Gray's Anatomy: Dissecting Scanning Activities Using IP Gray Space Analysis

Yu Jin, György Simon, Kuai Xu, Zhi-Li Zhang, Vipin Kumar
Department of Computer Science, University of Minnesota
{yjin,gsimon,kxu,zhzhang,kumar}@cs.umn.edu

***Abstract*— In this paper, we study the scanning activities towards a large campus network using a month-long netflow traffic trace. Based on the novel notion of "gray" IP space (namely, collection of IP addresses within our campus network that are not assigned to any "active" host during a certain period of time), we identify and extract potential outside scanners and their associated activities. We then apply data mining and machine learning techniques to analyze the scanning patterns of these scanners and classify them into a few groups (e.g., focused hitters, random address scanners, and blockwise scanners). The goal is to infer the scanning strategies of the scanners so as to provide some assessment of the potential harmfulness of these scanning activities – for example, whether the observed scanning activities are simply part of background radiation of global random scanning or more focused scanning targeted at our campus network. This is an on-going work; we report some preliminary, yet promising results obtained so far.**

I. Introduction

Cyber attackers often resort to scanning for reconnaissance – looking for certain services or hosts with certain vulnerabilities to attack or compromise – and for spreading malware such as worms, viruses or spams. From the perspective of a *campus* or *enterprise network*, while many scanning activities observed on such a network may simply reflect the "background radiation" [1]–[3] of various Internet-scale random scanning activities, remnants of past worm/virus outbreaks, or other malware activities on the *global* Internet at large, some could be results of more *targeted* reconnaissance or stealthy attack activities aimed specifically at the said campus/enterprise network. The latter is particularly worrisome and warrants closer scrutiny. In either case, monitoring and analyzing scanning activities is a crucial component in network intrusion and prevention, both to protect individual networks against malicious outside attacks and to mitigate and stop global outbreaks of worms, viruses and other malware in their early stages. An important goal of the analysis of scanning activities is to infer and uncover the scanning strategies and intentions of cyber attacks, the knowledge of which can guide us devise more effective monitoring, detection and defense mechanisms against cyber attacks. This, clearly, is a daunting task, given the very limited information we have regarding the observed scanning activities.

This paper constitutes a modest step towards this challenging goal. We apply data mining and machine learning techniques to identify and classify various scanning activities observed on our campus network, and explore a number of features to characterize and infer scanning patterns and strategies of outside scanners that might reveal some useful clues about the "intentions" of the scanners. In particular, we are interested in answering the following question: are observed scanning activities likely mere remnants of some global "background radiation," or reflections of malicious actions that include our own network as part of more specific targets?

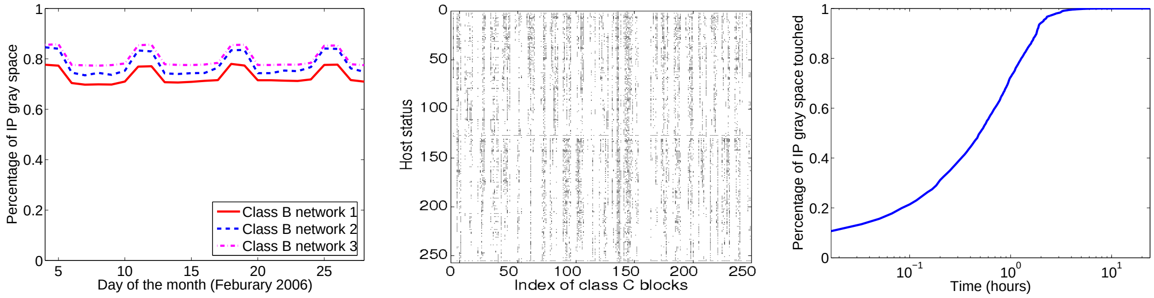
To address this question, we propose a novel technique, *IP gray space analysis*, based on the notion of "gray" IP addresses. Intuitively, gray IP addresses are those within a (campus/enterprise) network that are not assigned to any live host for the entire duration of a given time period, say, a particular day, the collection of which is referred to as the IP *gray space* of the network. By definition, any incoming traffic towards gray IP addresses is "unwanted," and thereby potentially "suspicious." Because gray IP addresses are in general randomly distributed within a network, it is extremely hard for an outside scanner to predict and thus avoid them. We use this key observation to develop several heuristics (see Section II) for identifying outside scanners that engage in "sustained" scanning activities. We then apply data mining/machine learning techniques to study their scanning behaviors. In Section III we classify scanners into three categories, *focused hitters*, *random address scanners* and *blockwise scanners*, by analyzing their address selection strategies. In Section IV we explore several features to further investigate the observed scanning patterns (or "footprints") of scanners, and attempt to infer whether the observed scanning activities are merely part of global "background radiation" or reflections of actions that likely target more specifically at our own network. The ultimate (and perhaps unattainable) goal is to infer the plausible "intentions" of scanners and assess the potential harmfulness of their actions. As part of on-going research towards this goal, we report some preliminary, yet promising results.

II. Identifying Scanners via Gray Space Analysis

In this section, we first introduce the novel notion of *gray* IP addresses, and present a simple heuristic to extract gray IP addresses – collectively referred to as the (IP) *gray space* (of a given network) – from our month-long netflow traces captured from our campus network. Then, we describe an algorithm for identifying potential scanning traffic by analyzing the activities on the IP gray space.

A. Extracting the IP Gray Space

We first present a formal definition of a *gray* IP address. Let I denote the collection of all IP addresses of a network



(a) Size of gray space over one month (b) Gray address distribution (c) Speed of gray space being touched

Fig. 1. IP gray space properties.

under consideration, t_0 the starting time of a time period of interest, and T the length of the period. We say that an (inside) IP address $g \in I$ is a *gray* (or inactive) address over the time period $[t_0, t_0 + T]$ if and only if no traffic *originating* from g is observed during $[t_0 - \tau, t_0 + T + \tau]$ for some fixed τ[1]. We use G to denote the collection of all gray IP addresses within the time period, i.e., the IP *gray space* of the network during the time period $[t_0, t_0 + T]$. The complementary set, $A = I - G$, is referred to as the *active space*. In other words, for any $a \in A$, there is traffic originating from a at some time during $[t_0 - \tau, t_0 + T + \tau]$; thus a is likely assigned to an active host during the time period. In this study, we set T to be 24 hours, t_0 the zeroth hour of a day, and τ one hour.

We apply the heuristic above to the netflow data collected at the border router of the University of Minnesota campus network during Feb. 2006. The data set includes all bi-directional traffic flows between inside hosts and outside hosts during the entire month.

Our campus network owns three class B (/16) IP address blocks, with a total $3 \times 2^{16} = 196608$ IP addresses. We found that each day in Feb. 2006, over 70% of the addresses are *gray* ("inactive") over the entire day. Fig. 1(a) depicts the percentage of gray IP addresses each day for the entire month of Feb. 2006. To illustrate how the gray IP addresses are distributed among the IP address blocks of the campus network, Fig. 1(b) illustrates the distribution of gray IP addresses in the 256 "class C" (/24) address sub-blocks within one of the class B (/16) address blocks. The x-axis represents each class C sub-block, while y-axis represents each host in a corresponding sub-block. A point on the graph stands for an active host on 2/6/2006. All the blank space belongs to IP gray space. (The gray IP address distribution looks similar also for the other two class B address blocks of our campus network.) We observe that the gray IP addresses spread over the entire class B address block, and they are unevenly distributed among different /24 address sub-blocks: although quite a number of /24 address sub-blocks are entirely gray, overall the *gray percentage* (i.e., the percentage of gray IP addresses in a /24 address sub-block) varies across the address sub-blocks, with a few having a gray percentage below 10%.

[1] In this definition, to be conservative, we require that there is also no traffic originating from g for a period of τ before and after the time period of interest to provide additional assurance that g is indeed unlikely to be assigned to any host over the said time period.

Although the size of the IP gray space constitutes over 70% of the total IP address space of the campus network, the IP gray space does vary from day to day: some gray IP addresses become active from one day to another, while others change from active to gray. About 57% of the IP addresses stayed gray in the entire one month of Feb. 2006. Furthermore, we observe that despite the fact that gray IP addresses do not generate any traffic to an outside host throughout an entire day, they invariably receive traffic from outside hosts. In fact, we observe that typically within a few hours from the start (zeroth hour) of a day, all gray IP addresses are "touched" by at least one outside host! Fig. 1(c) shows the percentage of gray IP addresses touched by at least one outside host as time goes by on Feb. 06, 2006 – in less than four hours all gray IP addresses are touched by an outside host. Moreover, on 2/6/2006, nearly 360K outside hosts touch at least one gray IP address inside our campus network!

B. Identifying Scanners and their Scanning Activities

Given the extracted IP gray space, we apply a *two-step* process to identify potential scanners and their associated scanning activities. The ideas behind this two-step process are as follows. Intuitively, without any knowledge of the IP gray space of a network, an outside scanner that generates sufficient scanning traffic would inevitably touch one or more gray IP addresses. Hence we consider any incoming flow that touches any gray IP address (referred to as a *gray flow*) as potentially suspicious and the outside host generating such a gray flow as a potential scanner. Among them, we narrow down to those with *sustained suspicious activities*–those that generate enough traffic towards our campus network, of which a considerable portion touching the IP gray space–and look for whether certain ports (either destination or source ports) are used *repeatedly* in those suspicious activities from an outside host. These ports, referred to as *dominant scanning ports* (DSPs), represent the likely services or exploits (i.e., ports with vulnerabilities) that the outside host is interested in and is thus scanning for. Using these DSPs, we can then separate the scanning activities of the said outside host from other (if any) traffic from the same host: this is done by excluding any incoming flow from the outside host that does not use any of the DSPs as the corresponding source/destination ports. The details of the two-step process are given below.

In this study, we regard outside hosts that generate at least 100 incoming flows over a day, 10% of which are gray flows as those with sustained suspicious activities[2]. We use the notation O_s to denote the collection of these hosts. (For example, for the day of 2/6/2006 $|O_s| = 7468$.) For each outside host $h \in O_s$, let $GF(h)$ denote the collection of gray flows generated by h. The destination ports (***dstPrt*** in short) used by gray flows in $GF(h)$ induce an empirical distribution: for each ***dstPrt*** i, $p_i := m_i/m$ where m_i is the number of gray flows in $GF(h)$ with ***dstPrt*** i, and m is the total number of gray flows in $GF(h)$, $m = |GF(h)|$. We apply an information theoretical metric, *Relative Uncertainty* (RU) [4], which provides a measure of variety, uniformity or randomness of a distribution, to determine and identify dominant scanning (destination) ports (if they exist). Using the above notations, $RU(dstPrt)$ is given by

$$RU(dstPrt) := \frac{-\sum_{i \in dstPrt} p_i \log p_i}{\log m} \in [0,1], \tag{1}$$

where $RU(dstPrt)$ close to 0 suggests one or a few ***dstPrt***'s dominate in the gray flows of an outside host h; while $RU(dstPrt)$ close to 1 signifies that there is no dominant ***dstPrt***'s. Similarly, we can define the relative uncertainty, $RU(srcPrt)$, for the source port (***srcPrt***) distribution of $GF(h)$.

To illustrate how $RU(srcPrt)$ and $RU(dstPrt)$ can be used to determine the existence of DSP's in the gray flows of an outside host, we use the flow data of 2/6/2006. Fig.2(a) shows a 3-D plot with $RU(srcPrt)$ and $RU(dstPrt)$ in the x-y plane, and the z-axis showing the number of hosts in O_s with a given RU pair (x, y). We see that nearly 99% of all hosts in O_s have either dominant scanning *destination* or *source* ports (with either $RU(dstPrt)$ or $RU(srcPrt) \leq 0.3$). Algorithm 2 presents a heuristic procedure for extracting DSP's from either the destination or source port distribution P_{prt} of host $h \in O_s$. The algorithm starts with an empty DSP set. It iteratively finds the port with the current highest probability, adds the port into DSP and removes all the flows associated with it from $GF(h)$. The algorithm terminates until either the number of the remaining $GF(h)$ is less than 10 or $RU(prt)$ in $GF(h)$ is greater than β_0 (we choose $\beta_0 = 0.7$). In other words, the algorithm stops when there are not enough flows left or the ports in the rest of the flows are nearly uniformly distributed. Fig.2(b)(c) show the top 20 DSP source ports and destination ports extracted using this algorithm and their frequency, which include ICMP scanning (port 0) and well-known exploit (UDP/TCP) ports such as 137, 139, 445, 1025, 1026, 1434 as well as a few popular service ports such as 25, 80, 443.

[2]Clearly these numbers are somewhat arbitrary. Our analysis shows that a small portion of outside hosts generate a large portion of gray flows. For example, on 2/6/2006, although only 2% of the outside hosts generate more than 100 flows, of which 10% touching the IP gray space, they contribute to 98% of the total gray flows.

Algorithm 1 Identifying dominant scanning ports

Parameters: $GF(h)$; $\beta = \beta_0$;
Initialization: $DSP := \emptyset$;
compute pro. dist. P_{prt} and $\theta := RU(prt)$ from $GF(h)$;
while $\theta \leq \beta$ and $|GF(h)| >= 10$ **do**
  find prt_i with highest P_{prt_i};
  $DSP := DSP \cup prt_i$;
  remove flows associate with prt_i from $GF(h)$;
  remove P_{prt_i} from P_{prt};
  compute $\theta := RU(prt)$ from $GF(h)$;
end while

Given the source or destination DSP's identified using the gray flows of an outside host, in the second step, we use them to separate incoming flows touching the active space (i.e., incoming *active* flows) from the same host that are likely involved in the same scanning activities from other incoming active flows of the host. The intuition here is that since an outside attacker does not know which IP addresses of a network are gray or active, the scanning flows he or she generates using the DSP's may also touch portion of the active space. For each $h \in O_s$, we consider any incoming active flow from h with any of the dominant scanning source or destination ports as part of the scanning activities of the outside scanner. For $h \in O_s$, we use $SF(h)$ to denote the set of the scanning flows of h, which include both the *active* and *gray* flows of h that use the ports in its DSP's. We use $OF(h)$ to denote the remaining *active* (only!) flows of h – referred to as *other flows* of h. We define the scanning flow ratio of h as $\gamma_s = |SF(h)|/(|SF(h)| + |OF(h)|)$, which indicates how dominant the scanning flows are in the outside host's interaction with the network in question. Fig.2(d) plots the cumulative distribution of γ_s for all 7468 hosts in O_s using the flow data of 2/6/2006. For nearly 80% hosts in O_s, $\gamma_s = 1$, suggesting that these hosts have no other meaningful interaction with hosts inside our campus network *except for generating scanning traffic touching both the gray and active IP addresses of the network*. The remaining 1166 hosts have a varying mixture of scanning flows and other flows.

III. Classifying Scanners

With the heuristic in the previous section, we extract totally 111,179 distinct scanners in one month (around 7000 every-day). In this section, we investigate their scanning behaviors in detail. Based on different address selection strategies, we classify the outside scanners into *focused hitters*, *random address scanners* and *blockwise scanners* as described below.

A. Focused Hitters vs. Random Scanners

An outside scanner could be scanning hosts inside our campus network randomly, searching for certain services or vulnerabilities; or he/she could be interested in a group of specific hosts (e.g., web or email servers) that were obtained through other information (e.g., DNS, URLs), and probe them for their aliveness or open ports. To distinguish these two types of scanning activities, we look at the distribution of the targets (i.e., inside hosts) touched by the scanning traffic in

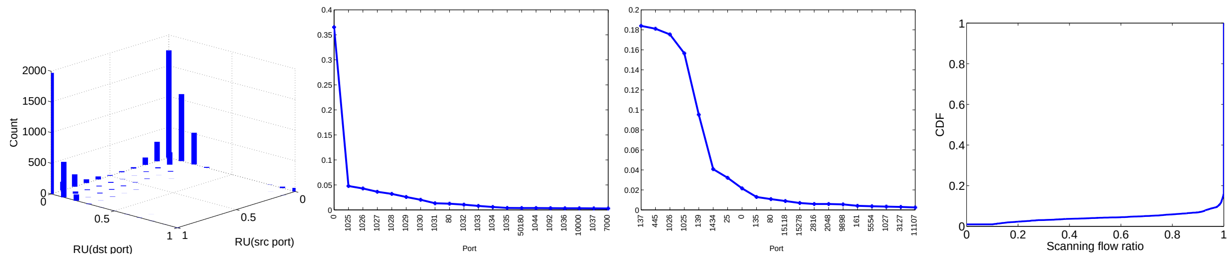


Fig. 2. (a) $RU(srcport)$ and $RU(dstport)$ distributions, (b) and (c) top 20 DSP source and destination ports, and (d) CDF of scanning flow ratios for outside hosts in O_s on 2/6/2006

$SF(h)$ for each outside scanner h. For this, we measure the *relative uncertainty* (RU) of the *dstIP*'s of flows in $SF(h)$, $RU(dstIP)$, defined analogously using Eq.(1).

Fig. 3 shows the distribution of $RU(dstIP)$ in $SF(h)$, where we see that a large majority of outside scanners have $RU(dstIP) >= 0.8$ (6650 out of 7468), namely, they randomly touch each target once or a few times. We refer to them as *random scanners*. For the remaining hosts (818), we find that most of these scanners repeatedly probe a small number (typically fewer than 20) of inside hosts, seemingly to test whether they are alive. We call these scanners *focused hitters*. From the DNS records, we find that many of the *gray* IP addresses touched by these focused hitters have hostnames indicating that they were email servers, web servers, etc., but for whatever reasons, were out of service on 2/6/2006. More in-depth investigation of these focused hitters (e.g., via the reverse DNS lookup and public spammer database queries) reveals that most of these scanners are email spammers probing for email servers, while others are web crawlers and p2p hosts. A detailed study of the focused hitters can be found in [5].

B. Random Address Scanners vs. Blockwise Scanners

For the remainder of this paper, we focus our attention on random scanners, analyzing their scanning strategies and attempting to infer their intentions, in particular, whether observed scanning activities are merely reflections of "background radiation," or something more sinister that target more specifically at our campus network. As a first step, in this subsection, we analyze the *address selection strategies* using the "footprints" of the random scanners. In particular, we are interested in determining how an outside scanner selects addresses inside our network for scanning.

We define the *gray ratio* of a network (e.g., the UMN network) as the total number of its gray addresses divided by the total number of its active addresses (over a given time period, say, a day). Given the random distribution of the gray IP addresses within our campus network, if an outside scanner randomly chooses addresses to scan, the *expected observed gray ratio* of the addresses appearing in the scanning sequence (or "footprints") is likely to be approximately the gray ratio of the entire UMN network, which is 2.83 on 2/6/2006.

Fig. 4 shows the total number of active hosts vs. gray hosts touched by each scanner on 2/6/2006. We see that while most of the points are close to the line $y = 2.83x$, there exist a number of "outliers", suggesting that in addition to purely random address selection, some other address selection strategies are also used. As will be shown below, we can separate random scanners primarily into two categories: *random address* scanners who select addresses randomly from a target address space (e.g., a class B address space of the UMN network as is used in this paper), and *blockwise* scanners who first select sub-blocks or subnets (say, between /32 and /16) within a target address space, and then either randomly or sequentially access addresses in each sub-block. In the followings we present a formal method for separating these types of random scanners. The basic idea is that if a random scanner chooses addresses randomly from the entire target space, then these addresses must also appear randomly selected from *any subspace* of the target space. We apply this idea to separate blockwise scanners from random address scanners.

Consider a target address space (say, a /16 class B address within the UMN network) with N addresses. For a fixed block size s, where $16 \leq s \leq 32$, let M_s denote the number of $/s$ address blocks (or $/s$ subnets) of the target space (given a class B target space and $s = 24$, there are $M_s = 256$ sub-blocks of /24). Denote an observed scanning sequence of a scanner as $a_1, a_2, \cdots, a_n$. For $i = 1, \ldots, n-1$, define $f(a_i, a_{i+1}) = 0$ if a_i and a_{i+1} belong to the same $/s$ block, 1 otherwise. Then the *observed average block difference* (w.r.t. s) of the scanning sequence is $D_s = \sum_{i=1}^{n-1} \frac{f(a_i,a_{i+1})}{n-1}$. Note that if a random scanner selects addresses randomly from the entire target address space, then for any s, $16 \leq s \leq 32$, the probability that two consecutive addresses a_i and a_{i+1} in the scanning sequence belong to the same $/s$ address block is $1/M_s$. Hence the *expected* average block difference of *random address* scanners is $E[D_s^{ras}] = 1 - 1/M_s = (M_s - 1)/M_s$. Hence, for a *random address* scanner, regardless of the block size s, the difference between D_s (the observed average block difference) and $E[D_s^{ras}]$ should likely be small. Therefore we define the *block difference deviation* (BDD) as follows and use it to separate random address scanners and blockwise scanners:

$$BDD = \sum_{s=16}^{32} \sqrt{\frac{(D_s - E[D_s^{ras}])^2}{16}}.$$

Fig. 5 shows the distribution of BDD's of random scanners on 2/6/2006, which is strongly *bi-modal* and clearly separates the random address scanners and blockwise scanners. From

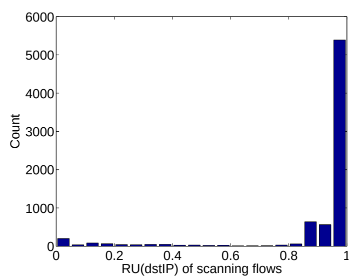


Fig. 3. Scanning flows $RU(dstIP)$ distribution

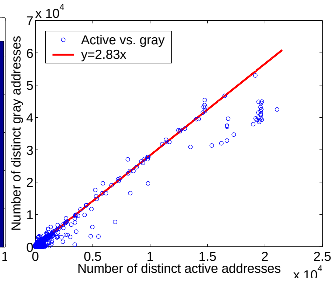


Fig. 4. No. of active addresses vs. no. of gray addresses

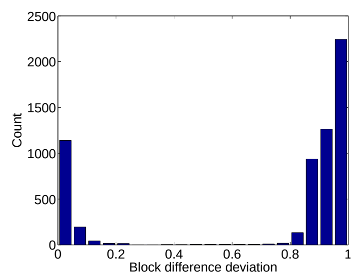


Fig. 5. Distribution of BDD

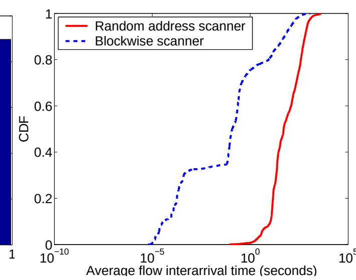


Fig. 6. CDF of average interarrival time

the figure, we can classify random scanners with $BDD \leq 0.7$ as random address scanners, and those with $BDD > 0.7$ as blockwise scanners. Replotting the gray ratio (numbers of active vs. gray hosts) for the random address scanners again as in Fig. 4 shows that we have an almost perfect fit with the line $y = 2.83x$, with nearly all outliers removed. Due to space limitation, the figure is not shown here.

IV. Inferring "Harmfulness" of Scanners

In this section, we try to infer whether an observed scanning activity is likely a mere reflection of "background radiation", or plausibly targets more specifically our campus network. We explore a number of features to help us address this question, and use them to indirectly assess the potential "harmfulness" of the scanning activities. This is still on-going research, here we report some preliminary yet promising results.

TABLE I

No. of /16 networks accessed by different types of scanners

	Total	#B=1	#B=2	#B=3
Random address scanner	1378	518	30	830
Blockwise scanner	4672	4033	492	147

Since the UMN campus network consists of three *separate* class B address blocks (128.101/16, 134.84/16, 160.94/16), we treat them as if they were three separate "telescopes" or vantage points for monitoring the global Internet activities. Intuitively, if a scanner performs sustained scanning activities targeting the global Internet (instead of the UMN network), we are more likely to observe them at more than one vantage point. Furthermore, for the random address scanners targeting the global Internet, the time lapse (i.e., interarrival time) between two consecutive observed scans (from the same scanner) would in general be relatively large, reflecting the speed of scanners. For blockwise scanners that target the global Internet, the time lapse between two consecutive visits of different address sub-blocks would be relatively large. In other words, the scanning sequence of a blockwise scanner can be separated into distinct *phases*, each touching one address sub-block, with relatively large time gaps between them.

Table I shows the number of class B subnets touched by different types of scanners on 2/6/2006. We see that a predominant majority of blockwise scanners touch only one class B subnet, while more than half of random address scanners touch all 3 of them. Fig. 6 shows the distribution of the average interarrival time for random address scanners and blockwise scanners. Almost all the random address scanners have an average interarrival time greater than 1 second, whereas only 20% of the blockwise scanners have an average interarrival time greater than 1 second. For blockwise scanners, we further check whether their scanning sequences can be separated into distinct subsequences or *phases*, with relatively large time gaps between them. For this, we model (each phase of) a scanning sequence as a Poisson process with its interarrival time represented by an exponential random variable X, where $P(X > t) = e^{-\lambda t}$, and λ represents the scanning speed. To determine and separate the phases, we do a hypothesis test on the interarrvial time X by choosing a threshold T s.t. $P(X > T) < \alpha$, where $\alpha = 0.01$. If $X > T$, then the chance of X being a normal interarrival time is below 1%. In other words, there is a statistically large gap between two consecutive scans, and thus we regard them as separation of two distinct scanning phases. (To eliminate possible false positives, we also ensure that those two scans belong to two different address sub-blocks.) Table II shows the total number of class B subnets touched by *single-phase* vs. *multi-phase* blockwise scanners. Almost all the scanners who access more than one class B networks are multi-phase scanners. Examining the distribution of the average time lapse between phases of multi-phase scanners (Fig. 6), we see that majority of them have an average *inter-phase time lapse* of more than 1 second, consistent with that of random address scanners.

TABLE II

No. of /16 networks accessed by different types of scanners

	Total	#B=1	#B=2	#B=3
Single-phase scanners	1367	1364	3	0
Multi-phase scanners	3305	2669	489	147

Combining the above observations, we find that observed scanning activities of random address scanners are likely reflections of background radiation. This conclusion is also borne out by the port information: nearly all ports used are known exploit ports such as slammer (UDP port 1434) or Dabber worm (TCP port 9898). Among the blockwise scanners, the multi-phase blockwise scanners are also likely reflections of background radiation, and again a large majority of the ports used contain known exploits. On the other hand, single-phase blockwise scanners almost always touch only one class B subnet, and are plausibly targeting only our network.

TABLE III

FEATURE DISTRIBUTIONS OF SCANNERS ON NON-BLOCKED PORTS

	Total	Responses	Other Flows	Overlapping Targets	Re-visits (2/7/2006)	Re-visits (2/8/2006)
Random address scanners	239	35 (14.6%)	92 (38.5%)	41 (17.2%)	126 (52.7%)	102 (42.7%)
Single-phase blockwise scanners	10	6 (60%)	7 (70%)	6 (60%)	1 (10%)	0 (0%)
Multi-phase blockwise scanners	96	46 (47.9%)	58 (60.4%)	50 (52.1%)	32 (33.3%)	32 (33.3%)

To further assess the plausible intentions and potential harmfulness of outside scanners, we consider several additional features that go beyond the analysis of the "footprints" of the scanners. These features include i) the existence of likely *responses* elicited by scanning flows, ii) (potential) *follow-up activities* as measured by both the existence of *other* flows (to active IP addresses) in addition to the scanning flows from an outside scanner and the existence of *overlapping targets* between the scanning flows and other flows of the scanner, and iii) *re-visits* of outside scanners in other days. For these features, we consider only observed scanning activities on *ports that are not blocked* by our campus network. Our campus network blocks 74 ports that solely corresponding to reported worms and other exploits, which in fact include a majority of scanning activities observed. In other words, scanning activities on these blocked ports will not elicit any response from any inside (live) host, and nor will any follow-up activity ensue as a result. There are totally 345 random scanners with sustained scanning activities on non-blocked ports that are observed on 2/6/2006. We hence attempt to infer and assess the potential harmfulness of these scanning activities. To measure the responses, we define a response to a TCP scan as an outgoing flow (from an active inside IP address) that matches the TCP scan, and contains at least 3 packets with the average packet size greater than 48 (we want to filter pure RST responses); while we define a response to a UDP scan as an outgoing flow matching the UDP scan.

Among the 345 random scanners on non-blocked ports, 239 of them are random address scanners, 106 are blockwise scanners, of which 96 are multi-phase blockwise scanners, and 10 being single-phase scanners. Table III tabulates the existence of responses, existence of other flows, existence of overlapping targets (of scanning and other flows), and the number of the scanners observed again (i.e., revisits) on 2/7/2006 and 2/8/2006 among the three types of random scanners. We see that random scanners on non-blocked ports are less likely to elicit responses, and they are also less likely to generate other flows (possibly "follow-up" activities) with overlapping targets than blockwise scanners. The latter are more likely to elicit responses, since they scan more extensively within one address sub-block, thus more likely to touch a live host. Likewise, they are more likely to generate other traffic to the same (live) hosts for possible "follow-up" activities. Hence, blockwise scanners appear to be more dangerous, thus warrant further scrutiny. In particular, the single-phase blockwise scanners seem to target specifically our campus network, with the highest likelihood of eliciting responses and generating follow-up activities with overlapping targets. In addition, random address scanners and multi-phase blockwise scanners tend to revisit our campus networks again during subsequent days scanning on the same or different ports, possibly due to the fact that these scanning hosts are infected with malware or part of botnets that perform repeated global scanning activities. Only one single-phase blockwise scanner was observed again on 2/7/2006 scanning different port, and none on 2/8/2006.

V. CONCLUSIONS AND FUTURE WORK

In this paper we studied the scanning activities towards a large campus network using a month-long netflow data, with the goal to infer the scanning strategies and "intentions" of scanners and thereby assess the "harmfulness" of their actions. Towards this goal, we introduced the notion of IP gray space, and developed a novel technique–IP gray space analysis–to identify potential scanners and study their scanning behaviors. In particular, we applied data mining/machine learning to analyze the scanning patterns of scanners and classify them into three categories: focused hitters, random address scanners, and blockwise scanners. We also explored several features to further investigate the observed scanning patterns (or "footprints") of scanners, and attempted to infer whether the observed scanning activities are merely part of global "background radiation" or are reflections of actions that are likely targeted more specifically at our own network. Our preliminary (yet promising) results suggest that 1) analysis of observed scanning behaviors can potentially reveal the plausible "intentions" of scanners; 2) scanning activities targeted perhaps more specifically at our own network are likely more harmful and thus warrant closer scrutiny. As part of our on-going research, we are exploring additional features and applying more sophisticated machine learning techniques to perform cross-feature correlation analyses, develop rule-based predictive models (see, e.g. [6]) for classifying scanners, and conduct more in-depth and long-term follow-up investigations of scanning and other related activities.

ACKNOWLEDGEMENT

This work was supported in part by the NSF grants CNS-0435444 and CNS-0626812, a University of Minnesota Digital Technology Center DTI grant, a Cisco gift grant and an IBM Faculty Partnership Award.

REFERENCES

[1] R. Pang, V. Yegneswaran, P. Barford, V. Paxson and L. Peterson, "Characteristics of Internet Background Radiation," in *IMC*, 2004.

[2] D. Moore, C. Shannon, G. M. Voelker, and S. Savage, "Network Telescopes," CAIDA, Tech. Rep., 2003.

[3] S. Staniford, V. Paxson, and N. Weaver, "How to Own the Internet in Your Spare Time," in *Proc. of USENIX Security Symposium*, 2002.

[4] K. Xu, Z.-L. Zhang and S. Bhattacharyya, "Profiling Internet Backbone Traffic: Behavior Models and Applications," in *SIGCOMM*, 2005.

[5] Y.Jin, K.Xu and Z.-L. Zhang, "Identifying and Tracking Suspicious Activities Through IP Gray Space Analysis," UMN, Tech. Rep., 2006.

[6] G. J. Simon, H. Xiong, E. Eilertson, and V. Kumar, "Scan detection - a data mining approach," in *SIAM SDM*, 2006.