Understanding Address Usage in the Visible Internet

USC/ISI Technical Report ISI-TR-656, February 2009

Xue Cai John Heidemann
USC/Information Sciences Institute, {xuecai,johnh}@isi.edu

ABSTRACT

Although the Internet is widely used today, there are few sound estimates of network demographics. Decentralized network management means questions about Internet use cannot be answered by a central authority, and firewalls and sensitivity to probing means that active measurements must be done carefully and validated against known data. Building on frequent ICMP probing of 1% of the Internet address space, we develop a clustering algorithm to estimate how Internet addresses are used. We show that adjacent addresses often have similar characteristics and are used for similar purposes (61% of addresses we probe are consistent blocks of 64 neighbors or more). We then apply this block-level clustering to provide data to explore several open questions in how networks are managed. First, the nearing full allocation of IPv4 addresses makes it increasingly important to estimate the costs of better management of the IPv4 space as a component of an IPv6 transition. We provide about how effectively network addresses blocks appear to be used, finding that a significant number of blocks are only lightly used (about one-fifth of /24 blocks have most addresses in use less than 10% of the time). Second, we provide new measurements about dynamically managed address space, showing nearly 40% of /24 blocks appear to be dynamically allocated, and dynamic addressing is most widely used in countries more recently to the Internet (more than 80% in China, while less then 30% in the U.S.).

Categories and Subject Descriptors

C.2.1 [**Computer-Communication Networks**]: Network Architecture and Design—*Network topology*; C.2.3 [**Computer-Communication Networks**]: Network Operations—*Network management*

General Terms: Measurement

Keywords: Internet address allocation, survey, pattern analysis, clustering, classification, availability, volatility

1. INTRODUCTION

Previous Internet topology studies focused on AS- and router-level topologies [3, 5, 7, 10, 21, 25, 26]. While this work explored the core of the network, it provides little insight into the edge of the Internet and the use of the IPv4 address space. The transition to classless routing [9] in the mid-1990s has made the edge opaque. Only recently have researchers begun to study edge-host behavior using server logs [32], web search engines on textual addresses [28], and ICMP probing [12].

Assumptions: In this paper we begin to explore the potential of *clustering of active probes to infer network address usage.* Our work makes three assumptions:

1. many active addresses will respond to probes,
2. contiguous addresses are often used similarly, and
3. patterns of probe responses suggest address usage.

While there are cases where these assumptions do not hold, we believe they apply to a large fraction of the Internet and so active probing can provide insight into address usage. Recent work has shown that active probes detect the majority of addresses in use, confirmed against large university and a random sample of the general Internet [12], supporting the first assumption.

We explore the rest assumptions in this paper. Our assumption about contiguous use follows from the traditional administrative practice of assigning blocks of consecutive addresses to minimize routing table sizes While there is no requirement that adjacent addresses be used for the same purpose, we will show that they often used similarly (Section 5.1).

Finally, we assume that repeated active probing with ICMP provides some information about how addresses are used. While ICMP provides only limited information (is the addresses responsive or not), repeated probing separates addresses in constant use from those used intermittently. We will show that these observations correlate with servers or always-on-desktop computers and dynamically assigned address block (Sections 4 and 6.1).

Figure 1 shows an example of what probing reveals, given one block of 256 addresses with prefix $p1$ where

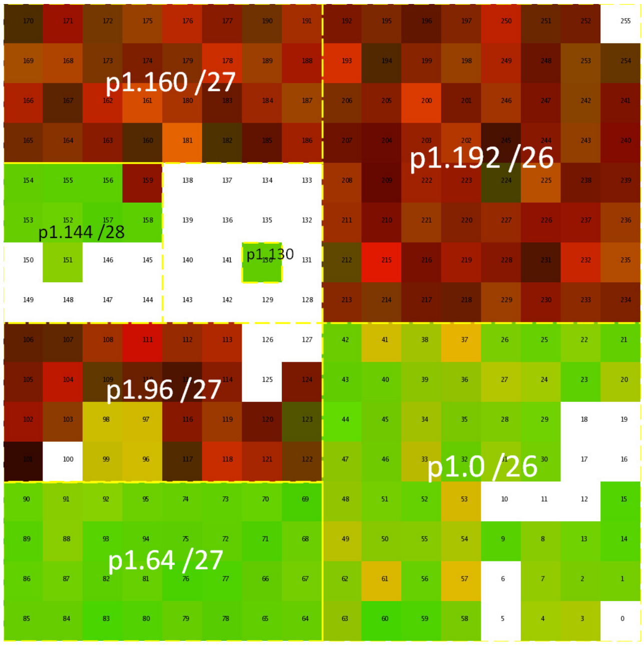

Figure 1: A /24 block (prefix is anonymized to $p1$) where probing suggests seven different regions. Addresses are on a Hilbert curve.

these assumptions apply[1]. Different shades indicate different ping response patterns by each address (green is availability and red is volatility, metrics we define later in Section 3.2). This block clearly shows two different patterns for seven address blocks. Manual examination shows these regions correspond with web servers and dial-up addresses. Hostnames suggest the lower left quarter of the block ($p1.0/26$) bottom right eighth ($p1.64/27$), and middle left ($p1.144/28$) are populated mostly with web servers; except for a few unoccupied addresses, these regions are almost always up and so show as light colors. Hostnames in the dark regions, upper-right ($p1.192/26$), upper-left ($p1.160/27$), and middle-left ($p1.96/27$) suggest use for dial-up, and probing shows they are only used infrequently and intermittently.

Approach: From these assumptions we develop new algorithms to identify blocks of addresses with consistent usage (Section 3). We start with Internet survey data, where each address in around 24,000 /24 address blocks is pinged every 11 minutes for one week [12]. From this dataset we derive several metrics about how each address is used. We then use these statistics to automatically identify blocks of consistent responsiveness. Ping responsiveness does not directly identify address use, so to get a better understanding of use we correlate our metrics with uses inferred from hostnames assigned to those addresses (Section 4).

Validation and Applications: Before applying these algorithms, we evaluate how often our assumptions hold. Our first question is therefore *are adjacent addresses used consistently* and *can we discover them reasonably accurately?* Before classless IP addressing [9] allocation strategies were aligned with externally visible address allocation, but since then there has been no way to easily evaluate how addresses are used. We explore these basic questions in Sections 5.1 and 6.1.1.

A first application of this approach is to understand how addresses are managed, beginning with what block sizes are typical(Section 5.1). We find that 2,529,216 addresses, or 61% of the probed address space, show consistent responses in blocks of 64 to 256 adjacent addresses (/26 to /24 blocks). And we observe that most addresses (around 55%) in the Internet are in /24 or bigger blocks.

A second application is to understand how effectively addresses are used (Section 5.2). We find that a significant number of blocks are only lightly used (about one-fifth of /24s show less than 10% utilization). These questions are of growing importance as the IPv4 address space nears full allocation; they help estimate the costs of improving IPv4 efficiency as compared to IPv6 transition.

Finally, we detect and quantify the use of dynamic address assignment (Section 5.3). Dynamic addresses are important for several reasons. Since they often represent poorly secured home computers, dynamic addresses factor in to some spam detection algorithms [32]. Identifying dynamic addresses is important to estimate the number of computers that do connect to the Internet [12]. We observe that nearly 40% of /24 blocks appear to be dynamically allocated to computers, and dynamic addressing is much higher in countries most recent to the Internet (more than 80% in China, while less then 30% in the U.S.).

The contribution of this paper is therefore to develop new approaches to classify Internet address usage and to apply those approaches to answer important questions in network management. As with many prior studies of the Internet, our approach is based on limited information and we do not claim perfect accuracy. However, we suggest the approach is promising and our preliminary results add important observations to what is currently known.

[1] Recall that IPv4 addresses are 32-bit numbers, usually written in the form $a.b.c.d$, where each component is an 8-bit portion of the whole address. Addresses are organized in *blocks* (sometimes called subnetworks) that are sized to powers of two. Blocks have a common *prefix*, the leading p bits of the address, written $a.b.c.d/p$. For example, 128.125.7.0/24 indicates a /24 block with 256 addresses in it of the form 128.125.7.x. We sometimes talk about blocks as p.0/24, where p represents the anonymized prefix.

2. RELATED WORK

A lot of work has been done to understand the Internet, most exploring on Internet topology [3, 5, 7, 10, 21, 25, 26]. Recent work has begun exploring edge host behavior [12, 28, 32]. Our work builds upon this prior work and specific work listed below.

Are contiguous addresses consistent and what are the typical block sizes? Although addresses are usually assigned as blocks and represented in prefixes by classful addressing [22] and classless addressing [9], there is no guarantee that contiguous addresses in the same block will be used in the same way. Huston's report has analyzed the common prefix lengths in BGP routing table [13]. But it cannot look at usage at granularities smaller than BGP prefixes. Our approach is able to look at these smaller block sizes through active probing.

Are allocated addresses being efficiently utilized? Several researchers have studied rates of IPv4 address consumption, predicting IANA will exhaust its allocation pool sometime between 2009 and 2016 [11, 13]. However, full allocation does not necessarily imply full use. Prior researchers have infer the address utilization by detecting allocated but not advertised prefixes in the BGP routing table [19]. As with allocation, what is routed may differ from what is actively used. Our work tries to track active use; and our study of individual addresses can reveal changes that happen to blocks inside an organization (smaller than are typically routed).

How many addresses are dynamically assigned? Xie et al. have begun to explore this question with a goal of identifying dynamic blocks to assist spam prevention [32]. Their work is based on passive collection of Hotmail web server logs, while our method uses a completely different approach by active probing and so can extend and corroborate their findings. Researchers at USC provide another perspective based on active probing with ICMP [12]. We make use of their datasets, while adding completely new analysis.

Active probing of the topology has long been carried out by CAIDA [2]. They traceroute to one address for each routed /24 address block. Our datasets differ, probing only a fraction of /24s, but probing all addresses in these blocks much more frequently. Probing /24s allows us to take the advantage of locality to study address usage. Because contiguous addresses are usually administrated together and used in the same way, analyzing the whole block instead of sampling one address from each block can provide information not previously available. In addition, our frequent sampling shows temporal changes useful for identifying dynamic address allocation.

3. METHODOLOGY

This section introduces our methodology: collecting raw data through an Internet survey, transforming that data into relevant observations, identifying blocks of consistent use, classifying blocks into ping-observable categories, in Section 4 relating these ping-observable categories to several hostname-inferred usage categories.

3.1 Data Collection: Surveying the Internet

To understand the Internet, we begin by collecting data about it. While we desire as much information about each Internet address or host as possible, we must balance that desire against today's security-conscious Internet culture. We chose to build on prior *Internet ICMP surveys* that ping each address of about 1% of the allocated Internet address space approximately every 11 minutes for one week or longer [12].

We believe 1% of the allocated address space represents a large enough fraction of the space to be representative. We follow the selection methodology outlined previously [12], selecting around 24,000 /24 blocks from blocks that were responsive in a prior census of all allocated addresses. We select blocks of addresses rather than individual addresses so we can study how addresses are allocated and used. Our choice of /24 blocks limits our ability to observe very large allocations, but allows the identification of blocks smaller than 256 addresses (Section 5.1). As with prior work, a half the selected blocks are kept consistent across multiple surveys and half are chosen randomly, enabling longitudinal studies while providing an subset that is selected with very little potential bias. We compare two surveys in Section 6.2;

Probes taken every 11 minutes limit our ability to detect very rapid churn of dynamic addresses, however prior studies of dynamic addresses placed typical use durations at 75 or 81 minutes [12, 17], suggesting we have reasonable precision. We use 1-loss repair to cope with singleton packet losses [12].

Observing the same addresses for at least a week allows us to detect several daily cycles, both for weekends and days of the work week. In the future we would like to expand to two-week surveys to provide some duplication of the weekly cycle.

Of course, use of ICMP for probing has significant limitations. The most serious is that large parts of the Internet are firewalled and choose not to respond to our probes. Some form of this bias is inherent in any study using active probing. Prior studies of a large university and a random sample of Internet addresses suggest ICMP probing undercounts hosts by a factor of 30–50%, and that ICMP is superior to TCP-based probing [12]. We recognize this limitation as fundamental to our methodology, but we know of no evidence or inference to suggest that the firewalled portions of the Internet use significantly different allocation strategies than the more open parts of the Internet. However, we are exploring additional ways to verify this assumption.

Name	Start Date (# days)	/24 Blocks probed	respond.	Use
IT17ws [30]	2007-06-01 (10)	22,367	20,849	all
IT17wrs	2007-06-01 (10)	17,366	16,295	§4
IT16ws [29]	2007-02-16 (6)	22,365	20,900	§6.2
VUSC s [31]	2007-08-13 (9)	768	299	§6.1
ISC-DS [15]	2007-01	hostnames		§4
RIR [23]	2007-06-13	block allocation		§5

Table 1: Datasets used in this paper.

Table 1 shows the datasets we use in our paper. We use two ICMP surveys taken by USC [12]: *IT17ws* and *IT16ws*; *IT17ws* is the main dataset used in this paper, while we use *IT16ws* for validation in Section 6.2. We collected *VUSC s* at our enterprise in order to compare our inferences with network operators as discussed in Section 6.1. Finally, we use a domain name survey from ISC [15] for training in Section 4, comparing that with *IT17wrs*, an overlapping subset of *IT17ws*.

3.2 Data Representation: Observations of Interest

Since one survey provides more than 5 billion observations, it is essential to map that raw data into more meaningful metrics that characterize address usage. We call this step *data representation*, and we define three metrics: *availability*, the fraction of time an address is responsive; *volatility*, a normalized representation of how many consecutive periods the address is responsive; and *median-up*, the median duration of all up periods.

To define these more formally, let

$$r_i(a) = 1 \text{ or } 0, \forall i \in [1, N_p]$$

be the positive or negative response of the i^{th} of the N_p probes to address a after 1-loss repair [12]. If each probe is made at time t_i, we can define the N_u up durations ($N_u < N_p$) of a survey as

$$\begin{aligned} u_j(a) \quad = \quad & t_{e_j} - t_{b_j}, \text{ where} \\ & r_i = 1, \forall i \in [b_j, e_j] \\ & \text{and } r_{(b_j)-1} = 0, r_{(e_j)+1} = 0 \\ & \forall j \in [1, N_u] \end{aligned}$$

(each up duration is a consecutive run of positive probes from b_j to e_j, inclusive). We can now clarify that availability, volatility, and median-up are given as:

$$\begin{aligned} A(a) &= \frac{1}{N_p}\left(\sum_{1}^{N_p} r_i\right) \\ V(a) &= N_u/(N_p/2) \\ U^*(a) &= \text{median}(u_j,\ \forall j \in [1, N_u]) \end{aligned}$$

Availability is normalized; it is the fraction of times a host is reachable. Volatility is normalized by $N_p/2$, the maximum number of states (alternating value each time). For example, if $N_p = 16$, and the responses r_i of address a are $[1, 1, 0, 0, 1, 1, 1, 1, 0, 0, 0, 1, 1, 1, 1, 1]$, then there are three up periods ($N - u = 3$) of lengths 22, 44, 55 minutes each. $A(a) = 11/16 = 0.688$, $V(a) = 3/(16/2) = 0.375$ and $U^*(a) = \text{median}(22, 44, 55) = 44$ minutes. (We also sometimes use un-normalized volatility, $V^*(a) = N_u$, simply the count of up periods.) We considered normalizing median-up to measurement duration, but chose not to because such normalization distorts observations about hosts that are not nearly always present. Finally, we often omit the (a) when the subject of the metric is clear.

While these metrics are not completely orthogonal, each has a purpose. Availability shows how effectively addresses are used. High volatility indicates addresses that are only intermittently used and often dynamically allocated. Median uptime suggests how long an address is used.

These estimates assume the r_i observations are correct and represent a single host. Because we know our data collection omits firewalled hosts (Section 3.1), we generally ignore addresses that do not ever respond. More troubling are addresses that are used by multiple computers at different times—such addresses actually represent *multiple* hosts. The purpose of dynamically allocated addresses is exactly to share one address with multiple computers, and we know dynamic assignment is common (see Section 5). If those hosts are used for different purposes (servers sometimes, and clients others), usage inference will be difficult and unreliable. However, we believe that it is relatively uncommon for a dynamic address to transition between client and server use, since servers usually require stable addresses. (There is some use of dynamic DNS to place services on changing addresses. We believe such use is rare for most of the world but plan to explore this issue in future work.)

3.3 Block Identification

We next use our observations about addresses to evaluate block size. To do this identification we develop a new clustering algorithm.

We assume blocks are allocated in sizes that are powers of two, so block identification is the process of finding a prefix where addresses in the block are used consistently. We find that some blocks are *not* used consistently, and different addresses show very different stability. In our analysis we will keep dividing these *mixed-use blocks* until they are consistent, if necessary devolving to a single address per block. Another challenge is that many blocks have gaps where a few addresses are used differently, or are not responsive, perhaps because they are unused or firewalled. Our algorithm weighs choice of larger blocks with some inconsistencies against smaller but more homogeneous blocks.

We only consider blocks of sizes /24 or smaller because current data collection method does not guarantee blocks larger than /24s. (Exploration of larger blocks is an area of potential future work.)

3.3.1 Clustering background

Our goal with clustering of address responsiveness is to determine whatever blocks that appear to be used consistently. We therefore use *partitional clustering*, one of the two general approaches to clustering in this well developed field [16]. Partitional clustering places each element into exactly one cluster; we choose it over the alternative, hierarchical clustering, which would place items into multiple, hierarchically nested clusters.

Jain describes partional clustering as: "Given n patterns in a d-dimensional metric space, determine a partition of the patterns into K groups, or clusters, such that the patterns in a cluster are more similar to each other than to patterns in different clusters" [16]. We build on the basic approaches of clustering for our method: a *pattern matrix*, *feature normalization*, and use of an *elbow criterion* to select the best choice.

Although we follow traditional clustering theory, Internet addresses impose a unique restriction. Addresses are only grouped into blocks that are contiguous, sizes of powers of two, and aligned at multiples of the size. For these reasons, we cannot directly use traditional algorithms such as K-means, but instead use components of existing clustering approaches. The most radical difference from traditional clustering is that addresses are only clustered with some number of immediate neighbors, not with arbitrary other addresses. We therefore find blocks of consecutive addresses by the definition of our algorithm, however the *size* of blocks it finds depends on the consistency of how the addresses are used.

A *Pattern Matrix* defines the features over the space being clustered. In our case, each address is defined by its three features (A, V, U^*), and the space is a number of disjoint /24 blocks. Each /24 block has a 256×3 pattern matrix x^*_{ij}, where j enumerates the three features, and i enumerates each address in a /24 block. From our 24,000 /24 blocks we get 24,000 pattern matrices in total.

Although our definitions of A and V are already normalized to the range $[0, 1]$, their distribution may be skewed, and U^* is not normalized. We therefore employ *feature normalization* to give each features equal weight. We define the normalized feature vector x_{ij}, given the mean and standard deviations m_j and s_j of each feature j:

$$x_{ij} = \frac{x^*_{ij} - \mu_j}{\sigma_j}$$

where μ_j and σ_j are the mean and standard deviation. We use Euclidean distance between two components of the feature vector to measure dissimilarity between two elements i and k over their features:

$$d(i, k) = \sqrt{\sum_{j=1}^{3} (x_{ij} - x_{kj})^2}$$

Many clustering algorithms, like K-means, require the number of clusters be chosen in advance. We cannot do that because clusters correspond to block size, a quantity we wish to discover. We also cannot simply minimize variance, because variance is trivially minimized in the degenerate case where each cluster is a singleton address.

We therefore employ an *elbow criterion*, a common rule of thumb to determine the number of clusters. We split each cluster into two whenever splitting adds significant information, and we stop when we pass the "elbow" of the curve and more clusters add little benefit. We measure information by the sum of variance in each cluster across the population—homogeneous clusters will have low variance; splitting them adds no new information. Heterogeneous clusters have high variance, and splitting them into two more self-consistent pieces reduces the sum of variance, increasing the amount of information.

3.3.2 Our Algorithm to Identify Block Sizes

Our algorithm follows the basic structure we outline above: we define a pattern matrix of addresses by features, normalize the features, then recursively search for clusters until reaching the elbow. We fill in the details next.

The algorithm is a recursive function, *BlockSizeId*, taking an address-feature matrix $256 \times (A, V, U^*)$ and a given prefix length P. Since the blocks in our survey are disjoint, we iterate over each /24 block in our survey separately, beginning with $P = 24$.

BlockSizeId then computes the sum of intra-block variance for all possible prefix lengths $p(P \leq p \leq 32)$ and selects smallest prefix length p_{elbow} where longer prefixes show minimal change. We define $vsum_p$ as sum of intra-block variance of sub-blocks with prefix length p:

$$n_p = 2^{p-P}, s_p = 2^{32-p}, \mu_{bj} = \frac{\sum_{i=(b-1)s_p+1}^{bs_p} x_{ij}}{s_p}$$

$$v_b = \sum_{j=1}^{3} \sum_{i=(b-1)s_p+1}^{bs_p} (x_{ij} - \mu_{bj})^2, 1 \leq b \leq n_p$$

$$vsum_p = \sum_{b=1}^{n_p} v_b, P \leq p \leq 32$$

where n_p is the number of sub-blocks with prefix length p, s_p is the size of sub-blocks (number of addresses) with prefix length p. For example, if $P = 24$ and $p = 27$, then $n_p = 8$ and $s_p = 32$. m_{bj} is the mean value of the j^{th} feature of addresses in the b^{th} sub-block. v_b is the intra-block variance of the b^{th} sub-block. In this example, it would be the intra-block variance of the b^{th} /27 sub-block.

We define minimal change in the elbow algorithm with an empirically selected constant threshold, ϵ. We select p_{elbow} as some p such that $vsum_{p+1} - vsum_p < \epsilon$. If $p_{elbow} = P$, then *no* division of this block reduces variance significantly and we terminate our recursive algorithm, declaring P the consistent block size. If this case does not hold, we have determined there are splits of the block that appear to be more consistent. We then split the block in half and recurse, calling *BlockSizeId* with the next longer prefix $P = p + 1$ on each half of the data. In principle, a block could be split repeatedly until it is composed on a single address and the algorithm terminates with zero variance. In practice, in Section 5.1 we show that the majority of the Internet addresses fall into larger blocks of consistent use.

3.3.3 A Block Identification Example

To illustrate *BlockSizeId* we next show analysis of an example /24 block taken from the Internet. Figure 2 shows the whole $p2.0/24$ surveyed block and the process of identifying the 4 consistently used blocks inside of /24 subnetwork. To a human observer, common patterns in the block are the /25 block on the left (red, indicating large volatility), a /27 block on the top right (dark red, indicating low availability and moderate volatility), a second /27 block on the bottom right (dark red), and the third /27 block on the bottom right (green, indicating high availability and low volatility). Unlike Figure 1, this block does not have reverse DNS entries and so we cannot confirm these assumptions with hostnames as the method shown in Section 4.

The graph immediately under the address plot in Figure 2 the first pass of *BlockSizeId* for the feature matrix for $p2.0/24$ and $P = 24$. In the graph, the y-axis shows variance for division of the block into each possible power-of-two smaller size. Here $p_{elbow} = 25$ and $p_{elbow} > P$, so we recurse on $p2.0/25$ and $p2.128/25$ with $P = 25$.

The second row of two graphs shows these two recursive invocations, $p2.0/25$ on the left and $p2.128/25$ on the right. First, considering $p2.128/25$ the graph on the right shows a consistent variance regardless of subdivision, and $p_{elbow} = P = 25$. This prefix appears to be consistently used and this branch terminates successfully. Second, for $p2.0/25$ (the graph on the left), a subdivision reduces variance and so we recurse again with $P = 26$.

The algorithm recurses until either $p_{elbow} = P$ or $P = 32$. In this example, the initial /24 block is divided into $p2.0/27$, $p2.32/27$, $p2.64/27$, and $p2.128/25$.

3.4 Ping-Observable Block Classification

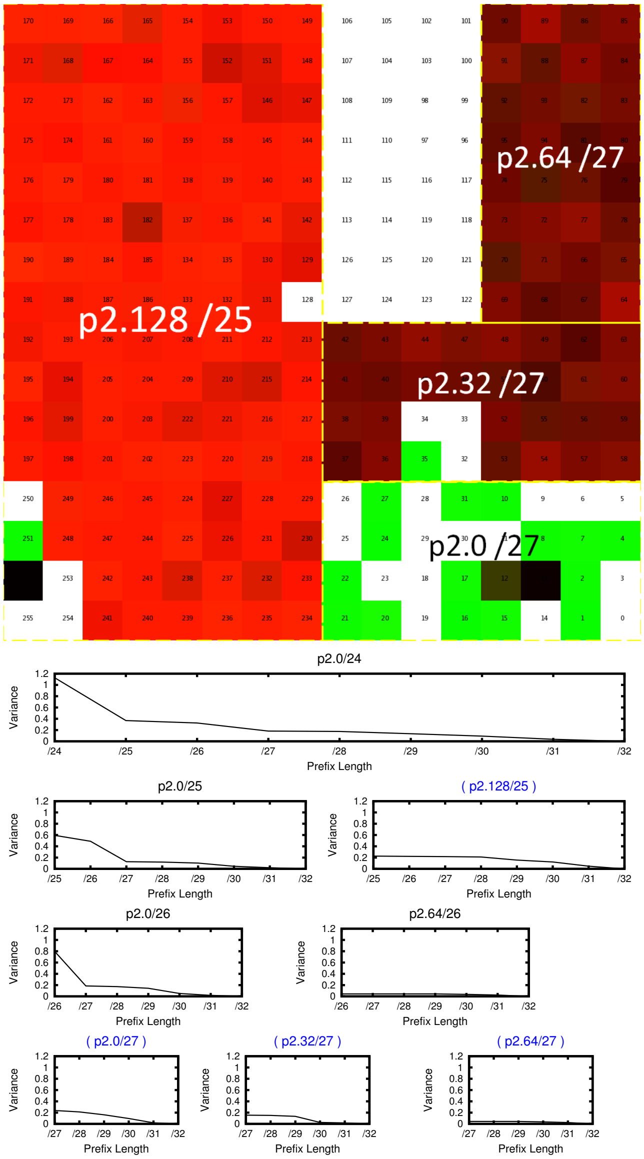

Figure 2: An example of *BlockSizeId* with threshold $\epsilon = 2.0$. A plot of the addresses is shown (top), while each row of graphs shows the variance at each recursion. Graphs of the four selected blocks are labeled with (parentheses).

We can now take remote measurements, convert them into observations, and use them to identify blocks of consistent neighboring addresses. We generalize our observations on addresses into observations about a block b by taking the median value of each observation:

$$(A(b), V(b), U^*(b)) = \text{median}(A(a), v(a), U^*(a)) \, \forall a \in b$$

We then classify these blocks into five *ping-observable categories*, using $(A(b), V(b), U^*(b))$. We use four thresholds, $\alpha_H = 0.95$, indicating high availability, $\alpha_L =$

0.10, indicating low availability, $\beta = 0.0016$, for low volatility, and $\gamma = 6$ hours, corresponding to a relatively long uptime. The specific thresholds we give are somewhat arbitrary, but were selected to provide reasonably good correspondence between these ping-observable categories and the hostname-inferred usage categories described next in Section 4. We examine sensitivity to our choices in Section 6.2.

Always-stable : highly available and stable.

$$(A \geq \alpha_H) \wedge (V \leq \beta)$$

Sometimes-stable : changing more often than always-stable, but frequently up continuously for long periods (high U^*).

$$(U^* \geq \gamma) \wedge (A \geq \alpha_L) \wedge (A < \alpha_H \vee V > \beta)$$

Intermittent : individual addresses are up for short periods (low U^*):

$$(U^* < \gamma) \wedge (A \geq \alpha_L) \wedge (A < \alpha_H \vee V > \beta)$$

Underutilized block : although addresses are occasionally used, they show low A values.

$$A < \alpha_L$$

Unclassifiable : we decline to classify blocks with few active responders. Currently we consider any block where fewer than 20% of addresses responding as unclassifiable.

We selected these categories to split the majority of the (A, V, U^*) space.

Appendix A shows how these terms divide the space.

4. TRAINING AND HOSTNAME-INFERRED USAGE CATEGORIZATION

Our methodology takes data about use of public addresses and produces five ping-observable categories. We would like to relate those categories to terms that are more meaningful to network operators, and to find what root causes correspond to and potentially cause blocks to be intermittent or underutilized.

Determining the operational characteristics of a network is quite challenging, however. In some cases we are able to discuss network policy with the operations staff to confirm our assumptions; we will use this approach to validate our conclusions against a large campus network in Section 6. However, such observations may be biased by the policies of a single institution. We would like to also draw data from the Internet at large, but it is infeasible to contact operations for large parts of the network. While tools such as nmap [18] can extract significant information from a network through sophisticated active probing, their use is easy confused with hostile network activity by many network operations.[2]

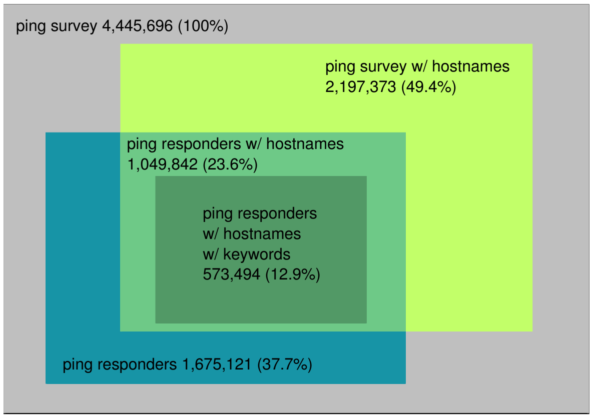

Figure 3: Our Investigation Targets: IP addresses ever responded in *IT17wrs* and have meaningful hostnames (with keywords). It is the middle part with 573,494 addresses in this figure.

Hostnames are a source of data that provides some information about how public computers are used—many hostnames contain keywords such as “www”, “dynamic”, or “dsl”. Wide hostnames collection is also feasible: many Internet hosts suggest reverse DNS lookup [20], reverse lookup occurs commonly as part of normal operation and so is unlikely to be seen as hostile. The Internet Systems Consortium has collected full tables of reverse DNS regularly since 1994 [15] and makes it available for a nominal fee.

We next describe how we map hostnames to 15 *hostname-inferred usage categories*, and how this data corresponds to our five ping-observable categories.

While one might study the Internet using hostname data alone, without ping data, we believe the information complements each other. About half addresses that are used lack reverse hostnames, and about 49% of hostnames lack meaningful keywords, and reverse names may not represent the computer’s true use (for hosts with multiple names, where often the reverse name is automatically assigned) so we think hostnames alone are not sufficient.

4.1 Hostname-inferred Usage Categories

Although hostnames are not perfect, we believe they provide a useful dataset to compare against our ping-observable categories. We use ISC survey 17 [14], taken slightly before the ping survey used for our primary analysis [30].

[2] Widescale nmap use would place us in contact with additional operations staff, but perhaps not on ideal terms.

group	category	keywords	count
allocation	static	static	28,137
	dynamic	dynamic, dyn	105,882
	dhcp	dhcp	14,290
	pool	pool, pond	66,009
	ppp	ppp	44,729
access link	dial	dial, modem	80,090
	dsl	dsl	208,682
	cable	cable	29,761
	wireless	wireless, wifi	910
	ded	ded, dedicated	733
consumer	biz	business, biz	12,999
	res	res, resident	25,847
	client	client	9,994
server	server	server, srv, svr, mx, mail, smtp, www, ns, ftp	12,568
	router	router, rtr, rt, gateway, gw	2,850

Table 2: Categories of hostname-derived usage.

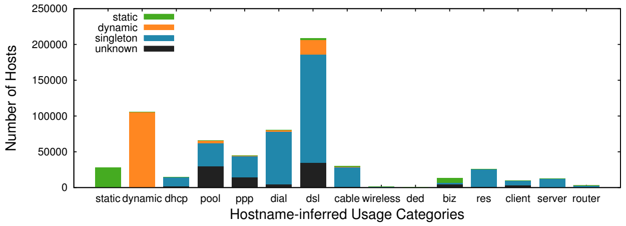

Figure 4: Numbers of hostname-inferred usage categories, with colors indicating those that also have allocation types.

Figure 3 shows the overlap of these datasets. shows our investigation targets. We begin with the 4.4M IP addresses probed the ping survey Nearly half of these (2.2M) have hostnames in the ISC reverse DNS survey. Of the 1.6M ping addresses that respond, we consider the 1.0M addresses that also have hostnames. We then focus on the 573,494 of those that have identifiable keywords in their hostname (12.9% of all addresses in the ping survey).

We follow recommendations that were proposed as standard naming conventions for Internet hosts [27] and that occur in 2000 or more hosts in our dataset. Although these were neither approved by the IETF, nor would the be mandatory even if approved, these terms do appear in about one-quarter of reverse hostnames. From their recommendations we define 15 *hostname-inferred usage categories* as shown in Table 2.

Figure 4 shows the count of hostnames in each category. The sum exceeds 573k because these categories partially overlap, and a single hostname may be in multiple categories. For example, some providers label DSL addresses with both DSL and static or dynamic. We see that access links keywords (DSL, dial, etc.) are very common, occurring in 51% of hostnames, and allocation types (static, dynamic, etc.) occur in about 22% of hostnames in *ping survey w/ hostnames*.

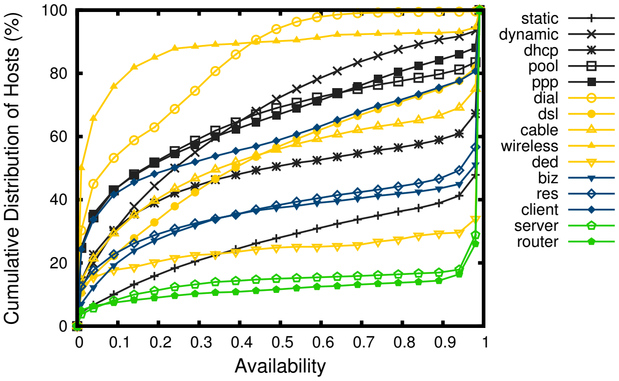

Figure 5: CDF of address availability (A) by hostname-inferred categories in *IT17ws*.

To provide some understanding the number of hostnames with multiple keywords, we subdivide each category by those that also contain static, dynamic, or any other additional keyword. Several groups types often have an additional indication of allocation type: while 10% of dsl are labeled dynamic (1.2% static), 50% of biz are labeled static. These secondary attributes reveal some technology trends: the ratio of dial also with static or dynamic types is around 1:17, while for DSL it is 1:8 suggesting increased use of static addresses in always-on DSL lines. For cable the ratio is 1:1, but the fraction of cables with an additional type is small enough that drawing conclusions may be risky.

4.2 Relating Hostname-Inferred to Ping-Observable Categories

Our goal in evaluating hostnames is to use them to understand and train our ping-observable categories. We next compare the two to see when of our observations (A, V, U^*) are correlated with hostname-inferred usage categories.

Figures 5, 6, and 7 show cumulative distribution functions of each observation against each hostname-inferred type. This data will prove essential to understand the root network causes of address underutilization and locations of dynamic addresses; we therefore defer a detailed discussion of this data to Sections 5.2 and 5.3.

Taken together, though, these graphs support the third assumption of our paper, that **patterns of probe responses can suggest address usage**. This assertion is supported because hostname-inferred categories (our approximation of usage) show fairly distinct distributions, particularly in availability (Figure 5) and median-uptime (Figure 7). As a specific example, Figure 5 shows that availability of more than 50% dial addresses is smaller than 0.1, while the A of more than 80% server addresses is larger than 0.95.

While the bulk of dial and server addresses are quite

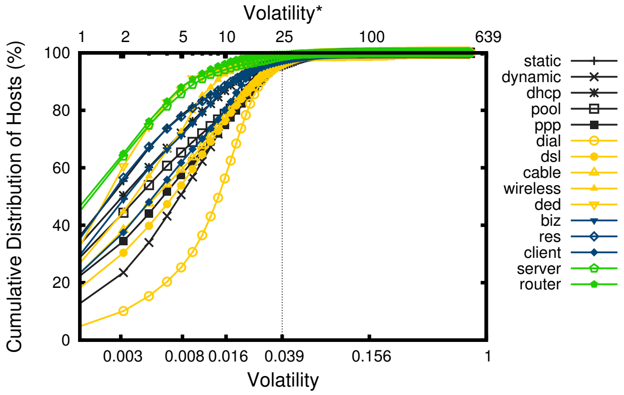

Figure 6: CDF of address volatility (V) by hostname-inferred category in *IT17ws*. Uptime counts (V^*) is shown along the top x-axis.

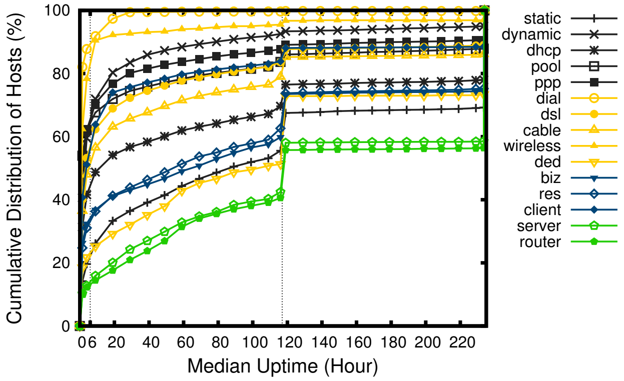

Figure 7: CDF of address median-up duration (U^*) by hostname-inferred category in *IT17ws*.

different, there are a few dial addresses with reasonably large A (5.4% have $A > 0.5$), and a moderate number of servers have poor availability (about 10% have $A < 0.15$). We conclude that, while ping-observable metrics are reasonable predictors of usage, they are far from exact, and any estimates will have fairly large error bounds. Perhaps this result is in keeping with previous observations about the great variability of the Internet [8].

Finally, we use the observations in these CDFs to set our thresholds for ping-observable classes (Section 3.4). The sharp knees at $A = 0.1$ in Figure 5 suggest $\alpha_L = 0.1$. Based on V in Figure 6, we select $\beta = 0.0016$ to separate most servers and stable uses from less stable. Finally, the sharp knee at around $U^* = 6$ hours in Figure 7 suggests this value for γ, This cutoff helps separate addresses which are not always-stable and not underutilized to two categories: sometimes-stable and intermittent.

Based on these thresholds, Figure 8 and Table 3 map the 15 hostname-inferred usage categories to the ping-observable categories (Section 3.4).

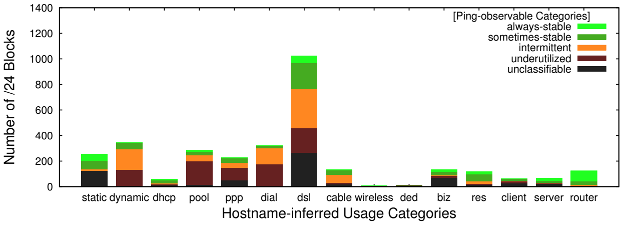

Figure 8: Relationship of ping-observed categories to hostname-inferred categories in *IT17ws*.

ping-observable category	hostname-inferred usage category
always-stable	router, server, (static), ded, (biz), (dhcp), (res)
sometimes-stable	res, static, biz, dhcp, (server), ded, (client), (cable), (dsl), (router), (ppp)
intermittent	cable, dynamic, dsl, (wireless), (dial), (ppp), (dhcp)
underutilized	pool, wireless, ppp, dial, client, (dynamic), (ded), (dsl)

Table 3: The mapping from the 15 hostname-inferred usage categories to 4 ping-observable categories. hostname-inferred usage category without (parentheses) is dominate.

5. APPLICATIONS

Having laid the groundwork for address analysis, we next use the data to explore several questions in network management: what are typical sizes of consistently used Internet address blocks? How effectively are they being used? And how prominent is dynamic addressing?

To assist in answering some of these questions we compare our observations with allocation data from the regional Internet registries (RIRs) [1]. This RIR data includes the time and country to which each address block is assigned. Although not completely authoritative, this data is the best publicly available estimate of address delegation of which we are aware. We collected data from each of the RIRs, selecting data dated June 13, 2007, to closely match our survey data.

5.1 Block Sizes

We begin by considering block sizes. Figure 9 and Table 4 show our data.

First, we observe that most addresses in the Internet are in /24 blocks. In fact, even though there more opportunities for small blocks, we find more /24 blocks than blocks of size /25 through /29. Since our data collection only probes consecutive runs of 256 addresses, this prevalence suggests we may need to probe larger consecutive areas to understand if even larger blocks are common but not seen in our survey.

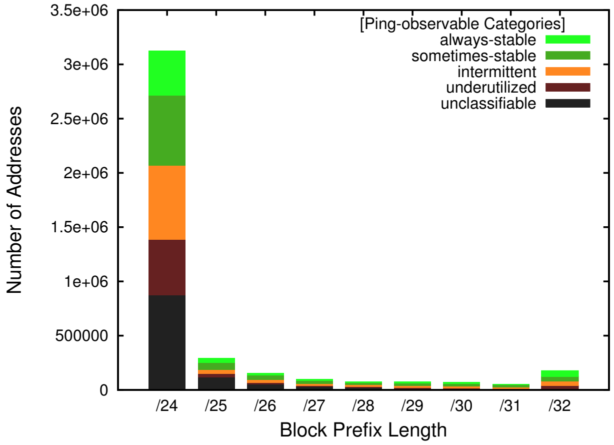

Figure 9: Number of addresses in each block size and ping-observable categories in *IT17ws*.

There are a very large number of the smallest blocks, with about as many /29s as /24s, and roughly twice as many /30s as /29s, and /31s as /30s. These results may be artifacts of our block discovery algorithm: it is statistically easier for an address to be consistent with a very few neighbors in a small block than with 128 neighbors in a /25. Finally, we can re-examine the second assumption underlying our work: are contiguous addresses often used similarly? If we define consistent usage as just the largest three block sizes (/24 through /26) that we successfully identify, we find 2,529,216 addresses are used consistently, or 44% of the probed address space.

While clearly defined, this percentage does not accurately present how much of the Internet is consistently used. Some of the probed address space is unclassifiable (with consistent usage but fewer than 20% of addresses responding), or completely non-responsive. We cannot say anything about blocks that fail to respond at all. The status of unclassifiable blocks is uncertain, but a conservative position is to declare them inconsistent. A more representative evaluation of the Internet is therefore compare how much is definitely used consistently (2.5M addresses in large blocks) against that is effectively inconsistent (the 506,178 addresses in small blocks) and the possibly inconsistent (the 1,087,472 addresses in unclassifiable blocks). This computation suggests that a lower bound of *61% of the responsive Internet is used consistently*, We believe this supports our second assumption: **the majority of contiguous addresses are used consistently**.

5.2 Address Utilization

Having characterized block sizes, we next evaluate how efficiently addresses are used. If IPv4 addresses are used inefficiently that represents an opportunity for improvement. However, greater efficiency comes with greater management cost; that cost must be weighed against simpler solutions such as IPv6.

5.2.1 *Quantifying underutilization and possible causes*

The *underutilized* ping-observable category is defined as a sequence of addresses that are used less than 10% of the time (Section 3.4). While one can imagine some circumstances where a public IP address that is used very infrequently might make sense (for example, perhaps a DTN satellite only infrequently in view [6]), large blocks of such addresses appear to represent a poor use of a public IP address. Moreover, these underutilized blocks are not simply public address space hidden behind a firewall (a common management practice to simplify routing), but large blocks where each address is visible, but only very infrequently.

The underutilized column of Table 4 shows that these blocks are quite common, accounting for 17–23% of blocks of each size, Although not shown in the table, the mean availability of addresses in /24 underutilized blocks is only 3.2% of our 10-day observation (*IT17ws*). Manual examination of addresses show the mean number of up periods is less than 5 ($V^*(b) = 4.6$), typically for around 1 hour ($U^*(b)$).

To understand *why* there are blocks of underutilized addresses we turn to our hostname–ping analysis from Section 4.2. Figure 8 shows that underutilization corresponds with several hostname-inferred usage categories, including large fractions of categories dial and pool, and large absolute numbers of ppp, dsl and dynamic. Our analysis of hostname categories supports this observation, where dial has low availability and median-uptime (Figures 5 and 7 and high volatility (Figure 6).

We hypothesize that this low utilization is tied to dial-up technology itself. Dial-up lines are often shared with voice communication, encouraging short, intermittent use. Yet dial-up POPs must be provisioned to handle peak loads. A secondary factor may be trends shifting customers from dial-up to higher speed connections. Perhaps old dial-up provisioned blocks are simply in lower demand than previously. Finally, while dial-up utilization is low, we cannot tell how many users each dial-up address serves. Perhaps address reuse is high enough to make these apparently underprovisioned addresses a bargain relative to supporting the same number of users with always-on connections.

Reversing the question, we can ask *which address blocks are well utilized*? Figure 8 shows that the categories of static, cable, biz, res, server, router have very few underutilized addresses. Static addresses are usually assigned to fixed-location desktops or businesses, and these computers tend to maintain Internet connection and occupy their address for a fairly long time. In addition, static addresses are often billed at a flat

								blocks	**addresses**
size			**sometimes-**			**classifiable**	**unclassifiable**	*[100%]*	
pfx	addrs	**always-stable**	**stable**	**intermittent**	**underutilized**	(100%)			
/24	256	1,603 (18%)	2,517 (29%)	2,673 (30%)	1,994 (23%)	8,787*	3,411 *[27%]*	12,198	3,122,688
/25	128	323 (23%)	523 (38%)	295 (21%)	237 (17%)	1,378*	920 *[40%]*	2,298	294,144
/26	64	346 (21%)	617 (38%)	378 (23%)	274 (17%)	1,615*	787 *[33%]*	2,402	153,728
/27	32	432 (20%)	855 (40%)	506 (23%)	361 (16%)	2,154†	872 *[29%]*	3,026	96,832
/28	16	759 (20%)	1,301 (34%)	993 (46%)	734 (19%)	3,787†	1,139 *[23%]*	4,926	78,816
/29	8	2,077 (21%)	3,190 (32%)	2,355 (24%)	2,227 (23%)	9,849†	0	9,849	78,792
/30	4	3,312 (19%)	5,656 (33%)	4,679 (27%)	3,707 (21%)	17,354†	0	17,354	69,416
/31	2	4,195 (16%)	9,867 (37%)	7,864 (30%)	4,566 (17%)	26,492†	0	26,492	52,984
/32	1	52,646 (30%)	42,847 (24%)	43,266 (25%)	36,707 (21%)	175,466†	0	175,466	
entire *IT17ws* dataset:		(1,603,086 addrs. in non-responsive blocks) + (4,122,866 in responsive blocks)						22,367	5,725,952

Table 4: Number of blocks of each size in *IT17ws*. Unclassifiable percentages relative to all blocks; other percentages relative to classifiable blocks.

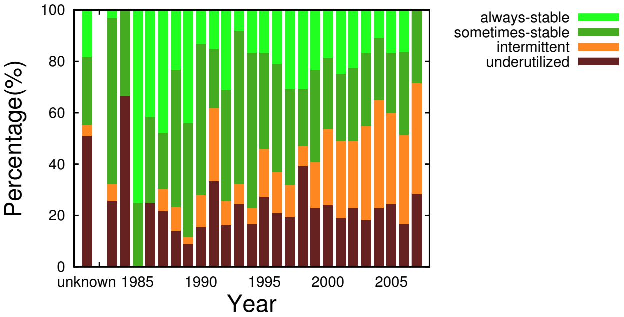

Figure 10: Trend of ping-observable category change in *IT17ws* /24 blocks

rate per month, while dynamic addresses may incur a time-metered charge.

5.2.2 *Locations and trends of underutilization*

Evaluating underutilization by country may highlight policy differences by regional registries or ISPs. After merging our data with RIR data, Table 5 shows utilization by country. We see that the United Kingdom and Japan have the largest fraction of underutilized blocks, 40–60%, suggesting potential local policy differences. We expected a large number of underutilized blocks in the U.S. because of wide deployment of dial-up. While the U.S. has the largest absolute number of underutilized blocks, its fraction is relatively low.

Table 6 shows that the fraction of underutilized blocks is fairly consistent by across all five RIRs, suggesting differences are likely due to country, not RIR policies.

Finally, the lower right graph in Figure 10 shows when underutilized blocks were allocated. The fraction of underutilized blocks by age seems fairly evenly distributed, except for peaks in very early allocations (1984 and unknown), where more than 60% of the blocks assigned are underutilized. Since that pre-dates widespread dialup, we do not have an immediate explanation for this peak.

5.3 Dynamic IP Addressing

DHCP's automatic address assignment [4] supports central assignment of IP addresses dynamically, requiring addresses only when users connect to the Internet. Although DHCP can be used to assign the same address to an always-up host, here we are interested in relatively assignments that change frequently, possibly for hosts that are only intermittently connected.

Dynamic assignment assignment of addresses allows ISPs to multiplex many users over fewer addresses. Dynamic addressing also provides ISPs the business opportunity of offering static addresses as a higher-priced service, and potentially makes it more difficult for users to operate bandwidth-consuming services.

To users, dynamic addressing has been promoted as a security advantage, on the theory that a compromised computer is more difficult to contact if its IP address changes. They impede users from some Internet activities, such as running services or accepting unsolicited inbound connections (for example, for incoming SIP calls). Wide use of dynamic addressing has promoted work-arounds to these problems such as STUN [24].

Dynamic addresses also complicate some network services, such reputation systems, and are correlated with spam sources. These reasons suggest a better understanding of dynamic addressing is important, and have prompted recent study [12, 28, 32]. We next show that our approach can identify dynamic addressees and suggest causes and trends that have been previously invisible.

5.4 Quantifying dynamic addressing

We believe that the *intermittent* and *underutilized* ping-observable categories correspond with the short-term dynamically assigned addresses we are interested in. This statement is supported by hostname data, where Figure 8 shows that intermittent blocks are prominent in hostnames that include dynamic, pool, ppp, dial, dsl, and cable, all of which often use short- or moderate-term dynamic addressing, and underutilized blocks are common to dynamic, pool, ppp, dial, and dsl.

code	country	always-stable	sometimes-stable	intermittent	underutilized	classifiable (100%)	unclassifiable	blocks *[100%]*
US	US	673 (27%)	**1,106 (45%)***	231 (9.3%)	472 (19%)	2,482	1,383 *[36%]*	3,865
CN	China	39 (4.1%)	117 (12%)	**615 (65%)***	171 (18%)	942	132 *[12%]*	1,074
JP	Japan	**383 (48%)***	50 (6.2%)	18 (2.2%)	**350 (44%)***	801	288 *[26%]*	1,089
DE	Germany	65 (10%)	125 (20%)	**388 (61%)***	62 (9.7%)	640	56 *[8.0%]*	696
KR	Korea	21 (4.6%)	131 (29%)	**237 (52%)***	68 (15%)	457	142 *[24%]*	599
FR	France	18 (4.1%)	**227 (52%)***	167 (38%)	28 (6.4%)	440	58 *[12%]*	498
GB	UK	39 (13%)	37 (12%)	52 (17%)	**179 (58%)***	307	180 *[37%]*	487
BR	Brazil	7 (3.9%)	35 (19%)	**86 (48%)***	52 (29%)	180	58 *[24%]*	238
all others		358 (14%)	689 (27%)	879 (35%)	612 (24%)	2,538	1,114 *[31%]*	3,652
/24 blocks in entire *IT17ws* dataset:						8,787	3,411 *[27%]*	12,198

Table 5: The distribution of /24 blocks in ping-observable categories of 10 countries.

registry	always-stable	sometimes-stable	intermittent	underutilized	classifiable (100%)	unclassifiable	blocks *[100%]*
RIPENCC	408 (14%)	798 (27%)	**1,084 (37%)***	661 (22%)	2,951	990 *[25%]*	3,941
APNIC	473 (18%)	422 (16%)	**1,091 (40%)***	716 (27%)	2,702	795 *[23%]*	3,497
ARIN	706 (27%)	**1,185 (45%)***	258 (9.7%)	512 (19%)	2,661	1,481 *[36%]*	4,142
LACNIC	13 (3.2%)	94 (23%)	**218 (53%)***	86 (21%)	411	120 *[23%]*	531
AFRINIC	3 (4.9%)	18 (30%)	**21 (34%)***	19 (31%)	61	19 *[24%]*	80
/24 blocks in entire *IT17ws* dataset:					8,787	3,411 *[27%]*	12,198

Table 6: The distribution of /24 blocks in ping-observable categories of 5 regional registries.

Table 4 shows that there are many dynamic addresses: 40–50% of classifiable blocks (depending on block size) appear to be dynamic. Even with wide deployment of always-on connectivity, nearly half of Internet addresses are used for short periods of time. For intermittent blocks, the mean availability is just under 30%, with nine use periods over the week and a the mean U^* around 2.5 hours.

5.5 Locations and trends for dynamic addressing

Analysis by country can suggest how political or cultural factors affect dynamic addressing. Table 5 shows that nearly two-thirds of Chinese blocks are intermittent, with Germany, Korean, and Brazil all nearly half or more. Several factors are potential causes for this use.

China has a very large population and is a relative latecomer to the Internet; from the beginning of commercial deployment in China ISPs have planned to make best use of relatively few IPv4 addresses per potential user. They have therefore promoted dynamic use to improve address utilization. An interesting direction for future work would be to evaluate how effective their utilization is. Unfortunately we only know address responsiveness, not the number of actual computers users per address needed to answer this question.

Time-metered billing is another reason for intermittent use. Parts of China and Germany employ metered billing, encouraging intermittent use even with broadband. Other potential reasons for intermittent use include turning off a router to conserve energy, or carrying over habits learned from dial-up use to broadband, and potentially continued use of dial-up connections shared with voice communication.

Evaluation of usage by regional registry in Table 6 presents even larger differences in use. We see that intermittent blocks are very prominent under APNIC and LACNIC (40–53%), five times more common than for ARIN in North America (9%). We believe these differences stem largely from policies of the countries the RIRs serve, not the RIRs themselves. We discussed Chinese practice above; several Latin American countries have limited choice in ISPs, with national providers adopting pricing schemes that strongly favor dynamic address assignment even for business use. We speculate that the large number of sometimes-stable blocks in ARIN is because of long DHCP lease times and always-on use by home users, enabled by relatively plentiful numbers of IPv4 addresses per user.

Finally we turn to trends in dynamic addressing. The lower left portion of Figure 10 shows intermittent blocks are increasingly likely in new address allocations. This observation is consistent with a growing recognition of eventual full allocation of the IPv4 address space and efforts to manage addresses in countries newer to the Internet. The rise in intermittent blocks matches a corresponding fall in always-stable blocks (top left, Figure 10). In addition to growing demand for dynamic addressing, this trend suggests most new addresses are added to provide service for home users, intermittently. While the absolute numbers of always-stable businesses and servers grows, its fraction of all addresses is shrinking.

6. VALIDATION

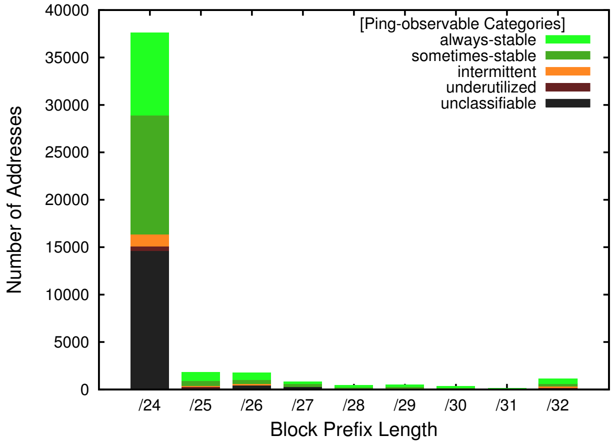

Figure 11: Number of addresses in each block size and ping-observable categories in *ITUSC s*.

The conclusions of this paper are based on the three assumptions we listed in the introduction: addresses respond to probes, adjacent addresses have similar use, and probes suggest use. Validation of the first assumption is the subject of prior work [12]; space prohibits revisiting that work here.

We have already presented data on the two final assumptions, showing that the majority of contiguous addresses are used consistently (Section 5.1) and probe responses can suggest use (Section 4.2). These conclusions are based on data taken from one survey *IT17ws* from the general Internet. While not biased, we cannot compare these results to the true network configuration that is distributed across thousands of enterprises. We next present two additional studies to further validate these assumptions. First we evaluate data taken from USC, a smaller and potentially biased dataset, but one where we have ground truth from the network operations staff. We then compare compare our Internet-wide results with a second dataset, *IT16ws*, to verify our conclusions do not reflect something unusual in a single time or set of addresses.

6.1 Validation within USC

We first compare our methodology against ground truth obtained from the administrators of our network at USC. This section uses dataset *ITUSC s* and applies the same analysis used on our general Internet dataset.

Figure 11 shows block sizes and classifications from our approach. USC shows many fewer intermittent and underutilized blocks compared to the Internet (Figure 9); we expect such variation across enterprises. We next use this data to evaluate how our assumptions affect our ability to accurately find of block size and consistency, then of block usage.

category:	**blocks**	**percentage**	
in routing table	243	100%	
false negative	105	43%	
not in use	19		
not responding	28		
few responding	12		
single-block multi-usage	46		
/25 to /27	9		
/28 to /32	37		
blocks identified	147		100%
correctly identified	138	57%	94%
false positive	9		6.1%
multi-block single-usage	9		

Table 7: Evaluation of block identification accuracy at USC to ground truth, with percentages relative to all blocks (left) and all identifications (right).

6.1.1 Block Size Validation

To validate our ability to determine block sizes, and that blocks at USC are used consistently, we compare our analysis with the internal routing table from our network administrators. This data helps quantify the accuracy of our approach, measuring the *false positive rate*, blocks that we detect but that do not actually exist, and the *false negative rate*, blocks that exist but we fail to detect.

Table 7 summarizes our comparison for all /24 blocks. (We do not evaluate smaller blocks because smaller subdivisions are usually handled at the department level and so are missing from our ground-truth routing table.) Relative to /24 blocks present, we find our approach correctly identifies 57% of all blocks in ground truth. Although we find the majority of blocks, we have a significant number of false negatives, failures to detect blocks. For this dataset, these false negatives show *our approach is somewhat incomplete*. On the other hand, if we evaluate our algorithm by what it says. We see very few false positives, correctly identifying 94% of all blocks we detect. For this dataset, almost no false positives show *our approach is quite accurate in what it asserts*.

To understand accuracy, we looked at when our approach incorrectly identifies blocks. All nine false positives are due to multiple blocks with common usage. We examined each incorrect block and found that USC administrators had placed two logically different blocks on adjacent addresses, but these administratively different blocks were used for similar purposes. Since our evaluation is based on external observations of use, we believe there is no way *any* external observer could determine these administrative distinctions.

Turning to false negatives, we found several sources of missed block identification. We found that many blocks were either in the routing table but not assigned to locations or services (19 not in use), in the routing table and assigned, but with no ping responses (28

category:	blocks	fraction
classified	138	100%
unclassifiable (false negative)	52	38%
incorrectly classified (false positive)	3	2.1%
always stable (dynamic)	3	
correctly classified (true positive)	83	60%
intermittent (dynamic)	4	
sometimes stable (dynamic)	5	
intermittent (VPN)	1	
underutilized (VPN/PPP)	2	
always stable (lab)	2	
sometimes stable (lab)	2	
always stable (building)	25	
sometimes stable (building)	42	

Table 8: Evaluation of block classification accuracy at USC to ground truth.

not responding), or filled with only a few responders (12 few responding). In all cases, our algorithm refuses to make usage assertions on unused or sparsely used space. Non- or few-responding blocks may be due to firewalls, reflecting a limitation of our probing method. Not-in-use blocks would be impossible for any external observer to confirm. In principle our algorithm could identify non-responsive blocks, but it is difficult for external observation to distinguish unused from firewalled space.

Finally, other other false negatives occur due to blocks that have been administratively assigned as /24s but then are used for different purposes. Nine of these show large, consistent patterns, possibly indicating delegation at the department level that is not visible to university-wide network administrators. If so, these represent incompleteness in our ground-truth data. Smaller mixed-use blocks represent violations of our assertion that adjacent addresses are used consistently.

6.1.2 *Block Usage Validation*

Table 8 shows the accuracy of our approach for the 138 blocks we classify. We declare 38% unclassifiable (false negatives); in these cases we have discovered the correct block size but decline to declare a ping-observable category because the block is only sparsely responsive. We correctly classify the majority of blocks, selecting ping-observable categories that are consistent with the use of 60% of blocks. We mis-identify three blocks (a 2% false positive rate), all reported as dynamically allocated but observed as always stable. These blocks perhaps represent DHCP-assigned addresses with very long lease times for computers that are always up.

6.2 Results Consistency Across Repeated Surveys

We next wish to understand if the parameters of our data collection or analysis have a disproportionate effect on our conclusions about Internet-wide address usage. To evaluate this, we compare our conclusions from *IT17ws* with the same analysis applied to second dataset, *IT16ws*, taken five months earlier and with half of blocks different. The survey selection methodology is described by its authors [12]. Half of the /24 blocks in the survey are consistent across each survey, and half are randomly chosen in each survey. This comparison therefore observes both if network changes alter observations of the same blocks, and if a different set of blocks show very different behavior.

We find that our estimates of the distributions of block sizes are almost identical in the two surveys. If we define s_p as the vector of number of blocks of prefix length p, the correlation coefficient of these vectors for the two surveys is 0.99998. We conclude that a random sample of 1% of the Internet is large enough that the block size observations are hardly affected if half of the sample is changed.

Our work assumes that contiguous addresses are often used consistently. Following our approach in Section 5.1, we consider blocks of size /24 through /26 as consistent, and size /27 through /32 as inconsistent. In *IT17ws*, 44% of probed Internet and 61% of responsive Internet is consistent, while *IT16ws* find 43% and 60%. We conclude that *IT16ws* and *IT17ws* both support our assumption.

Finally we consider the consistency of our ping-observable classification between *IT16ws* and *IT17ws*. Initially we found the correlation of the number of blocks in each category to be generally good but not great across all block sizes—it ranged from 0.663 to 0.938 for blocks smaller than /29, but it the correlation for /24 blocks was only 0.349. Examination of the data showed that around 500 blocks were shifting between always- and sometimes-stable. This shift occurred because of a change in volatility and our selection of the always-stable requirement that $V \leq \beta$ and $\beta = 0.0016$. For very stable hosts, a few outages can change V^* significantly. Examining our datasets, showed that *IT16ws* and *IT17ws* are of different duration (6 and 10 days). A longer duration makes it easier to distinguish between sometimes- and always-stable blocks. When we keep the observation duration the same by considering only a 6-day subset of *IT17ws*, the correlation coefficient for /24 classification rises to 0.626. We conclude that most ping-observable classifications are good, but there the separation between sometimes- and always-stable categories is somewhat sensitive.

7. CONCLUSION

In this paper we have developed a new approach to identify how Internet addresses are used from active probing. Our work assumes many addresses respond to active probes (as evaluated previously [12]), contiguous addresses are often used similarly, and probes can reveal that use. We validate the two new assumptions with multiple datasets of randomly selected Internet blocks and with data from USC. We then use our approach and data to answer important questions in network management including common block sizes for address management, efficiency of address utilization, and the extent and trends in dynamic address allocation.

Acknowledgments

We would like to thank Brian Yamaguchi and Maureen Dougherty for the valuable USC campus network administrative information. We thank Yuri Pradkin for the probing software and ping data collection used here, and Paul Vixie and ISC for providing their DNS data.

8. REFERENCES

[1] American Registry for Internet Numbers. RIR statistics exchange format. Technical report, ARIN, Sept. 2008. (Retrieved January, 2009).

[2] K. Claffy, Y. Hyun, K. Keys, M. Fomenkov, and D. Krioukov. Internet mapping: from art to science. In *IEEE DHS CATCH*, Washington, US, Mar. 2009. IEEE.

[3] X. Dimitropoulos, D. Krioukov, M. Fomenkov, B. Huffaker, Y. Hyun, kc claffy, and G. Riley. AS relationships: Inference and validation. *ACM Computer Communication Review*, 37(1):29–40, Jan. 2007.

[4] R. Droms. Dynamic host configuration protocol. RFC 2131, Internet Request For Comments, Mar. 1997.

[5] B. Eriksson, P. Barford, and R. Nowak. Network discovery from passive measurements. In *Proc. of ACM SIGCOMM Conference*, pages 291–302, Seattle, Washigton, USA, Aug. 2008. ACM.

[6] K. Fall. A delay-tolerant network architecture for challenged internets. In *Proc. of ACM SIGCOMM Conference*, pages 27–34, Karlsruhe, Germany, Aug. 2003. ACM.

[7] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the Internet topology. In *Proc. of ACM SIGCOMM Conference*, pages 251–262, Cambridge, MA, USA, Sept. 1999. ACM.

[8] S. Floyd and V. Paxson. Difficulties in simulating the Internet. *ACM/IEEE Transactions on Networking*, 9(4):392–403, Aug. 2001.

[9] V. Fuller, T. Li, J. Yu, and K. Varadhan. Classless inter-domain routing (CIDR): an address assignment and aggregation strategy. RFC 1519, Internet Request For Comments, Sept. 1993.

[10] L. Gao. On inferring automonous system relationships in the internet. *ACM/IEEE Transactions on Networking*, 9(6):733–745, Dec. 2001.

[11] T. Hain. A pragmatic report on IPv4 address space consumption. *The Internet Protocol Journal*, 8(3), 2004.

[12] J. Heidemann, Y. Pradkin, R. Govindan, C. Papadopoulos, G. Bartlett, and J. Bannister. Census and survey of the visible internet. In *Proceedings of the ACM Internet Measurement Conference*, pages 169–182, Vouliagmeni, Greece, October 2008. ACM.

[13] G. Huston. IPv4 address report. `http://bgp.potaroo.net/ipv4/`, June 2006.

[14] Internet Software Consortium. Internet domain survey. web page `http://www.isc.org/solutions/survey` accessed January 2007, Jan. 2007.

[15] Internet Software Consortium. Internet domain survey. web page `http://www.isc.org/solutions/survey` accessed January 2008, Jan. 2008.

[16] A. Jain and R. Dubes. *Algorithms for clustering data*. Prentice Hall, Englewood Cliffs, NJ, 1988.

[17] M. Khadilkar, N. Feamster, M. Sanders, and R. Clark. Usage-based DHCP lease time optimization. In *Proc. of 7thACM Internet Measurement Conference*, pages 71–76, Oct. 2007.

[18] G. Lyon. nmap. computer software at `http://insecure.org/nmap/`, Sept. 1997.

[19] X. Meng, Z. Xu, B. Zhang, G. Huston, S. Lu, and L. Zhang. IPv4 address allocation and the BGP routing table evolution. *ACM Computer Communication Review*, 35(1):71–80, Jan. 2005.

[20] P. Mockapetris. Domain names—concepts and facilities. RFC 1034, Internet Request For Comments, Nov. 1987.

[21] W. Mühlbauer, O. Maennel, S. Uhlig, A. Feldmann, and M. Roughan. Building an AS-topology model that captures route diversity. In *Proc. of ACM SIGCOMM Conference*, pages 195–204, Sept. 2006.

[22] J. Postel. Internet protocol. RFC 791, Internet Request For Comments, Sept. 1981.

[23] Regional Internet Registry. Resource ranges and geographical data. web page `ftp://ftp.afrinic.net/pub/stats/afrinic/`, `ftp://ftp.apnic.net/pub/stats/apnic/`, `ftp://ftp.arin.net/pub/stats/arin/`, `ftp://ftp.lacnic.net/pub/stats/lacnic/`, `ftp://ftp.ripe.net/ripe/stats/`, June 2007.

[24] J. Rosenberg, J. Weinberger, C. Huitema, and R. Mahy. STUN—simple traversal of user datagram protocol (UDP) through network address translators (NATs). RFC 3489, Internet Request For Comments, Dec. 2003.

[25] R. Sherwood, A. Bender, and N. Spring. Discarte: A disjunctive internet cartographer. In *Proc. of ACM SIGCOMM Conference*, pages 303–314, Seattle, Washigton, USA, Aug. 2008. ACM.

[26] L. Subramanian, S. Agarwal, J. Rexford, and R. H. Katz. Characterizing the Internet hierarchy from multiple vantage points. In *Proc. of IEEE Infocom*, pages 618–627, June 2002.

[27] M. Sullivan and L. Munoz. Suggested generic DNS naming schemes for large networks and unassigned hosts. Work in progress (Internet draft draft-msullivan-dnsop-generic-naming-schemes-00.txt, Apr. 2006.

[28] I. Trestian, S. Ranjan, A. Kuzmanovic, and A. Nucci. Unconstrained endpoint profiling (Googling the Internet). In *Proc. of ACM SIGCOMM Conference*, pages 279–290, Seattle, Washigton, USA, Aug. 2008. ACM.

[29] USC/LANDER project. Internet Addresses Survey dataset, PREDICT ID USC-LANDER/internet_address_survey_reprobing_it16w-20070216. web page `http://www.isi.edu/ant/lander`, Feb. 2007.

[30] USC/LANDER project. Internet Addresses Survey dataset, PREDICT ID USC-LANDER/internet_address_survey_reprobing_it17w-20070601. web page `http://www.isi.edu/ant/lander`, June 2007.

[31] USC/LANDER project. Internet Addresses Survey dataset, PREDICT ID USC-LANDER/survey_validation_usc-20070813. web page `http://www.isi.edu/ant/lander`, Aug. 2007.

[32] Y. Xie, F. Yu, K. Achan, E. Gillum, M. Goldszmidt, and T. Wobber. How dynamic are IP addresses? In *Proc. of ACM SIGCOMM Conference*, Kyoto, Japan, Aug. 2007. ACM.

APPENDIX

A. EXAMINING THE (A,V,U*) SPACE

Section 3.4 defined our ping-observable categories based on the (A, V, U^*) values of blocks. To develop an understanding of how these metrics help categorize the Internet, Figure 12 shows the density plot of (A, V, U^*) space separated in three planes. For each plot, we create 100 bins for each of two parameters, then count the number of /24 blocks identified in *IT17ws* that fall into that bin with any value of the third parameter.

All of the planes show blocks with many different values, providing no definitive clusters. However, there are concentrations in some areas of some planes, even though there are a few blocks in between those concentrations. The (A, V) plane shows two concentrations, with a portion of blocks tend to gather around $(A, V, U^*) = (0.975, 0.005, *)$, showing highly available and highly stable behavior. We classify most of them into always-stable blocks. Another portion of blocks tend to gather around $(A, V, U^*) = (0.050, 0.005, *)$ which exhibit highly underutilized behavior. We classify them

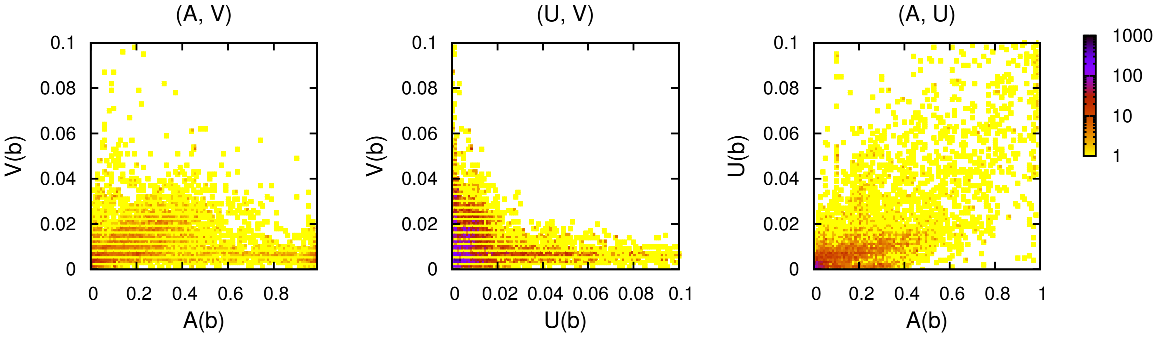

Figure 12: Density plots of /24 blocks in *IT17ws* across each of the A/V, U/V, A/U planes.

into underutilized blocks. The rest blocks are distributed between $(A, V, U^*) = (0.100 - 0.400, 0.075 - 0.022, *)$, with no obvious boundary to differentiate sometimes-stable and intermittent blocks on (A, V) plane. Instead, we inspect the (A, M) and (M, V) planes to split these apart. Even there, we do not see a sharp boundary. However, we place a line at $U = 0.026$ ($U^* = 6$ hours) to classify sometimes-stable ($U^* \geq 6$ hours) and intermittent ($U^* \leq 6$ hours) blocks.