Guilt by Association-based Discovery of Botnet Footprints

Alper Caglayan, Milcord LLC
Mike Toothaker, Dan Drapeau, Milcord LLC
Dustin Burke, Milcord, LLC
Gerry Eaton, Milcord LLC

acaglayan [@] milcord [dot] com

ABSTRACT

In this paper, we describe a Guilt-by-Association approach to determining botnet footprint starting from a subset of known domains belonging to a specific botnet, and demonstrate our approach using recent botnets. Our empirical results leverage the botnet database that we have collected over a period of 12 months with our real-time fast flux network detection algorithm [1]. Botnets exploit a network of compromised machines (zombies) for illegal activities such as Distributed Denial of Service (DDoS) attacks, spam campaigns, phishing scams and malware delivery using DNS record manipulation techniques. Our results, which build upon our behaviour [2] and social network analysis [3] results, show that it is possible to identify a large portion of a botnet once a small segment of that botnet is identified through manual means, and to explain the differences in botnet footprint prediction using our proposed connectivity metric.

1.0 INTRODUCTION

Although botnets are detected and classified using automated means [1], the labeling of a specific botnet (e.g. Storm, Waledac, Zeus, Avalanche, etc.) is a tool enhanced manual process. Typically, the labeling process involves building several honeypots for catching spam, capturing the malware code used for botnet infection, executing the captured bot in a sandbox to detect the malicious URLs visited, reverse engineering the malware code to uncover additional domains used for second stage downloads, command and control, beacon sites with the aid of automated static and dynamic analysis tools. One of the most comprehensive studies conducted in this area is Spamanalytics [4] research in the Sep. 2009 issue of ACM Communications, which analyzes the conversion rate of three spam campaigns comprising over 469 million emails. By infiltrating the Storm botnet using 8 proxy bots, the researchers convinced Storm to modify a subset of the spam it already sends, thereby directing any interested recipients to Web sites under their control to study the click-through rates. In terms of response rates, India, Pakistan, and Bulgaria have the highest response rates than any other countries. The United States, although a dominant target and responder, has the lowest resulting response rate of any country, followed by Japan and Taiwan. An overview of the challenges involved in executing malware in sandboxes can be found in Miwa et. al [5].

Our paper answers the following question: Given a small subset of a botnet identified through reverse engineering, is it possible to predict the full footprint of this botnet? Our solution is based on the Guilt by Association analysis of the network of nameservers, domains, and IPs. Using our botnet database, we analyze if other (both flux and non-flux) domains have used the labeled nameservers. Similarly, we analyze if other domains have used the IPs of labeled domains. Given a subset of the labeled botnet, we investigate the coverage (the percent of the botnet footprint that can be derived by the labeled subset) of our Guilt by Association approach. We demonstrate our approach comparing the results for the Avalanche, Conficker, Gumblar, Pushdo, Waledac and ZeuS botnets. Our results show that the Guilt by Association approach can predict the botnet footprint with a moderate sample size, and the prediction coverage depends on the number of clusters and connectivity of each cluster in the botnet graph.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE
NOV 2010

2. REPORT TYPE
N/A

3. DATES COVERED
-

4. TITLE AND SUBTITLE
Guilt by Association-based Discovery of Botnet Footprints

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Milcord LLC

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release, distribution unlimited

13. SUPPLEMENTARY NOTES
See also ADA564697. Information Assurance and Cyber Defence (Assurance de l'information et cyberdefense). RTO-MP-IST-091

14. ABSTRACT

In this paper, we describe a Guilt-by-Association approach to determining botnet footprint starting from a subset of known domains belonging to a specific botnet, and demonstrate our approach using recent botnets. Our empirical results leverage the botnet database that we have collected over a period of 12 months with our real-time fast flux network detection algorithm [1]. Botnets exploit a network of compromised machines (zombies) for illegal activities such as Distributed Denial of Service (DDoS) attacks, spam campaigns, phishing scams and malware delivery using DNS record manipulation techniques. Our results, which build upon our behaviour [2] and social network analysis [3] results, show that it is possible to identify a large portion of a botnet once a small segment of that botnet is identified through manual means, and to explain the differences in botnet footprint prediction using our proposed connectivity metric.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**SAR**	**16**	

ICANN describes [6] fast flux as ‘rapid and repeated changes to host and/or name server resource records, which result in rapidly changing the IP address to which the domain name of an Internet host or name server resolves’. While fast flux methods do have a legitimate use as a load balancing technique for high availability and high volume Web sites, its malicious use enables concealment of the Command and Control server using compromised machines (‘zombies’) that are used in DDoS, spam, phishing, malware delivery. There are three main variants of fast flux hosting: (1) basic fast flux hosting where IP addresses of malicious web sites are fluxed, (2) Name Server (NS) fluxing where IP addresses of DNS name servers are fluxed, and (3) double flux, where IP addresses of web sites and name servers are fluxed [6].

Published works on fast flux databases include the ISOC Network and Distributed System Security Symposium (NDSS) paper [7] on measuring and detecting fast flux service networks, the FluXOR paper [8] on detecting and monitoring fast-flux service networks, and our CATCH paper detecting and classifying fast flux service networks in real time. The ISOC paper uses a direct DNS monitoring approach over seven weeks of collected data, and is based on building a linear classifier using a flux score, which is a function of number of unique A records in all lookups, number of NS records in a single lookup, and number of unique ASNs (Autonomous System Number). The FluXOR method collects domains from spam emails in honeypots, monitors their DNS over a period of 3 hours and uses a trained Naïve Bayes classifier to classify as benign or fast-flux.

Our approach detailed in [1] complements current data collection research by focusing on the real time detection and classification of fast flux service networks using both active and passive DNS monitoring. We employ a Bayesian classifier that fuses multiple indicators including fast flux activity index, network footprint index, TTL, guilt by association, and others. In our earlier paper [2], we analyzed the short-term, long-term, organizational and operational fast flux service networks. In this paper, we analyze the structural relationships (domain, nameserver, IP connectivity) of fast flux botnets, identify recurrent structural clusters across different botnet types, and demonstrate the guilt by association knowledge encoded in these structures. For instance, for a new suspicious domain, having the IP it resolves to or having one of its nameservers in our database tagged as fast flux speeds up the detection process. Using a social network connectivity metric, we show that {Command and Control and phishing}, {malware and spam botnets} have similar structural scores with this metric.

Other related research includes the correlation of multiple DNS returns [9] at the University of Melbourne. This research has shown that the correlation of evidence from multiple DNS servers offers substantial speed up in the detection of fast flux botnets. In our approach, the use of different DNS servers corresponding in our active and passive monitoring subsystems offers the same advantage. Research at Indiana University [10] focused on the longevity of phishing botnets lasting less than 10 days to see if fast flux evasion increases the expected lifespan. This research shows that double flux increases the lifespan of phishing botnets. In our research, we focused on the lifespan of spam, CnC, malware and phishing networks. Research at Georgia Tech [11] analyzed the rate of change in DNS records for spam botnets using fast flux evasion techniques. In particular, scam domain change on shorter time intervals than their TTL values.

2.0 EXPERIMENTAL BACKGROUND

2.1 Data Collection

We collected our fast flux database using our Fast Flux Monitor (FFM); a Web service application designed to detect whether a domain exhibits fast flux (FF) or double flux (DF) behaviour. The primary technical components of FFM include: (1) sensors which perform real-time detection of FF service networks using behavioral analysis that examine various indicators, (2) a database of known FF service networks – zombie IPs used for domain names, nameservers, and (3) analytical knowledge harvested from the database, which can include: (i) the fast flux service network's size and growth rate estimates , (ii) the social network of a fast flux service network where IPs are shared by different fast flux service networks, (iii) the footprint of a fast flux service network for a given enterprise, (iv) the footprint of a fast flux service network for a given ISP, and (v) the footprint of a fast flux service network for a given country.

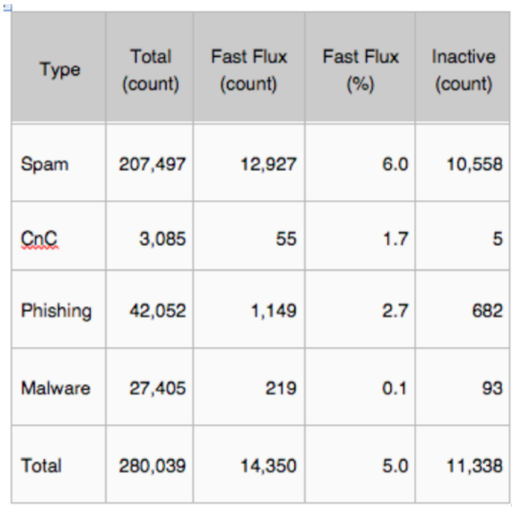
We have employed multiple sensors for our FFM active sensors: (1) FF Activity Index, (2) Footprint Index, and (3) Time To Live (TTL), and (4) Guilt by Association Score. In active monitoring, we perform DNS lookup with dig, and record the A records returned with each query. For nameservers, we perform dig in order to resolve a set of nameservers. For each nameserver, we perform an nslookup in order to resolve the set of IP addresses associated with the nameservers. We then query our database to see if any of the resultant IP addresses have been associated with other domains that we have been monitoring.

Table 1: Botnet database domain coverage

Entities	Total	Fast Flux	Active
Domains	506,764	22,495	8,042
IPs	547,246	231,825	119,288
Nameservers	231,825	29,452	2,802
Total	1,285,835	283,772	182,132

Table 1 shows the coverage of our Fast Flux Botnet database, which began monitoring in March 2008. This database is the input into the pattern analysis reported in this paper. As of Nov. 30 2009, our botnet database contains over 500,000 domains. About 5% of these domains have been classified as fast flux by our real-time fast flux detection algorithm. Fast flux spam botnets constitute the largest category in our database followed by fast flux phishing, Command and Control (CnC), and malware delivery botnets. There are a significant number of inactive fast flux domains, enabling us to study the long-term behaviour of fast flux botnets.

Table 2: Botnet attributes in fast flux database



Type	Total (count)	Fast Flux (count)	Fast Flux (%)	Inactive (count)
Spam	207,497	12,927	6.0	10,558
CnC	3,085	55	1.7	5
Phishing	42,052	1,149	2.7	682
Malware	27,405	219	0.1	93
Total	280,039	14,350	5.0	11,338

Table 2 shows the coverage of fast flux network attributes - domains, domain IPs, and nameservers for a labeled subset. Our database contains over 8,000 active fast flux domains with over 119,000 IPs and 2,800 nameservers. When inactive ones are taken into account, there are over 465,000 fast flux domain, IP, and nameserver entities in our fast flux botnet database.

In contrast to the database restricted to phishing botnets in [12] and the database restricted to spam botnets, having multiple types of botnets for spam, phishing, malware and Command and Control enables us to compare the lifespan, size, and structural connectivity of different fast flux botnet types. Our analysis answers these questions for fast flux service networks: *Are spam botnets larger in size than phishing botnets? Is the lifespan of malware botnets longer than that of phishing botnets? Do malware botnets use more nameserver fluxing than phishing botnets? Does (the number of domains/the number of nameservers) change across botnet types? Is there sufficient guilt by association knowledge encoded in domain/nameserver/IP graphs that can be exploited for botnet detection?*

2.2 Botnet Operational Behaviour

Figure 1 shows the lifespan distribution for the general population of botnets. We also analyzed the lifespan distribution of fast flux botnets used for spam campaigns, phishing scams, and malware delivery separately. In terms of the fast flux botnet lifespan, spam botnets outlive malware botnets, which, in turn, outlive phishing botnets. Phishing fast flux botnets tend to die out very quickly, most living less than a week. In contrast, malware delivery and spam botnets live up to 30 and 90 days, respectively. Referring back to Figure 1, the humps at 20 and 70 days represent the large scale botnets that is the focus of our paper.

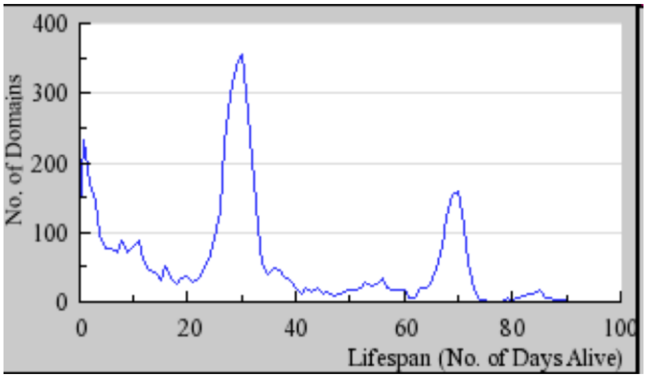


Figure 1: Lifespan distribution of fast flux botnets

We analyzed each botnet’s distribution across different countries, and Autonomous System Numbers (ASNs). Figure 2 shows the number of countries where fast flux botnets operate. While there are a large number of botnets operating in a few (less than 5) countries, most of the botnets operate in between 20 and 40 countries. The number of botnets operating in more than 40 counties falls off sharply.

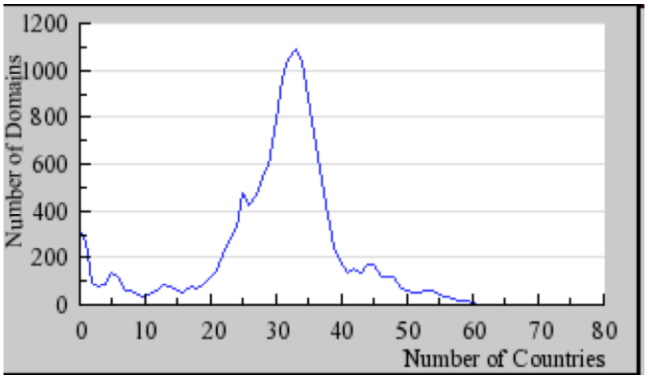


Figure 2: Operations in Multiple Countries

We also analyzed the distribution of fast flux botnet across ASNs. Figure 3 shows the number of ASNs in which fast flux botnets operate. While there are a large number botnets (100) operating in less than 10 ASNs, fast flux botnets operating in 10 – 250 ASNs have a bimodal distribution with sizable mass at modes correspondiung to 100 and 175 ASNs. We believe that the bimodal distribution is due to the presence of two large botnets (e.g. Waledac) in our database. Removal of the large botnets from consideration yields a uniform distribution between 10 and 250 ASNs. The ASN distribution has a sharp fall off after 250 ASNs with practically no botnets operating in more than 375 ASNs. We suspect that his behavior is due to the limited size of the botnet ASN targets, namely, ISPs and universities.

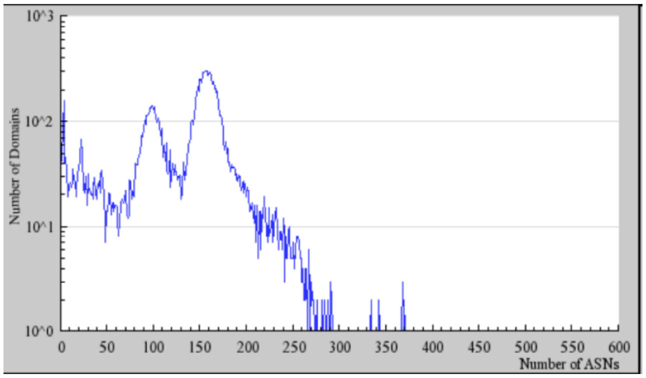


Figure 3: Operations in Multiple ASNs

2.3 Use Cases: Avalanche vs. ZeuS

We used the Avalanche and ZeuS botnets to study the effectiveness of our Guilt by Association approach. Figure 4 shows Top-10 country distribution for Avalanche bots from our Fast Flux Monitor database. The database contains 14,765 IPs from 74 countries that are associated with 1,951 Avalanche domains, of which 110 are currently fluxing and 585 with a history of fluxing. Six of the countries in the Top-10 are NATO members. Poland, Romania and Hungary lead the high number of bots relative to the total number of IPs issued.

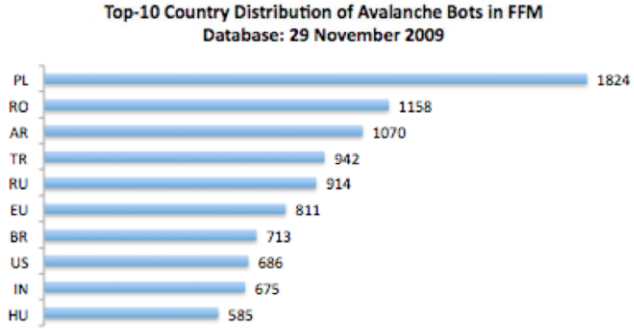


Figure 4: Avalanche Country Distribution

A significant trend in cyber crime in 2009 has been the emergence of the fast-flux Avalanche phishing kit or gang. [13],[14]. In the first half of 2009 estimates of its prevalence range from 24% to 28% of all phishing attacks, with even higher use estimated in Q3. It is further estimated that during this period Avalanche domains represented 43% (4,000) of all malicious domain name registrations, which were paid for with stolen credit cards. A distinctive characteristic of Avalanche domains was the disproportionately high use of .EU and .BE TLDs (top-level domains). Avalanche targets large financial institutions, the US Internal Revenue Service, job search providers, for the purpose of conducting financial fraud. In addition to social engineering or phishing attacks, many of the Avalanche domains were also used to launch drive-by downloads (HTTP attacks). The volume of take-down requests from anti-phishing providers overwhelmed the ability of many registrars to execute the domain take-down requests.

One of the uses of Avalanche is as a delivery mechanism for Zeus (AKA Zbot), a proven and resilient botnet and malware development framework that has been used to exfiltrate data from financial institutions, government agencies, ISPs (Internet Service Providers), and social networking sites. Zeus malware can be distributed through phishing and drive-by downloads. Its rise in activity in 2009 correlates with a rise in Avalanche attacks.

3.0 GUILT BY ASSOCIATION

3.1 Guilt Association Knowledge

Analytic sensors are derived from our cumulative collection of observed activities. We have developed a number of ‘Guilt by Association’ sensors. These analytical sensors include domains sharing a known guilty nameserver, domain names resolving to a known guilty IP, and nameserver domains resolving to a known guilty IP. In addition to helping real-time detection, such guilt by association relationships generate a rich social network view of the fast flux domain networks, which we used to analyze and cluster fast flux botnet domains.

Figure 5 shows the number of shared IP addresses across fast flux domains using a log-log scale. The plot shows a linear trend in this scale in that the number of IP addresses shared decreases with increasing number of fast flux domains. For instance, there are 100 botnet domains sharing 100 IP addresses whereas there are only 10 botnet domains sharing 1,000 IP addresses, resulting from having more small botnets than large ones.

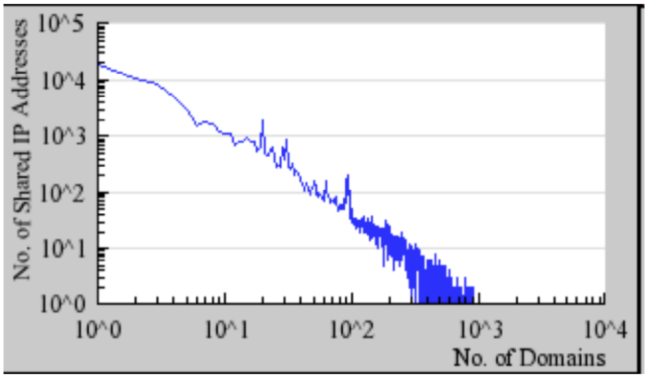


Figure 5: Distributions of IP Shared

Figure 6 shows the social network of a large phishing botnet where the red nodes represent domains and green nodes signify nameservers. The structural connectivity is important in predicting the footprint of a botnet. Phishing botnets like the one in Figure 5 typically employ large number of nameservers, thus reducing the expected footprint prediction coverage due to the low connectivity of the botnet graph.

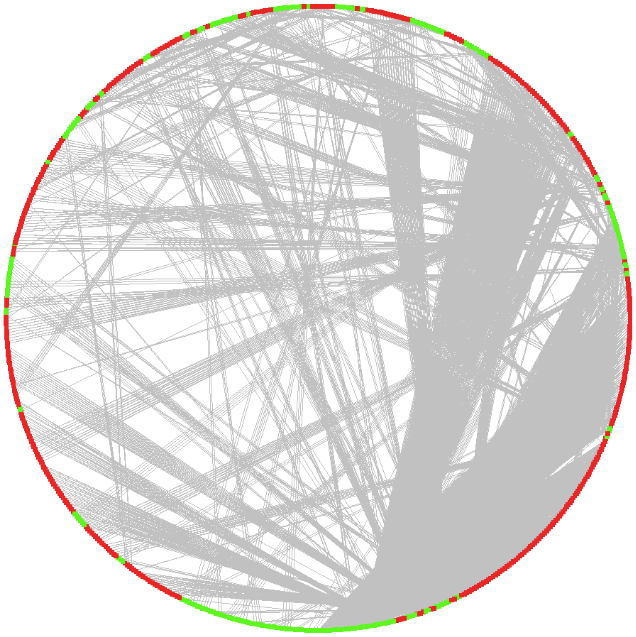


Figure 6: Structural Connectivity of a Phishing Botnet

In contrast, Figure 7 shows the social network graph of the Waledac botnet used for spam where the red nodes represent the domains, and green nodes signify the nameservers. As seen from the figure, Waledac network cluster has an overwhelming number of domains than nameservers, thus increasing the expected footprint prediction coverage of our Guilt by Association approach.

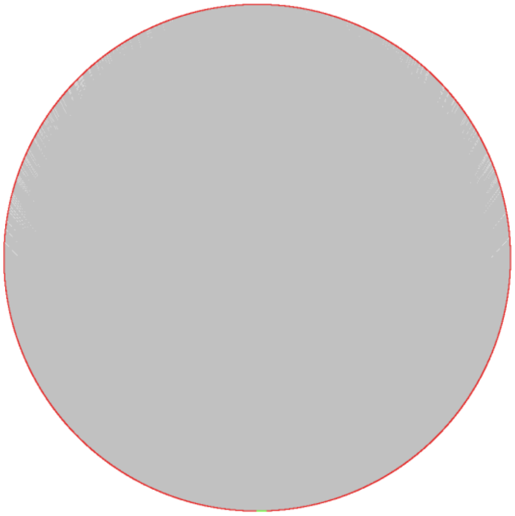


Figure 7: Structural Connectivity of a Spam Botnet

3.2 Guilt by Association Belief Function

We used an empirical belief function based classifier to classify a domain. Given a set of test domains for a botnet, we compute the belief function B for each domain in that set by:

$B = (E_f – E_a) / (E_f + E_a)$

where Ef is the evidence for the tested domain being in the botnet, and Ea is the evidence for the tested domain against being in the botnet. We compute the evidence using the following set:

I_b = {the set of all IPs that the botnet domains not in the test set are associated with}

For each domain’s associated IPs, we then compute the belief function using the number of IPs in the set Ib as evidence for and the number of IPs not in the set Ib evidence against. This empirical belief function returns a number between -1 and 1. We classify a domain as being predicted to be a part of the botnet footprint if its belief function is greater than -0.95.

4.0 RESULTS

4.1 Botnet Footprints

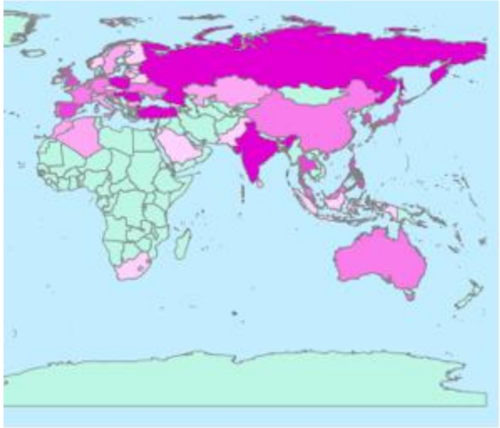


Figure 8: Avalanche Country Distribution

Figure 8 shows the Avalanche bots in our Fast Flux Monitor database for Europe, Asia, Australia and Africa. Countries in dark pink have higher concentrations of Avalanche botnets, while countries in lighter shades of pink have lower concentrations. Areas in green have no presence.

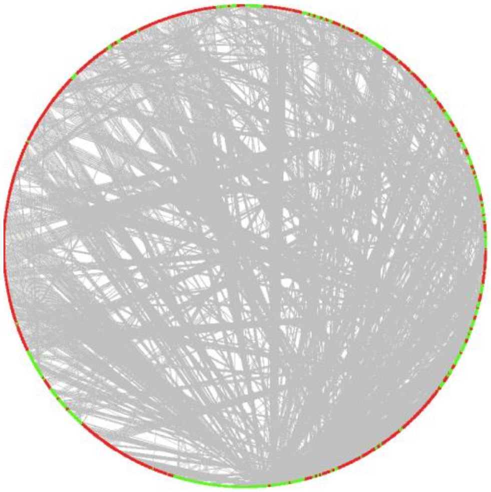


Figure 9: Avalanche Domain-Nameserver Graph

Figure 9 shows the domain (red) nameserver (green) graph. As expected, Avalanche has a typical connectivity for a phishing class of botnets. Figure 10 shows the same graph for the ZeuS botnet. In contrast to Avalanche, there are multiple clusters in the ZeuS domain nameserver graph. In fact, there over 200 more small clusters of the Zeus network not shown in Figure 10. Based on the visual analysis, we would expect our Guilt by Association approach do better for Avalanche as opposed to ZeuS in predicting the botnet footprint coverage, which is discussed next.

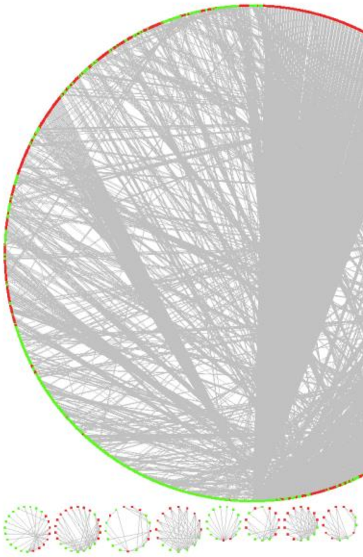


Figure 10: ZeuS Domain Nameserver Clusters

4.2 Footprint Predictions

Tables 3 and 4 show the footprint prediction coverage for the ZeuS and Avalanche botnets. In table, 3 the first row represents the prediction accuracy of footprint coverage for the case when 95% of the botnet is known and 5% is unknown. The third and fourth columns show the mean of correct predictions in %, and the standard deviation of correct predictions in %. Our Guilt by Association approach can predict the unknown 5% of the Zeus botnet with a 67% accuracy where the standard deviation of error is 6%. We obtained the empirical means and standard deviations by taking multiple 5% samples from the population, and computing the Guilt by Association classification. For the same conditions, Table 4 shows that our Guilt by Association approach can predict the unknown 5% of the Avalanche botnet with a 99% accuracy where the standard deviation of error is 1%.

As expected, when the known segment of a botnet is reduced, the footprint prediction accuracy gets degraded.

Table 3. ZeuS Footprint Prediction Performance

Known Botnet (%)	Unknown Botnet (%)	Mean of Correct Predictions (%)	Standard Deviation of Predictions (%)
95	5	67	6
50	50	57	1
25	75	43	1
10	90	17	1
5	95	0	0

Table 3 shows that the prediction accuracy is 57% when only half of the botnet is manually labeled. The prediction accuracy is respectable at 43% when only 25% of the botnet is manually labeled. However, for the ZeuS botnet, knowing only 5% of the botnet does not result in any number of correct predictions with 10% being the cut-off for meaningful footprint prediction performance.

Table 4. Avalanche Footprint Prediction Performance

Known Botnet (%)	Unknown Botnet (%)	Mean of Correct Predictions (%)	Standard Deviation of Predictions (%)
95	5	99	1
50	25	98	0
25	50	91	3
10	90	97	1
5	95	96	1

Table 4 shows the same analysis for the Avalanche botnet. In contrast to ZeuS, the prediction performance is better for the Avalanche botnet. For instance, when 95% of the botnet is known, Guilt by association can predict the remaining 5% of Avalanche with 99% accuracy while only with 67% accuracy for ZeuS.

Table 5. Conficker Footprint Prediction Performance

Known Botnet (%)	Unknown Botnet (%)	Mean of Correct Predictions (%)	Standard Deviation of Predictions (%)
95	5	99	1
50	25	97	1
25	50	97	1
25	75	97	1
10	90	95	1
5	95	91	2

Table 5 shows the same analysis for the Conficker botnet. In contrast to ZeuS, the prediction performance is better for the Conficker botnet. For instance, when 5% of the botnet is known, guilt by association can predict the remaining 95% of Conficker with 91% accuracy while only with 0% accuracy for ZeuS.

Table 6. Gumblar Footprint Prediction Performance

Known Botnet (%)	Unknown Botnet (%)	Mean of Correct Predictions (%)	Standard Deviation of Predictions (%)
95	5	100	0
50	25	98	3
25	50	98	1
25	75	97	1
10	90	97	1
5	95	96	0

Table 6 shows the same analysis for the Gumblar botnet. In contrast to ZeuS, the prediction performance is better for the Gumblar botnet. For instance, when 25% of the botnet is known, guilt by association can predict the remaining 75% of Gumblar with 97% accuracy while only with 43% accuracy for ZeuS.

Table 7. Pushdo Footprint Prediction Performance

Known Botnet (%)	Unknown Botnet (%)	Mean of Correct Predictions (%)	Standard Deviation of Predictions (%)
95	5	100	0
50	25	100	0
25	50	100	0
25	75	99	2
10	90	99	1
5	95	100	0

Table 7 shows the same analysis for the Pushdo botnet. In contrast to ZeuS, the prediction performance is much better for the Pushdo botnet. For instance, when 10% of the botnet is known, guilt by association can predict the remaining 90% of Pushdo with 99% accuracy while only with 17% accuracy for ZeuS.

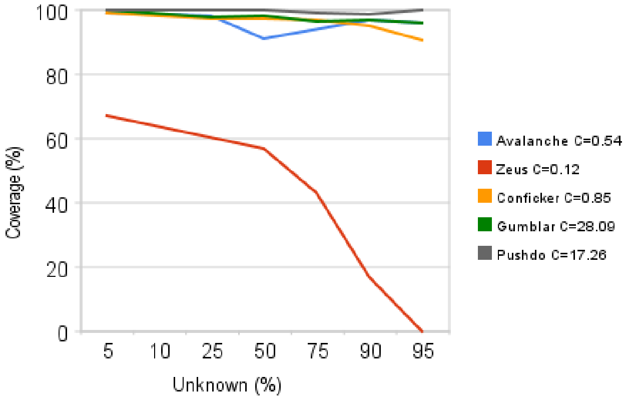


Figure 11. Botnet Footprint Prediction

Given a small subset of a botnet threat identified through reverse engineering, is it possible to predict the full footprint of this botnet? Based on Guilt by Association analysis of the network of botnet nameservers, domains, and IPs, we investigated the percent of the botnet footprint that can be derived by the labeled subset. Figure 11 summarizes the coverage vs. known label size for the Avalanche, Conficker, Gumblar, Pusdhdo, and ZeuS botnets. As seen from the figure, if 10% of the botnet is known, then 95% of the total footprint can be predicted for the Avalanche, Conficker, Gumblar and Pushdo botnets. In contrast, if 10% of the botnet is known, then only 20% of the total footprint can be predicted for the ZeuS botnet. The results are due to topological attributes of the two types of botnets, which is discussed next.

4.3 Structural Connectivity

In order to explain the differences between the prediction differences, we developed a computationally inexpensive connectivity metric for botnet structures. There are other connectivity metrics [17] but they are computationally more expensive, and the metric semantics does not apply to the problem that we have. One such metric is the graph clustering coefficient, which is the average of the densities of the neighborhoods of all of the nodes likes domains and IPs. In order to compute this density, we need to come up with a distance so that we can put a buffer around each node, which is a hard thing to do in a nameserver, domain and IP space.

For our analysis, we found the following Connectivity Metric to be sufficient. For a network of nodes with two types (e.g. {domains, nameservers}, or {domains, IPs}), we define:

Connectivity Metric (C) = 100 [c/(n*m)]

Where c is the number of connections, n is the number of nodes of the first type, and m is the number of nodes of the second type.

Table 8. Botnet Connectivity Metric Comparison

Botnet	Domain to NS Connectivity (C)	Domain to IP Connectivity (C)
Avalanche	1.1	0.5
Conficker	2.8	0.9
Gumblar	1.6	28.1
Pushdo	32.7	17.3
Waledac	1.3	0.6
ZeuS	0.2	0.1

Table 8 shows the Connectivity Metric for the Avalanche, Conficker, Gumblar, Pushdo, Waledac and ZeuS botnets in our database. The Connectivity Metric explains the predicted footprint coverage for Pushdo to be the best, and for ZeuS to be the worst. Our results summarized in Figure 11 validate this analysis. Figure 12 shows a sample of the small ZeuS domain nameserver clusters omitted in Figure 10. These clusters consist of domains and nameservers that do not exhibit fast flux behaviour. Whether the small clusters represent the discreet probe of networks by large criminal organizations, or small operator hosting set-ups that downloaded free phishing kits, the ZeuS botnet is stealthier than the others by relying on a large number of smaller clusters used for attack campaigns such as Kneber.

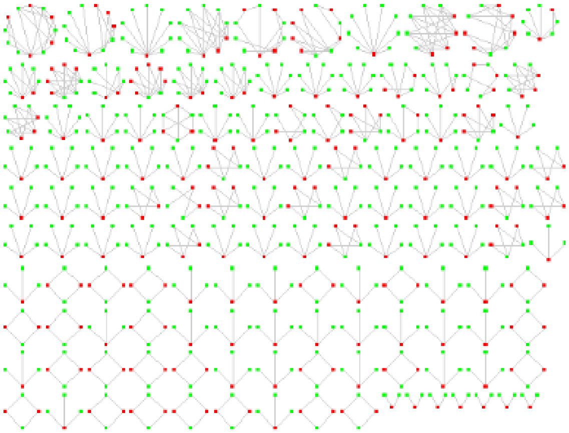


Figure 12. Small ZeuS Domain Nameserver Clusters

5.0 CONCLUSIONS

In this paper, we presented a Guilt by Association approach to predicting the footprint of a botnet given a sample using our botnet database collected over 12 months. Our results show that it is possible to effectively predict the footprint of a botnet starting from varying sample sizes of known bots. In particular:

- Guilt by association prediction coverage effectiveness depends on the network clusters and the connectivity in each cluster.
- The lower the number of clusters in the botnet graph, the higher the predicted footprint coverage.
- Guilt by association prediction coverage is larger for fast flux spam networks compared to fast flux phishing networks.
- Our proposed connectivity metric serves explains the effectiveness of Guilt-by-Association based footprint prediction.
- As expected, the footprint prediction coverage increases with increasing sample sizes of labeled bots.
- ZeuS network has the worst footprint prediction performance compared to Avalanche, Conficker, Gumblar, Pushdo, Waledac botnets.
- Small domain and nameserver, and domain and IP clusters that do not flux help the stealthy operation of botnets like ZeuS in attacks such as Kneber.

Automated prediction of a botnet's footprint based on the domain, nameserver, and IP connectivity using the manually created subset of botnet domains is essential in understanding the threat posed by botnets. Although our results focused on botnets used in commercial cyber crime, our results have military significance as similar botnets can be used for asymmetric military warfare. Our Guilt by Association approach enables analysts to form a total view of the threat by helping them to unify what seems like multiple threats into a single one, thus easing the mitigation effort.

As we have seen in the kinetic realm of asymmetric warfare, adversaries have demonstrated their ability to narrow our technology advantage by building powerful weapons from underground, open source components. This ability to narrow our weapons advantage is even greater in the cyber-realm, as cyber armies employ the same state-of-the-art tools and methods used by organized cyber criminals. In the 2009 cyber threat report, VeriSign's iDefense notes that the most advanced cyber criminal organizations are focusing on development of 'bulletproof hosting' infrastructure, of which fast-flux hosting is a principal component because it provides them with stealth, resilience, and 'infinite scalability' [15]. The GhostNet study of alleged Chinese espionage tools and methods against Tibet, details the use of small bots against high-value targets in embassies, news media, and NGOs, using primarily two attack vectors, spear-phishing and drive-by downloads, to deliver payloads for exfiltration [16]. Regardless of whether the adversaries are profit-motivated criminals, cyber terrorists, patriotic hackers, or military sponsored, NATO cyber defenders need tools and methods for detecting and monitoring botnet infrastructure.

6.0 ACKNOWLEDGEMENTS

This research was supported by Department of Homeland Security, Science and Technology Directorate Cybersecurity R&D program.

7.0 REFERENCES

[1] Caglayan, A., Toothaker, M., Drapeau, D., Burke, D., and Eaton, G., “Real-time Detection and Classification of Fast Flux Service Networks”, Cybersecurity Applications and Technology Conference for Homeland Security (CATCH), March 3 - 4, 2009, Washington, DC.

[2] Caglayan, A., Toothaker, M., Drapeau, D., Burke, D., and Eaton, G., “Behavioral Analysis of Fast Flux Service Networks”, Cyber Security and Information Intelligence Research Workshop (CSIIRW-09), April 13 - 15, 2009, Oak Ridge, TN.

[3] Caglayan, A., Toothaker, M., Drapeau, D., Burke, D., and Eaton, G., “Behavioral Patterns of Fast Flux Service Networks”, Cyber Security and Information Intelligence Track, Hawaii International Conference on System Sciences (HICSS-43), January 5-10, 2010, Kauai, HI.

[4] C. Kanich, C. Kreibich, K. Levchenko, B. Enright, G. Voelker, V. Paxson, and S. Savage. “Spamalytics: An Empirical Analysis of Spam Marketing Conversion”, 15th ACM Conference on Computer and Communications Security (CCS), 27-31 October 2008, Alexandria, VA.

[5] Miwa, S., Miyachi, T., Eto, M., Yoshizumi, M, and Shinoda, Y, “Design and Implementation of an Isolated Sandbox with Mimetic Internet Used to Analyze Malwares”, DETER Community Workshop on Cyber Security Experimentation and Test 2007, August 5-8, 2007, Boston, MA.

[6] ICANN. GNSO Issues Report on Fast Flux Hosting, March 2008.

[7] ICANN Security and Stability Advisory Committee. SAC 025: SSA Advisory on Fast Flux Hosting and DNS, March 2008.

[8] Holz, T. Gorecki, C. Rieck, C. Freiling, F. “Measuring and Detecting Fast-Flux Service Networks.” Presented at NDSS Symposium (2008).

[9] Passerini, E. Paleari, R. Martignoni, L. Bruschi, D. “FluXOR: detecting and monitoring fast-flux service networks.” Detection of Intrusions and Malware, and Vulnerability Assessment (2008), pp. 186-206.

[10] Zhou, C. V., Leckie, C. and Karunasekera, S., “Collaborative Detection of Fast Flux Phishing Domains”, Journal of Networks, Vol. 4, No. 1, February 2009.

[11] McGrath, D. K., Kalafut, A., Gupta, M., “Phishing Infrastructure Fluxes All the Way”, IEEE Security and Privacy Magazine Special Issue on Securing the Domain Name System, September/October 2009.

[12] Konte, M., Feamster, N. , and Jung, J., “Dynamics of Online Scam Hosting Infrastructure”, Proceedings of Passive and Active Measurement Conference (PAM), Seoul, Korea, April 2009.

[13] ICANN. ICANN Situation Awareness Note 2009-10-06.

[14] Anti-Phishing Working Group (APWG). An APWG Industry Advisory – Global Phishing Survey: Trends and Domain Name Use in 1H2009, October 2009.

[15] iDefense. An iDefense Topical Research Report: 2009 Cyber Threats and Trends. Dec. 12, 2008.

[16] Deibert, R., Manchanda, A., Rohozinski, R., Villeneuve, N., Walton, G. "Tracking GhostNet: Investigating a Cyber Espionage Network," March 2009.

[17] Wasserman, S. and Faust, K., 1994. Social Network Analysis: Methods and Applications. Cambridge: Cambridge University Press.

[18] Cox, A., and Golomb, G., "The Kneber Botnet", NetWitness Corporation, Herndon, VA, Feb. 17, 2010.