Cross-Analysis of Botnet Victims: New Insights and Implications

Seungwon Shin, Raymond Lin, and Guofei Gu

SUCCESS Lab, Texas A&M University,
College Station, Texas, USA
{swshin,rlin,guofei}@cse.tamu.edu

Abstract. In this paper, we analyze a large amount of infection data for three major botnets: Conficker, MegaD, and Srizbi. These botnets represent two distinct types of botnets in terms of the methods they use to recruit new victims. We propose the use of cross-analysis between these different types of botnets as well as between botnets of the same type in order to gain insights into the nature of their infection. In this analysis, we examine commonly-infected networks which appear to be extremely prone to malware infection. We provide an in-depth passive and active measurement study to have a fine-grained view of the similarities and differences for the two infection types. Based on our cross-analysis results, we further derive new implications and insights for defense. For example, we empirically show the promising power of cross-prediction of new unknown botnet victim networks using historic infection data of some known botnet that uses the same infection type with more than 80% accuracy.

1 Introduction

Recent botnets use several methods to find and infect victims. Among these methods, most botnets have mainly employed two infection techniques [9] [7] [6]:

- Bots automatically propagate themselves (auto-self-propagating, *Type I*). To do this, bots usually employ network scanning techniques to find vulnerable hosts and exploit them. This approach is active and aggressive in infecting victims. Conficker [3] is a good example of this kind of botnets [9].
- Bots spread themselves with the help of people or other methods (non-auto-self-propagating, *Type II*). In this case, since bots cannot find new victims automatically, malware writers should employ other techniques. They install a malicious binary into compromised web sites and trick people into downloading it (i.e., drive-by-download [12]) or they ask other malware owners, who have pre-installed malware, to distribute their malware (i.e., pay-per-installation (PPI) [17] [28]). This approach seems to be relatively passive because the operation sequence of this approach may depend on human actions or other tools. The MegaD [7] and Srizbi [5] botnets, which are known

as spam botnets, are representative examples of this type of botnet [7] [17] [6].

Both auto-self-propagating and non-auto-self-propagating botnets have become serious threats to the Internet. For example, some of them have infected millions of victims [4] and some are infamous for generating a significant amount of spam emails [8]. Analyzing and understanding them is thus becoming an important and urgent research task in order to design more effective and efficient defenses against them.

In this paper, we start our research with a simple yet important question: are there any similarities/differences in infection patterns (e.g., the distribution of victims) between these two types of botnets? We believe the answer to this question can greatly deepen our understanding of the nature of these botnets and enable us to develop more accurate/targeted Internet malware monitoring, detection, prediction techniques, strategies and systems. Since both types of botnets have quite different infection approaches, i.e., auto- and non-auto-self-propagating, we could predict that their infection patterns are likely also different. To understand whether this hypothesis is right or wrong, one needs to collect and cross-analyze both types of botnets. However, although there are several previous measurement/analysis studies that have made significant efforts to understand botnet infection characteristics [16] [13] [14] [6], they mainly focus on only one specific botnet, rarely providing cross-analysis of different (types of) botnets. This is probably due to many reasons, for example practical difficulties on data collection: (a) collecting a good amount of real-world botnet data is hard; (b) collecting multiple different (types of) real-world botnet data is even harder.

In this work, we have collected a large amount of real-world botnet infection data, including millions of Conficker victims and several hundred thousands of MegaD and Srizbi victims. They cover the two representative infection techniques mentioned before with reasonably large amount of samples and thus are suitable for our study. We perform an in-depth cross-analysis of different botnet types and show what similarities/differences exist between them. Slightly contradictory to the hypothesis we made above, we find that both types of botnets have a large portion of victims overlapped and the overall victim distributions in IPv4 space are quite similar. However, they do show several interesting characteristics different from each other. To obtain a fine-grained understanding of these similarities and differences, we further perform an in-depth set of large-scale passive and active measurement studies from several perspectives, such as IP geographical location, IP address population/density, networks openness (remote accessibility), and IP address dynamism. Our results reveal many interesting characteristics that could help explain the similarities/differences between the two botnet infection types.

Furthermore, from our measurement results, we have further derived new implications and insights for defense. We found that due to the heavily uneven distribution of botnet victims, we can observe strong neighborhood correlation in victims. Although it is intuitive that *Type I* malware (specifically scanning

malware) tends to infect neighbor networks and thus neighborhood watch could be a useful prediction technique [2], it is unknown whether this applies to the case of *Type II* malware. For the first time in the literature we show with empirical evidence that *Type II* botnet victims also exhibit this similar property. More interestingly, we have empirically discovered that even if we only know some information of one botnet (e.g., past botnet data), we could predict unknown victims of another botnet (e.g., a future emerging botnet) with reasonably high accuracy, given that both botnets use the same infection type. This sheds light on the promising power of cross-analysis and cross-prediction.

In short, the contributions of this paper are as follows.

- We collect a large amount of real-world botnet data and provide the first cross-analysis study between two types of botnet infections to the best of our knowledge. This kind of study is useful to understand the nature of malware infection and help us gain insights for more effective and efficient defense.
- We perform a large-scale passive and active measurement study for a fine-grained analysis of similarities/differences in two botnet infection types. We study several aspects such as IP geolocation, IP address population/density, IP address dynamism, and network openness (remote accessibility). We have many interesting findings. To name a few (incomplete) examples, (a) different countries are likely prone to different types of malware infections while some countries such as Turkey are extremely vulnerable to both infection types; (b) malware infection seems to have very interesting correlation with geopolitical locations; (c) IP address dynamism and network openness are likely to cause more malware infections (for certain type). And they have different effect on different types of botnet infections.
- Based on our cross-analysis result, we further derive new implications and insights for defense. We perform an empirical test to predict unknown victim networks of non-auto-self-propagating botnets by looking up their neighbor information. We further extend it to cross-predict unknown victim networks of a new botnet using existing knowledge of botnets with the same infection type and we show that the prediction accuracy can be reasonably high (more than 80%).

2 Data Collection and Term Definition

In this section, we provide information of data that we have analyzed and we define several terms used in this work.

Data Collection To understand the characteristics of different types of botnets, we have collected data for three major botnets: Conficker, MegaD, and Srizbi. Conficker [3] is a recent popular botnet known to have infected several million Internet machines. It propagates automatically through network scanning. It first scans random networks to find new victims and if it infects a host successfully, it scans neighbor networks of the host to find victims nearby [9]. Thus it is a

representative example of *Type I botnets*. The MegaD [7] and Srizbi [6] botnets are two recent botnets known for sending large volume of spam since 2008. In particular, it is mentioned that MegaD was responsible for sending about 32% of spam worldwide [7] and Srizbi was responsible for sending more than half of all the spam in 2008 [1]. They are representative examples of *Type II botnets* because they spread by drive-by-download [7, 6] or pay-per-install methods [17].

The Conficker botnet data has been collected by setting up sinkholing servers because Conficker uses domain-fluxing to generate C&C domain names for victims to contact [3]. With the help of *shadowserver.org*, we have collected a large dataset of Conficker infection including about 25 million victims [2]. The *shadowserver.org* has set up several sinkhole servers and registered the domain names same as the Conficker master servers to redirect queries of the Conficker bots to the sinkhole servers. Then, the sinkhole servers capture the information of hosts contacting them and the hosts can be considered as the Conficker infected victims.

Botnet	*Data Source*	*Main Infection Vector*	*# of Victims*	*Collection Date*
Conficker	Sinkhole server [20]	network scanning	24,912,492	Jan. 2010
MegaD	Spam trap [19]	drive-by-download or PPI	83,316	Aug. 2010
Srizbi	Spam trap [19]	drive-by-download	106,446	Aug. 2010

Table 1. Data summary of collected botnets.

The MegaD and Srizbi botnet data has been collected through the *botlab project* [19], of which spam trap servers were used to gather information of hosts sending spam emails. The detailed summary information regarding our collected data is presented in Table 1. The *botlab project* captures spam emails from spam-trap servers and further investigates the spam emails through various methods such as crawling URLs in the spam emails and DNS monitoring. From correlating the investigation results, the *botlab project* finally reports which hosts are considered as infected by spam-botnets such as MegaD and Srizbi.

Term Definition Before we perform cross-analysis on the data, there are several important issues to be addressed which can bias our result. The first thing is the *dynamism* of the IP address of a host. Many ISPs use dynamic IP address re-assignment to manage their assigned IP addresses efficiently [10]. This makes it hard to identify each host correctly. This may cause some biases in measuring the population or characteristics of the botnet [11]. Second, we are not likely to collect the *complete* data of certain botnets but only parts of the data (e.g., MegaD and Srizbi), and this can also cause some biases.

To account for these issues, instead of basing our analysis unit granularity on the individual IP address level, we generalize our analysis to examine at the network/subnet level by grouping adjacent IP addresses. This will help mitigate

the effect of *dynamism*, because it is common that dynamic IP addresses of a host come from the same address pool (subnet). Also, we believe that it is sufficient to examine subnets because even if only one host in the network is infected, the neighbor hosts are likely to be vulnerable or be infected soon [2].

In this work, we define our base unit for analyzing, i.e., *"infected network"*, as the /24 subnet which has at least one malware infected host. Thus, if a subnetwork is infected by a *Type I botnet*, we call the subnet a *Type I infected network* and a similar definition is also applicable to *Type II infected networks*. In addition, we define a *Common infected network* as an *infected network* which has victims of both types of botnets. There may be some *infected networks* that are exclusively infected by either *Type I* or *Type II*, which are defined as *Type I EX* or *Type II EX infected networks*, respectively.

In our data set, we found 1,339,699 *infected networks* in the case of the Conficker botnet, 71,896 for the MegaD botnet, and 77,934 for the Srizbi botnet. Thus, we have data for around 1,339,699 *infected networks* for the *Type I botnet* and 137,902 *infected networks* for the *Type II botnet*[1]. From this we have identified 97,290 *Common infected networks*.

3 Cross-Analysis of Botnet Victims

In this section, we provide detailed cross-analysis results of two types of botnets.

3.1 Point of Departure

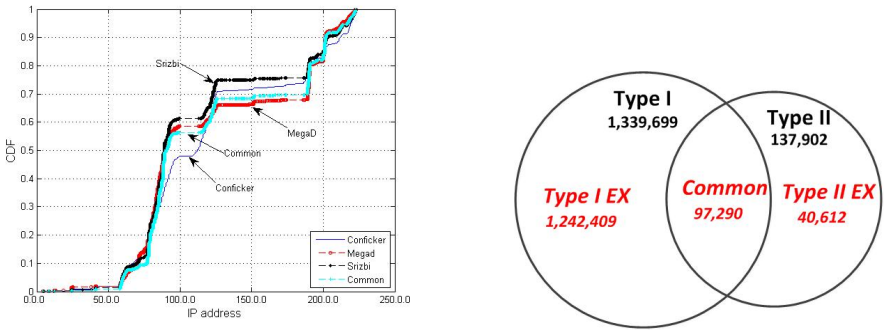
We start our analysis with the following *Hypothesis 1* that we proposed in Section 2.

Hypothesis 1. *Since the two types of botnets have very different infection vectors, they may exhibit different infection patterns (e.g., distributions of their infected networks).*

To verify this hypothesis, we measure how many *infected networks* are shared by both types of botnets and how they are different from each other. The basic measurement results are shown in Figure 1. Figure 1(a) shows the distribution for *infected networks* of each type of botnet over the IP address spaces (*Type I (Conficker), II (MegaD and Srizbi), and Common infected networks*). Interestingly, the distributions of *Type I* and *Type II botnets* are very similar to each other. Specifically, the IP address ranges of (77.* - 96.*), (109.* - 125.*), and (186.* - 222.*) are highly infected by both types of botnets and their shared regions (*Common*) are also distributed in the similar ranges.

To investigate how many *infected networks* are *"really"* shared between them, we draw a diagram which represents the number of *infected networks* of each type of botnet and networks that they share in common in Figure 1(b). There are

[1] There are 11,928 *infected networks* in common between MegaD and Srizbi.



(a) Infected network distributions over IP address spaces (b) Infected network diagram

Fig. 1. Infected network distributions and diagram.

97,290 *Common infected networks*, 1,242,409 *Type I EX networks*, and 40,612 *Type II EX networks*.

Contrary to our expectation, the two types of botnets are distributed over similar IP address ranges and there are many *Common infected networks* between them. However, this observation is only about the distribution over the IP address space and it is very hard to find semantic meanings such as their physical locations from this result. For instance, even though we know a /24 subnet *111.111.111/24* is an infected network, we may not understand who are using the subnet and where the subnet is located. More importantly, why is the subnet more likely to be infected by certain type (or both types) of botnets? In addition, the ranges are too broad to comprehend clearly. We show range (77.* - 96.*) is highly infected, but that does not mean that all IP addresses in the range are infected, we need more fine-grained investigation. Besides that, we also find that there are some differences between them (i.e., *Type I EX and II EX infected networks* are still significant) and they also need to be understood, because they can show which ranges are more vulnerable to which type of botnet. Only considering IP address ranges might not clearly show these differences.

Thus, we are motivated to consider more viewpoints that provide us some understandable meanings with fine-grained level semantic information. We have selected four interesting viewpoints (we call them *categories*): (i) geographical distribution of infected networks, which lets us identify more (or less) vulnerable locations and their correlation with certain types of infections, (ii) IP address population/density, which helps us understand relationships between the number of assigned IP address to the country and the number of infected networks of the country, (iii) remote accessibility of networks, which shows us how open (and thus possibly prone to infection) the networks are and whether there is a correlation with certain infection types, and (iv) dynamism of IP addresses, which tells us whether vulnerable networks use more dynamic IP addresses and the correlation with infection type. In each category, we build a hypothesis based on some intuition and then we perform a large scale passive or active measure-

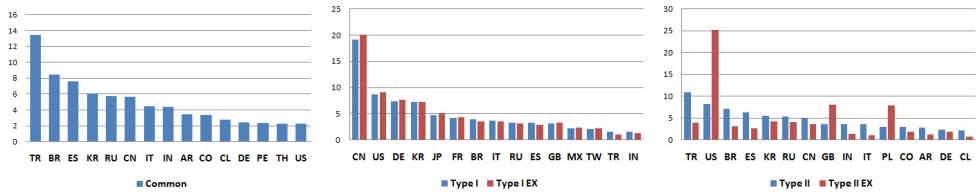
ment to verify the hypothesis and gain some insights.

Insight 1. *Interestingly, the two types of botnets are distributed in similar IP address ranges despite of their different infection types. In addition, the ranges are continuous and it might imply that vulnerable networks are close to each other. More fine-grained analysis over the ranges might help us find new results and insights.*

3.2 Geographical Distribution of Infected Networks

In our first test, we have observed that two types of botnets seem to have similar distributions over the IP address space. Thus, we could infer that the distributions of two different types of botnets over geographical locations are similar to each other. From this intuition, we make the following hypothesis.

Hypothesis 2. *Type I and Type II infected networks are mainly distributed over similar countries.*



(a) Common (b) Type I and I EX (c) Type II and II EX

Fig. 2. Infected network distributions over the countries (x-axis for country code, y-axis for percentage)

To verify this hypothesis, we investigate how each type of infected network is distributed over countries. When we observe the overall distribution of each type of botnet over the countries, we find that all *Common, Type I, Type I EX, Type II, and Type II EX infected networks* spread all over the world (with the exception of Africa), but there are some concentrated areas. To analyze the result in detail, we select the top 16 countries of each case and show their distributions in Figure 2. Results are sorted by the number of *infected networks* of the countries. Here, X-axis represents the country code and Y-axis represents the percentage of each infection type, e.g., if there are 100 *Common infected networks* overall and 14 *infected networks* are located in Turkey (its country code is TR[2]), the percentage of Turkey is 14%.

[2] Each country code represents followings; AR Argentina, AU Australia, BR Brazil, CA Canada, CL Chile, CN China, CO Colombia, DE Germany, ES Spain, FR France, GB Great Britain, IN India, IT Italy, JP Japan, KR South Korea, MX Mexico,

In Figure 2(a), *Common infected networks* are mainly distributed in Asia (e.g., Turkey, Korea, Russia, China, and India) with more than 35%. Figure 2(b) also presents that *Type I and I EX infected networks* are mainly distributed over Asia. The distributions of *Type I EX infected networks* are quite similar to that of *Type I*. The distributions of *Type II and II EX infected networks* are shown in Figure 2(c). Here we still observe more than 30% as being located in Asia.

From the observations, we find two interesting things. First, the set of countries that are highly infected are not very different for each type of botnet (i.e., if some countries are highly infected by *Type I botnet*, they are also likely to be infected by *Type II botnets*). This implies that these countries are more prone to be infected regardless of infection methods. Second, there are some countries that are highly vulnerable to one type of botnet over the other. China is a good example of this. China has a lot of *Type I infected networks*. However, it has relatively small portions of *Type II infected networks*. We presume that most of the networks in China are accessible from remote scanning botnets because *Type I botnets* usually use network scanning techniques to find new victims. We will test this in section 3.4 and show whether our presumption is correct.

Insight 2. *There are some countries which are prone to be infected by both types of botnets. However, some other countries are more likely to be infected by one type of botnet. Management policies of networks (e.g., network access control) could affect malware infection of the country.*

3.3 IP Address Population

From the previous result, we know that the *infected networks* of each type of botnet are concentrated mainly within several countries but the infection rates between them are different. Why is the infection rate between them different? Are there any possible answers or clues that might explain this? To find out some clues, we first focus on the number of IP addresses assigned to each country.

IP addresses are not assigned evenly over networks or locations [22] [21]. In terms of the IPv4 address space, there are some IP address ranges which have not been assigned to users but registered only for other purposes, e.g., (224.* - 239.*) is assigned for multicast addresses [22]. In addition, IP addresses have been assigned differently over locations, e.g., more than 37% of IP addresses are assigned to the United States, while Turkey only has less than 0.5% [21]. From this fact, we can easily infer that countries that have more IP addresses could have more chances to be infected by malware leading to *Hypothesis 3*. Here, we will use the term of *IP address population* to represent the number of assigned IP addresses and we define *high IP address population country* as the country ranked in the top 30 in terms of the number of assigned IP addresses, and *low*

NL Netherlands, PE Peru, PL Poland, RO Romania, RU Russian Federation, SE Sweden, TH Thailand, TR Turkey, TW Taiwan, US United States, VN Vietnam

IP address population country as the country ranked below 30. All ranking information is based on [21].

Hypothesis 3. *Countries with more IP addresses (high IP address population countries) might contain more of both types of infected networks than low IP address population countries.*

To verify this hypothesis, we compare the number of *infected networks* of each type of botnet with the number of IP addresses assigned to each country. The comparison results are shown in Figure 3. We can see that the number of *infected networks* of the *Type I, II, I EX, II EX botnets* are relatively proportional to the *IP address population* (i.e., the more IP addresses a country has, the more *infected networks* it contains). However, in the case of *Common infected networks*, they are *NOT* proportional to *IP address population*. On the contrary, they are mainly distributed over some *low IP address population countries*.

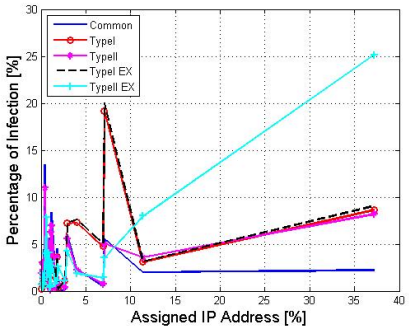


Fig. 3. Infected network distribution versus IP address population (x-axis for percentage of assigned IP addresses to a country, y-axis for percentage of infection of each type of botnet in the country)

Intuitively, countries with more IP addresses have more chances to be infected. Thus, we can easily accept the results of *Type I, II, I EX, II EX*. However, why do some *high IP address population countries* have less *Common infected networks* while some *low IP address population countries* have more? There may be several possible reasons for this. For example, the security education/knowledge of people may play a role. People may open some vulnerable services or click suspicious URLs without serious consideration, if they do not have enough education/knowledge of security in some countries. Another possible reason is in regards to network management. If networks in a country are well managed and protected very carefully, it is harder for malware to find chances to infect the networks. Thus, malware infection rate would not be proportional to the number of IP addresses in the country.

The other interesting point is the *percentage of infected networks over all networks of the country* (e.g., if a country has 100 networks and if 10 networks among them are infected, the percentage of *infected networks* of the country is 10%). We have observed that *high IP address population countries* are likely to have more infected networks. However, it does not mean that most (or a high percentage) of networks in the country are infected. For example, even though the United States has more number of *Type II infected networks* than other countries (except Turkey), the *infected networks* may only cover small percentage of all networks in the United States, because the country has around 38% of IP addresses of the world. This can reveal some *low IP address population countries* whose networks are more vulnerable (in terms of percentage) than other countries and they could be ignored if only considering the absolute number of *infected networks*.

To investigate the percentage of *infected networks* of each country, we have used the data from the *IP2Location.com* report [21]. In the report, we find that 2,505,141,392 IP addresses have been observed in the world. This may not cover all observable IP addresses in the world. However we believe that it is close to the real value. Their report also shows the percentage of IP addresses that each country has out of all observed IP addresses.

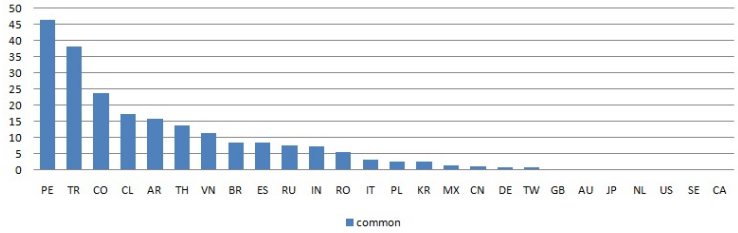
We use this data to calculate the number of IP addresses assigned to each country. Then, we calculate the number of /24 sub-networks of each country by dividing the number of IP addresses assigned to the country by 256. At this time, we make an assumption that *"IP addresses are assigned to each country with the minimum unit size of /24 subnet"* to make our calculation easy. And we calculate the ratio of *infected networks* in each country with it and the number of infected /24 subnets. This scenario can be formalized as follows.

- Θ = the number of all IP addresses in the world (i.e., 2,505,141,392)
- ϵ_j = the percentage of assigned IP addresses to the country j
- α_j = the number of /24 subnets in country j
- γ_i = the number of *infected networks* of type i botnet (e.g., γ_1 represent the number of *infected networks* of *Type I botnet*)
- η_i = the percentage of *infected networks* of type i botnet in each country

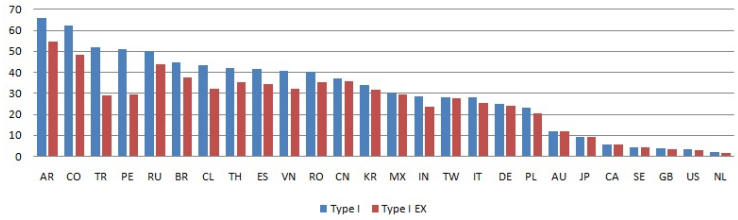
Our goal is to calculate the value of η of each country, and this can be obtained by the following formula (here $j \in \{1, 2, ..., 240\}$, and 240 denotes the number of countries which have observable IP addresses).

- $\alpha_j = \frac{\Theta}{256} * \epsilon_j$
- $\eta_i = \frac{\gamma_i}{\alpha_j} * 100$, where $i \in \{1, 2\}$

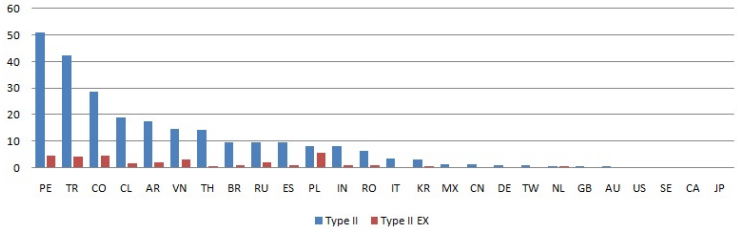
The distribution of the values of η over some selected countries are shown in Figure 4. This result is quite different from the previous result (in Figure 2). In the case of *Common* (Figure 4(a)), some top ranked countries in Figure 2 show quite low η values. Russia, Korea, China, and the United States are examples of this case, however Turkey still represents high η value. From the results, we can understand which countries are more vulnerable (i.e., high η value). Peru is an



(a) Common



(b) Type I and Type I EX



(c) Type II and Type II EX

Fig. 4. η values of selected countries (x-axis for country code, y-axis for η value)

interesting case. It has not been known as a country containing large number of *infected networks* in our previous results. However large portions of its networks in the country seem to be infected. *Type I, I EX, II, and II EX* also show similar characteristics to the *Common* case and the results are shown in Figure 4 (b) and (c). Based on these results, we may focus on some vulnerable countries (e.g., Turkey and Peru) to study infection trends of botnets. They may be good candidates for monitoring in order to comprehend the infection trends of botnets.

We try to reveal the reason why Turkey and Peru show high η values. From our investigation, we find a possible reason. It can be caused by *geopolitical reasons*. Some previous work pointed out that Turkey has been suffered from large cyber attacks generated by its neighbor countries such as Russia [24]. This explanation is also applicable to Peru, because it is surrounded by several countries that have a lot of malware infected networks such as Brazil and Mexico.

Insight 3. *To understand malware distributions, we might put our focus on not only high IP address population countries with large number of infected networks, but also some low IP address population countries where large portions of their networks seem to be infected. Malware infection of these low IP address population countries could be affected by geographical neighbors.*

3.4 Remote Accessibility

Another category that we consider is the network openness or remote accessibility (i.e., whether a host can be directly accessed from remote hosts or not). As we described in the previous section, one major scheme of finding new victims of the *Type I botnet* is scanning remote hosts (or networks). Enterprise networks are usually protected by several perimeter defending systems such as firewalls, in an attempt to block malicious threats from remote hosts. However, *not* all networks are protected as such and if they are not protected, malware can infect internal unguarded hosts more easily. From this intuition, we build the following hypothesis.

Hypothesis 4. *Networks that are more open (more directly accessible from remote hosts) might have more infected networks of Type I botnets than that of Type II botnets.*

We have tested the network accessibility by sending several *Ping* packets (i.e., five ICMP echo request packets per host in our test) to several randomly selected hosts in a network. If any of our *Ping* queries is successful in selected hosts, we regard that the network is reachable from remote hosts, otherwise we regard that the network is unreachable. This test has been already used before to understand the network reachability by previous work [23]. Note that this test may only show the *lower bound* of reachable networks, because some perimeter defending systems (e.g., firewalls) block incoming ICMP packets, or our randomly selected hosts may be not alive during testing. In this test, we assume that each /24 subnet have the same network access control policy (i.e., if one of the host in the same /24 subnet is accessible from the remote host, we consider that all hosts in the same /24 subnet might also be accessible).

In our test, we can access 54.32% of *Type I infected networks*, which is more than half. This indeed shows that *Type I infected networks* are more open (remote accessible). It confirms our hypothesis, although we presume this ratio could be higher for *Type I*. This could be probably explained by (a) our network reachability test is only a low-bound estimation, and (b) more networks are aware of malware scanning attacks and thus more (previously open) networks installed firewalls. In the case of the result for *Type II*, it shows 46.85% networks are accessible, which is much less than *Type I*. This is probably because the infection vectors of *Type II botnets* do not depend on remote accessibility.

The result of *Common* is interesting, because it shows more than 60% of networks are accessible. This implies that remote accessible networks are much more vulnerable to malware attacks. It might be reasonable, because even though

network accessibility may not help *Type II botnets* infect hosts, at least it helps *Type I botnets*.

In addition, we measure the remote accessibility of networks of three countries: Turkey, China and the United States. These countries show somewhat interesting patterns (e.g., China has a lot of *Type I infected networks*, but has relatively small number of *Type II infected networks*). In our measurement, we find that 64.09% of networks in China are accessible from remote hosts. This corresponds with our previous prediction (i.e., networks in a country that has a lot of *Type I infected networks* might be more accessible from remote hosts) in section 3.2. We discover that 51.8% of networks are accessible in the case of Turkey and 40.92% of the United States. This result seems to be reasonable, because these countries are more vulnerable to *Type II* than *Type I botnets.*

Insight 4. *Open (remote accessible) networks are more likely to be infected, particularly by Type I infection. However, it does not mean that inaccessible networks are much more secure, because malware (Type II infection) can still infect hosts in protected networks by several smart attack methods such as social engineering.*

3.5 Dynamism of IP Address

Previous work has shown that a lot of bots used dynamic IP addresses [10]. We want to investigate whether the networks with more dynamic IP addresses are more vulnerable than those with static IP addresses for both types of botnet infections.

Hypothesis 5. *Places (or networks) with more dynamic IP addresses are more prone to be infected by both types of botnets.*

To understand this, we have analyzed how many infected networks are using dynamic IP addresses. For the analysis, we apply the technique of finding dynamic IP addresses proposed by Cai et al. [23]. In their analysis, they used reverse DNS PTR records of each host. They believed that the reverse PTR record can represent the status of a host and if some keywords of a reverse PTR record represent dynamism of IP address, the host is likely to use dynamic IP address. For instance, if a reverse PTR record of a host A is *dynamic-host.abcd.com*, it is very likely for the host A to use dynamic IP address, because its reverse PTR record has a keyword of *dynamic-host*. Note that this test only shows the lower bound of dynamic networks due to the limitation of reverse DNS lookup and selected keywords. Even though this test can not show all networks using dynamic IP addresses, it could give us information of which type of botnet has more dynamic IP addresses. Based on this idea, we use the same keywords mentioned in [23] to find hosts (and finally networks) which are likely to use dynamic IP addresses. If we find any host in a subnet using keywords representing the dynamism, we simply consider that the subnet uses dynamic IP addresses.

Type	*Dynamic IP*	*Static IP*
Common	62%	38%
Type I	50.1%	49.9%
Type II	58.4%	41.6%
Type I EX	49.08%	50.92%
Type II EX	51.87%	48.13%

Table 2. Comparison of the percentage of dynamic or static IP addresses of each type.

We have measured how many *infected networks* use dynamic IP addresses and the results are summarized in Table 2. The results are quite interesting. In the case of *Type I, I EX, and II EX* we find that around 50% of *infected networks* use dynamic and other 50% of *infected networks* use static IP addresses. However, in the case of *Common and Type II*, *infected networks* use more dynamic IP addresses than static IP addresses.

The result of *Common* matches the previous result [10] which mentioned dynamic IP addresses are more vulnerable. However, the result of *Type I* does not fully match the previous result, i.e., *Type I botnet* infection does not have noticeable preference on networks with more dynamic addresses. This is actually reasonable because *Type I botnets* locate a remote victim by scanning the IP address space regardless whether the target address is dynamic or static. In the case of *Type II botnet infection*, we do observe infection preference on networks with more dynamic addresses. This is also reasonable because there are probably more home users in these (dynamic) address space who have less security awareness and potentially more vulnerable computers and web browsing patterns.

Insight 5. *Networks with more dynamic IP addresses are more vulnerable to malware attacks. This is more noticeable in the case of Type II botnet infection than Type I.*

4 Neighborhood Correlation of Botnet Victims

In this section, we provide a prediction approach based on our insights obtained in the previous section.

4.1 Watch Your Neighbors

Insight 1 in Section 3.1 points out that both types of botnets have heavily uneven distributions of infected networks and there are several heavily (continuous) infected areas in some part of the IPv4 space. This implies that *infected networks* of both types of botnets might be close to each other, i.e., it is very likely for them to be located in the same or similar physical locations and neighbor networks (e.g., belonging to the same /16 networks). This intuition has already been discussed before and verified in some previous work for some *Type I botnet*

[9] [13] [2]. An interesting thing is that one of the previous work provides an approach of predicting unknown victims based on the intuition and it predicts unknown victims with more than 90% accuracy with only employing a simple method (e.g., K-Nearest Neighbor classification) [2]. However, this work has only focused on the case of *Type I botnets*.

The reason for strong neighborhood (network) correlation of *Type I botnets* is intuitive, because *Type I botnets* will very likely scan neighbor networks to recruit new victims. Then, can we apply a similar prediction approach to *Type II botnets*? At first glance, this might not be the case because *Type II botnets* have very different infection vectors/types from *Type I botnets*. However, we have also shown in the previous section that the distributions of both types of botnets are continuous and seems to be close to each other (in Figure 1(a)). Thus, it is hard to immediately draw a conclusion whether similar neighborhood correlation could be found in *Type II botnets* or not. Next, we plan to empirically verify this myth.

The previous work [2] has used the K-Nearest Neighbor (KNN) classifier which is a very popular machine learning algorithm and it uses neighbor information for classification. We also apply the KNN algorithm and select the same features for the KNN classifier used in [2]: /24 subnet address and physical location of *infected networks*.

To perform this experiment, we first prepare data for representing the class of *benign* and *malicious* networks. At this time, the *infected networks* of *Type II botnets* can be used to represent the *malicious class*. However, since we do not have data for the *benign* class, we also collect many (at the same scale as malicious networks) clean networks[3] to represent it. When we collect benign networks, we intentionally choose those which are close to *infected networks* in terms of the IP address and physical location, and they could be also neighbors of *infected networks*.

After the preparation, we divide each *Type II botnet* data (MegaD and Srizbi) into two sets for training/testing. And then, we apply the KNN classifier to predict unknown *infected networks*.

As shown in Table 3, the prediction results are quite interesting. Even though the prediction accuracy is lower than the case of *Type I botnet* (i.e., [2] reported around 93% of accuracy), our predictor for *Type II botnet* (in both MegaD and Srizbi cases) shows more than 88% accuracy with some reasonably small number of false positives.

The results imply that *Type II botnets* also have the similar characteristics as *Type I botnets* (i.e., if a host is infected, its neighbors are also likely to be infected). Then, why does this happen? It may be very hard to find concrete answers or clues for this question (unlike the intuitive explanation for *Type I* infection).

From our investigations, we could provide a possible answer. It may be caused by its infection media. As we described before, one promising infection method of *Type II botnets* is drive-by-download, which typically uses spam emails contain-

[3] We checked whether they are clean or not by looking up several DNS blacklists.

Botnet	K	Prediction Accuracy	False Positive Rate
MegaD	1	88.35%	7.35%
	3	88.25%	7.36%
	5	88.14%	7.54%
Srizbi	1	88.20%	6.23%
	3	87.70%	6.04%
	5	88.30%	5.77%

Table 3. Botnet prediction results.

ing links to compromised web sites, to trick people into downloading malicious binaries. Thus, the infection pattern of *Type II botnet* might highly depends on who receives spam emails. We find articles describing how spammers harness email addresses [26] [27], and they point out that collecting mailing lists is one of their main tasks. It is likely for mailing lists to contain email addresses belonging to similar locations (e.g., same company and same university). It implies that spam emails are delivered to people who are likely to be close to each other and thus victims infected by spam emails might also be close to each other.

4.2 Cross-Bonet Prediction

We have shown that if a host is infected by a *Type II botnet*, its neighbor networks are also likely to be infected by this *Type II botnet*. When we perform this test, we treat data of MegaD and Srizbi separately. However we know that these two botnets are very similar in terms of infection vectors. To confirm the similarity of their infected networks, we calculate a *manhattan distance* between the distribution of the two types of botnets. The *manhattan distance* between two items is the sum of all feature value differences for each of the all features in the item, and it is frequently used to denote whether two data distributions are similar or not (e.g., if a distance between data distributions of A and B is smaller than between that of A and C, A and B are closer to each other than C). It can be formalized as the following equation (assuming that there are two items/distributions of x and y, and they both have n elements).

$$Manhattan\ Distance = \sum_{i=1}^{n} |x_i - y_i|$$

We use the probability distributions of infected networks of Conficker, MegaD, Srizbi over IP address spaces to measure the *manhattan distance* and we find that the *manhattan distance* between Conficker and MegaD is 1.1427, Conficker and Srizbi is 1.1604, and MegaD and Srizbi is 0.8404. From the results, we can easily see that the distance between the *Type I* and *Type II botnet* distributions is larger than the distance between the two *type II botnets* distributions. This result shows that the distributions of infected networks with the same infection type are closer to each other than that of different types of botnet (i.e., infected networks of botnets in the same type show very similar distribution patterns).

This result gives us another insight that *if two botnets share the same infection vectors (i.e., they are of the same type), we might predict unknown infected networks of one botnet (e.g., a future botnet) with the help of the information of the other botnet (e.g., historic data).* This insight can be verified with a similar test that we have done before. We can perform a test by simply changing the training and testing data set to cross botnets. In the previous test, we extract the training and testing data from the same botnet. However in this case, we use data from botnet A for training and data from botnet B for testing. For example, when we predict (unknown) *infected networks* of the Srizbi botnet, we use data of the MegaD botnet for training.

Botnet	K	Prediction Accuracy	False Positive Rate
MegaD(train), Srizbi(test)	1	87.80%	7.41%
	3	86.75%	7.49%
	5	86.45%	7.69%
Srizbi(train), MegaD(test)	1	84.09%	6.53%
	3	83.89%	6.31%
	5	83.65%	5.09%

Table 4. Botnet cross-prediction results.

The cross-prediction results are quite surprising. As denoted in Table 4, this approach can predict unknown *infected networks* of the other botnet with more than 83% accuracy. This prediction accuracy is slightly less than what we observed previously. We believe that these results show us that even if we have no knowledge of some botnets (e.g., a future emerging botnet), if we have some information of a botnet whose infection vector is very similar to them[4], we may be able to predict unknown *infected networks*. To show a realistic example of application of the neighborhood correlation, let us first assume that a network administrator knows historic infected networks by Srizbi botnets. Then, he gets to know that the MegaD botnet starts spreading but he does not have any information of which networks are and will be infected. In this case, he can use the information of Srizbi botnet information (e.g., victim distribution). Based on the physical location and IP address of victims of Srizbi, he can predict future victim networks that will possibly be infected by MegaD with a reasonably high probability.

5 Limitations and Discussions

Like any measurement/analysis work, our empirical study has some limitations or biases. Even though we have collected a large amount of Conficker botnet

[4] Note that this is a very reasonable assumption because fundamental infection types of botnets are very limited and do not change frequently.

data, we have a relatively smaller amount of data for the MegaD and Srizbi botnets. This might cause some bias in our measurement results and subsequent analysis. In addition, the dynamism of IP addresses may lead to some over-estimation from the collected data. To reduce some of the side effects, we generalize our analysis over a network consisting of several adjacent IP addresses (i.e., measuring/analyzing over /24 subnets instead of each individual host).

To discover interesting insights, we leverage some previous work. For example, we use previous work to obtain how dynamic IP addresses are distributed over countries, but the information is not complete, i.e., it does not cover all countries. However, the provided information may help to uncover interesting cases (e.g., countries which are highly infected by botnets), hence the information is still useful.

When we perform the test to find networks with dynamic IP addresses through looking up reverse DNS PTR records of hosts in the networks, we may not collect reverse PTR records from all hosts because registration of a reverse PTR record is not always necessary. However previous work already verified the feasibility of such kind of test [23], lending credibility to these results (at least providing a good low-bound estimation).

6 Related Work

There are several studies of measurement or analysis of the *Type I botnet* victims. CAIDA provides basic information about the victim distribution of the Conficker botnet in terms of their IP address space and physical location [14]. In [13], Krishnan et al. conducted an experiment to detect infected hosts by Conficker. Weaver [15] built a probabilistic model to understand how the Conficker botnet spreads via network scanning. These studies provided useful and interesting analysis of the Conficker botnet. Shin et al. provided a large scale empirical analysis of the Conficker botnet and presented how victims are distributed [2]. However, our work is different from them in that we perform cross-analysis of different botnets and propose an early warning approach based on cross-prediction. Even though [2] observed neighbor correlation in Conficker, this work differs in that we empirically verified similar neighborhood correlation in *Type II botnets*. In addition, we have proposed and verified cross-botnet prediction techniques to predict unknown victims of one botnet from the information of the other botnet if they have similar infection vectors.

Measurement studies of the *Type II botnet* were also conducted. In [6], Mori et al. performed a large scale empirical study of the Srizbi botnet. John et al. set up a spam trap server to capture botnets sending spam emails [16]. This work also showed the distribution of victims in terms of their IP addresses. Even though these studies provided detailed analysis of some *Type II botnet(s)*, they still differ from our work in that they concentrate on a single (or one type of) specific botnet.

Some interesting studies from the analysis of *Type II botnets* have been also proposed. In [17], Cho et al. analyzed the MegaD botnet and showed how it

works. Caballero et al. provided an interesting technique to infiltrate the MegaD botnet and performed an analysis of its protocol [18].

Cai et al. measured how IP addresses are distributed over the world through several interesting sampling techniques [23]. Our work leverages some of its results but is different from their work in the main purpose.

7 Conclusion and Future Work

In this paper, we have collected a large amount of real-world botnet data and performed cross-analysis between different types of botnets to reveal the differences/similarities between them. Our large scale cross-comparison analysis results allow us to discover interesting findings and gain profound insights into botnet victims. Our results show fine-grained infection information and nature of botnet victims. They show some interesting relationships between geopolitical issues and malware infection, which might be the first work shedding light on this correlation. This study can guide us to design better botnet prediction or defense systems.

In our future work, we will study new approaches to explain relationships between geopolitical locations and malware infection more clearly with some realistic examples. In addition, we will collect more botnet data and investigate more diverse categories to discover correlations with different malware infection types.

References

1. Pauli, Darren: Srizbi Botnet Sets New Records for Spam: PC World. Retrieved 2008-07-20
2. Seungwon Shin and Guofei Gu: Conficker and Beyond: A Large-Scale Empirical Study. In: Proceedings of 2010 Annual Computer Security Applications Conference (ACSAC'10) (2010)
3. Microsoft Security Techcenter, Conficker Worm, http://technet.microsoft.com/en-us/security/dd452420.aspx
4. UPI, Virus strikes 15 million PCs, http://www.upi.com/Top_News/2009/01/26/Virus-strikes-15-million-PCs/UPI-19421232924206/
5. Symantec, Trojan.Srizbi, http://www.symantec.com/security_response/writeup.jsp?docid=2007-062007-0946-99
6. McAfee, Srizbi Infection, http://www.mcafee.com/threat-intelligence/malware/default.aspx?id=142902
7. SecureWorks, Ozdok/Mega-D Trojan Analysis, http://www.secureworks.com/research/threats/ozdok/?threat=ozdok
8. m86security, Mega-d, http://www.m86security.com/trace/i/Mega-D,spambot.896.asp.
9. Eric Chien, Downadup: Attempts at Smart Network Scanning, http://www.symantec.com/connect/blogs/downadup-attempts-smart-network-scanning
10. Yinglian Xie and Fang Yu and Kannan Achan and Eliot Gillum and Moises Goldzmidt and Ted Wobber: How Dynamic are IP Addresses?: Proceedings of ACM Special Interest Group on Data Communication (SIGCOMM) (2007)

11. Moheeb Abu Rajab and Jay Zarfoss and Fabian Monrose and Andreas Terzis: My botnet is bigger than yours (maybe, better than yours): why size estimates remain challenging: Proceedings of the first conference on First Workshop on Hot Topics in Understanding Botnets (2007)
12. Manuel Egele1 and Peter Wurzinger and Christopher Kruegel and Engin Kirda: Defending Browsers against Drive-by Downloads: Mitigating Heap-spraying Code Injection Attacks: Proceedings of the Sixth Conference on Detection of Intrusions and Malware and Vulnerability Assessment (DIMVA) (2009)
13. Srinivasan Krishnan and Yongdae Kim: Passive identification of Conficker nodes on the Internet: University of Minnesota - Technical Document (2009)
14. CAIDA, Conficker/Conflicker/Downadup as seen from the UCSD Network Telescope, http://www.caida.org/research/security/ms08-067/conficker.xml
15. Rhiannon Weaver: A Probabilistic Population Study of the Conficker-C Botnet: Proceedings of the Passive and Active Measurement Conference (2010)
16. John P. John and Alexander Moshchuk and Steven D. Gribble and Arvind Krishnamurthy: Studying Spamming Botnets Using Botlab: Proceedings of the Annual Network and Distributed System Security (NDSS) (2009)
17. Chia Yuan Cho and Juan Caballero and Chris Grier and Vern Paxson and Dawn Song: Insights from the Inside: A View of Botnet Management from Infiltration: Proceedings of the USENIX Workshop on Large-Scale Exploits and Emergent Threats (LEET) (2010)
18. Juan Caballero and Pongsin Poosankam and Christian Kreibich and Dawn Song: Dispatcher: Enabling active botnet infiltration using automatic protocol reverse-engineering: Proceedings of ACM Computer and Communications Security (CCS) (2009)
19. BOTLAB, A Study in Spam, http://botlab.org/
20. Shadowserver, Botnet Measurement and Study, http://shadowserver.org/wiki/
21. IP2Location, IP2Location Internet IP Address 2009 Report, http://www.ip2location.com/
22. IANA, IANA IPv4 Address Space Registry, http://www.iana.org/assignments/ipv4-address-space/ipv4-address-space.xml
23. Xue Cai and John Heidenmann: Understanding Address Usage in the Visible Internet: USC/ISI Technical Report ISI-TR-656 (2009)
24. Heather Alderfer and Stephen Flynn and Bryan Birchmeier and Emilie Schulz: Information Policy Country Report: Turkey: University of Michigan School of Information Report (2009)
25. Nicholas Ianelli and Aaron Hackworth: Botnets as a Vehicle for Online Crime: CERT/CC Technical Report (2005)
26. Uri Raz, How do spammers harvest email addresses ?, http://www.private.org.il/harvest.html
27. FAQs.org, FAQ: How do spammers get people's email addresses ?, http://www.faqs.org/faqs/net-abuse-faq/harvest/
28. Juan Caballero and Chris Grier and Christian Kreibich and Vern Paxson: Measuring Pay-per-Install: The Commoditization of Malware Distribution: Proceedings of USENIX Security Symposium (2011)