For the second problem, after calculating the actual scanning rate, active time and quantity of peers etc., appropriate distribution used to estimate the live population can be chosen. The common drawback of this type of approaches is that there is false positive more or less.

A typical method based on correlation of multiple bases is BotHunter [18]. The basic idea is: by capturing the data exchange, generated in the process of spread and attack of botnet, between inside and outside of network border, "chain of evidence" of botnet activity will be formed through correlating the captured data exchange according to the botnet working process. Five behaviors can be used as "evidences" of botnet activities: (1), vulnerability scans from outside to inside of network; (2), vulnerability explorations from outside to inside of network; (3), download requests of bot program from inside to outside of network; (4), C2 dialogue from inside to outside of network; (5), attacks or scans from inside to outside of network. Typically, correlation can reduce the false rate greatly. (Take the BotHunter for example, experimental results show that 95.1% of detection rate can be reached).

With the different detection basis and data sources, above methods have different detection accuracy. Generally, misuse detection based on C2 features possesses higher precision, but the scalability and adaptability responding to C2 change of botnet is poor. Therefore, the misuse detection is proper used in the occasion C2 features are clearly known. Instead, anomaly detection methods based on C2 communication patterns has lower accuracy and proper in the occasion that C2 features are not grasped accurately. In terms of botnet size, common existing problems include: (1), disambiguation of NAT address, that is, if results of detection contain NAT address, how to get the collection of hosts behind the NAT address; (2), if the basis of detection contains botnet C2 channel ID, how to know the actual infected host behind the channel ID, for C2 channel ID sometimes do not correspond with the infected host one by one. If a single bot experiences multiple C2 IDs, the botnet size will be overestimated; (3), for the multiple sets of bots, how to determine the real botnet belonged to.

III. MEASUREMENT OF BOTNET FOOTPRINT

Methods discussed in the previous section are about live population measurement of botnet. For the purpose of comprehensive understanding of botnet, botnet footprints need to be calculated. The most accurate method to calculate botnet footprint is, of course, to determine whether hosts are infected by botnet through host-based misuse or anomaly detection firstly, and then take count of infected hosts, but the feasibility of this practice is very small. Therefore, statistical inference is usually the only choice to calculate

botnet footprint, and there is no doubt that results are not accurate.

Till this survey is written, there are no literatures which focus on accurate estimation of offline hosts in observed network. Intuitively, it can be inferred from statistical data of the total population and online ratio of hosts in the network, with which footprint of botnet can be calculated. In detail, there are two problems need to be solved: (1), how to determine the number of offline hosts at a time based on online hosts within the observed network? (2), in offline hosts, how many of them are bots with high probability? For the first problem, if considering large scale network, the statistical data [19] can be referred to determine the offline portion of network population at a time. Instead, if the observed network is smaller in range, the same statistical data can be gotten by monitoring the network for a period of time. For the second problem, with "uncleanliness" [20] of network, the infected part in offline hosts can be inferred. The basic idea of uncleanliness is: the situation of infection in one network is a network property in nature, which is depended on the security status of the network and having nothing to do with the attacker. Therefore, in space, infected hosts will focus on "dirty" network; in time, infection will happen on the same collection of hosts repeatedly. According to the theory of uncleanliness, offline infected hosts could be estimated more accurately, with which footprint of botnet can be inferred.

IV. Dynamic Tracking of Botnet Size

Tracking dynamic changes of botnet size includes three aspects: (1), means or patterns of botnet propagation, for the pattern of botnet propagation determines the law of dynamic botnet propagation and future botnet size; (2), obfuscation produced by some botnet activities, such as botnet clone and botnet migration etc, which brings a series of challenges to track botnet size; (3), dynamic model of botnet.

Botnet experiences versatile propagation ways. One of the most important is vulnerability scanning [5]. Since botnet owns strong ability to control bots, the scanning of botnet is very flexible. All the scanning manner of botnet can be roughly classified as worm class and non-worm class. Worm class scanning is a way using more primitive style through which botnet has large scanning volume and holds a large amount of infected hosts in short time. Non-worm class scanning is integrated with a variety of scanning algorithm, including scanning on a network segment, hit-list and random scanning. Though the amount of infections caused by non-worm scanning is less than worm class scanning, the detection is more difficult for higher stealthiness.

Discovering botnet clone, migration and other activities, in essence, is to determine the ownership of collections of bots. Some studies about botnets' similarity are for this problem. For Example, in [6], known multiple IRC bots collections detected by communication between C2 controller and bots, due to the existence of botnet clone, migration and hierarchical management, these collections might belong to a same botnet. The algorithm uses "distance of communication's characteristics" and "overlap rate of bots" as the metrics of botnet similarity. Communication's characteristics includes: (1), traffic volume, i.e. the number of IPs which have communications in a certain period of time, reflects the habits of bots' online time); (2), the frequency of communication, i.e. the volume of single bot's traffic, reflects the habits of bot master's activities and version of the bot program. In the calculation of these characteristics, using of NAT technology would result in a consequence that activities of all bots concentrated on few public IPs. Thereby, the frequency of communication which from NAT public IP is much higher than the actual frequency of botnet communication (this can be used as the basis to judge NAT's public IP) and it should be removed. The algorithm uses "IP aggregation" to calculate "overlap rate of bots". "IP aggregation" is an operation of getting 24-bit prefix of bot's IP. The idea behind is that overlap rate calculated by bots' IP does not always equal to overlap rate of infected hosts for existence of dynamic assigned IP. Although "IP aggregation" can not be representative of infected hosts accurately, the error is depressed. The experiments prove that the accuracy of algorithm was 89%, i.e. more accurate footprint can be obtained by this method for 89% of IRC botnets. The advantage of the method is considering and attempts to resolve migration, clone and hierarchical management issues of IRC botnet. Disadvantages include: multiple collections of bots should be known before using the algorithm and it is only applicable for the botnets with centralized structure. For similarity judgment on the botnets, except for the"distance of communication's characteristics" and "overlap rate of bots", judgment by DNS [21] analyzing is also feasible, for bots belonging to current controller would issue DNS query to the target controller for performing migration.

In conclusion, unresolved problems of tracking botnet size include: (1), dynamic IP addresses and NAT addresses; (2), how to track entire life cycle while every stages in life cycle have different characteristics; (3), how to identify botnets detected at different time in the existence of botnet clone and migration.

V. Area Issue of Botnet Size

Note that, like live population, footprint of botnet also have regional issues (local or global), it is different between using local live population to calculate the local footprint and using local live population to calculate global footprint. Empirical estimation of local and global footprints also belongs to different issues. The main difference is that the global footprints need to consider the impact of time zones. There are usually two approaches to calculate the global size of botnet and they are statistical inference and empirical estimation.

For the statistical inference of global footprint based on live population of botnet, there is an example on the basis of propagation model [3]. The basic idea is: considering a situation that occurred on time t and in any one of time zones, the number of infected hosts on t is the difference between the number of original infected hosts and immune hosts. After taking derivation of time, it can be obtained that differential equation of botnet propagation in the time zone, and the final propagation model with full time zones can be

inferred through expanding from single time zone to all. Using DNS redirection method and recording communication between C2 controller and bots, the validity of the propagation model has been verified. With this model, live population and footprint of botnet in any time can be predicted. The most significant feature of this model is to take multiple time zones into account, but there is a premise that the original amount of infected hosts, ratio between online and offline of hosts and scanning rate, immunization rate of botnet in a particular time zone should be known in advance. Moreover, some metrics do not always remain invariable; therefore the great challenges to use model still exist.

For the empirical estimation of global footprint of botnet, there is an example, which once used in [5] and is based on cache hit of DNS. The basic idea is: it will lead to get large number of DNS reply that sending DNS requests of C2 controller's domain name to DNS servers as much as possible. Analyzing the TTL in the DNS replies can speculate the existence of bots in the network DNS server located. If the value of TTL is small, prove that there is at least one bot in the network DNS server located. Global footprint of botnet can be measured by similar study of cache hit on all of DNS server at large. The advantage of this method is easy to implement; disadvantages are: (1), the domain name of botnet's C2 controller should be known in advance; (2), the result of this method is only the lower bound of botnet's actual footprint; (3), it is difficult to determine the detection intervals for botnet which uses DDNS technology.

VI. SUMMARY

Based upon the discussion of this survey, conclusion can be drawn that the measurement of botnet size is not an isolated problem. It is related closely with capturing of bot programs, botnet detection and behavior analysis of botnet etc. Moreover, just as blind men touching an elephant, each way to measure botnet size reflects only a perspective of observation. If the objective is to get an accurate and comprehensive description of the botnet, it is necessary to consider various factors, such as dynamic IP assignment, NAT, botnet migration, botnet clone and timeliness of model etc. That is to say, for a complete definition of botnet size, following constraints should be taken into account: (1), the network area of concern, for it is very difficult to examine the infections of botnet in whole internet; (2), the exact meaning of a particular botnet, for example, it is different between a botnet controlled by a bot master and a botnet belonging to a botnet type; (3), time, botnet size is a metric with dynamic changes over time, thereby factor of time should be considered in definition of botnet size.

For the measurement of botnet size, further research could be: (1), limited by experimental conditions, most data sources are derived from local information in observed network. Hence a statistical inference model is needed for getting global size, but till this paper is written, without a recognized statistical inference model to estimate global size of botnet is in use. (2), most of the methods are only against a section of botnet's life cycle. There is no a model or tracking means for full cycle of botnet, and, there is no model to retort the changes caused by update of bot program. (3), most of methods detect botnets with IRC or C2 type of centralized structure. Few studies performed for botnet with decentralized C2 structure, such as P2P botnet; (4), though there are some solutions to the problems of dynamic IP, NAT and other factors in the face of measuring botnet size, large error still exists.

In addition, it is foreseeable that in the future more and more modern botnet will emerge from internet along with more stealthy C2 communications, more intelligent control style. Aimed at this trend, it should be considered to integrate multiple information and methods to improve the accuracy of the measurement of botnet size. Furthermore, for the limited scope of observation, distributed coordination mechanism between multiple organizations should be sought actively.

ACKNOWLEDGMENT

This paper is supported by National Basic Research Program of China under Grant Nos.2003CB304804, 2009CB320505 (973); the National Key Technology R&D Program of China under Grant No.2008BAH37B04

REFERENCES

[1] Symantec Security Response team. Symantec Global Internet Security Threat Report Trends for 2009. Volume XV, Published April 2010.

[2] M. Fabian, M. A. Terzis, "My botnet is bigger than yours (maybe, better than yours): Why size estimates remain challenging[C]," in Proc. of the 1st Workshop on Hot Topics in Understanding Botnets (HotBots 2007), 2007.

[3] D. Dagon, C. Zou, W. Lee, "Modeling botnet propagation using time zones[C]," in Proc. of the 13th Annual Network and Distributed System Security Symp. (NDSS 2006), 2006.

[4] Nazario, J. Holz, T. As the net churns: Fast-flux botnet observations. Malicious and Unwanted Software, 2008. MALWARE 2008. 3rd International Conference on. Publication Date: 7-8 Oct. 2008. On page(s): 24-31.

[5] Moheeb Abu Rajab, Jay Zarfoss, Fabian Monrose, Andreas Terzis. A multifaceted approach to understanding the botnet phenomenon[C]," In Proceedings of ACM SIGCOMM/USENIX Internet Measurement Conference (IMC), Oct., 2006. Rio de Janeiro, Brazil.

[6] Li Runheng, Wang Minghua, Jia Yan. Modeling Botnets' Similarity based on Communication Feature Extraction and IP Assembly, Chinese Journal of Computers, vol. 33, pp. 45-54, 2010.

[7] Hyunsang Choi; Hanwoo Lee; Heejo Lee; Hyogon Kim; Botnet Detection by Monitoring Group Activities in DNS Traffic. Computer and Information Technology, 2007. CIT 2007. 7th IEEE International Conference on. 16-19 Oct. 2007 Page(s):715 - 720.

[8] Villamarin-Salomon, R.; Brustoloni, J.C. Identifying Botnets Using Anomaly Detection Techniques Applied to DNS Traffic. Page(s): 476-481. Digital Object Identifier 10.1109/ccnc08.2007.112

[9] Anirudh Ramachandran, Nick Feamster, and David Dagon. Revealing Botnet Membership with DNSBL Counterintelligence. Conference on Botnet Detection - Countering the Largest Security Threat Arlington, VA, JUN 22-23, 2006, 2008 Page(s): 131-142.

[10] Goebel J, Holz T. Rishi: Identify bot contaminated hosts by IRC nickname evaluation. In: Proc. of the 1st Workshop on Hot Topicsin Understanding Botnets (HotBots2007).2007.

[11] Shirley, B. Mano, C.D. Sub-Botnet Coordination Using Tokens in a Switched Network. Global Telecommunications Conference, 2008. IEEE GLOBECOM 2008. IEEE. Publication Date: Nov. 30 2008-Dec. 4 2008. On page(s): 1-5

[12] Yinglian Xie, Fang Yu, Kannan Achan, Rina Panigrahy, Geoff Hulten, Ivan Osipkov. Spamming Botnets: Signatures and Characteristics. in ACM SIGCOMM 2008, Seattle, WA, August 2008

[13] Sang-Kyun Noh, Joo-Hyung Oh, Jae-Seo Lee, Bong-Nam Noh, and Hyun-Cheol Jeong. Detecting P2P Botnets using a Multi-Phased Flow Model. 2009. ICDS '09. Third International Conference on1-7 Feb. 2009 Page(s):247 – 253.

[14] Su Chang; Linfeng Zhang; Yong Guan; Daniels, T.E.; A Framework for P2P Botnets. Communications and Mobile Computing, 2009. CMC '09. WRI International Conference on. Volume 3, 6-8 Jan. 2009 Page(s):594 – 599

[15] Livadas C,Walsh B,Lapsley D,Strayer T. Using machine learning techniques to identify botnet traffic. In Proc. of the 2nd IEEE LCN Workshop on Network Security.2006.967-974.

[16] Hu Jun, Li Zhitang, Yao Dezhong. Measuring Botnet Size by Using URL and Collaborative MailServers. The 5th International Conference on Networking and Services, Valencia, Spain ,2009

[17] Weaver R. A Probabilistic Population Study of the Conficker-C Botnet. 11th International Conference on Passive and Active Measurement, APR 07-09, 2010 Zurich, SWITZERLAND

[18] Gu G, Porras P, Yegneswaran V, Fong M, Lee W. BotHunter: Detecting malware infection through IDS-driven dialog correlation. In: Proc. of the 16th USENIX Security Symp. (Security 2007). 2007.

[19] World Population Statistics, http://www.internetworldstats.com/

[20] M. Patrick Collins, Timothy J. Shimeall,et al. Using Uncleanliness to Predict Future Botnet Addresses. Oct. 2007 Proceedings of the 7th ACM SIGCOMM conference on Internet measurement.

[21] Zang Tianning, Yun Xiaochun, Zhang Yongzheng, Men Chaoguang. A Botnet Migration Analyzer Based on the C-F Model. Geomatics and Information Science of Wuhan University, 2010, (5).