Identifying Dynamic IP Address Blocks Serendipitously through Background Scanning Traffic

Yu Jin, Esam Sharafuddin, Zhi-Li Zhang
University of Minnesota

ABSTRACT

Today's Internet contains a large portion of "dynamic" IP addresses, which are assigned to clients upon request. A significant amount of malicious activities have been reported from dynamic IP space, such as spamming, botnets, etc.. Accurate identification of dynamic IP addresses will help us build blacklists of suspicious hosts with more confidence, and help track the sources of different types of anomalous activities. In this paper, we contrast traffic activity patterns between static and dynamic IP addresses in a large campus network, as well as their activity patterns when countering outside scanning traffic. Based on the distinct characteristics observed, we propose a scanning-based technique for identifying dynamic IP addresses in blocks. We conduct an experiment using a one-month data collected from our campus network, and instead of scanning our own network, we utilize identified outside scanning traffic. The experiment results demonstrate a high classification rate with low false positive rate. As an on-going work, we also introduce our design of an online classifier that identifies dynamic IP addresses in any network in real-time.

1. INTRODUCTION

Knowledge of IP address assignments, e.g., whether IP addresses within an address block are dynamically or statically assigned, can provide valuable information and hints in managing and securing one's network. For instance, on the Internet at large, a significant amount of malicious activities have been reported (see, e.g., [1–5]) from dynamic IP addresses, such as spamming, botnets, and so forth. Information regarding the source IP addresses of suspected malicious activities (e.g., email spam) not only provides us with more confidence in classifying such malicious activities, but also allows us to associate multiple instances of such activities from the same dynamic address block over time to better track the origins of attackers. Within a campus or enterprise network, dynamic addresses are typically assigned to mobile devices (e.g., laptops) which tend to roam and be used in unprotected networks (e.g., the wireless hotspot in a coffee shop or at home), thus are more likely to get infected with malware. Hence, knowledge of such address blocks can assist network operators/security analysts of a campus/enterprise network in focusing additional scrutiny to suspicious activities on these blocks, detecting and preventing attacks from inside (compromised) hosts. For the purpose of profiling the activities and behavior of hosts within a network [6, 7], knowledge of dynamic and static addresses is also important in building and associating behavior models to appropriate hosts for anomaly detection and behavior tracking.

Information regarding whether an IP address is dynamic or not may not be readily available, even for those within one's own network. This is particularly true for large networks with decentralized management, where large blocks of addresses are allocated and delegated to sub-organizations which control and manage how these addresses are assigned and utilized. While it is possible to infer whether an IP address is dynamic or static by its DNS name, such an approach may not always be feasible nor accurate for a variety of reasons. Not all IP addresses have DNS names assigned or registered. Furthermore, from the DNS name, it may not be completely clear whether an IP address is dynamic or static. In addition, DNS records are not always kept up-to-date. Hence, alternative methods for accurately classifying IP addresses, in particular for identifying dynamic IP addresses, are needed.

In this paper, we investigate the feasibility of classifying IP addresses based on "usage patterns" or "traffic activities" on a large campus network. More specifically, we consider the following problem setting. Suppose that at a certain vantage (e.g., a border router of a campus network), we can passively observe – and if necessary, inject active probes – traffic coming into or going out of a particular address block (of an appropriate size, say, /24 or /28). Is it possible to infer and classify the said address block as dynamic or static based solely on such observations? Here, in accordance within common practice, we assume that the addresses within the whole contiguous block, typically in size of 2^k, for some (relatively) small k, e.g., $k = 3, 4, \ldots, 8$, are assigned as *dynamic* (i.e., allocated to hosts via DHCP with a limited lease time), or *static* (i.e., allocated to hosts "permanently"). To answer this question, we extract and analyze the traffic activities of dynamic and static address blocks of a large campus network with diversified user population and usage patterns, utilizing a month-long netflow data collected at the campus border router.

As the basis for our study, we first perform an exhaustive DNS look-up to extract the registered DNS name, if available, of each IP address of a class B address block within the campus network. We develop a simple name-based heuristic to classify individual IP addresses into four groups, *Dynamic* and *Static*, as well as *NoName* which contains IP addresses with no registered DNS names, and *Undecided* which contains those IP addresses we cannot classify with high confidence whether they are static or dynamic based on their

DNS names alone. Using the classification of individual IP addresses, we then examine and infer the *block structures* of the address assignments to group individual dynamic or static IP addresses into (contiguous) *address blocks* of appropriate sizes. This outcome of name-based classification process is used for two purposes: they provide us with a set of sample dynamic and static address blocks that are used for our subsequent analysis of traffic activity patterns of dynamic and static address blocks; they also serve as training and test datasets for the evaluation of a simple *scanning-based* dynamic address classifier we have designed.

In analyzing the usage patterns of dynamic and static address blocks, we introduce a simple apparatus, *(traffic) activity matrix*, to succinctly represent the (incoming and outgoing) traffic activities of an address block, and put forth several metrics to mathematically characterize their properties. By examining the *overall* traffic activity patterns on an address block, we find that while there are some discernible differences between dynamic and static address blocks, they are unlikely to yield a useful and robust classifier to distinguish dynamic and static address blocks. The most striking feature of the overall traffic activity patterns lies in the strong difference between incoming traffic activities and outgoing traffic activities, regardless of the types of address blocks. This striking difference is caused by the prevalence of wide-spread scanning activities, which typically elicit different responses from dynamic and static address blocks, and thus can be serendipitously exploited in classifying dynamic and static address blocks. Based on this key observation, we develop a simple *scanning-based dynamic address classifier* consisting of two hypothesis tests on the responses to scanning traffic of an address block. We also explore the crucial parameters for implementing such a classifier in practice. Extensive evaluation shows that this simple classifier is capable of identifying dynamic address blocks with fairly high accuracy and relatively low false positive rate.[1]

Our study not only provides an affirmative answer to the question posed earlier, but also shows that we can utilize the prevalence of outside scanning traffic to our advantage: by focusing on outside scanning traffic and the responses they elicit, we can serendipitously gain certain knowledge about the behavior of our own network, e.g., dynamic and static address assignments, and use such knowledge to better defend our own network. We are currently exploiting such knowledge to generalize the gray space analysis and host profiling techniques for rapid and high-fidelity detection of scanning and other malware activities. To the best of our knowledge, our paper is the first study that investigates the traffic activity patterns of dynamic and static address blocks; without relying on DNS names, our scanning-based dynamic address classifier is the first classifier based solely on direct observation and analysis of traffic behavior.

The remainder of the paper is organized as follows. We describe the DNS name-based classification mechanism and its results in Section 2. In Section 3, we define the activity matrix and introduce different metrics to characterize the traffic activity patterns. Section 4 focuses on the patterns of scanning traffic activities. The scanning-based dynamic address classifier is explained and evaluated in Section 5. We conclude the paper in Section 6.

[1]Part of the classification errors can in fact be attributed to either the imprecision in the name-based classification or "anomalous" usage patterns of certain address blocks, e.g., static address blocks assigned to computers in a lab which are turned on only during the business hours.

2. DNS NAME-BASED CLASSIFICATION

In this section we first devise a simple DNS-name based heuristic for classifying *individual* IP addresses into dynamic, static and other groups. Using this classification, we also investigate the block structures of dynamic and static address assignments, based on which we extract dynamic and static address blocks of appropriate sizes. The outcomes of this name-based classification process will be used in the subsequent sections both for analysis of the activity patterns of dynamic and static address blocks, and for design and evaluation of a scanning-based dynamic address block classifier.

Dataset. In this study, we utilize a month-long archive of netflow records collected at the border router of our campus network. The campus network owns three class B (/16) IP blocks with a total of 196608 IP addresses. The collected netflow includes all bidirectional flow traffic between inside and outside hosts for one whole month. Unless otherwise specified, the study uses the netflow records of traffic to and from one of the 3 class B address blocks.

2.1 Classifying Individual IP Addresses using DNS Names

DNS names of hosts are in general chosen using certain (unwritten) convention. For instance, for hosts assigned with *static* IP addresses, users or network operators typically pick names that are mnemonics (e.g., www for web servers, mail for email servers, and various (typically) proper nouns such as place or person names for desktops, etc.). In contrast, IP addresses within a dynamic address block, if they are named at all, are typically named with a keyword such as "dhcp", "dip", "dialup" and often affixed with a number or (part of) its IP address, e.g., `dhcp-11` or `dip.101.31`. Taking advantage of these common naming practices, we devise a simple name-based heuristic for IP address classification.

We perform an exhaustive look-up for all the addresses of one of the three class B address blocks within our campus network, using Reverse DNS (rDNS) and whois database [8]. Based on the results of this exhaustive lookup, we classify the individual IP addresses into four categories. For those IP addresses that the lookup fails, we put them in the *NoName* category. This category constitutes 35.6% of all IP addresses. For the remaining IP addresses for which rDNS returned DNS names, we classify them based on the keywords contained in the DNS names. We identify a list of keywords in which we have high confidence that they are associated with dynamic IP addresses, such as "dhcp", "dip", "dynamic", "wireless", etc. We put these addresses into the category *Dynamic*. This category accounts for 27.6% of all IP addresses. The third category, *Undecided*, contains those IP addresses (roughly 9.9%) based on the names of which we cannot infer whether they are static or dynamic with high confidence. The common keywords contained in their names are "ej" (likely standing for "Ethernet jack"), "x", or other similar keywords. Depending on how they are allocated to users and the way users utilize them, these IP addresses can resemble dynamic IP addresses in some cases, and static IP addresses in other cases. For all the remaining addresses (26.9%), we place them into the *Static* category. Manual inspection of their DNS names shows that nearly all of them contain keywords such as "www", "mail", or some forms of

user-chosen mnemonics for client machines.

From the above results, IP address classification using DNS names is a heuristic that may not always work. We see that for a significant portion of IP addresses, the classification either fails or is indecisive. While a large portion of the IP addresses within the *NoName* category are unused or "dark" (judging based on the month-long netflow records), there are a non-negligible number of them that indeed are part of the "used" IP address space, i.e., assigned to "live" hosts (either dynamically or statically) for some period of time during the one-month period under study. These IP addresses, together with those in the *Undecided* category, comprise at least 15% of a class B address block. In addition, as mentioned in the introduction, DNS records may not always be kept up-to-date. We also note that such an approach can be quite laborious, especially for a large address space, because of its need for some level of manual inspection: this is because it is impossible to have a complete list of keywords for classifying dynamic and static IP addresses. Nonetheless, due to its use of "domain knowledge" (e.g., the naming conventions of an organization), the name-based classification heuristic enables us to classify a *subset* of IP addresses with fairly high confidence. It thus serves as a good starting point for our study.

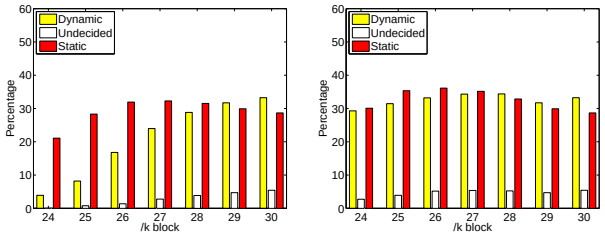
2.2 Block Structures of Address Assignments

Using the classification of individual IP addresses, we now examine and infer the block structures of IP address assignments, with the goal to determine the appropriate block sizes for grouping dynamic and static IP addresses into contiguous address blocks. This is motivated by the fact that in general, IP addresses are allocated as a block of dynamic or static addresses – this is particularly true for dynamic addresses which are allocated to and assigned by DHCP servers. For this reason, we examine blocks of contiguous IP addresses with varying block sizes, and investigate the percentage of IP addresses belonging to the same category as well as the likely mixture of different categories.

First, we observe that a high percentage of IP addresses in the *noname* category tend to interleave with one of the other 3 categories. This is not surprising – network administrators in general allocate addresses in blocks, a subset of which may be initially assigned to hosts or used for dynamic address assignment via DHCP. As a result, only a portion of these addresses were given DNS names, with the rest left unnamed. Based on this observation, we consider *noname* IP addresses interleaved with IP addresses from another category to belong to that category (dynamic, static or undecided). Henceforth, we consider only blocks of IP addresses in three categories: dynamic, static and undecided, treating *noname* IP addresses that are interleaved with one of these three categories as part of that category[2].

Fig. 1 shows the percentage of dynamic, static and undecided IP addresses that are allocated in various block size, where the x-axis represents the block size (/24 through /30) and the y-axis represents the percentage of IP addresses in blocks (out of the total number of the class B IP addresses). In particular, Fig. 1[a] shows the percentage of IP addresses in complete blocks, i.e., blocks for which all the IP addresses

[2]We find that almost all blocks (of size 16 or larger) that contain only *noname* IP addresses are "dark" or unused, and thus are uninteresting from the perspective of our study. So we do not consider them here for ease of exposition.



(a) Percentage of complete blocks (b) Percentage of 90% blocks

Figure 1: Percentage of IP addresses in blocks

belong to the same category, whereas Fig. 1[b] shows the percentage of each category type for which 90% of the IP addresses within the block belong to the same category.

From Fig. 1[a] we see that dynamic blocks account for only 4% of /24, but with decreasing block sizes, their percentage increases up to 29% of /28 and slightly more than 30% of /30 block sizes. Whereas, static IP address blocks tend to be allocated in large block sizes, accounting for 22% and 32% for /24 and /28 block sizes, respectively. When using the 90% addresses belonging to one category as the criteria to group and classify address blocks, Fig. 1[b] shows both dynamic and static blocks have similar percentage across all block sizes. Together they account for the majority of their corresponding individual IP addresses within the class B address space, as the *undecided* blocks comprise only a small percentage in both the cases of complete and 90% blocks.

From Figs. 1[a] and [b], we see that using the 90% addresses belonging to one category as the criteria to group and classify address blocks, the percentage of dynamic IP address blocks increases significantly, especially for larger block sizes. Careful inspection reveals that this is mostly because dynamic address blocks tend to contain a few "static" IP addresses, the names of which do not contain the keywords used for dynamic address classification. These names almost invariably contain certain keywords, indicative of special servers (e.g., DHCP or DNS servers) or devices (access points, switches/routers), e.g., "ac" for access points in a dynamic wireless address block. In contrast, the percentage of static address blocks does not change significantly. We find that occasionally, the static address blocks would contain IP addresses that fall into the *undecided* categories. In addition, there are a small number of blocks with more than 10% mixture of other types, containing especially a fair portion of the *undecided* IP addresses. These are "mixed-usage" address blocks, or address assignments that do not fall on the conventional boundary of a power of 2.

2.3 Sample Dynamic and Static Blocks

Based on the results from the previous two subsections, we extract a (sub)set of dynamic and static address blocks which will be used as *sample* datasets for our study of usage or traffic activity patterns of dynamic and static address blocks, as well as for the design and evaluation of a scanning-based dynamic IP address block classifier. Balancing between the block size (or the number of IP addresses within the block) and number of blocks available, we choose to consider two block sizes /24 (with 256 addresses within each block) and /28 (with 16 addresses). In choosing these

dynamic and static address blocks, we also exclude those address blocks containing a large portion of *noname* IP addresses, which do not generate any traffic within the month, thus are "dark" or unused. In other words, these address blocks tend to have only a small number of assigned IP addresses, *with little traffic activities*. They are therefore not very useful for our study. In the following, we describe the criteria we use for selecting the sample dynamic and static address blocks.

The dynamic and static address blocks are selected using the following two conditions: 1) there are at least 40% of the IP addresses within the block in which each IP send at least one outgoing traffic during the entire month of February 2006 (i.e., the corresponding IP address is *not* "dark"), 2) at least 90% of these IP addresses satisfying 1) are either exclusively dynamic or exclusively static. Using this criteria, for /24 block size, we obtain 35 dynamic and 10 static /24 blocks, and for /28 block size, we obtain 1034 dynamic blocks and 289 static blocks. These sample address blocks are used for our study in the subsequent sections.

3. ACTIVITY PATTERNS OF DYNAMIC VS. STATIC BLOCKS

In general, dynamic and static address blocks are allocated for different usages. For example, static addresses are typically assigned to "fixed" or "long-lived" machines (e.g., servers or desktops) on a network, while dynamic addresses are assigned to "mobile" hosts (e.g., laptops) that come and go. This is particularly true for a campus network. Hence, intuitively, we would expect to see differing traffic or activity patterns on different types of address blocks, reflecting the usages and roles of the machines that are "active" on the address blocks over time. In this section, we first introduce a simple *activity matrix* to represent the overall (either *incoming* or *outgoing*) activity patterns of an address block over time, namely, when there is incoming traffic towards or outgoing traffic from certain IP addresses within the address block. We then put forth several metrics to characterize and compactly summarize the overall activity patterns of an address block. Using the sample dynamic and static address blocks identified using the name-based heuristic in the previous section, we study and analyze the activity patterns of these address blocks.

3.1 Activity Matrices

We study the "activity patterns" of an address block (say, a /24 or /28 address block) by examining when there is incoming traffic towards or outgoing traffic respectively from some addresses within the address block over a certain observation period T, say, a day or a week. For simplicity, we divide the observation period T into discrete time slots of length τ. Unless otherwise specified, in the remainder of the paper we choose T to be one day (from 0th hour to 24th hour), and τ to be 5 minutes. This gives us $n := T/\tau = 288$ 5-minute time slots in a day. Let m (a power of 2) be the size of the address block. To succinctly represent the activity patterns of the incoming and outgoing traffic of the address block, we introduce two matrices, $IA := [ia_{i,t}]_{m\times n}$ and $OA := [oa_{i,t}]_{m\times n}$, referred to as the *incoming traffic activity matrix* and *the outgoing traffic activity matrix*, respectively. For IA, we define $ia_{i,t} = 1$ if we observe incoming traffic towards the ith IP address of the block at any time within the tth time slot[3]; the entries for OA are similarly defined.

Fig. 2[4] displays the scatter plots of the IA and OA matrices for four (two static and two dynamic) sample /24 address blocks using the netflow data collected at our campus border router on 02/08/2006. One striking observation from the scatter plots is the sharp difference between the IA and OA activity matrices, regardless whether it is a dynamic or static address block: a) incoming traffic is far more "active" (i.e., with more 1's) than outgoing traffic; and b) while OA's of different address blocks are at times fairly distinct, IA's of all address blocks look remarkably similar. This visual observation suggests that a vast majority of incoming traffic activity towards each address block is largely independent of and agnostic of the nature, roles or "liveness" (i.e., whether an IP is currently assigned to a live host) of the hosts on the address block. In contrast, the outgoing traffic activity matrix reveals more information regarding the address block: three of them have a clear "time-of-day" pattern with most activities concentrated during the business hours; the fourth one (residential hall) is active nearly all day except during the wee hours of the morning. Comparing the OA's of static and dynamic address blocks, however, there are few "outstanding" features that distinguish the dynamic address blocks from the static ones, except that the OA's of the static blocks tend to have quite a few vertical bars, indicating that nearly all machines on the block are active at certain time slots.

3.2 Characterizing Traffic Patterns

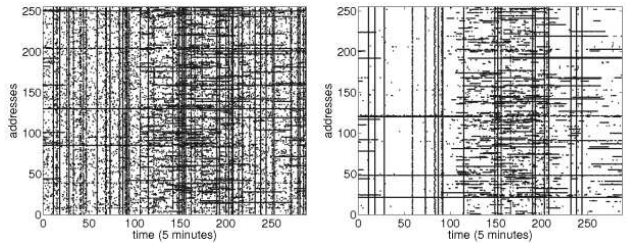
We introduce several metrics to compactly and mathematically characterize the properties of the activity matrices. Given an incoming traffic activity matrix $IA = [a_{i,t}]_{m\times n}$, let $\hat{a} =: \sum_{i=1}^{m}\sum_{t=1}^{n} a_{i,t}$ represent the total number of 1's in IA, or the total amount of incoming traffic "activities" towards the address block during the time period T. Define the *density* of IA, $d(IA) := \hat{a}/(m \times n)$ (we will drop the reference to IA when the context is clear), which measures the "average" activity per address per time slot. Intuitively, for incoming traffic, $d(IA)$ tells us *on the average* how likely we may see activity to a random address at a random time slot. For each row (i.e., IP address) i, let $p_{i\cdot} = (\sum_t a_{i,t})/\hat{a}$, be the percentage of incoming traffic activities towards the address i; and for each column (i.e., time slot), let $p_{\cdot t} = (\sum_i a_{i,t})/\hat{a}$, be the percentage of incoming traffic activities occurring at time slot t. We define the *address diversity* (*AD(IA)*) and *time diversity* (*TD(IA)*) as follows:

$$AD(IA) := \frac{-\sum_{i=1}^{m} p_{i\cdot}\log p_{i\cdot}}{\log m} \text{ and } TD(IA) := \frac{-\sum_{t=1}^{n} p_{\cdot t}\log p_{\cdot t}}{\log n}.$$

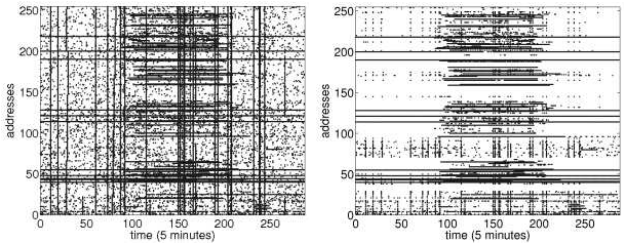
By definition, $AD(IA)$ and $TD(IA)$ are the normalized entropies (or *relative uncertainty* [6]) of the distributions $\{p_{i\cdot}\}$ and $\{p_{\cdot t}\}$. Intuitively, the address diversity $AD(IA)$ reflects how random or uniform the incoming traffic touches the addresses within a block, while the time diversity $TD(IA)$ reflects how the incoming traffic activities are spread out over

[3]Namely, we observe at least one netflow with the said IP address as the destination, an outside IP address as the source address, and a beginning time stamp within the time slot.

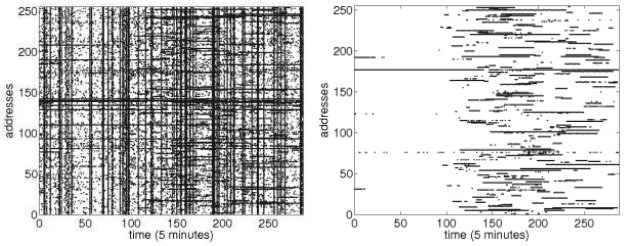
[4]Due to privacy concerns, we have randomized the order of the addresses within each block.



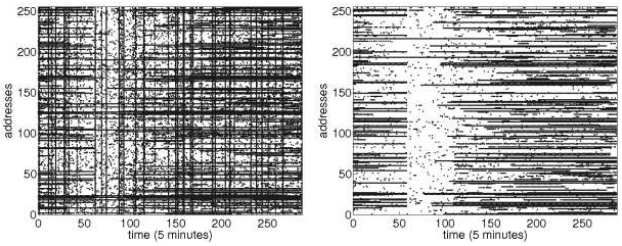
(a) Static address block (lab network) IA vs. OA



(b) Static address block (department network) IA vs. OA

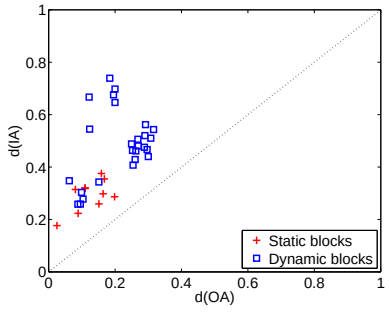


(c) Dynamic address block (campus wireless network) IA vs. OA

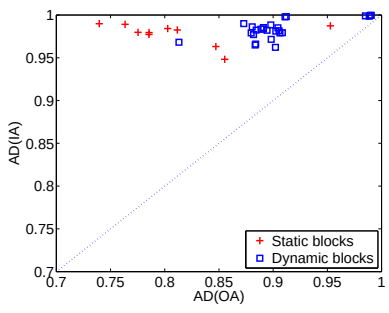


(d) Dynamic address block (residential hall network) IA vs. OA

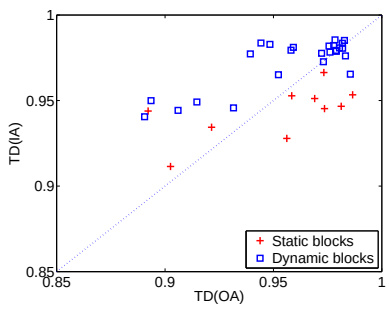
Figure 2: IA and OA matrices (static address blocks vs. dynamic address blocks)



(a) $d(IA)$ vs. $d(OA)$



(b) $AD(IA)$ vs. $AD(OA)$



(c) $TD(IA)$ vs. $TD(OA)$

Figure 3: IA and OA metrics (static address blocks vs. dynamic address blocks)

time. For example, $AD(IA)$ closer to 1 (thus $p_{i\cdot}$ being approximately equal) means that among those receiving any incoming traffic, the incoming traffic touches each address roughly equally (albeit perhaps at different times). Likewise, $TD(IA)$ closer to 1 means that among the time slots where there is incoming traffic, the incoming traffic activities spread over these time slots roughly equally (albeit perhaps touching different addresses). For an outgoing traffic activity matrix OA, the density $d(OA)$, address diversity $AD(OA)$ and $TD(OA)$ are defined in the same fashion, and can be similarly interpreted.

In Fig. 3, we plot these metrics of the incoming and outgoing traffic activity matrices IA and OA for the sample /24 dynamic and static address blocks identified in the previous section, using the netflow data collected on 02/06/2006. Comparing $d(IA)$ (y-axis) and $d(OA)$ (x-axis) of each block shown as a point $(d(OA), d(IA))$ in Fig. 3(a), we see that the incoming traffic activities are much denser than the outgoing traffic activities for each block, whether it is dynamic or static. In terms of address diversities, we see that $AD(IA) > AD(OA)$ for all address blocks. In particular, incoming traffic touches each address with nearly equal frequencies; for outgoing traffic, static address blocks in general have a lower $AD(OA)$ than dynamic address blocks, indicating that activities on the static address blocks tend to be less equally distributed among the (active) addresses, while dynamic ones are likely more equally distributed. In terms of time diversities, both $TA(IA)$ and $TA(OA)$ are larger than 0.9, indicating that traffic activities tend to be more or less equally distributed among all the time slots *that are active* (i.e., with incoming or outgoing traffic).

All in all, we see that by examining the *overall* traffic activity patterns on an address block, while there are some subtle but discernible differences between dynamic and static address blocks, they are unlikely to provide us with a useful and robust classifier to distinguish dynamic and static address blocks. The most striking feature of the overall traffic activity patterns lies in the strong difference between incoming traffic activities and outgoing traffic activities, regardless of address blocks. As will be explained in the next section, the culprit here is the prevalence of wide-spread scanning activities on the Internet. In the next section we zero in on these scanning activities, and investigate whether there are significant differences in responses from dynamic and static addresses to such scanning activities. In section 5, we will exploit these differences to design a robust classifier for identifying dynamic and static address blocks.

4. IMPACT OF OUTSIDE SCANNING ON ACTIVITY PATTERNS

Based on our observations in the previous section, there is a strong difference between the densities of incoming and outgoing traffic activities in all the sample address blocks. There are far more incoming traffic activities than outgoing traffic activities. Intuitively, this can be caused by a large amount of unproductive incoming traffic, which gets few responses from the targeted network. In this section, we study one of the major sources of such unproductive traffic, the scanning traffic. We first describe our method for extracting scanning traffic and show the impact of scanning traffic on incoming traffic activity. We then use the three metrics defined in the previous section to characterize the incoming and outgoing activities of scanning traffic. The distinct outgoing scanning traffic activity pattern from dynamic blocks provides us with the intuition of identifying dynamic address blocks through scanning.

4.1 Impact of Scanning Traffic

In terms of scanning traffic, the incoming activities are referred to as the scanning traffic itself, while the outgoing activities consist of all the responses to scanning traffic. To study the impact of scanning traffic, we apply the IP gray space analysis technique described in [9] to the netflow dataset and identify a set of 6050 scanners on 02/08/2006. Meanwhile, we identify all the ports that those scanners target, and extract all the flows from the identified scanners towards those targeted ports and refer to it as scanning traffic. To single out the responses toward scanning, we match each incoming scanning flow within a time window T, say, 30 minutes. We consider an outgoing flow from our campus network as a response to scanning, if, within T, either 1) it matches the 5-tuple[5] of a previously observed TCP or UDP scanning flow, or 2) it is an ICMP flow and matches the 2-tuple[6] of a precedent incoming ICMP or UDP scanning flow.

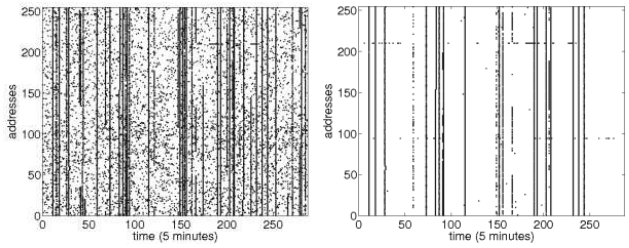
To describe the activity patterns of scanning, we define the activity matrices IA and OA for scanning traffic similarly as in section. 3. Fig. 4 illustrates the scatter plots of IA and OA matrices of scanning traffic for two static address blocks and two dynamic IP blocks corresponding to those in Fig. 2. Comparing with Fig. 2, we observe even a larger difference between incoming and outgoing activities. Furthermore, we find that vertical bars still exist in the IA matrices of scanning traffic. Investigation on those vertical bars suggests that they are from coordinated scanners (different scanners that cooperatively scan a particular network) or blockwise scanners (scanners that choose target on the basis of blocks and scan all the addresses within each block). In contrast to incoming activities which look similar in all the blocks, the outgoing activities of scanning traffic reveal much distinction between static address blocks and dynamic address blocks. In OA matrices of static blocks, we observe a number of vertical bars corresponding to those in IA matrices, which indicates responses from all the static addresses within the same block towards those coordinated scanning or blockwise scanning. In contrast, even though there are quite a number of vertical bars in the IA matrices of dynamic blocks, no vertical bar has been observed in their OA matrices, instead, the responses from dynamic blocks depict a more random manner.

4.2 Characterizing Scanning Traffic Patterns

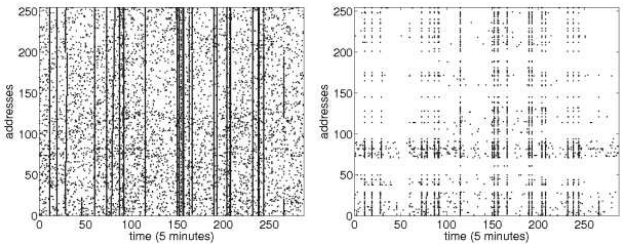
To characterize our observations from IA and OA matrices of scanning traffic, we apply those metrics, density (d), address diversity (AD) and time diversity (TD) defined in the previous section to the IA and OA matrices of scanning traffic. In Fig. 5, we plot these metrics of the incoming and outgoing scanning traffic activities for the same /24 sample blocks in the previous section. In Fig. 5(a), comparing with Fig. 3(a), all the points $(d(OA), d(IA))$ move away from the diagonal line, which indicates a much more significant difference between incoming and outgoing activities of scanning

[5]5-tuple is referred to as source IP, destination IP, source port, destination port and protocol.

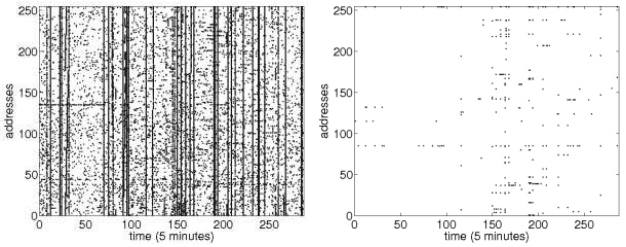
[6]2-tuple is defined as the pair of source IP and destination IP.



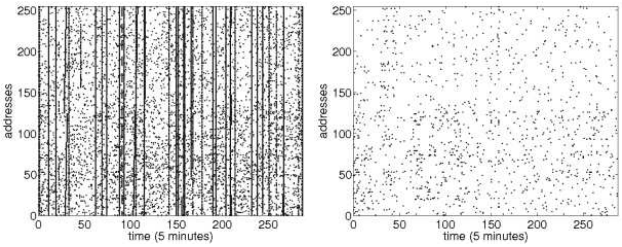
(a) Static address block (lab network) IA vs. OA



(b) Static address block (department network) IA vs. OA

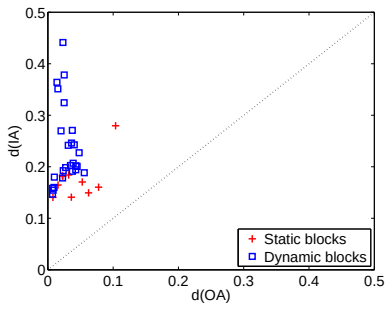


(c) Dynamic address block (campus wireless network) IA vs. OA

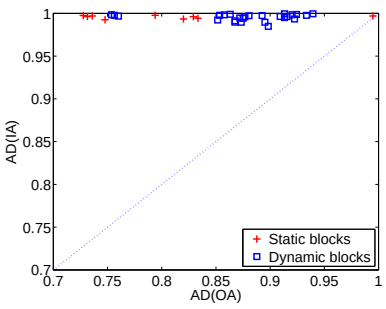


(d) Dynamic address block (residential hall network) IA vs. OA

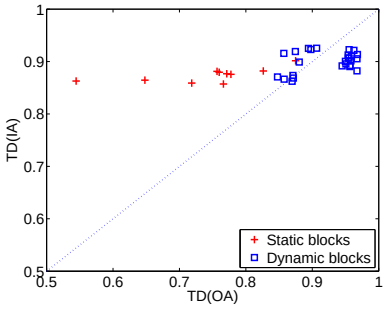
Figure 4: IA and OA matrices of scanning traffic (static address blocks vs. dynamic address blocks)



(a) $d(IA)$ vs. $d(OA)$



(b) $AD(IA)$ vs. $AD(OA)$



(c) $TD(IA)$ vs. $TD(OA)$

Figure 5: IA and OA metrics of scanning traffic (static address blocks vs. dynamic address blocks)

traffic. In other words, the difference between incoming traffic activities and outgoing traffic activities demonstrated in Fig. 5(a) is largely caused by scanning traffic. Furthermore, we observe that points corresponding to dynamic address blocks are further away from the diagonal line than static ones, meaning that in dynamic blocks, the incoming scanning traffic activities are much denser than outgoing scanning traffic activities. Except for the existence of firewalls which block a portion of the scanning traffic, the difference between static and dynamic blocks in terms of traffic densities is caused by their different address allocation strategies. A static IP address is usually associated with a live end host for most of the time; hence, the probability of observing a response from a static address to scanning traffic is quite large. On the contrary, a dynamic IP is assigned to clients upon request and released frequently; therefore, there is an extended period of time everyday that a dynamic IP address remains unassigned; thus, the chance of observing responses from a dynamic address is relatively small.

Fig. 5(b)(c) depict the address diversity (AD) and time diversity (TD) of static address blocks vs. dynamic address blocks, respectively. Despite the fact that incoming scanning traffic is random ($AD(IA)$ and $TD(IA)$) are both close to 1) across all the blocks regardless of static or dynamic, we notice that $AD(OA)$ and $TD(OA)$ values of static address blocks are generally smaller than those of dynamic address blocks. In other words, the responses from dynamic blocks are more random or uniform than those from static blocks. This can be explained by the different usages of static addresses and dynamic addresses. The addresses within a particular dynamic block are usually assigned to clients following a specific IP assignment policy which balances the workload among all the dynamic IP addresses. Meanwhile, at each time interval, there is usually only a small portion of dynamic IP addresses corresponding to live end hosts, so within each time slot, we observe little variation in the number of responses; albeit the existence of coordinated or blockwise scanning activities. On the contrary, static addresses within the same block respond simultaneously towards coordinated or blockwise scanning traffic, which decreases the time diversity of outgoing scanning traffic activities. Meanwhile, the existence of hosts with different workloads, such as servers, lab machines, etc. makes the outgoing scanning activities less random among all the static addresses within the same block.

Intuitively, in order to distinguish static and dynamic address blocks, we can utilize the strong difference in the outgoing scanning traffic activity patterns between static and dynamic address blocks. We can launch a number of scanning sequences towards a particular address block and measure the response patterns from that block. If less responses are observed and those responses are quite random, we consider the block to be a dynamic address block. Utilizing this idea, in the next section, we describe our design for a classifier which identifies dynamic address blocks through sequences of scanning.

5. A SCANNING-BASED DYNAMIC ADDRESS CLASSIFIER

As indicated in the previous section, when facing the same scanning traffic, static blocks and dynamic blocks demonstrate distinct response patterns. In this section, we first present an (ideal) statistical model for characterizing the responses from an IP address block. We then propose two hypothesis tests for classifying whether an address block is dynamic based on the response model. We devise a dynamic address block classifier by combining these two hypothesis tests and evaluate its performance using the data from our campus network. Finally, we discuss several important issues in implementing such a classifier in practice.

5.1 Modelling Responses from Address Blocks

We model the responses from a particular address block (either static or dynamic) towards scanning traffic as follows. Suppose there is a scanning sequence at time t towards an entire address block with size m. For any IP address i within the block, assume it has a fixed probability to respond to the scanning traffic, say, p_i, then its response to the scanning traffic can be treated as a Bernoulli random variable, which we denote as x_{it}, where $x_{it} = 1$ and $x_{it} = 0$ stand for the cases of response and nonresponse, respectively.

Now we consider the address block level, let $\vec{x}_t = [x_{1t}, x_{2t}, \cdots, x_{mt}]^T$ denote the vector of responses from all the m IP addresses within the block at time t, or we call it a *block response vector*. The block response vector $\vec{x}_t$ is considered as a multivariate Bernoulli random variable with mean vector $\vec{p} = [p_1, p_2, \cdots, p_m]^T$ and covariance matrix Σ_0, where each entry σ_{ij} in Σ_0 equals $E[(x_i - p_i)(x_j - p_j)]$.

Assume there are n independent scanning sequences towards the block, then we will observe n i.i.d response vectors, $\vec{x}_1, \vec{x}_2, \cdots, \vec{x}_n$. Let $\vec{y} = \sum_{t=1}^{n} \vec{x}_t$, then $\vec{y}$ follows a multivariate binomial random distribution with mean vector $n\vec{p}$ and covariance matrix $n\Sigma_0$. Given a large sample of response vectors, the multivariate binomial random variable $\vec{y}$ can be approximated by a m-dimensional multivariate normal random variable with mean vector $n\vec{p}$ and covariance matrix $n\Sigma_0$.

We denote the *response rate vector* $\vec{\bar{x}}$ as $\vec{\bar{x}} = \vec{y}/n$, then $\vec{\bar{x}}$ follows a multivariate normal distribution with mean vector $\vec{\mu}$ and covariance matrix Σ, where $\vec{\mu} = \vec{p}$ and $\Sigma = \frac{1}{n}\Sigma_0$.

In order to estimate the parameters $\vec{p}$, Σ_0, we either actively launch or passively observe n scanning sequences at different times towards the same address block, from which we can obtain n response vector samples. The unbiased estimator for $\vec{\mu}$ will be the sample mean of the n response vectors, which is $[\bar{x}_1, \bar{x}_2, \cdots, \bar{x}_m]^T$, and the unbiased estimator for Σ_0 is the sample covariance matrix S_0, with each entry given by $s_{ij} = \frac{1}{n-1} \sum_{k=1}^{n} (x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j)$. Thus, the unbiased estimator for Σ is $S = \frac{1}{n} S_0$.

5.2 Testing of Responses from Dynamic Blocks

Using this model, we can describe the distinct properties of responses from dynamic address blocks into hypothesis tests, which can help us identify dynamic address blocks. From our studies of outgoing activity patterns of scanning traffic in the previous section, the two properties that characterize responses from dynamic address blocks are: 1) IP addresses within the same dynamic address block tend to have equal response rates given a long-term observation. 2) The majority of the addresses within a dynamic address block are likely to have low response rates. In this section, we interpret how we model each property into a hypothesis test and how we choose parameters for those tests. At the end of this section, we discuss how to build a classifier by combining those two hypothesis tests to achieve the best

classification performance.

1) Test of level mean vector

Given a long observation time period, the response rates of different dynamic IP addresses within the same block will likely be equivalent. We specify this assumption as $\mu_1 = \mu_2 = \cdots = \mu_m$. Intuitively, we classify an address block to be dynamic if we have enough confidence to believe that the response vectors from that block fit a m-dimensional multivariate model with a level mean vector, which can be described using the following hypothesis test:

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_m \; vs. \; H_1 : otherwise$$

To perform the hypothesis test, we first construct a comparison matrix:

$$C = \begin{pmatrix} 1 & -1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & -1 & \cdots & 0 & 0 \\ & & \cdots\cdots\cdots\cdots & & & \\ 0 & 0 & 0 & \cdots & 1 & -1 \end{pmatrix}_{(m-1)\times m}$$

Hence H_0 is equivalent to $C\vec{\mu} = \vec{0}$. Given the assumption $\vec{\bar{x}} \sim N_m(\vec{\mu}, \Sigma)$, we know $C\vec{\bar{x}} \sim N_m(C\vec{\mu}, C^T \Sigma C)$. Using the Hotelling's T^2 statistic [10], which is

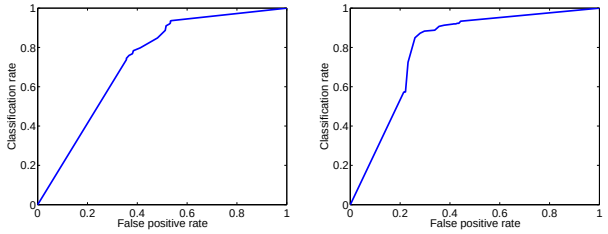
$$t^2 = n(C\vec{\bar{x}})^T (CSC^T)^{-1} C\vec{\bar{x}}$$

t^2 follows the F-distribution, hence we reject $H_0 : C\vec{\mu} = \vec{0}$ if

$$t^2 > \frac{(n-1)(m-1)}{(n-m+1)} F_{m-1,n-m+1}(\alpha)$$

where $F_{m-1,n-m+1}(\alpha)$ is the upper 100α percentile of an F distribution with $m-1$ and $n-m+1$ degrees of freedom.

Selecting α values from 0.1 to 0.001, we show the ROC curve for the level test in Fig. 6(a).



(a) level mean vector test (b) low response rate test

Figure 6: ROC curves for two hypothesis tests

2) Test of low response rate

The response rate of each IP address within a dynamic address block is usually smaller than that of a static address block. Recall that from our response model, that the response vector $\vec{\bar{x}} \sim N_m(\vec{\mu}, \Sigma)$. To describe property of low response rates of most addresses within a dynamic address block, we introduce a parameter $0 < \beta$, $\vec{\mu}$ of a dynamic address block should be closer to $\vec{0}$ than to $\beta\vec{1}$. We specify it in the following hypothesis test:

$$H_0 : \vec{\mu} = \vec{0} \; vs. \; H_1 : \vec{\mu} = \beta\vec{1}$$

If we accept H_0, it means that the likelihood that an observed response vector $\vec{\bar{x}}$ comes from a multivariate normal distribution with $\vec{\mu} = \vec{0}$ is larger than the likelihood that it is from a distribution with $\vec{\mu} = \beta\vec{1}$, which means we are more confident that most of the addresses within the block have a relatively smaller response rate compared with β, hence we identify it as a dynamic block.

Under the normality assumption, comparison of those two confidence levels is equivalent to comparing of two Hotelling's T^2 statistics, $t_0(\vec{\bar{x}}) = n(\vec{\bar{x}} - \vec{0})^T S^{-1} (\vec{\bar{x}} - \vec{0})$ and $t_1(\vec{\bar{x}}) = n(\vec{\bar{x}} - \beta\vec{1})^T S^{-1} (\vec{\bar{x}} - \beta\vec{1})$, which stands for the statistical distance between $\vec{\mu}$ and $\vec{0}$ and between $\vec{\mu}$ and $\beta\vec{1}$, respectively. Consequently, when $t_0(\vec{\bar{x}}) < t_1(\vec{\bar{x}})$, we believe the block to be dynamic. Choosing different β values (from 0 to 1), We show the ROC plot for the overall response rate test in Fig. 6(b).

3) Combining two hypothesis tests

Because the two hypothesis tests characterize different aspects of dynamic address blocks in terms of their responses towards scanning traffic, the dynamic address blocks will be able to pass both hypothesis tests. In a word, under these two hypothesis tests, our rule for identifying dynamic address blocks is: *An address block is considered as dynamic if it passes both of the two hypothesis tests.* If a block fails any of those two hypothesis tests, we have enough confidence to believe it is not a dynamic block.

5.3 Experiment Design

In our experiment, we choose block size to be 16 (/28 blocks) [7]. Based on DNS names, we select totally 1323 blocks for our experiment, in which 1034 blocks are identified as dynamic blocks, and 289 of them are static blocks.

We collect scanning traffic targeted at those blocks, and partition all the scanning flows into 5-minute intervals, or scanning phases, based on their arrival time. Notice that due to the huge amount of available scanning traffic, within each scanning phase, each block is likely to be scanned multiple times by different scanners. To maximize the chance of an active IP address responding to scanning traffic, we consider that there is a response from that address if it responds to at least one of all the scanning flows touching it.

For each block, we have collected a sequence of response vectors towards different scanning phases. However, to fulfil the model assumptions, in our initial experiment, we choose 4 hours as the interval length between two chronologically adjacent scanning sequences to assure the i.i.d assumption of contiguous response vectors, meanwhile, we use all the available scanning sequences within a month to guarantee a large sample size so as to fulfil the normality assumption. We will discuss the impact of different lengths of interval and different number of scanning sequences on our detection results.

In order to use the classifier to identify dynamic address blocks, we need to combine the results of those two hypothesis tests. Obviously, the combined classifier will reduce both the classification rate and the false positive rate. However, due to the fact that two hypothesis tests characterize different aspects of responses from dynamic address blocks, we expect to see a small decrease in classification rate but a large decrease in false positive rate; hence, we select parameters α,β such that they provide a high classification rate as well as a moderate false positive rate. From the ROC curves (Fig. 6(a)(b)), we choose $\beta = 1$ which gives 92.3% classification rate and 42.9% false positive rate, and $\alpha = 0.1$ which produces a classification rate of 93.4% and a false positive rate of 53.3%.

[7]For /29 and /30 blocks, the classifier also has equivalent performance as /28 blocks. For blocks larger than /28, we can classify it in an iterative way

5.4 Results

We apply the classifier to the 1323 block samples and we obtain 90.1% classification rate and 24.9% false positive rate. In other words, the classifier correctly predicts 90.1% of the dynamic address blocks and 75.1% of the static address blocks. An investigation of the misclassified blocks reveals different situations when errors may occur.

The major causes of misclassifying static blocks into dynamic blocks are: 1) Due to firewall or other reasons, some static machines do not response to majority of the scanning traffic, this situation accounts for majority of the misclassifications. This also triggers our study of maximizing response rate by choosing appropriate scanning ports, which we will introduce at the end of this section. 2) Hosts that only respond to scanning for a short period of time each day. Investigation on their flow data indicates there is no outgoing traffic from those hosts when they do not respond to scanning traffic. Their domain names illustrate that they are likely to be machines in student labs which are only turned on for a short period each day.

For those misclassified dynamic address blocks, there are two major causes: 1) Due to special dhcp assignment policies which permanently assign most of the dhcp addresses within a block to clients. If majority of the dhcp addresses within a block are unchangingly assigned to clients, the classifier will misclassify it into static blocks. 2) Heavily used wireless network, such as computer science wireless network, where almost all the IP addresses are fully utilized throughout the day.

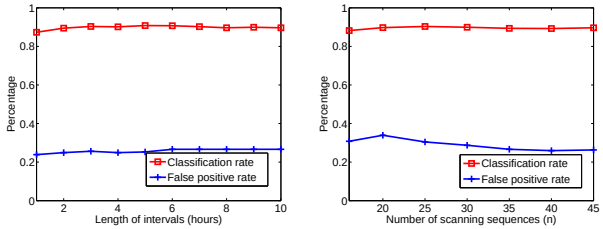
In a different perspective to evaluate the performance of the classifier, we also obtain a list of IP addresses from the network operator, which indicates the IP allocation information in different types of internal networks, e.g. dormitory network and two department networks. Our classifier has identified 82.7% blocks that belong to dorm networks as dynamic, compared to 61.5% dynamic blocks and 53.9% dynamic blocks identified in two different department networks, respectively. It agrees with our domain knowledge that dorm networks tend to have larger portion of dynamic IP addresses.

Investigation on the classification results indicates that the two-facet hypothesis testing method can accurately differentiate static address blocks and dynamic address blocks. The results also suggest that the majority of the errors are caused by those outliers in static blocks and dynamic blocks, whose activities and address allocation strategies contradict the common definitions of static or dynamic addresses. Because we are implementing a classifier to identify dynamic address blocks outside the campus network, in our next step, we discuss some important implementation issues regarding building a real-time classifier.

5.5 Choosing Implementation Parameters

From our previous experiment, we have obtained the parameters α, β based on the ROC curves. As our goal is to identify dynamic address blocks in outside-campus networks, there is usually no existing scanning sequences to utilize; instead, we need to scan the network multiple times. When we launch scanning against a particular network, we need to consider two parameters, the interval between adjacent sequences and the total number of scanning sequences. This is the case, because a large enough interval between two contiguous scanning sequences satisfies the i.i.d assumption of different samples in the response model, meanwhile, a large number of response vector samples is needed to fulfil the requirement of the normality assumption in the model.

To select a proper interval length between two contiguous scanning sequences, we test different intervals from 1 hour to 10 hours on the training dataset using one month scanning traffic, Fig. 7(a) demonstrates the classification accuracy for dynamic blocks vs. accuracy for static blocks.



(a) Given different intervals between scanning sequences (b) Given different numbers of scanning sequences

Figure 7: Classification results

Initially, we set the interval to be one hour and obtain 87.3% classification rate and 23.8% false positive rate. As we increase the length of the interval until 5 hours, we observe an increasing trend in the classification rate and a decreasing trend of false positive rate. After that, the performance of the classifier becomes a little worse (89.9% classification rate and 26.6% false positive rate) because as the interval length gets larger, the number of available scanning sequences becomes limited. In a word, the classifier has the best performance when the interval length is larger than 3 hours. It also indicates that the majority of usage time per user are less than 3 hours for dynamic IP addresses.

Another important question in practice is how many times we should scan an address block before we determine whether it is a dynamic block or not. To answer this question, we choose 4-hour interval between scanning sequences, and we look at the change of the classification performance as we increase the number of scanning sequences from 16 to 45 (Fig. 7(b)).

Starting at 16 sequences, the classifier has 88.2% classification rate and 30.8% false positive rate. From 20 to 45 sequences, we observe a slight increase of the classification rate (from 89.3% to 90.3%), and a large decrease in the false positive rate (from 30.4% to 25.9%). Therefore, the classifier reaches a good performance with only a small sample size, which is very helpful in real-time implementation, because it does not require a large amount of scanning activity.

5.6 Selecting Scanning Ports

For the purpose of dynamic and static address classification, selection of appropriate scanning ports is another decision that needs to be carefully considered in practice, whether we apply the scheme in a passive monitoring environment (e.g., for classifying our own network) or for active probing (e.g., for classifying a remote network by sending scanning probes). This is because firewalls are often installed to block certain scanning activities, in particular, those associated with known malware. In a large network (e.g., a large campus network such as ours), there may be multiple tiers of firewalls installed at various of levels of the network with diverse policies. For instance, at our campus network we have a campus-level firewall which blocks all out-

side traffic on well-known malware ports such as 137, 139, 445 and so forth. At various subnets within the campus, firewalls may be also installed at internal gateway routers, which often deploy filtering policies that are specific to the role of the corresponding subnets. For example, some departmental subnets may block certain peer-to-peer (P2P) ports, which are typically allowed in dorm subnets. On the other hand, some subnets (e.g., dorms) may block certain service ports (e.g., web and ftp ports).

Hence, if not carefully selected, blocked scanning traffic may skew the response rate and thus affect the accuracy of the classification. The diversity of firewall policies also complicates the task of scanning port selection. On the other hand, *we can in fact take advantage of the fact that different subnets (i.e., address blocks) may "favor" (i.e., let through) different scanning traffic to enhance the effectiveness and accurate of dynamic address block classification by judiciously selecting scanning ports that are address-block-specific.*

Date	Ports
02/03/2006	80,3372,4501,6129,5900
02/08/2006	4899,6129,5900,4000,8080
02/23/2006	80,4899,22,6000,5900

Table 1: Ports of top 5 response rates

	02/08/2006	02/23/2006
Static blocks	ICMP,80,443,4899,8080	22,80,4899,5900,6000
Residence Hall	ICMP,7000,7001	ICMP,1024,3072
Campus Wireless	ICMP,443,4000	ICMP,80,5900
Dial-up	1025,1026,1027,4899	1024,1025,1026,1027

Table 2: Ports with high response rates in different address blocks

To illustrate the above points, we conduct a detailed study of scanning traffic activities and their resulting responses (or lack thereof) on various address blocks using the month-long netflow records. Table 1 lists the top five ports with the highest response rates on three different days of February 2006, from the perspective of the *entire* campus network. We see that the top five ports tend to contain some well-known service ports such as HTTP (80 or 8080), ssh (22), X11 (6000), as well as some ports providing special remote services or applications that have also some known vulnerabilities such as 4899 (radmin – remote administrator default port), 3372 (Microsoft distributed transaction service coordinator for window 2000), 5900 (vnc – virtual network computer), 6129 (dameware remote admin), and so forth. On the other hand, Table 2 shows that the ports that elicit highest response rates on several static blocks as well as dynamic address blocks of three different types on two different days can be quite different from those top five ports viewed from a "global" (entire campus network) standpoint. These results demonstrate that different types of address blocks may "favor" different scanning traffic, due to the diversity in their firewall policies as well as the nature of the machines (and the applications) running within the blocks.

To evaluate the impact of scanning port selection on the effectiveness and accuracy of IP address classification, we perform series of experiments by selecting scanning traffic using different scanning ports. Recall that in the previous subsections we used *all* scanning traffic within some test periods (20 minutes) that are randomly selected and spaced apart with a certain minimum threshold (the interval length), say, 5 hours. As these results are the baseline results, we conduct two series of experiments for each address block: in the one series we select fixed sets of 5 ports randomly chosen from the top 20 ports that elicit highest response rates from the perspective of the entire campus network; in the other series we select the top 5 ports that elicit highest response rates from the perspective of a specific address block under testing.

We find that for the first series of experiments using the randomly selected five ports, the classification rates either stay approximately the same for some blocks, or drop to nearly 75% for some blocks. Likewise, the false positive rates are either not affected significantly, or increase to nearly 35%. Detailed investigation reveals that for those blocks with decreased classification rates or increased false positive rates, there are two factors in play: first and the main factor is that using the randomly selected scanning ports, the number of scanning sequences in each period may be drastically reduced (sometimes to none), thereby significantly skewing the testing results; and second, in the cases of scanning sequences do exist in a test period, some of the randomly selected ports do not elicit any response (although the target IP addresses are "live"), perhaps because these ports are blocked, or no services on these ports are running on the machines, but the ICMP responses are blocked. In contrast, when using the *address-block-specific* scanning port selection (namely, the top five ports with highest responses), the results display noticeable improvement over the baseline results, with the classification rates increased by an average of 3%, and false positive rate reduced by an average of 5%. Due to space limitation, we do not include the detailed results here.

In summary, our results show that carefully selecting scanning ports is important for IP address classification; moreover, by using the address-block-specific selection strategy – e.g., by selecting the scanning ports with the highest response rates over an extended observation period, we can further improve the classification rate and reduce the false positive rate of our proposed scanning-based dynamic address classifier. Such a selection strategy is feasible and practical in a passive monitoring environment for serendipitously learning and classifying internal IP address blocks using "background" scanning traffic, as is the case we focus on in this study. To apply the proposed dynamic IP address classification to an active probing environment to learn and classify remote IP addresses, the issue can be more complicated: either a "learning" phase is used ahead of time for learning the appropriate scanning ports for classification, or a fairly large of number of ports (or "vertical port scanning") is used in the classification process, and top ports that elicit highest rates are included in the hypothesis testing. Due to the intrusive nature of active probing, we do not conduct such experiments in this study. Evaluation and testing of such "active-probe" based dynamic address classification will be left to a future paper.

6. CONCLUSIONS AND FUTURE WORK

In this paper, we investigated methods for identifying dynamic address blocks within a large campus or enterprise network. Using DNS name-based method, we first extracted a number of static and dynamic address blocks to study their traffic activities and we defined activity matrices to charac-

terize their incoming and outgoing traffic activity patterns. Next, we focused on scanning traffic towards those blocks and illustrated significant different patterns of responses to scanning traffic between static and dynamic address blocks. Based on our observations, we designed a classifier which accurately identifies dynamic address blocks based on the response model of two hypotheses tests. Finally, we discussed key issues for building a real-time classifier. Our current work is to implement an online classifier to identify dynamic address blocks on the Internet.

7. REFERENCES

[1] Y. Xie, F. Yu, K. Achan, E. Gillum, M. Goldszmidt and T. Wobber. How Dynamic are IP Addresses. In *Proc. of ACM SIGCOMM*, 2007.

[2] A.V.Ramachandran, N.Feamster. Understanding the Network-level Behavior of Spammers. In *Proc. of ACM SIGCOMM*, September 2006.

[3] M. Casado and M. Freedman. Peering Through the Shroud: The Effect of Edge Opacity on IP-Based Client Identification. In *Proc. of ACM/USENIX NSDI*, 2007.

[4] L. Gomes, C. Cazita, J. Almeida, V. Almeida, and W. Meira. Characterizing a spam traffic. In *Proc. of ACM SIGCOMM IMC*, 2004.

[5] J. Jung, E. Sit. An Empirical Study of Spam Traffic and the Use of DNS Black Lists. In *Proc. of ACM SIGCOMM IMC*, 2004.

[6] K. Xu, Z.-L. Zhang and S. Bhattacharyya. Profiling Internet Backbone Traffic: Behavior Models and Applications. In *Proc. of ACM SIGCOMM*, August 2005.

[7] T. Karagiannis, K. Papagiannaki and M. Faloutsos. BLINC: Multilevel Traffic Classification in the Dark. In *Proc. of ACM SIGCOMM*, August 2005.

[8] Whois.net-Domain Research Tools. http://www.whois.net/.

[9] Y.Jin, G. Simon, K.Xu, Z.-L. Zhang and V. Kumar. Gray's Anatomy: Dissecting Scanning Activities Using IP Gray Space Analysis. In *SysML07*, 2007.

[10] R. Johnson, and D. Wichern. *Applied Multivariate Statistical Analysis*. Prentice Hall, 2007.