Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.12785/amis/072L07

Combining the Capture-Recapture Method and Simple Linear Regression Analysis of the Malicious Domains Estimation

Tung-Ming Koo[1], *Hung-Chang Chang*[2]

[1]Graduate Institute of Information Management, National Yunlin University of Science and Technology, Taiwan
[2]Graduate Institute of Information Management, National Yunlin University of Science and Technology, Taiwan

Received: 7 Oct. 2012, Revised: 8 Jan. 2013, Accepted: 10 Jan. 2013
Published online: 1 Jun. 2013

Abstract: Fast-flux service networks (FFSNs) are currently the greatest threat encountered in the computer networking field. This technique hides attackers behind a network of proxy servers (agents), thereby avoiding detection by security personnel. FFSN benefits criminal parties because it protects their Web sites and extends Web site life span. FFSN is becoming more dangerous, and estimating the size of FFSN-agents is becoming increasingly difficult. Additionally, because flux-agents may represent bot nodes, we can estimate the scale of FFSNs to determine the extent of threats. This study primarily estimates the population size of FFSNs. The flux-agent population size was estimated using the joint hypergeometric maximum likelihood estimator (JHE) of capture-recapture methods (CRMs). The results showed that the JHE and CRM estimated the population size more rapidly compared to general survey approaches.

Keywords: Capture-recapture method (CRM), fast-flux service network (FFSN), population estimation

1 Introduction

In recent years, numerous hackers have employed a new application of a DNS technique, that is, fast-flux service networks (FFSNs). FFSNs incorporate a round robin DNS (RR-DNS) technique [1] that uses DNS records to induce continuous and rapid changes in proxy nodes (agents) to evade detection by information security personnel[2]. Network security has been the greatest concern of Internet users for many years. Currently, people have adopted the habit of posting details of their lives on networks, causing hackers to continuously develop new methods to commit online crimes and obtain substantial illegal profit. Thus, network security is a constant battle between illegal organizations and information security personnel.

Currently, larger networking Web sites (such as Google and Yahoo) rapidly associate their domain names with various physical machines to balance Web-based loads. Fast flux applies this concept similarly, but instead associates its domain names with agents that comprise the victims computers. Hackers subsequently employ FFSNs to use compromised machines as a springboard to protect their illegal online activities, such as hosting phishing Web sites, malware downloading sites, and spamming e-mail content sites [3]. Hackers use the compromised computer as a front-end proxy, hiding behind numerous proxy nodes. Therefore, even if being detected by information security personnel, only the IP address of the proxy node can be detected. This technique protects attackers from being exposed and extends the duration of criminal activities. This study focuses on estimating the population size of malicious domains provided by Web sites such as ATLAS and Malware Domain List. After conducting a general survey of malicious domains, we employed the joint hypergeometric maximum likelihood estimator (JHE) of capture-recapture methods (CRMs) to push back and estimate the size of the flux agent. We then employed linear regression analysis to establish a regression prediction database and thereby construct a complete prediction model and reduce prediction time. The second chapter of this paper includes a brief discussion of literature on related techniques investigated in this study. The third chapter describes the empirical results of the prediction model constructed using JHE and

* Corresponding author e-mail: g9823811@yuntech.edu.tw

linear regression analysis. The fourth chapter presents a discussion of the empirical data analysis and results. The fifth chapter explains our conclusions and possible future research developments.

2 Literature Review

2.1 Round Robin DNS (RR-DNS)

With the Internet becoming increasingly convenient, the demand for Internet access continues to grow. Numerous companies are also progressively evolving into Web-based and electronically operated companies. Therefore, the degree of dependency on the Internet has also increased; thus, single servers can no longer manage data loads. However, this problem can be resolved by providing more servers to manage loads and increasing the Internet connection quality.

RR-DNS is a technique that associates numerous IP addresses with a single domain name. A DNS server then provides IP addresses using a cycle method when clients computers transmit requests. This technique enables the equal distribution of Internet access to each server to ultimately achieve a balanced server load [1].

2.2 Botnet

Bots were initially created to facilitate the management of the IRC channel. Bots can conduct pre-IRC channel management and be controlled by commands established in advance [4]. Since IRC communication protocols were employed to control the SebSeven V2.1 backdoor program in 1999, bots have been widely adopted by criminals for illegal purposes [5]).

Botnets are also known as zombie networks, which are defined as a group of computers that have been infected by malicious software. Thus, criminals can control and perform specific attacks on computers over the Internet[6]. Controllers of botnets are commonly known as botmasters. Using the bot program, a botmaster can control bot program-infected computers from a distance [7]. To expand the population of botnets, criminals infect host computers through spam e-mails, social networking sites, software defects, or by embedding malicious programs into Web pages.

The most commonly used attack method is a distributed denial-of-service (DDoS) attack. The botmaster commands the infected computers to simultaneously send requests to a specific Web site. These numerous requests overload the Web site, rendering it unable to provide the normal services. With their Web site services offline, Internet service providers may lose a significant amount of money. Botnets are also employed for sending spam e-mails. The average annual income of a spammer ranges[8] between 50,000 and 100,000 U.S. dollars. Additionally, botnets are also used for click fraud to obtain illegal profit [9]. Click fraud occurs when criminals control numerous infected computers and click on Web advertisements to earn advertising revenues. Furthermore, criminals also log the keys struck[10] on the host computer keyboard (known as keylogging) to acquire personal information and credit card details.

2.3 Fast-Flux Service Networks (FFSNs)

The FFSN technique allows criminals to hide behind a group of proxy servers (agents) to avoid detection and tracking by information security personnel. This technique is a new method for hiding malicious Web sites. FFSNs use the DNS exchange mechanism and also combine mechanisms for Web site load balancing and proxy server redirecting. Thus, this technique extends the existence of malicious Web sites, attracts and affects more users, and consequently creates a larger botnet.

According to the definition provided by the Honeynet Project research organization [11], the primary objective of a FFSN is to obtain a legitimate fully qualified domain name (FQDN) for the distribution to numerous IP addresses. These IP addresses then employ the round robin and time-to-live (TTL) mechanisms to accelerate the IP address replacement process, thereby achieving the goal of an FFSN. Subsequently, when browsers connect to these Web sites, this method instead connects the browsers to various infected computers.

FFSNs can be divided into two sections: the front-end proxy, and the back-end, which is the real attacker (MotherShip). The front-end proxy is composed of infected physical machines (agents). The infected machines have low TTL values and are used for proxy redirecting. The DNS or HTTP requests sent by clients are actually sent to the attackers. The back-end (attackers) primarily aims to receive requests from infected physical machines, after which the infected machine responds by providing the client with the requested information. By employing this method, information security personnel cannot detect the actual address of the MotherShip.

2.4 Capture-Recapture Method

The capture-recapture method (CRM) is a statistical method commonly used to estimate the parent population size of living organisms, such as fish and animals. The CRM is also employed in human social studies and can be used to estimate and monitor population size when population estimation is difficult using general survey approaches [12].

This statistical method is categorized into two model types. If the parent population does not give birth, die, immigrate, and/or emigrate during the investigation period (that is, the parent population size did not change),

then the model is considered a closed CRM. A classic example of a closed CRM is the Lincoln-Petersen method. Conversely, if the parent population does give birth, die, immigrate, and/or emigrate during the investigation period, then the model is considered an open CRM. Open CRMs acquire different estimations at different times and locations and allow changes (birth, death, immigration, and/or emigration) in the parent population during the investigation period. A frequently employed open CRM is the Jolly-Seber method [13].

CRM involves random sampling at least twice. During the capturing process, each captured sample is marked and then released into their habitat. During the second capturing process, only the marked samples are recorded, and the unmarked samples are marked and then released. This process is repeated several times and each capture is recorded. Population estimation is then conducted according to the recorded data [12].

Sandeep et al. [14] estimated the population size of a P2P Internet network based on the characteristic[15] of PSP networks where each node partially protects neighboring nodes. Their findings verified that CRM can accurately predict the population of P2P Internet networks.

Weaver et al. [16] employed CRM to estimate the number of phishing Web sites (Netcraft and Castlecops) within the Internet at that time. Their results showed that phishing Web sites tend to be focused on specific netblocks [16].

Cao [17] suggested two sampling methods based on CRM. Using a P2P Internet network with 100,000 nodes as the estimated size, the experiment results verified that increasing the number of nodes in the capture-recapture process results in more precise estimation results. Additionally, Holz et al. [18] also suggested that CRMs can be employed to estimate the number of flux agents within a fast-flux domain.

Koo [19] proposed a model for estimating the size of P2P botnets. In P2P botnets, each node possesses data of partial botnet members. This model used this characteristic by collecting peer lists and recording the added P2P botnet node data. These node data were then used to collect additional neighboring node data. By combining this process with a CRM, the overall size of P2P botnets could be estimated [19].

2.5 Population Estimation

The Lincoln-Petersen method has been employed to establish superior population estimation methods. Regarding closed CRMs, the Lincoln-Petersen and Program CAPTURE method are most frequently used. Regarding open CRMs, the Jolly-Seber method is the most frequently employed. Scholars have also developed numerous estimation software programs such as SURGE, POPAN, NOREMARK, MARK, and CARE-2 [20].

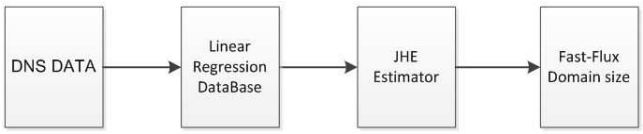

Fig. 1 Figure 1 Prediction module

3 Research Methodology

3.1 Prediction System Framework and Module

This study adopted a self-developed prediction system for FFSNs. This system obtains quantitative data by automatically processing malicious domain lists. Subsequently, prediction and empirical analysis of the quantitative data is conducted to determine the parameters of relevant prediction modules and validation data. These relevant data are then stored in the database. To predict other unknown malicious domains using the prediction module shown in Fig. 1, the regression database can be referenced to confirm the predicted linear conformity. Furthermore, according to the predicted curve with the highest conformity, and substituting the prediction value into the JHE prediction module, we can obtain the actual quantity of malicious domain predictions.

3.2 Joint hypergeometric maximum likelihood estimator (JHE)

FFSN is a technique that enables attackers to hide behind a group of proxy servers (agent) and evade detection by information security personnel. This technique prevents the exposure of malicious Web sites and extends the life span of malicious Web sites, attracts more users, and ultimately increases the size of the malicious domain. The harm caused by FFSNs is becoming increasingly severe; however, current academic studies on the size of victim populations are limited. We applied JHE as the primary calculation method to estimate the population size of FFSN victims in this study. The self-developed FFSN prediction system was also employed. The structure of this method is shown in Fig. 1. The FFSN prediction system determined the A record by analyzing the malicious domain, and subsequently applied the JHE estimation model to calculate and estimate the entire victim population size.

The JHE method adopts the following three assumptions:

1. Assumption 1: The population is closed (no births, deaths, immigrations, or emigrations) during the investigation period.
2. Assumption 2: The probability each individual being recaptured is equal.

3.Assumption 3: No shedding or misjudgment of the marks/tags has occurred.

If the FFSN population corresponds to these three assumptions, the FFSN conformities can be described as follows:

1.No changes in the FFSN agent size occurred during a certain period.
2.Because FFSNs operate using the RR-DNS technique, the probability rate of DNS A records being captures from the dig (domain information proper) query tool for querying domain name servers are equal.
3.Because the estimate size of FFSNs is automatically recorded by the experiment systems and stored to the database, no shedding or misjudgment occurs. Therefore, we selected the JHE estimation method for estimating the entire FFSN population size in this study.

According to the fast-flux domain obtained from the A record node information that was provided in response to the DNS query after a long duration, the system analysis of the fast-flux agent information was conducted as follows: The system calculated and analyzed the actual cumulative quantity of the single fast-flux domain and the repeatability of agent nodes, each A record was then marked, and the node from each capturing event was calculated. Because each capture time interval must be the TTL time of the DNS query, the interval unit was set as the TTL frequency (that is, the TTL unit). When combined with the required JHE estimation, we obtained the three computational parametersT_i,m_i, and n_i, where$T_i$represents the quantity of markers before the i^{th} time, m_i represents the quantity of repeated markers before and after the i^{th}, and n_i represents the unrepeated quantity after the i^{th} time.

The primary principle of JHE estimation assumes hypergeometric distribution for each capture and sampling process. Therefore, the probability of the i^{th} capture and m_i marked animal to be observed is

$$P_r\left(m_i|NT_in_i\right)=\frac{\binom{T_i}{m_i}\binom{N-T_i}{n_i-m_i}}{\binom{N}{n_i}}i=1\ldots..k \quad (1)$$

where N represents the total parent population (unknown).

Thus, the joint likelihood function is

$$\mathrm{L}\left(\mathrm{N}|T_in_im_i\right)=\prod_{i=1}^{k}\frac{\binom{T_i}{m_i}\binom{N-T_i}{n_i-m_i}}{\binom{N}{n_i}}i=1\ldots..k \quad (2)$$

Using this likelihood function to calculate the maximum likelihood estimator (MLE) is known as the JHE closed population model estimation method. However, if k approaches 1, that is, only one capturing and sampling process exists, applying the deviation correction model suggested by Chapman [21] (

$$\widehat{N}$$

: estimation method) can be expressed as

$$\widehat{N}=\frac{(n_i+1)(T_i+1)}{m_i+1}-1i=1,\ldots\ldots,k \quad (3)$$

This correction method is effective for reducing the deviation caused by the limited number of markers during the observation period.

3.3 Simple Linear Regression

The values predicted using regression models are typically the regression surface or corresponding surfaces. When the number of independent variables included in the model increases, the corresponding surface increases in complexity. The regression model occurs in many forms; however, regardless of the form, the aptness of the model must be determined before application. Model aptness is generally assessed using graphical representations (scatter plots), although a statistical test can also be applied. Additionally, various variable conversion techniques can be employed to ensure coordination between the data and model. Because users cannot predetermine whether the data and model can be coordinated in actual field applications, and conversion techniques are difficult to control, various assumptions of regression analysis must be examined.

Regression models comprise both linear and non-linear models. The non-linear regression analysis models commonly used include the quadratic curve model (QUA), composite model (COM), growth model (GRO), logarithmic model (LOG), cubic curve model (CUB), S equation model (S), exponential model (EXP), inverse model (INV), power (POW), and logistic distribution (LGS) model. In this study, we conducted simple linear regression analysis as follows: (1) The variables used for this study were differentiated into dependent and independent variables; (2) dependent variable functions were transformed into independent variable functions following relevant theories; (3) the obtained sample data was used to estimate the model parameter [22]; and (4) regression analysis was performed using the two or more collected variables as the basis to determine a regression equation between the variables. Step 4 is also known as parameter estimation. The most commonly used estimation method for regression analysis is the least squares method.

In this study, regression was conducted using sample data between variables and the least squares method; this approach is known as estimated regression. However,

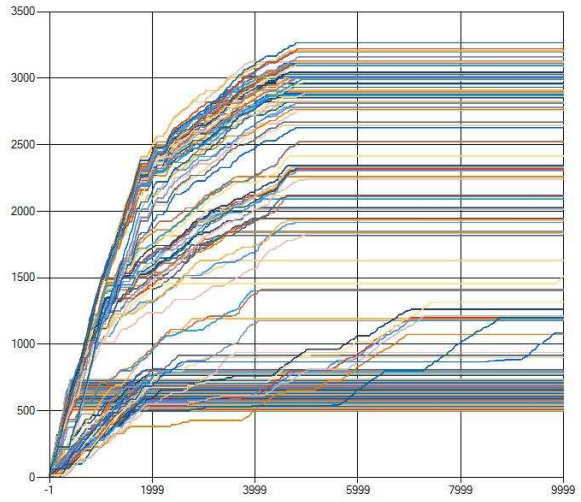

Fig. 2 Figure 2 IP-predicted distribution curve

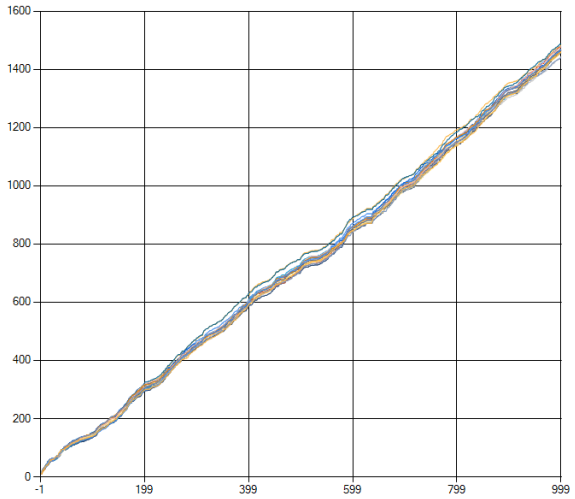

Fig. 3 Figure 3 Initial prediction curve

estimated regression does not actually a regress the parent population because of random sampling errors. For example, in linear regression analysis, the i^{th} observation value can be decomposed into the following theoretical error equation:

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon_i \quad (4)$$

The error term ε_i for this equation typically assumes a normal distribution $ND(0, \sigma_2)$, where Y represents the predicted value of the contingency number, X_1 represents the observation value of the independent variable, β_0 and β_1 represent the regression parameters, and ε_i represents the random error term. If ε_i = 0, then the estimated regression is the regression of the parent population.

$$\mathrm{Y} = \beta_0 + \beta_1 X_1 \quad (5)$$

In this study, we obtained TTL units and the IP predicted distribution curve from the DNS query data system, as shown in Fig. 2. We found that the curve distribution exhibit straight linear growth during the initial prediction period. Yang [23] contended that the prediction ability of the regression model is superior when the model assumes that the data corresponds to the simple linear regression model, and when the sample is limited and fits the normal linear regression model. As shown in Fig. 3, we selected the linear regression analysis method to construct the prediction model developed in this study.

4 The Empirical Results

4.1 Origin of the Data

The fast-flux domain used in this study was obtained from ATLAS Arbor Networks [24], which is a Web site that primarily collects and sums the FFSN domains detected, and the well-known Web site MalwareDomainList [25]. The study period was from May 2010 to August 2012. The number of domain groups and pieces of DNS query data used in this study was 1,340 and 4,525,237, respectively.

4.2 System Environment

The system environment was established in an actual network environment. We tested the system environment using IP addresses from various countries, and subsequently captured and recaptured the A record of the FFSN domain agent. We found that captured A records exhibited no differences, which verified that A records have no regional and transnational problems. Consequently, we used a DNS Server (IP: 8.8.8.8) provided by Google to collect data for analysis.

4.3 The Empirical Data

4.3.1 The JHE Estimation Method

The JHE estimated method was conducted with the following specifications: a tolerable maximum error probability of 0.05 and a prediction accuracy of 95%, with the TTL increasing from *10 to *400 in increments

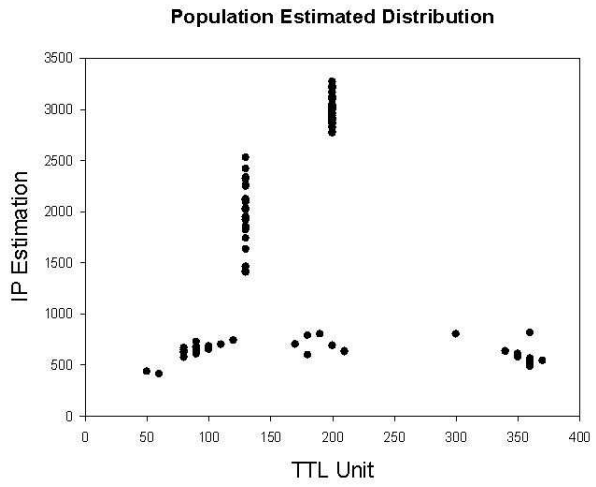

Fig. 4 Figure 4 Graphical representation of size estimation analysis

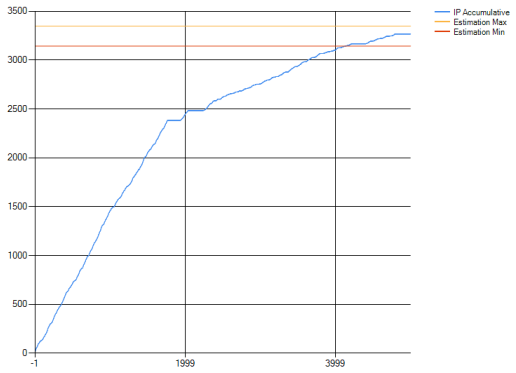

Fig. 5 Figure 5 Domain estimation map (erosocialka.ru)

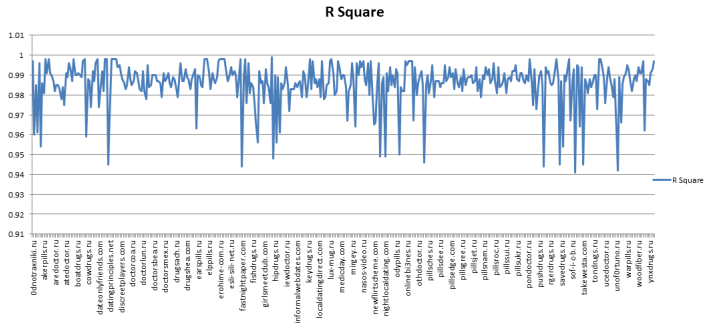

Fig. 6 Figure 6 R-square distribution)

of 10. We substituted these specifications into the JHE estimation method and performed size estimations. We obtained the actual accumulated quantity of fast-flux domains through system analysis, and cross-validated the calculated values with TTL (when = 0.05) to obtain the analytical data (Fig. 4). As shown in Fig. 4, predictions of IP quantity and TTL unit distribution using the JHE method are categorized into the following three major groups:

1. When the IP estimation ranged between 2766 and 3266, the number of TTL units was 200.
2. When the IP estimation ranged between 1402 and 2525, the number of TTL units was 130.
3. When the IP estimation ranged between 409 and 813, the number of TTL units ranged between 50 and 370.

Empirical data obtained from erosocialka.ru is shown in Fig. 5. The maximum and minimum error of the actual population size was 2.040221% and -2.349344%, respectively. This indicates that substituting 200 TTL units into the JHE estimation method provided an accurate estimation of the number of erosocialka.ru values.

4.3.2 Simple Linear Regression Analysis

In this section, we substituted statistical software into the empirical result of the JHE method obtained in Section 4.3.1. We used simple linear regression analysis to calculate the linear regression parameters of the individual domain, and subsequently produced a regulated database for future prediction analysis.

The empirical results of the JHE method showed that when the IP quantity ranged between 2,766 and 3,266, the data before 200 TTL points were used for analysis. When the IP quantity ranged between 1402 and 2,525, the data before 130 TTL points were used for analysis. When the predicted IP quantity range was less than 813, all the TTL point data were used for analysis. Figure 6 shows the distribution of the coefficient of determination R-squared (R2) for regression analysis (the statistical software possessed a confidence level of 95%).

During regression analysis, R2 typically indicates the proportion of variance that can be explained through regression analysis. Therefore, R2 can be employed as an accuracy indicator by using x to predict y. For example, when employing age (x) to predict the duration (in hours) of sleep (y), an R2 = 0.76 suggests that 76% of the total variance in sleep duration can be explained by age. Larger R2 values typically indicate higher accuracy. The 1,340 groups of domain experimental data that underwent regression analysis generated R2 values between 0.94 and 0.99. This suggests that 94% to 99% of the domain data can be explained by TTL. This finding further indicates that simple linear regression analysis generates extremely accurate data.

Fig. 7 Figure 7 Diagram of FFSN relationship linkage)

4.3.3 Visualization of the Network Connection Relationship

We imported the 4,525,237 DNS queries and 1,340 domain groups into image visualization software GEPHI. We then generated layout images with interconnections using the softwares layout algorithm function (the relationship linkage and strengthening algorithm). As shown in Fig. 7, three distinct populations were observed, which corresponds with the empirical analysis data described in Section 4.3.1.

4.4 Origin of the Data

The empirical results obtained in this study suggest that the size of FFSN populations can be estimated by predicting the size of fast-flux domains using the JHE estimation method of CRMs within a fixed TTL unit time. Furthermore, simple linear regression analysis can also be employed for FFSN size estimation to accelerate preliminary prediction analysis and obtain a regression analysis explanatory power ranging between 94

The estimation rules of the CRM method are as follows:

1. The number of agents can be predicted accurately when the estimated IP quantity range between 2,766 and 3,266, and the TTL units are 200 (TTL = 10 s, for a total of 2,000 s).
2. The number of agents can be predicted accurately when the estimated IP quantity ranges between 1,402 and 2,525, and the TTL units are 130 (TTL = 10 s, for a total of 1,300 s).
3. When the estimated IP quantity ranges between 409 and 813, the TTL units range between 50 and 370 (TTL = 300 s 3,600 s). Therefore, the JHE method cannot be used to accurately estimate the number of agents.

The method for estimating the size of FFSN populations within a fixed TTL unit time presented in this study was developed and verified after collecting numerous fast-flux domains.

During the empirical and data analysis processes, we observed a specific phenomenon in the interconnected layout diagram. We found numerous domain names with the same A record; for example, the IP address 122.194.5.110 corresponded to two domains, namely dartzofmybpull.ru and shokoladdeath.ru. From Fig. 7, we can speculate that the domains with the same A record nodes may have originated from the same attacker or botnet.

5 Conclusion and Future Prospects

The primary objective of this study was to estimate the size of FFSN populations. The field of computer networking faces increasing threats from FFSNs because this technique enables attackers to hide behind numerous proxy servers, increasing the difficultly of tracking and detecting attackers. This is also why current scholars are actively searching for FFSN characteristics that can be used to identify the primary attacker. Furthermore, FFSNs comprise numerous proxy servers that are infected by bots. If the infected server population increases, the threat to network security also increases. Therefore, the size of FFSN populations must be estimated to facilitate the effective control of FFSN development.

The results of this study showed that the size of FFSN populations can be estimated using the JHE method of CRMs. We also suggested that the estimation time can be reduced by applying a fixed TTL unit and using regression analysis. The general survey approach for estimating the fast-flux domain size requires at least 14 days to complete. Therefore, the method proposed in this study can perform comparatively rapider estimations.

During the empirical process, we found that the A records corresponded to numerous domain names. Additionally, based on observations of the interconnected layout image, we speculated that domains with the same A record nodes may originate from the same attacker or botnet.

Based on statistical data, we found that the FFSN proxy node points with more than 3,000 IP addresses contained 276 domains. This result demonstrates the significant population size. If preventive measures are not adopted, this population will damage network security.

References

[1] Hsiao, Y.L., Chen, Y.C., and Huang, C.H., On the Availability of Network Services Based on RR-DNS, Taiwan Academic Network Conference ,(2003)

[2] Bo Li, Yuhong Li, A Bi-Directional Security Authentication Architecture for the Internet of Vehicles, Applied Mathematics and Information Sciences, Special Issues. Nov., 821-827, (2012)
[3] Tseng, C.C., DNS Old Technique New Gameplay, Journal of Information Security, 66, 74-78, (2009)
[4] Scharrenberg, P., Analyzing Fast-Flux Service Networks, Diploma Dissertation, RWTH Aachen University, Aachen, North Rhine Westphalia, Germany, (2008)
[5] Bacher, P. et al., Know your enemy: Tracking botnets. The Honeynet Project, (2005)
[6] Y. Shang, Optimal Attack Strategies in a Dynamic Botnet Defense Model, Applied Mathematics and Information Sciences, Volume 6. Jan., 29-33, (2012)
[7] Saha, B., and Gairola, A., Botnet: An Overview, (2005)
[8] Ming-Yang Su, Chen-Han Tsai, A Prevention System for Spam over Internet Telephony, Applied Mathematics and Information Sciences, Special Issues. Apr., 579S-585S, (2012)
[9] Daswani, N., and Stoppelman, M., The anatomy of Clickbot.A, HotBot 07 Proceedings of the first conference on First Workshop on Hot Topics in Understanding Botnets, (2007)
[10] A. Jemai, A. Mastouri, H. Eleuch,Study of key pre-distribution schemes in wireless sensor networks: case of BROSK (use of WSNet), Applied Mathematics and Information Sciences, Volume 5. Sep., 655-667, (2011)
[11] Project, T.H., Know Your Enemy: Fast-Flux Service Networks, Available from: http://www.honeynet.org/papers/ff., (2007)
[12] Wild Animals Research Tutorial Manual, College of Agriculture, National Pingtung University of Science and Technology, Pingtung, Taiwan, (2006)
[13] Jolly, G. M., Explicit estimates from capture-recapture data with both death and immigration-stochastic model, Biometrika, 52, (1965)
[14] Sandeep Mane, Sandeep Mopuru, Kriti Mehra, and Jaideep Srivastava. Network Size Estimation In A Peer-to-Peer Network. Technical Report TR 05-030, Department of Computer Science - University of Minnesota, (2005)
[15] Xibin Zhao, A Mathematical Characterization of System Design and Modeling, Applied Mathematics and Information Sciences,Volume 6. May., 345-356, (2012)
[16] Weaver, R., and Collins, M., Fishing for phishes: Applying capture-recapture methods to estimate phishing populations, In Proceedings of 2nd APWG eCrime Researchers Summit, (2007)
[17] Cao, C., Research About Size Estimation Methods in P2P Network, Computer Engineering and Applications, 48(29), 99-101, (2008)
[18] Holz, T., Gorecki, C., Rieck, K., and Freiling, F.C., Measuring and Detecting Fast-Flux Service Networks, In Symposium on Network and Distributed System Security, (2008)
[19] Koo, T.M., Chang, H.C., Liao, W.C., Estimating the size of P2P botnets, International Journal of Advancements in Computing Technology, 4(12), 386-395, (2012)
[20] Hu, W.Y., Population Estimate of Dunlin (Calidris alpina) at Changhua Coastal Area, Masters Dissertation, Department of Environment Science, Tunghai University, Taiwan, (2006)
[21] Chapman, D. G., Some properties of the hypergeometric distrution with applications to zoological censusesUniv, Calif. Public. Stat, 1, 131-60, (1951)
[22] Dielman, T. E., Applied Regression Analysis for Business and Economics, PWS-Kent Pub. Co., Boston, MA., USA, (1991)
[23] Yang, Y.Y., A Comparison on the Prediction Performance of Regression Analysis and Artificial Neural Networks, Masters Dissertation, Department of Statistics, National Chengchi University, Taipei, Taiwan, (2002)
[24] ATLAS.Global Fast Flux ,Available from: http://atlas.arbor.net/summary/fastflux,(2011)
[25] Malware Domain List, Available from: http://www.malwaredomainlist.com/, (2009)

Tung-Ming Koo received the Ph.D. degrees in Computer Science from Oklahoma State University, USA. He is currently an associate professor in Department of Information Management, National Yunlin University of Science and Technology, Taiwan. Most of his research areas are Information Security, Data Compression and Algorithm.

Hung-Chang Chang is a PhD student in Department and Graduate Institute of Management Information System at the National Yunlin University of Science and Technology, Taiwan. He was a Supervisor in the Formosa Advanced Technologies Company from 2001-2012.