Using uncleanliness to predict future botnet addresses

M. Patrick Collins
CERT Network Situational Awareness Group
5000 Forbes Avenue
Pittsburgh, PA 15213
mcollins@cert.org

Timothy J. Shimeall
CERT Network Situational Awareness Group
5000 Forbes Avenue
Pittsburgh, PA 15213
tjs@cert.org

Sidney Faber
CERT Network Situational Awareness Group
5000 Forbes Avenue
Pittsburgh, PA 15213
sfaber@cert.org

Jeff Janies
CERT Network Situational Awareness Group
5000 Forbes Avenue
Pittsburgh, PA 15213
janies@cert.org

Rhiannon Weaver
CERT Network Situational Awareness Group
5000 Forbes Avenue
Pittsburgh, PA 15213
rweaver@cert.org

Markus De Shon
CERT Network Situational Awareness Group
5000 Forbes Avenue
Pittsburgh, PA 15213
mdeshon@cert.org

ABSTRACT

The increased use of botnets as an attack tool and the awareness attackers have of blocking lists leads to the question of whether we can effectively predict future bot locations. To that end, we introduce a network quality that we term uncleanliness: an indicator of the propensity for hosts in a network to be compromised by outside parties.

We hypothesize that unclean networks will demonstrate two properties: spatial and temporal uncleanliness. Spatial uncleanliness is the tendency for compromised hosts to cluster within unclean networks. Temporal uncleanliness is the tendency for unclean networks to contain compromised hosts for extended periods.

We test for these properties by collating data from multiple indicators (spamming, phishing, scanning and botnet IRC log monitoring). We demonstrate evidence for both spatial and temporal uncleanliness. We further show evidence for cross-relationship between the various datasets, showing that botnet activity predicts spamming and scanning, while phishing activity appears to be unrelated to the other indicators.

1. INTRODUCTION

Botnets are a common attack tool due to the anonymity and flexibility that they provide attackers. Modern bots can be used for DDoS, spamming, network infiltration, keylogging and other criminal acts [5, 15]. Past research, notably by Mirkovic *et al.* [18], has shown that botnet based attacks can be divided into distinct phases of acquisition and use.

We expect that bot acquisition is opportunistic [2]: while attackers may avoid certain networks [24], in the majority of cases, attackers have no interest or knowledge about targets except that the target is vulnerable. With automatically propagating attack tools, an attacker may not know about the existence of a target until after he compromises it.

Specific variants of bots such as Gaobot can spread themselves using network shares, AOL Instant Messenger, and multiple Windows vulnerabilities[1]. Given the sheer population and variation in worms, and the virulence of common attacks, it is now reasonable to expect that any publicly accessible host on the Internet will be attacked by every common method within a short period[2].

If we assume that an attacker cannot distinguish between the hosts within a network, then he is equally likely to attack any of them. In addition, with no advance knowledge of what a target is vulnerable to, an attacker will use all attacks available to him. Finally, given that the population of attackers is so large, individually attacker preferences become less relevant: while one worm may opt not to use a particular vulnerability, another dozen will. Consequently, the probability that a machine will be compromised during some period is not a function of that host's attacker. We hypothesize that the probability of compromise is instead a property of the host's *defenders*.

We characterize a network's defensive posture by its *uncleanliness*, which is an indicator of the propensity for hosts within that network to be compromised. Consider two institutions, A and B. Institution A maintains an aggressive firewall policy, disables all email attachments, maintains all files on a central server and reimages all hosts on the network from a fully patched and maintained master image each night. Institution B has no inventory of hosts on its network, runs a variety of hardware and software installations that administrators might not even be aware of, has a large number of self-administered machines and has no firewall. We expect that institution A would be less vulnerable to attack, that compromised hosts would be quickly detected,

[1] `http://www.symantec.com/enterprise/security_response/writeup.jsp?docid=2006-052712-1744-99&tabid=2`

[2] A report of the expected time between attacks for specific vulnerabilities is available at `http://isc.sans.org/survivaltime.html`; the interval between attacks for the average address is on the order of 20 minutes

and that those compromised hosts would be restored to an uncompromised state quickly. Conversely, machines in institution B would be reached by a larger number of attacks, and compromised hosts may not be noticed or repaired until long after the compromise has taken place. Institution B's network is unclean.

We can observe the uncleanliness of a network by examining its result. If a host is compromised, we expect that the attacker will use it to, among other activities, spam, scan and DDoS other hosts. If uncleanliness is a network-specific property, we expect that compromised hosts will cluster in specific networks, which we can identify via the phenomena of *spatial* and *temporal* uncleanliness. We emphasize that uncleanliness is a network property: hosts are *compromised*, networks are *unclean*.

We define *spatial uncleanliness* as the tendency for compromised hosts to cluster in unclean networks. Spatial uncleanliness implies that if we see a host engaged in hostile activity (such as scanning), we have a good chance of finding another IP address in the same network engaged in hostile activity. We will test for spatial uncleanliness by examining the clustering of addresses within networks. If our hypothesis about spatial uncleanliness is correct, then we would expect a set of compromised addresses to be reside in fewer equally sized networks than addresses chosen at random from a population reflecting the structure of the Internet.

We define *temporal uncleanliness* as the tendency for compromised hosts to repeatedly appear in the same networks over time. Temporal uncleanliness implies that if a host is compromised, then other hosts within that network will be compromised in the future. We will test for temporal uncleanliness by examining the ability of unclean networks to predict future host compromises. If our hypothesis about temporal uncleanliness is correct, then networks containing compromised hosts will predict future compromised hosts more accurately than equally sized networks chosen at random.

Figure 1 explains our intuition for spatial and temporal uncleanliness. This figure shows two plots: the upper counts the number of unique hosts scanning a large network from January to April, 2006. The lower plot is a plot showing how many of these scanning addresses were also present in a botnet reported during the first week of March, 2006. This plot contains two lines: one counts the number of unique addresses from the bot report that were also identified scanning; the second counts the number of unique addresses from the bot report that were present in a 24-bit CIDR block where at least one address was also scanning.

First note that these reports resulted from two different detection methods: the bot data was collected by observing IP addresses communicating on IRC channels, while the scanning data was collected using a behavioral scan detection method deployed on an observed network [6]. There is a strong intersection between the two sets: at its peak, 35% of the addresses reported as belonging to the botnet are scanning the observed network.

Second, we observe that using the /24's comprising the botnet identifies more scanners than the botnet addresses alone. We demonstrate in §4 that this result is significant.

Finally, as this figure shows, abnormal scanning (and therefore botnet compromise) occurs over several weeks. If bots take several weeks to be identified and removed, we expect that an unclean network will remain unclean for some time, and therefore we can predict future hostile activity from the same network over the term of the lifetime of a particular compromise.

The primary contribution of this paper is a study of the properties of uncleanliness and whether they can be used effectively to predict future activity. To do so, we test for the existence of spatial and temporal uncleanliness by comparing the traffic from various reports of hostile activity. We demonstrate that compromised hosts are both more densely clustered than normal traffic and predict future unclean activity. In addition, we show that scanning, spamming and bots show evidence of cross relationship, such as the scanning observed in Figure 1. We also show that these phenomena do not predict future phishing sites, but that past phishing sites do. We therefore demonstrate that temporal uncleanliness holds for all four indicators. We then test the strength of this predictive mechanism by evaluating its suitability to block traffic crossing a large network. We demonstrate that limited predictive blocking is feasible, due to the impact of locality [17] evident in network traffic.

The remainder of this paper is structured as follows: §2 outlines relevant previous work in reputation management and identifying hostile groups by past history. In §3, describe and classify the data sources that we use in this paper. §4 examines the spatial uncleanliness hypothesis, and §5 examines the temporal uncleanliness hypothesis. §6 examines the impact of blocking unclean networks and §7 discusses the results.

2. PREVIOUS WORK

Researchers initially studied botnets due to their use in DDoS attacks. Mirkovic *et al.* [18] identified DDoS attacks which used two distinct phases: acquiring hosts to use for the DDoS and using those hosts to conduct an attack. Freiling *et al.* [5] identify a variety of other attacks that botnets can conduct efficiently. Collins *et al.* [2] examined attacks as conducted by opportunistic attackers: that is, the attacker has no interest or knowledge of the target except that the target is exploitable. Our work uses these concepts to study the impact of largely automated acquisition and its impact on network defense.

Botnet demographics have been studied using Honeypots and by actively probing bot networks [8, 9, 21]. Rajand *et al.*'s [21] analysis is particularly relevant due to the extended period during which they observed network traffic, allowing them to identify not only botnet demographics but activity. Our work differs from these analyses by comparing multiple observed phenomena and using this information to predict future activity.

In operational security, blacklists are commonly used to identify and block hosts that are already assumed to be hostile. Examples of such blacklists include Spamhaus' ZEN list [20] and the Bleeding Snort rule set [23]. Researchers such as Levy [16] note that spammers increasingly rely on the use of occupied hosts to generate spam messages - these approaches are more attractive to spammers because they offload processing requirements from the spammer (as noted by Laurie *et al.* [15]) and because they hide the attacker's identity [4].

In addition, researchers have studied the impact of blacklists on spamming and other hostile activity. Jung *et al.* [12] compare spamming blacklists against spam traffic to MIT

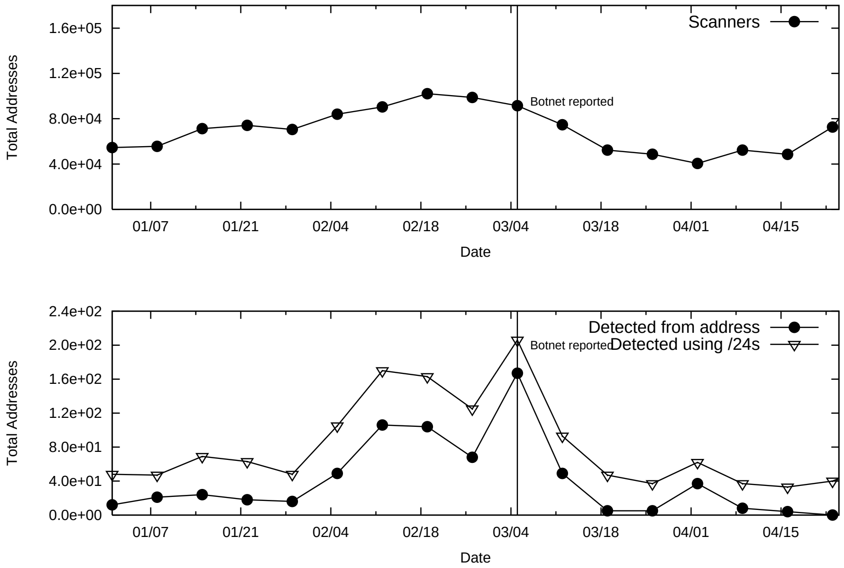


Figure 1: Relationship between scanning and botnet population. Addresses in the botnet scan the observed network for approximately a month before the report, while activity drops noticeably after the report.

in 2000 and 2004, finding that in 2004, 80% of spammers were identified by blacklists. Ramachandran *et al.* [22], examine blacklist abuse by botnet owners. Ramachandran notes that botnet owners appear to place a higher premium on addresses not present on blacklists. Since uncleanliness is intended to predict future hostile addresses, this may impact the costs noted by Ramachandran.

McHugh *et al.* [17] use locality to characterize normal network behavior and differentiate attacks. Krishnamurthy *et al.* [14] group IP addresses into heterogeneously sized *network aware clusters* in order to characterize target audiences for networks, and demonstrate that many sites have common audiences. Jung *et al.* [10] use network aware clustering for DDoS defense. These methods of blocking are predicated on the assumption that attack traffic differs from normal traffic due to a limited and clustered audience for any normal service. Our filtering approach differs from the past history used in these cases by developing a set of explicitly untrusted networks.

3. SOURCE DATA

We demonstrate evidence of uncleanliness by showing that address distributions from unclean data sets show specific qualities. In order to do so, we must collate information from various sources, many of which use different collection methods. In this section, we describe a taxonomy and notation scheme for managing this data. This section is divided as follows: §3.1 explains the taxonomy and notation for reports, and §3.2 describes the individual reports.

3.1 Model

In order to estimate the uncleanliness of a network, we must compare data from multiple sources. For example, an attacker may initially use a bot for scanning, then for spamming. We call these sources *reports*, each of which consists of a set of IP addresses describing a particular phenomenon over some period. Reports differ by the *class* of data reported, the period covered by the report, and the method used to generate that data.

We use four classes of unclean data for this paper:

Bots: An IP address identified as hosting some form of bot software or communicating with a botnet command and control host.

Phishing: An IP address identified as hosting a phishing site to fraudulently acquire private user information.

Scanning: An IP address identified as scanning using the methods developed by Gates *et al.* [7] and Jung *et al.* [11].

Spamming: An IP address identified as spamming using a behavioral spam detection technique [3].

These reports all describe phenomena associated with compromised hosts. Scanning and spamming are both common botnet uses, and phishing requires setting up a fraudulent web site.

We further divide reports as either *provided* or *observed.* Provided reports are collected from external parties, and

[3] This spam detection method is currently under review.

Unclean reports					
Tag	Type	Class	Valid Dates	Size	Reporting method
bot	Provided	Bots	2006/10/01-2006/10/14	621,861	Bot addresses acquired through private reports from a third party
phish	Provided	Phishing	2006/05/01-2006/11/01	53,789	Addresses from a Phishing report list
scan	Observed	Scanning	2006/10/01-2006/10/14	151,908	IP addresses scanning the observed network
spam	Observed	Spam	2006/10/01-2006/10/14	397,306	IP addresses spamming the observed network
Reports for hypothesis testing					
bot − test	Provided	Bots	2006/05/10	186	Botnet addresses acquired through private communication
control	Observed	N/A	2006/09/25-2006/10/02	46,899,928	Control addresses acquired from the observed network

Table 1: Table of tags used to analyze spatial and temporal uncleanliness.

can use different methodologies to observe the same effects. For example, a phishing list can acquire IP addresses by using spam traps [19] or by collecting user reports, (e.g., the submission form at the CastleCops PIRT service [1]). For the analyses within this paper, we use only one source per report and assume that the source's collection methodology is consistent over the report period.

In contrast to provided reports, observed reports are generated from network traffic logs reporting traffic covering a large edge network. Because we generate observed reports, we are able to collect observed reports at any time, which gives us greater flexibility in picking data than in the case of provided reports.

Each report is differentiated by a *tag* which, for this paper, summarizes the period and source for the report. We express this using the notation $\mathcal{R}_{\mathsf{T}}$, where T is the tag (e.g., scan). A list of reports is provided in Table 1; this list is used for testing uncleanliness properties. Another list, given in Table 2, will be used for the analysis in §6.

Because we expect uncleanliness to be a network property, we approximate distinct networks by using identically sized CIDR blocks. We define a CIDR masking function $C_n(i)$. The CIDR masking function evaluates to the unique CIDR block with prefix length n that contains the IP address i (e.g., $C_{16}(127.1.135.14) = 127.1.0.0/16$). For convenience, when the CIDR masking function is applied on a report S, the result is set-valued and returns the set of all n-bit CIDR blocks in that set, that is:

$$C_n(S) \equiv \bigcup_{i \in S} C_n(i) \tag{1}$$

When determining whether or not an IP address resides within a set of CIDR blocks, we will use a CIDR inclusion relation, $\sqsubset$, to indicate that an IP address is resident in one of a set of CIDR blocks:

$$i \sqsubset S \rightarrow \exists n \text{ s.t. } C_n(i) \in C_n(S) \tag{2}$$

With all sets and reports, we use bars to indicate cardinality, i.e., $|S|$ is the number of elements in the set S.

3.2 Reports

Table 1 is an inventory of the reports used in this paper to test spatial and temporal uncleanliness. Recall that provided reports have been given to us by other parties and that we generate observed reports using traffic logs from the observed network. Because of this, the dates that we can test for temporal uncleanliness are constrained by the times that the provided reports cover.

The observed network is composed of over 20 million distinct IPv4 addresses and contains several servers that are heavily used by clients across the Internet. Given the size and activity of the observed network, we assume that IP addresses from the Internet crossing into it are a representative sample of the Internet as a whole.

All reports have been filtered to only include addresses that are outside of the observed network and are not otherwise reserved (*e.g.*, all addresses specified in RFC 1918 have been removed from reports). This filtering step is intended to remove selection bias from our observed reports; given our familiarity with the observed network and its size, we may identify more of a particular phenomenon than the provided reports may identify.

We classify four of the reports in this list as *unclean reports*. These are the reports we use as ground truth for identifying the four classes described in §3.1: bots, phishing, scanning and spamming. During the two week period of October 1st–14th, 2006, we have both provided and observed reports on all classes of unclean activity, consequently we use this period to test temporal uncleanliness.

The next set of reports are used specifically to test the spatial and temporal uncleanliness hypotheses. The bot − test report describes a small botnet from five months before all the other activity analyzed in this paper, bot − test is used as an extreme case for prediction: if a five-month old report can accurately predict current unclean activity, then a recent report should be more effective.

The control report consists of 47 million unique IP addresses observed during the week of September 25th, 2006. We compare the data from our other reports against randomly generated subsets of control in order to determine whether or not these reports exhibit spatial or temporal uncleanliness. We use the control report to more accurately reflect the structure of IPv4 space than we would using purely randomly chosen IP addresses. The report consists of IP addresses observed to engage in payload-bearing TCP activity, which reduces the risk of the address being spoofed. Furthermore, as noted in §3.1, the observed network includes a variety of servers used by hosts throughout the Internet, and by focusing exclusively on the IP addresses of the hosts without using any criteria apart from the unspoofed criterion, we expect the resulting report to approximate a random sample of active IP addresses on the Internet.

4. SPATIAL UNCLEANLINESS

We define *spatial uncleanliness* as the propensity for occupied addresses (bots) to be clustered in unclean networks. In this section, we formulate and test the *spatial uncleanliness hypothesis*.

This section is divided as follows: §4.1 describes the methodology used to test for spatial uncleanliness. §4.2 describes the results of our tests and shows evidence for spatial uncleanliness.

4.1 Model and Methodology

Recall our assumption that the likelihood of a host being compromised is a network property: if a network is unclean, then its administrators will not identify compromised machines or rapidly repair them. Consequently, we expect that multiple hosts within an unclean network will be compromised, and that compromised addresses will cluster within unclean networks. In order to test this hypothesis, we will compare the expected population of compromised hosts within equally sized CIDR blocks.

Throughout this paper, we use homogeneously sized CIDR blocks to model individual networks. Given that we lack accurate information on network populations, we make a *ceteris paribus* assumption that equally sized blocks should have equivalent populations. In comparison, heterogeneous partitioning such as network-aware clustering [14], can result in network populations that differ in size by several orders of magnitude. Based on DDoS filtering work done by Collins and Reiter [3], we expect that CIDR prefix lengths above 16 bits will be too imprecise for effective filtering and detection. Consequently, we limit our block sizes to between 16 and 32 bits.

To test for spatial uncleanliness, we begin with a measurement for comparative density. If we have two sets, S_1 and S_2, and $|S_1| = |S_2|$, then we say that S_1 is *denser at n-bits* if $|C_n(S_1)| < |C_n(S_2)|$. That is, if the total number of n-bit CIDR blocks containing S_1 is smaller than the set containing S_2.

In §1, we stated that spatial uncleanliness implies that if a host is compromised, there is a good chance another host on the same network will be compromised. Consequently, if we had a set of compromised host addresses, and a control set of randomly selected addresses with equal cardinality, we would expect that the compromised address set was *at least* as dense at all CIDR prefix lengths.

We therefore summarize the spatial uncleanliness hypothesis as follows: if we have a report which selects unclean traffic from the Internet, $\mathcal{R}_{\mathsf{unclean}}$, then the IP addresses within that report will be more densely packed than a set of IP addresses with equal cardinality randomly selected from the control set.

To test the spatial uncleanliness hypothesis, we use the formulation given in Equation 3. Assume two reports: $\mathcal{R}_{\mathsf{unclean}}$ which reports on unclean traffic, and $\mathcal{R}_{\mathsf{normal}}$, a control group. If both reports are of equal cardinality, then:

$$\forall n \in [16, 32]\ |C_n(\mathcal{R}_{\mathsf{unclean}})| \leq |C_n(\mathcal{R}_{\mathsf{normal}})| \qquad (3)$$

Recall from above that we have limited our block sizes to between 16 and 32 bits.

4.2 Analysis

In order to test the spatial uncleanliness hypothesis as formulated in Equation 3, we compare the population of addresses per n-bit CIDR blocks for an unclean report against the expected population for n-bit CIDR blocks across the Internet as a whole.

As discussed in §3.2 we model the population of $\mathcal{R}_{\mathsf{normal}}$ by randomly selecting IP addresses from $\mathcal{R}_{\mathsf{control}}$. Kohler *et al.* [13] observe that IP addresses are not evenly distributed across IPv4 space; as a consequence, a purely random model will result in an artificially depressed density estimate. We test two population estimates to compensate for this. The first, *naive*, estimate selects addresses evenly from across all /8's which are listed as populated by IANA[4]. The second, *empirical*, estimate draws addresses $\mathcal{R}_{\mathsf{control}}$ In the empirical estimate, we create 1000 randomly generated subsets of $\mathcal{R}_{\mathsf{control}}$ and group the resulting addresses.

Figure 2 plots the number of blocks observed for CIDR block prefix lengths of 16 to 32 bits. This plot compares the botnet density, $\mathcal{R}_{\mathsf{bot}}$, against both the empirical and naive density estimates of equal size (621,861 addresses, as per Table 1), with the intent of comparing the effectiveness of the estimates against the population actually observed. As this figure shows, the number of distinct blocks containing bots is less than or equal to the number of blocks for either the empirical or naive estimates. Of particular note is that as the block size decreases, moving from left to right on the graph, the number of blocks observed for both the empirical estimate and the botnet report do not proportionately increases. If addresses were evenly distributed, as is the case with the naive estimate, then we would expect the number of blocks observed to double with each unit increase in prefix length.

Based on the results from Figure 2, we use empirical estimation throughout the rest of this paper. A note about Figures 2 and 3: the empirically estimated populations are plotted using boxplots, however the variation in the number of blocks at a particular prefix length relative to the total number of blocks is very small and consequently not visible.

Figure 3 compares control data (empirically estimated populations) against the chosen reports for each class: spamming, scanning, botnets and phishing. As with the population plot in Figure 2, these plots represent the total number of n-bit blocks observed for that population. Since each population is of equal size, the lowest line will have the highest density. For each plot in Figure 3, the control data consists of 1000 random subsets of $\mathcal{R}_{\mathsf{control}}$ and plotting the resulting distribution as a boxplot. Again, we note that the variation in block counts for the empirical data is very narrow and generally not visible in these plots.

Figure 3(i) is a plot of the comparative volume for $\mathcal{R}_{\mathsf{bot}}$. As this plot shows, the population of $\mathcal{R}_{\mathsf{bot}}$ is more densely packed than the expected population drawn from $\mathcal{R}_{\mathsf{control}}$. Figure 3(ii) plots the volume of $\mathcal{R}_{\mathsf{phish}}$ reported from May to October, 2006. We use a five month sample due to the

[4] http://www.iana.org/assignments/ipv4-address-space

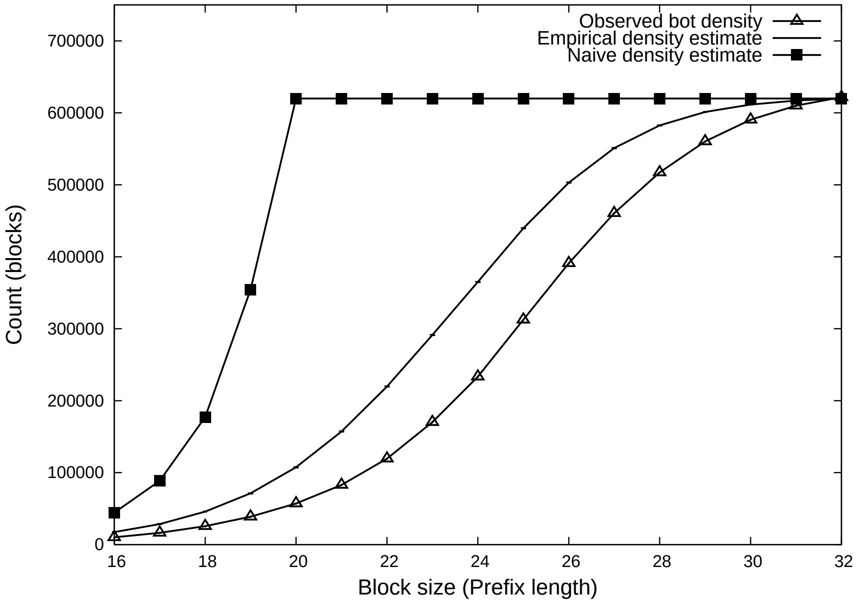


Figure 2: Comparison of density estimation techniques (naive and empirical) against actual botnet density. Note that the number of blocks estimated using the naive technique is considerably higher than the other two.

smaller size of the phishing reports in comparison to the other reports. As shown in Table 1, the six month phishing report is approximately an order of magnitude smaller than the other unclean reports. As with Figure 3(i), addresses in the phishing report are more tightly packed than addresses selected from the control report.

Figure 3(iii) plots the volume of $\mathcal{R}_{\text{spam}}$ from October 1st to 14th, 2006. Figure 3(iv) plots the volume of $\mathcal{R}_{\text{scan}}$ for the same period. Each of these reports is more tightly packed than the comparative control reports.

As Figures 2 and 3 show, unclean reports have an n-bit density greater than or equal to or greater then the n-bit density of the control reports for all values of n. Consequently, this data supports the spatial uncleanliness hypothesis: compromised hosts are disproportionately concentrated in certain networks.

5. TEMPORAL UNCLEANLINESS

We now address temporal uncleanliness: the propensity for networks to remain unclean for extended periods of time. In order to test for temporal uncleanliness we compare the ability of a report of unclean addresses to predict future compromised addresses; in particular, whether or not a report of bot addresses can predict future bots, spamming, scanning and phishing.

This section is divided as follows: §5.1 describes our method for measuring the presence of temporal uncleanliness, and §5.2 shows the results.

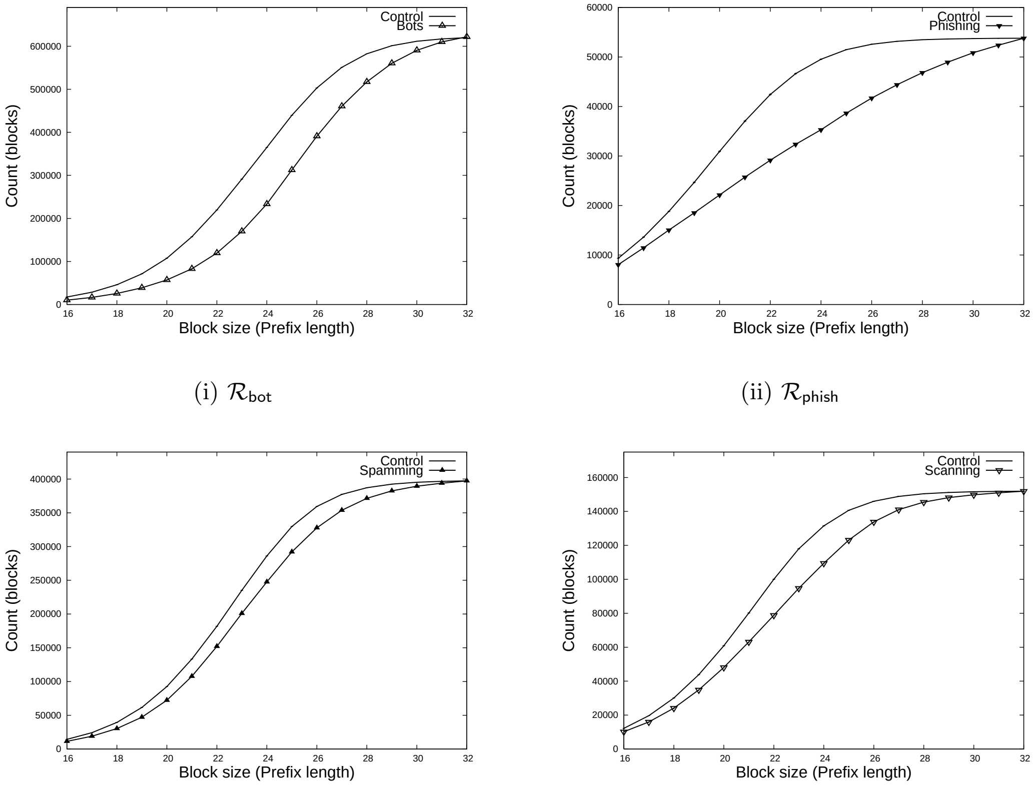
5.1 Model and Methodology

To observe temporal uncleanliness, we examine the *predictive* capacity of reports of unclean data. Consider three reports:$\mathcal{R}_{\text{event}-\text{past}}$, which reports on some event in the past; $\mathcal{R}_{\text{normal}-\text{past}}$, which reports on past activity without any particular criterion, and $\mathcal{R}_{\text{event}-\text{present}}$, which describes the same event's population in the present. If $\mathcal{R}_{\text{event}-\text{past}}$ and $\mathcal{R}_{\text{normal}-\text{past}}$ are of equal cardinality, then $\mathcal{R}_{\text{event}-\text{past}}$ is a better predictor of the report $\mathcal{R}_{\text{event}-\text{present}}$ at prefix length n if:

$$\begin{aligned} |C_n(\mathcal{R}_{\text{event}-\text{past}}) \cap C_n(\mathcal{R}_{\text{event}-\text{present}})| & > \\ |C_n(\mathcal{R}_{\text{normal}-\text{past}}) \cap C_n(\mathcal{R}_{\text{event}-\text{present}})| & \end{aligned} \quad (4)$$

If temporal uncleanliness exists, then we expect that unclean reports will consistently be better predictors of future unclean reports than a control report. However, we note that due to spatial uncleanliness, an unclean report will populate fewer equally sized blocks than an equivalent control report. As a consequence, as block size increases, the control report will have a larger number of imprecise successes. Therefore, there will be some prefix length below which the unclean report will be a worse predictor.

For testing, we use the form of the temporal uncleanliness hypothesis given in the equation below. Given that $\mathcal{R}_{\text{unclean}-\text{past}}$ and $\mathcal{R}_{\text{normal}-\text{past}}$ have equal cardinality, then



(i) $\mathcal{R}_{\mathsf{bot}}$

(ii) $\mathcal{R}_{\mathsf{phish}}$

(iii) $\mathcal{R}_{\mathsf{spam}}$

(iv) $\mathcal{R}_{\mathsf{scan}}$

Figure 3: Comparative density of unclean blocks against addresses selected from $\mathcal{R}_{\mathsf{control}}$. Note that in each case, the expected number of blocks in $\mathcal{R}_{\mathsf{control}}$ is higher than the observed values, indicating that unclean addresses are more densely packed in those blocks than randomly selected addresses.

$$\begin{aligned}&\exists n \in [16, 32] \text{ s.t.}\\ &\quad |C_n(\mathcal{R}_{\mathsf{unclean-past}}) \cap C_n(\mathcal{R}_{\mathsf{unclean-present}})| >\\ &\quad |C_n(\mathcal{R}_{\mathsf{normal-past}}) \cap C_n(\mathcal{R}_{\mathsf{unclean-present}})|\end{aligned} \tag{5}$$

That is, there exists a prefix length where a previously generated report of unclean activity is more predictive of present unclean activity than a control report of equal cardinality. As with spatial uncleanliness, we limit our analyses to blocks with a CIDR prefix length of at least 16 bits.

5.2 Analysis

We now test the temporal uncleanliness hypothesis formulated in Equation 5. To do so, we use $\mathcal{R}_{\mathsf{bot-test}}$ as $\mathcal{R}_{\mathsf{unclean-past}}$ and then compare against each of our unclean reports collected during the period of October 1st-14th, 2006. Recall that we don't control the dates for which we receive provided reports. During this period, we have data from each of the provided reports and could generate observed reports for the same period. By using a five month gap in time, we also test an extreme case: if past activity can effectively predict future activity five months in advance, then we should be able to predict future activity over shorter periods.

Figure 4 shows the relative predictive capacity of $\mathcal{R}_{\mathsf{bot-test}}$ against future unclean reports; for these figures, $\mathcal{R}_{\mathsf{phish}}$ is a sub report of $\mathcal{R}_{\mathsf{phish}}$ from Table 1. This report is considerably smaller than the others, containing 2302 addresses. This results in a smaller degree of intersection with the randomly generated reports from the control report.

As in §4.2, we generate the reference line by plotting a boxplot showing the variance of 1000 randomly selected test reports drawn from $\mathcal{R}_{\mathsf{control}}$. In contrast with Figure 3, the small cardinality of $\mathcal{R}_{\mathsf{bot-test}}$ ensures that the variations observed by the boxplot are visible. We consider t $\mathcal{R}_{\mathsf{bot-test}}$ to be a better predictor than $\mathcal{R}_{\mathsf{control}}$ if the cardinality of its intersection with the corresponding unclean report is higher than the intersection with randomly selected addresses in 95% of the observed cases.

As Figure 4 shows, $\mathcal{R}_{\mathsf{bot-test}}$ is a better predictor than $\mathcal{R}_{\mathsf{control}}$ for botnets, spamming and scanning at various prefix lengths. Also of note is the impact of spatial uncleanliness: in these three figures, $\mathcal{R}_{\mathsf{bot-test}}$ is a better predictor for prefix lengths of approximately 19-20 bits and longer. At shorter prefix lengths, randomly selected addresses become better predictors. Using the 95% threshold, $\mathcal{R}_{\mathsf{bot-test}}$ is a stronger predictor of future botnet activity between 20 and 25 bits, spamming between 19 and 32 bits, and scanning between 20 and 24 bits. For prefix lengths longer than these values, the two reports are equally predictive due to the low probability of seeing CIDR blocks from either report intersect.

Figure 4(ii) plots the predictive capacity of $\mathcal{R}_{\mathsf{bot-test}}$ against $\mathcal{R}_{\mathsf{phish}}$. In contrast to the other plots in Figure 4, this plot indicates that $\mathcal{R}_{\mathsf{bot-test}}$ is not a good predictor of future phishing activity in comparison to randomly selected control sets.

We have two hypotheses as to why this occurs for phishing data: Ramachandran *et al.* [22] describe how botnet owners place a higher premium on addresses that have not yet been identified as bots. Because phishing sites need to be publicized, a phishing IP address becomes public knowledge, marked on blacklists and consequently highly unattractive for the owner of a botnet.

An alternative explanation is that, in contrast to botnets, phishing sites are generally hosted on web servers, and a phisher may prefer to host phishing sites in a actual datacenter to ensure robustness during a flash crowd. At the minimum, a phishing site must be publicly accessible, while a useful bot can exist behind a NAT or a firewall. Therefore, phishers may prefer sites that are already hosting web servers and have the resources to handle a high traffic load.

In order to determine whether the temporal uncleanliness hypothesis does hold for phishing, we now consider a test that uses phishing data exclusively. Figure 5 plots the intersection of $\mathcal{R}_{\mathsf{phish-test}}$ against the same phishing set as in Figure 4(ii). In this case, $|\mathcal{R}_{\mathsf{phish-test}}| = 1386$. We note that this figure shows strong evidence for temporal uncleanliness in phishing.

Since these results show that five month old reports can be used to more effectively predict the population of future reports than randomly selected IP addresses from a week before, we conclude that the temporal uncleanliness hypothesis is supported by this data. Furthermore, in Equation 5, we chose a range of IP blocks arbitrarily, we can now establish a lower limit for the prefix length of 20 bits, an an upper limit in excess of 24 bits.

We have also shown that phishing activity and botnet activity are not related in the way that bots, scanning and spamming are. As noted elsewhere [21, 15], scanning and spamming are commonly implemented with botnets, so we would expect that $\mathcal{R}_{\mathsf{bot}}, \mathcal{R}_{\mathsf{scan}}$ and $\mathcal{R}_{\mathsf{spam}}$ are related. However, the inability of $\mathcal{R}_{\mathsf{bot-test}}$ to predict future phishing activity suggests that a measurement for uncleanliness will have to be multidimensional: phishing sites are still taken over, but it may be that phishers have different criteria for the machines they occupy than botnet owners.

6. BLOCKING TESTS

The spatial and temporal uncleanliness hypotheses together provide a method for identifying the risk that traffic from a particular network originates from a compromised host. We now address the issue of whether unclean networks can be *effectively* blocked; that is, whether or not blocking a set of unclean networks will adversely affect legitimate traffic entering an active network.

To determine whether we can effectively block traffic, we conduct a limited experiment to show the impact of blocking a set of unclean networks would have on incoming traffic to a live network. The remainder of this section is structured as follows:§6.1 describes our analytical method, and §6.2 discusses the results.

6.1 Method

To determine whether we can productively block traffic from unclean networks, we examine traffic logs from a live network and compare the intersection between incoming traffic, the $\mathcal{R}_{\mathsf{bot-test}}$ and other uncleanliness reports from the same observation period as the incoming traffic.

We begin by collecting traffic logs of all traffic that crosses the observed network from all IP addresses $i \sqsubset C_{24}(\mathcal{R}_{\mathsf{bot-test}})$ for the observation period of October 1st–14th 2006. This report, $\mathcal{R}_{\mathsf{candidate}}$, consists of all IP addresses observed in traffic crossing the observed network that share a /24 in common with any of the IP addresses in $\mathcal{R}_{\mathsf{bot-test}}$. This allows us to test the effectiveness of filtering from the /24 to the /32 range; we pick this range because, as seen in Figure 3, 24 bits is the minimum block size at which $\mathcal{R}_{\mathsf{bot-test}}$ is an unambiguously better predictor of future uncleanliness than control data. We further constrain $\mathcal{R}_{\mathsf{candidate}}$ to those addresses that generate at least one TCP record during this period.

The traffic data used in this analysis consists CISCO NetFlow[5] V5 records. NetFlow records are a representation of approximate sessions consisting of a log of all identically addressed packets within a limited time. Flow records are a compact representation of traffic, but do not contain payload.

Consequently, our analysis includes a degree of uncertainty because we cannot validate what any session was engaged in. To compensate for this , we differentiate addresses by membership in one of the unclean reports and by behavior observed in the flow records. We partition the addresses in $\mathcal{R}_{\mathsf{candidate}}$ into three reports: $\mathcal{R}_{\mathsf{unknown}}$, $\mathcal{R}_{\mathsf{hostile}}$ and $\mathcal{R}_{\mathsf{innocent}}$. A full inventory of the reports used in this analysis is given in Table 2.

$\mathcal{R}_{\mathsf{hostile}}$ consists of any IP address in $\mathcal{R}_{\mathsf{candidate}}$ that is also present in the unclean reports (i.e., scanning, spamming, phishing or botnet membership). The hostile set is identified purely by intersecting these reports, and once an IP address is identified as hostile it cannot be present in the remaining two reports.

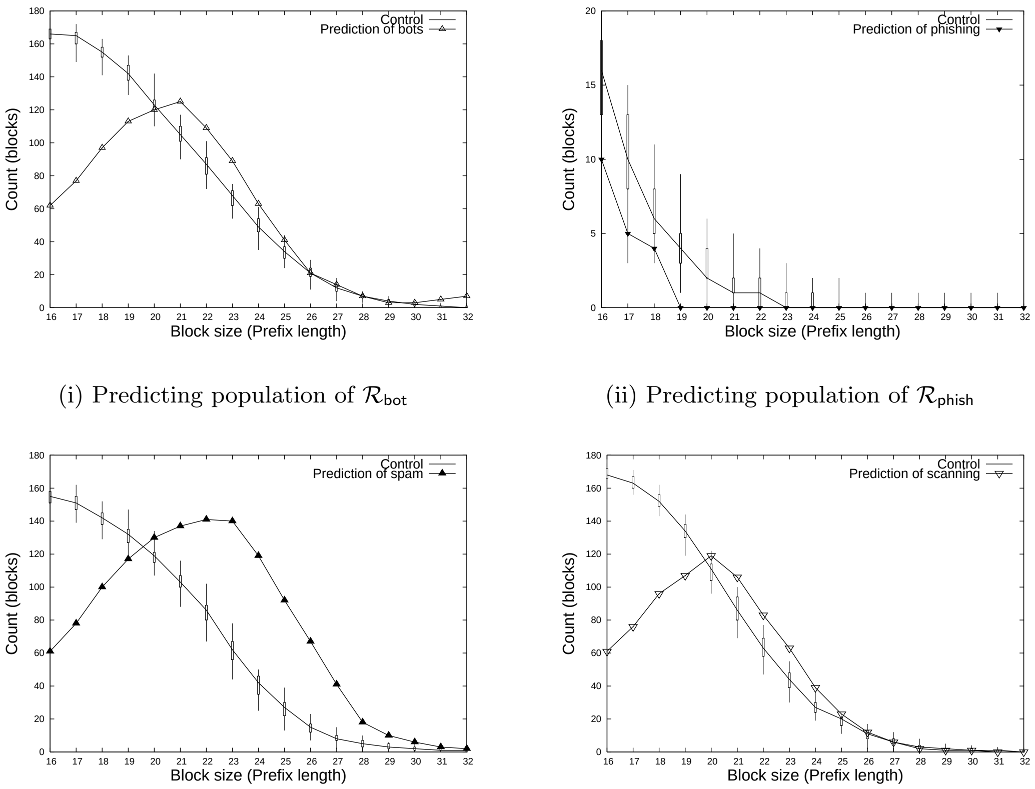
$\mathcal{R}_{\mathsf{unknown}}$ consists of the addresses in $\mathcal{R}_{\mathsf{candidate}}$ address that are *not* present in one of the unclean reports, but have no *payload bearing* flows. We define a flow as payload-bearing if it is a TCP flow with at least 36 bytes of payload and at least one ACK flag. Due to TCP options, a 3-packet SYN scan will often have 36 bytes of payload, even though this data is still part of the TCP handshake. Hand-examination of the flow logs found multiple examples of 36-byte SYN-only scans to apparently randomly selected ports on diverse targets.

The IP addresses in $\mathcal{R}_{\mathsf{unknown}}$ are not proven to be hostile but are highly suspicious. Due to the lack of payload in flow data, we cannot definitively categorize members of this report into either of the other two reports and consequently we remove them from the false positive calculations. For this analysis, we consider the false negative rate to be effectively zero, as we are only considering addresses that we have opted to block.

The population of $\mathcal{R}_{\mathsf{innocent}}$ consequently consists of any IP address that does conduct payload-bearing TCP activity and is not present in any of the unclean reports.

Our prediction scenario assumes that the network blocks $C_n(\mathcal{R}_{\mathsf{bot-test}})$ for some value of $n \in [24, 32]$. The success

[5] `http://www.cisco.com/go/netflow`



(i) Predicting population of $\mathcal{R}_{\mathsf{bot}}$

(ii) Predicting population of $\mathcal{R}_{\mathsf{phish}}$

(iii) Predicting population of $\mathcal{R}_{\mathsf{spam}}$

(iv) Predicting population of $\mathcal{R}_{\mathsf{scan}}$

Figure 4: Comparative predictive capacity of $\mathcal{R}_{\mathsf{bot-test}}$ against control data. Note that $\mathcal{R}_{\mathsf{bot-test}}$ is a better predictor than control data for everything except phishing.

of this defensive mechanism is based on how many hostile and innocent addresses are blocked by the attack mechanism (as noted above, while the unknown population is calculated and analyzed in this exercise, it is not scored). The score for the defensive mechanism is the relative success, measured in true and false positives of the filter as a function of n. We define a false positive as a member of $\mathcal{R}_{\mathsf{innocent}}$ blocked by the filter, and true positive as a member of $\mathcal{R}_{\mathsf{hostile}}$ blocked by the filter.

6.2 Results

We now calculate the success of our blocking mechanism. We emphasize that this is a test of a virtual blocking capacity; we did not actually block addresses, but instead observed the activity engaged in by candidate addresses and evaluated the impact if we had blocked them. To evaluate the efficacy of our defensive mechanism, we use ROC analysis: we compare true positive rates and false positive rates against an operating characteristic of the prefix length used to characterize the networks in $\mathcal{R}_{\mathsf{bot-test}}$.

For these analysis, the true positive rate, is the percentage of blocked IP addresses which were reported in suspicious activity during the observation period. The false positive rate consists of all addresses which were engaged in meaningful activity that were not reported as hostile. Note that, as discussed in §6.1, we define a third category for addresses which communicate with the observed network, but do not exchange any payload.

To calculate the true and false positive rates, we define a membership function, m:

$$m(i,S) = \begin{cases} 1 & C_{32}(i) \sqsubset C_{32}(S) \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

For any prefix length n, we calculate the population as a function of n by summing the unique IP addresses that appear within the $\mathcal{R}_{\mathsf{bot-test}}$

$$\text{pop}(n) = \sum_{i \sqsubset C_n(\mathcal{R}_{\mathsf{bot-test}})} m\left(i, \mathcal{R}_{\mathsf{candidate}} \cap (\mathcal{R}_{\mathsf{innocent}} \cup \mathcal{R}_{\mathsf{hostile}})\right) \tag{7}$$

As noted above, this calculation explicitly avoids the use of $\mathcal{R}_{\mathsf{unknown}}$. We calculate the true positive and true negative values by calculating a similar value over the various reports:

$$\text{TP}(n) = \sum_{i \sqsubset C_n(\mathcal{R}_{\mathsf{bot-test}})} m(i, \mathcal{R}_{\mathsf{candidate}} \cap \mathcal{R}_{\mathsf{hostile}}) \tag{8}$$

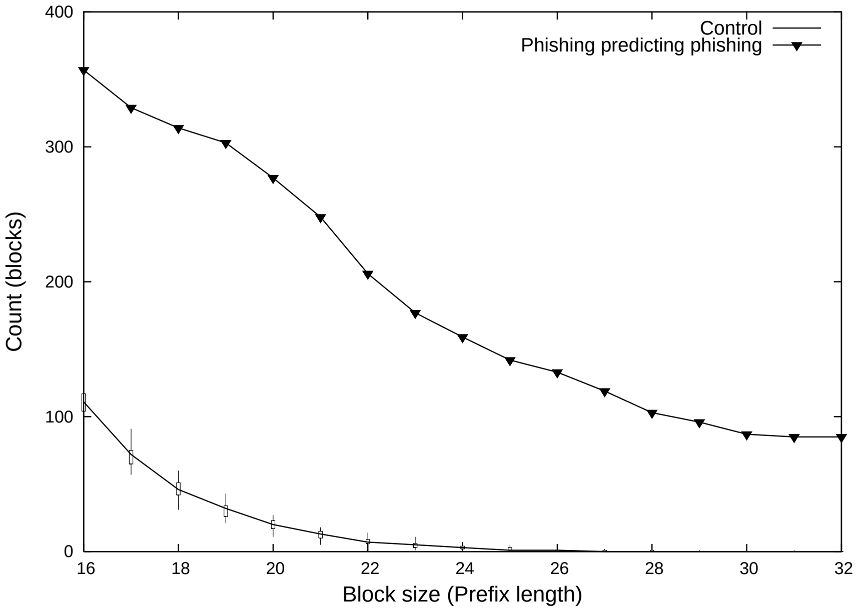


Figure 5: Comparative predictive capacity of phishing reports. Note that this data does effectively predict future traffic, like the bots in Figure 4(i),(iii) and (iv).

$$\mathrm{FP}(n) = \sum_{i \sqsubset C_n(\mathcal{R}_{\mathsf{bot-test}})} m(i, \mathcal{R}_{\mathsf{candidate}} \cap \mathcal{R}_{\mathsf{innocent}}) \qquad (9)$$

Table 3 summarizes the effectiveness of this prediction method. As this table shows, all three populations increase as the bit length increases. At $n = 24$, 90% of the incoming addresses are correctly identified as hostile. If we assume that unknown addresses are hostile, the true positive rate is 97%. Furthermore, the false positive rate remains relatively low until $n = 26$.pa

n	$TP(n)$	$FP(n)$	pop(n)	$\mathcal{R}_{\mathsf{unknown}}$
24	287	35	322	708
25	172	22	194	344
26	81	1	82	200
27	38	1	39	105
28	18	0	18	60
29	7	0	7	29
30	1	0	1	14
31	1	0	1	7
32	1	0	1	0

Table 3: Observed true and false positive counts

Of note with this dataset are the volume of uncertain addresses (i.e., the population of $\mathcal{R}_{\mathsf{unknown}}$). At a 24 bit prefix length, $|C_{24}(\mathcal{R}_{\mathsf{bot-test}}) \cap C_{24}(\mathcal{R}_{\mathsf{unknown}})|$ yields approximately 700 addresses. We first note that unknown addresses have engaged in TCP communications, but have not exchanged payload - consequently, blocking these addresses does not impact traffic.

Of more concern is that all of the addresses in $\mathcal{R}_{\mathsf{unknown}}$ engage in some form of suspicious behavior (that is, suspicious apart from trying to connect with the network and not exchanging payload). Hand examination found many addresses trying to open communications from ephemeral ports to ephemeral ports or engaged in slow scanning. The latter addresses did not appear in our scanning report because the scan detection mechanism is calibrated to identify scans that take place over an hour, while scans observed in this dataset would often contact less than 30 addresses per day over the observation period.

The strength of this blocking method is predicated on the relatively sparse amount of traffic issuing from these blocks. As Table 3 shows, 1030 IP addresses were blocked when n was set to 24 bits. $|C_{24}(\mathcal{R}_{\mathsf{bot-test}})| = 173$, which yields a potential set of 44,288 address that can be blocked. Consequently, less than 2% of the total IP addresses available in those /24s communicated with the observed network during this time.

Some of the effectiveness of this method may be attributed to the demographics of the botnet and the observed network $\mathcal{R}_{\mathsf{bot-test}}$ consists primarily of addresses outside the English-speaking world, with 70% of the addresses coming from Turkey. Despite its size, the observed network an edge network; all traffic at its border is either originating from an address within that border or going to an IP address within that border.

We therefore conclude that our test results indicate the

Reports used for prediction testing					
Tag	Type	Class	Valid Dates	Size	Reporting method
unclean	Provided	Special	2006/10/01-2006/10/14	1,158,103	The union of the four unclean reports, note that there is overlap
candidate	Observed	N/A	2006/10/01-2006/10/14	1030	IP Addresses crossing the network border and that are in the same /24's as $\mathcal{R}_{\mathsf{unclean}}$
hostile	Observed	N/A	2006/10/01-2006/10/14	287	Members of $\mathcal{R}_{\mathsf{candidate}}$ also present in $\mathcal{R}_{\mathsf{unclean}}$
unknown	Observed	N/A	2006/10/01-2006/10/14	708	Members of $\mathcal{R}_{\mathsf{candidate}}$ not in $\mathcal{R}_{\mathsf{unclean}}$, but engaged in suspicious activity
innocent	Observed	N/A	2006/10/01-2006/10/14	35	Members of $\mathcal{R}_{\mathsf{candidate}}$ not present in $\mathcal{R}_{\mathsf{hostile}}$ or $\mathcal{R}_{\mathsf{unknown}}$

Table 2: Table of reports used for prediction test.

feasibility of blocking hostile addresses, but that this approach is best used in conjunction with other traffic analysis mechanisms in order to determine the best practices for individual networks.

7. CONCLUSION

In this paper, we have demonstrated that it is possible to effectively predict future hostile activity from past network activity. To do so, we have defined a network-based quality of uncleanliness, which is an indicator of how likely a network is to contain compromised hosts.

As an initial work in this field, we have focused on testing basic hypotheses about uncleanliness, which we have defined with the spatial and temporal uncleanliness hypotheses. Using reports of network activity and traffic logs of a large network we have shown evidence of spatial and temporal uncleanliness. We have also shown that an uncleanliness measure may involve multiple dimensions, such as botnets and phishing.

Finally, we have demonstrated that spatial and temporal uncleanliness, coupled with the limited audience of an edge network, can be effectively used to block hostile traffic in the future. Given the demographics issues noted in §6, uncleanliness may best be used as a risk indicator – by showing that a network is demonstrating unclean behavior, security personnel can evaluate whether the risk of hostile activity from the network is worth the benefit of receiving commerce and communication from that network under normal circumstances.

Our immediate goal following this work is to develop a more rigorous and precise uncleanliness metric. In particular, a multidimensional uncleanliness metric to measure the aggregate probability that an address is occupied. The elements of this metric involve the components discussed in this work as well as other predictive indicators of vulnerability (communication with botnet C&C nodes).

We also believe that spatial uncleanliness has useful implications for network log analysis. If we know that a host from one network is attacking, scanning or otherwise interfering with the traffic on an observed network, it is reasonable to examine other traffic from that network to see if there is coordinated hostile activity.

8. ACKNOWLEDGEMENTS

We would like to thank the referees and our shepherd for their insightful comments in preparing this paper.

9. ADDITIONAL AUTHORS

Additional author: Joseph B. Kadane (CMU Department of Statistics, email: `kadane@stat.cmu.edu`)

10. REFERENCES

[1] CastleCops. Castlecops phishing incident reporting & termination (PIRT) squad. Accessible at http://www.castlecops.com/pirt, fetched on January 29th, 2007.

[2] M. Collins, C. Gates, and G. Kataria. A model for opportunistic network exploits: The case of P2P worms. In *Proceedings of the 2006 Workshop on Economics and Information Security*, 2006.

[3] M. Collins and M. Reiter. An empirical analysis of target-resident DoS filters. In *Proceedings of the 2004 IEEE Symposium on Security and Privacy*, 2004. May 9 – 12, 2004.

[4] D. Cook, J. Hartnett, K. Manderson, and J. Scanlan. Catching spam before it arrives: domain specific dynamic blacklists. In *ACSW Frontiers '06: Proceedings of the 2006 Australasian workshops on Grid computing and e-research*, Darlinghurst, Australia, Australia, 2006.

[5] F. Freiling, T. Holz, and G. Wicherski. Botnet tracking: Exploring a root-cause methodology to prevent distributed denial-of-service attacks. In *Proceedings of the 2005 European Symposium on Research in Computer Security*, 2005.

[6] C. Gates, J. McNutt, J. Kadane, and M. Kellner. Detecting scans at the ISP level. Technical Report CMU/SEI-2006-TR-005, Software Engineering Institute, 2006.

[7] C. Gates, J. McNutt, J. Kadane, and M. Kellner. Scan detection on very large networks using logistic regression modeling. In *ISCC '06: Proceedings of the 11th IEEE Symposium on Computers and Communications*, Washington, DC, USA, 2006.

[8] T. Holz. Learning more about attack patterns with honeypots. In *Sicherheit 2006: Sicherheit - Schutz und Zuverlässigkeit, Beiträge der 3. Jahrestagung des*

Fachbereichs Sicherheit der Gesellschaft für Informatik e.v. (GI), 20.-22. Februar 2006 in Magdeburg, 2006.

[9] T. Holz, S. Marechal, and F. Raynal. New threats and attacks on the world wide web. *IEEE Security & Privacy*, 4(2), 2006.

[10] J. Jung, B. Krishnamurthy, and M. Rabinovich. Flash crowds and denial of service attacks: Characterization and implications for CDNs and web sites. In *Proceedings of the International World Wide Web Conference*, May 2002.

[11] J. Jung, V. Paxson, A. Berger, and H. Balakrishnan. Fast Portscan Detection Using Sequential Hypothesis Testing. In *IEEE Symposium on Security and Privacy 2004*, Oakland, CA, May 2004.

[12] J. Jung and E. Sit. An empirical study of spam traffic and the use of DNS black lists. In *IMC '04: Proceedings of the 4th ACM SIGCOMM conference on Internet measurement*, New York, NY, USA, 2004.

[13] E. Kohler, J. Li, V. Paxson, and S. Shenker. Observed structure of addresses in IP traffic. *IEEE/ACM Transactions on Networking*, 14(6), 2006.

[14] B. Krishnamurthy and J. Wang. On network-aware clustering of web clients. In *Proceedings of the 2000 ACM Special Interest Group in Communications SIGCOMM Conference*, 2000.

[15] B. Laurie and R. Clayton. Proof-of-work proves not to work. In *Proceedings of the 2004 Workshop on Economics and Information Security*, 2004.

[16] E. Levy. The making of a spam zombie army: Dissecting the sobig worms. *IEEE Security and Privacy*, 1(4), 2003.

[17] John McHugh and Carrie Gates. Locality: A new paradigm for thinking about normal behavior and outsider threat. In *Proceedings of the 2003 New Security Paradigms Workshop*, Ascona, Switzerland, 2003. August 18 – 21, 2003.

[18] J. Mirkovic, G. Prier, and P. Reiher. Attacking DDoS at the source. In *ICNP '02: Proceedings of the 10th IEEE International Conference on Network Protocols*, Washington, DC, USA, 2002.

[19] K. Plößl, H. Federrath, and T. Nowey. Protection mechanisms against phishing attacks. In *Proceedings of the second annual conference on Trust, Privacy and Security in Digital Business*, volume 3592 of *Lecture Notes in Computer Science*, August 2005.

[20] The Spamhaus Project. Zen blocklist. Available at `http://www.spamhaus.org/zen`, Fetched on January 29^{th},2007.

[21] M. Rajand, J. Zarfoss, F. Monrose, and A. Terzis. A multifaceted approach to understanding the botnet phenomenon. In *Proceedings of the 2006 ACM Internet Measurement Conference*, 2006.

[22] A. Ramachandran, N. Feamster, and D. Dagon. Revealing botnet membership using DNSBL counter-intelligence. In *Proceedings of the 2006 USENIX workshop on steps for reducing unwanted traffic on the internet (SRUTI)*, 2006.

[23] Bleeding Edge Threats. Bleeding snort ruleset. Available at `http://www.bleedingsnort.com/index.php/about-bleeding-edge-threats/all-bleeding-edge-threats-signatures/`, Fetched on January 29th, 2007.

[24] P. Walt. Agencies feel botnets' light footprint. *Government Computer News*, January 2007.