Do You See Me Now? Sparsity in Passive Observations of Address Liveness

Jelena Mirkovic, Genevieve Bartlett, John Heidemann, Hao Shi, Xiyue Deng
USC Information Sciences Institute Marina Del Rey, CA 90292
E-mail: {$sunshine, bartlett, johnh, shihao, xiyueden$}@isi.edu

***Abstract*—Accurate information about address and block usage in the Internet has many applications in planning address allocation, topology studies, and simulations. Prior studies used active probing, sometimes augmented with passive observation, to study *macroscopic* phenomena, such as the overall usage of the IPv4 address space. This paper instead studies the completeness of *passive* sources: how well they can observe *microscopic* phenomena such as address usage within a given network. We define *sparsity* as the limitation of a given monitor to see a target, and we quantify the effects of interest, temporal, and coverage sparsity. To study sparsity, we introduce *inverted analysis*, a novel approach that uses complete passive observations of a few end networks (three campus networks in our case) to infer what of these networks would be seen by millions of *virtual* monitors near their traffic's destinations. Unsurprisingly, we find that monitors near popular content see many more targets and that visibility is strongly influenced by bipartite traffic between clients and servers. We are the first to quantify these effects and show their implications for the study of Internet liveness from passive observations. We find that visibility is heavy-tailed, with only 0.5% monitors seeing more than 10% of our targets' addresses, and is most affected by *interest sparsity* over temporal and coverage sparsity. Visibility is also strongly bipartite. Monitors of a different class than a target (e.g., a server monitor observing a client target) outperform monitors of the same class as a target in 82-99% of cases in our datasets. Finally, we find that adding active probing to passive observations greatly improves visibility of both server and client target addresses, but is not critical for visibility of target blocks. Our findings are valuable to understand limitations of existing measurement studies, and to develop methods to maximize microscopic completeness in future studies.**

I. Introduction

Accurately measuring Internet address and block usage (*liveness*) is of growing importance with many applications. Such information is vital for understanding use trends and identifying under-used portions in the Internet and in improving the efficiency of measuring network topologies [10], Internet outages [19], and DHCP allocation strategies [20]. It can also provide data for network simulations [14], [17]. But these applications may need accurate liveness data not just at the *macroscopic*, Internet-wide scale, but also at *microscopic* scales—for specific network blocks or organizations.

Prior studies of liveness have used active probing [5], [12], [19], sometimes in conjunction with passive observations [8], [23], [20] to provide an Internet-wide view of address and block usage. Both active probing and passive observation, however, will miss some addresses and blocks. These errors may be inconsequential at the *macroscopic* level (across the whole Internet address space), but they can introduce a systematic measurement bias at the *microscopic* level, when considering specific networks. For example, where missing thousands of addresses is just measurement noise on the entire address space, if these missing addresses are all cloud servers or embedded devices, their omission would bias studies of server traffic or IoT security. Prior work sought to quantify sources of measurement error, at the macroscopic level [12], [23], [20] and we compared the effectiveness of passive and active for service discovery in a campus network—one example of a microscopic view [2]. This paper complements these past efforts by studying the *specific factors* which influence the ability of passive data sources to accurately observe *specific host and network populations*, measuring microscopical activity. Our findings can help researchers improve collection strategies, interpret and refine observations, and clarify sources of imprecision in Internet measurements.

The first contribution of our paper is *inverted analysis*, a new measurement methodology that helps us assess the completeness of microscopic passive observations. We assume a *monitor*, placed at some vantage point, assesses liveness for a given *target* network, through the traffic that the target sends. Inverted analysis uses complete passive observations at edge networks (three large U.S. universities, in our case), and treats these networks as our measurement targets. We then place "virtual" monitors at *all other network blocks*, allowing us to estimate what *each virtual monitor* would see of our targets, and to reason about causes of incompleteness of these microscopic observations.

Our second contribution is to identify types of *sparsity*—the properties of the monitor and target that limit visibility. *Interest sparsity* reflects how much users near the monitor care about content hosted by the target, and vice versa. *Temporal sparsity* follows from the finite duration of any observation, which may miss infrequently used addresses. *Coverage sparsity* occurs when a monitor does not observe some links or when traffic is down-sampled to handle high line rates. While intuition suggests that macroscopic visibility of popular monitors will be high, we are the first to quantify this effect at a microscopic level. We find that visibility is heavy-tailed, with only 0.5% monitors seeing more than 10% of addresses at our three university targets. We further find that visibility is *bipartite*—most networks host primarily clients or servers, which leads to bipartite traffic and hinders complete observations between networks of the same type. While this intuitive as well, we are the first to quantify these effects. We find that, when observing

a randomly chosen set of addresses, 99% of the time server monitors outperform client monitors when observing client addresses, and 82% of the time client monitors outperform server monitors when observing server addresses. Finally, we find that interest sparsity has a dominating effect on visibility, while temporal and coverage sparsity only attenuate this effect.

Our third contribution is to identify the implications of interest sparsity on existing Internet measurement studies. While prior studies recognized the importance of using multiple data sources [8], [7], [20] to study liveness, and the importance of observations at popular servers, ours is the first work that sheds light on causes of reduced visibility and populations that may be poorly observed by a given monitor. We find that due to bipartite traffic and heavy-tail popularity *any single observer, large or small, will systematically miss certain populations*. In fact, we find in §V-C that even small observers can outperform large ones on certain populations. Our work provides guidelines that suggest who will be missed by a given set of observations, and how to select the best additional sources to fill these gaps. One of the additional sources we consider is active probing, as it was used in many recent studies of Internet liveness [8], [23], [20]. We find that active probing and passive observation discover complementary sets of address information, while their visibility into blocks is comparable. Active outperforms passive for 99% of server addresses, and passive at popular servers outperforms active for almost 100% of client addresses.

II. Problem Statement: Liveness and Sparsity

We first frame the problem we study, defining liveness, visibility and sparsity. With this background, we then move on to describe our inverted analysis approach in §III-A.

Liveness. Liveness estimation can be *cumulative*, denoting a target as live if it is active in any available data source during some long time interval [7], [23], [12], or it can be *instantaneous*, giving a snapshot of live addresses at a given time [9]. Liveness can further be assessed via *counts* of live addresses, or by learning their *exact identities*. One may also study liveness of *blocks* of adjacent addresses. This paper examines *cumulative counts* of live addresses and /24-prefix blocks ("blocks" for short) using passive observations.

Components of Passive and Active Measurement. Internet liveness was studied through passive observations, active probing [5], [12], [9], and their combination [2], [8], [23], [20]. Active measurement sends probes (ICMP echo or TCP SYN) and recognizes addresses that reply as live. Addresses could be probed once (census) or repeatedly (survey) in a measurement period. Passive observations denote sources of traffic as live as seen by a *monitor*, and recorded as packets, flows (pcap, netflow, and Argus are common formats) or host/server log entries. Both techniques have limitations, since passive monitors are vulnerable to spoofing (forged source addresses), and active measurement may encounter honey pots [1].

Visibility and Sparsity. We define the *visibility* $V_{m,t}$ of a monitor m with respect to a given target t as the percentage of t's live addresses or /24 blocks that are observed by m. $V_{m,t}$ is the fraction of ground truth m is able to learn about t. *Sparsity* is a limitation of the monitor that reduces its visibility.

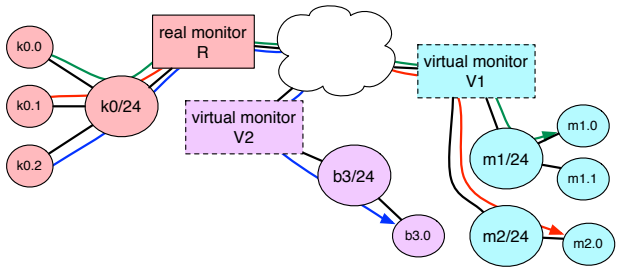


Fig. 1: A real monitor at a Known Network and two virtual monitors.

We identify three types of sparsity. **Interest sparsity** occurs when a monitor does not observe the target because the lack of user interest leads to a lack of observable traffic. This could happen because end clients from a target network are not interested in the content served by networks close to the monitor, and vice versa. **Temporal sparsity** occurs when a monitor's observation is short and thus it misses targets that are intermittently active. **Coverage sparsity** occurs when a monitor does not observe some links or when observations are down-sampled to reduce the load. In this paper we examine the impact of these three types of sparsity on visibility.

III. Methodology

We next describe *inverted analysis*, our novel approach to studying passive monitors, and describe our data sources.

III-A Inverted Analysis

We propose *inverted analysis* to study the limitations of Internet-wide passive monitoring, as illustrated in Fig. 1. We start from observations of a passive monitor *R* placed at the edge of some organizations' network (the *Known Network*). This placement allows the monitor to observe most or all traffic sent by this network, including regular client/server traffic due to user interest, responses to external scans and scans sent by the Known Network. Thus R's observations are a good source for estimation of the *ground truth* for the Known Network.

The idea behind inverted analysis is that we can use traffic seen by a real monitor (colored lines in Fig. 1 seen by the monitor R) to project what any *virtual* monitor (e.g., V1 and V2), located near the traffic's destinations, would see of the ground truth for the Known Network. In our example R sees bi-directional traffic between k0.0 and m1.0, k0.1 and m2.0 and k0.2 and m3.0. From this R can conclude that there are three live addresses in the Known Network—our ground truth. It can also project that if monitors were deployed at positions illustrated by V1 and V2, V1 would observe two live addresses and V2 would observe only one live address from the Known Network. This enables us to study the visibility (as defined in §II) of our three university targets (§III-B), whose traffic we can passively observe, by many possible monitors in the Internet.

While we assume a monitor at a network's edge, one could also use an aggregate of network service logs or host logs to infer liveness. We have two such aggregate monitors in our datasets (§III-B). Further, either a real or a virtual monitor could be placed at backbone links instead of a network's edge [4], [6]. Due to asymmetric routing, real monitors on backbone links may suffer from large amounts of coverage sparsity and thus cannot be used for inverted analysis to infer ground truth about any target. However, if inverted analysis were enriched with accurate routing information, one could place virtual monitors on backbone links. We leave this for future work.

For discussion purposes we group virtual monitors by their *power*, the fraction of the ground truth they observe. We consider four regions of visibility: *low* ($< 1\%$), *medium* (1-10%), *high* (10-50%) and *near-complete* (50-100%).

III-B Data Sources

We use five passive datasets in our study, from four sources: two Web server log summaries from a major U.S-based CDN (aggregate monitors) and three network traces, each a week long, capturing traffic at the edge of three US universities (real monitors at Known Networks). We know of no other available, non-anonymized sources of packet data. In §V-C, we also use an active source, the union of ISI Internet censuses [12] from Los Angeles, Colorado, and Japan (*it60all*) that overlaps our measurement period. We use our university networks as both targets (§V-B), to explore interest sparsity through inverted analysis, and as monitors (§V-C), to generalize our results. The CDN dataset is used only as a monitor because it contains information only about the client addresses accessing this CDN, but not about the CDN's addresses. *CDNloc* summarizes all log entries from all servers at the Los Angeles and Chicago PoPs, continuously over one week. *CDNglob*, covers all PoPs (more than 30) but only for 1 peak hour per day, for a week.

Tab. I summarizes our datasets. Four of our datasets cover the same week in June 2014, allowing for comparison of what each sees of the Internet (§V-C). Jointly, our passive datasets see 5.1 M blocks and 700 M addresses, and are thus comparable to sources in other recent work [12], [19], [8], [23], [20]. Because our sources predominantly have IPv4 traffic, our analysis focuses only on this traffic, but our methodology could easily be applied to IPv6.

Anonymization. All addresses in our sources have their lowest 8 bits scrambled with CryptoPAN [22]. Anonymization is consistent within each dataset but not across sources. This allows for comparison of counts in /24 blocks across datasets, but not of individual addresses. We compare address visibility across N datasets $D_1, ..., D_N$ in the following manner. For each block b visible by a subset of datasets $D_i, ..., D_j$, we adopt the highest address count $a = max(a_{D_i}, a_{D_j})$ as the ground truth for b. We obtain the total address count for a network as the sum of address counts in all its blocks. This method underestimates the actual address counts, but it is necessary because of inconsistent anonymization.

Our three universities all have around 30,000 students, and collect Argus-format flow data. Our *UGA* dataset does not capture ICMP traffic, and many local addresses are NATted within *UGA*. We exclude NATted traffic from our analysis.

Filtering Spoofing. Our university datasets may contain spoofed traffic with external addresses, which would skew our analysis in §V-C. We filter spoofed traffic using statistical filtering as proposed in [23].

Limitations. Our data sources have some limitations. Sources of our edge network traffic all come from university networks of moderate sizes (65k addresses, with 5–30k active addresses). Thus our observations about how well others see our edge networks and how well our edge networks observe the world may be biased if our visibility is specific to some property of universities. We find, however, that most of the visibility in our edge datasets comes through traffic that local clients exchange with popular Web servers. We believe this pattern generalizes to client-heavy networks and is not specific to universities, so our conclusions should hold for other networks of similar size and activity, provided they are client-heavy.

Visibility of our edge networks may be skewed if they were more responsive to scanners (they would then appear more visible than typical). However, our three networks are diverse: *USC* is responsive to scanners, while *CSU* and *UGA* are mostly closed (details are in our technical report [15]). This diversity of our Known Networks supports robust conclusions from our data about edge network visibility. Our university datasets further include a mix of clients and servers, allowing us to study the bipartite nature of traffic and visibility.

Finally, although "only" a week long, our time-synchronized datasets suffice to study what different observers see of the same targets at the same time, and why.

Validation. For the large networks we study, there is no complete ground truth about which addresses are active over time. (Our discussions with network operators of *USC* suggest that allocation and activity is decentralized and so even they do not have ground truth.) Thus, we cannot independently validate our findings about *what fraction* of a network is observable by a remote passive monitor. However, our findings pertain more to the *relationship* between visibility available to different monitors, based on the monitor popularity and monitor and target type. These findings will hold regardless of the ground truth for the target's liveness. Further our real monitors in our three Known Networks see all the traffic between selected address ranges (Tab. I) and the Internet. Thus we are confident that our passive observations of liveness for these networks are very close to the ground truth.

III-C Labeling Flows, Addresses, and Networks

When studying visibility of our Known Networks by virtual monitors, we also seek to understand *why* those virtual monitors see our Known Networks, and *what types of hosts* they see. We now explain labeling of flows, addresses and ultimately networks, which helps us understand causes of visibility.

Labeling flows. We start by classifying each flow by the *role of the address from a Known Network* and the flow's *purpose*. We summarize here only relevant classification rules.

#	Monitor location		Observation				Known Net size	
	org	prefixes	format	start	duration	all flows	addrs	blocks
Aggregate Monitors								
1	*CDNloc*	LAX and ORD POPs	logs	2014-06-17	7 days	266 B	—	—
2	*CDNglob*	all POPs	logs	2014-06-17	7 days	200 B	—	—
Known Networks								
3	CSU	129.82/16	Argus	2014-06-17	7 days	2.2 B	17,732	186
4	USC	128.125/16, 68.181/16	Argus	2014-06-17	7 days	461 M	31,997	492
5	UGA	128.192/16	Argus	2016-02-06	13 days*	682 M	5,243	198
Active Sources								
6	it60all [21]		census	2014-06-19	32 days	—	—	—

(* UGA source captures traffic every second day.)

TABLE I: Datasets used in this paper.

A TCP or a UDP flow is labeled as a *client* flow if an address from a Known Network sends traffic from a non-service port to a service port (as defined by IANA [13]), and we define a *server* flow in a similar manner. We label TCP/UDP flows that exchange payload in both directions as *payload*, and we label TCP flows that do not go past the 3-way handshake as *scans*. Flows that contain scan or ICMP Echo responses from a Known Network are labeled as *responder*.

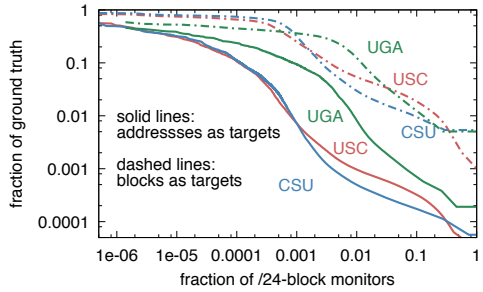
Labeling addresses. We label each address by aggregating address-role labels of its flows. Addresses that only have client (but not server) flows are labeled as *client*, those with only server (but not client) flows are *servers* and those with both are *client-servers*. Addresses that only have responder (but not client or server) flows are labeled as *responders*.

Labeling networks. We manually label autonomous systems (ASes) as client- or server-heavy using the following methodology. We first select a random subset of ASes that appear in our dataset. Our labeling starts by using WHOIS information to identify the owner of an AS. We then examine the owner's web pages, and any information about the owner on Bloomberg [3] and PeeringDB [18]. *Server-heavy* organizations primarily host content, and the majority of their addresses are servers. We label ASes as server-heavy if they are owned by hosting providers (e.g, Fastly), CDNs (e.g., Akamai), content providers (e.g., Facebook), or enterprises (e.g., banks). *Client-heavy* organizations primarily provide connectivity to users, and thus the majority of addresses in these networks are clients. We label ASes as client-heavy if they are owned by connectivity-providers (e.g., T-Mobile), research networks (e.g., CENIC) or universities. When classification is unclear we follow the primary service presented on the owner's web page.

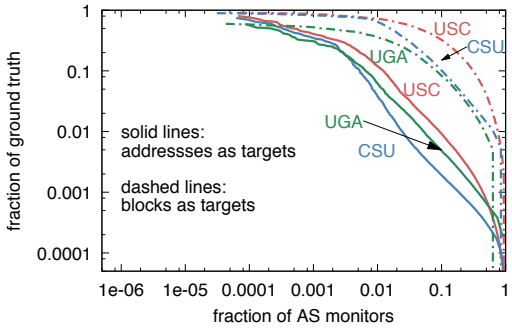
We could not fully automate inference of an AS's label as either a connectivity (client-heavy) or a content provider (server-heavy). The loose structure of several tasks requires human intelligence: (1) there is no consistent way to infer the primary Web page for the AS's owner from the AS's name or number, (2) information about an AS's purpose is often distributed over several Web pages and often must be translated from another language, (3) for ASes that offer multiple services, manual inspection is needed to establish a primary service. Our list of labeled ASes is available at https://ant.isi.edu/datasets/sparsity/.

IV. Interest Sparsity

We now show that interest sparsity significantly influences visibility of our Known Networks by virtual monitors, and



(a) From /24 block monitors



(b) From AS monitors

Fig. 2: Visibility of addresses and blocks from all possible virtual monitors at the level of /24 blocks, and entire ASes.

that visibility is heavy-tailed. We explore the causes of *interest sparsity* in §V.

Visibility is Heavy-Tailed. We place a virtual monitor at each /24 block that receives traffic from our three university targets (Known Networks), and evaluate the percentage of the target's addresses and blocks that the monitor sees. Fig. 2(a) shows the log-log complementary CDF of address and block visibility. Both graphs show an inflection point, where a handful of monitors see half of each target, then a near linear region where visibility falls off as a heavy tail over about three orders of magnitude. Also, many Internet blocks see nothing of our targets and are not shown on graphs. There is no interaction between 93% of routable blocks and *UGA*, 58% and *CSU*, and 41% and *USC*.

Whole ASes improve visibility. When we look at monitors placed at /24 blocks, visibility is heavy-tailed. Many organi-

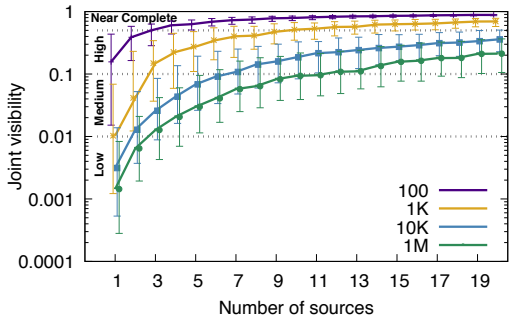
zations run networks that are larger than 256 addresses, and many CDN-blocks may not see our targets due to geographical content distribution. We next consider monitors that cover an entire AS. We evaluate this as a thought experiment, since some ASes may be large and difficult to fully monitor.

We map blocks that see our targets into ASes using MaxMind, and identify other blocks that belong to the same AS using WHOIS. Fig. 2(b) shows the CCDF of visibility of our targets from AS-sized monitors. Visibility of both addresses and blocks improves when compared to that of block-sized monitors, but the heavy tail remains: only 0.5% of organization-level monitors see more than 10% of our target addresses.

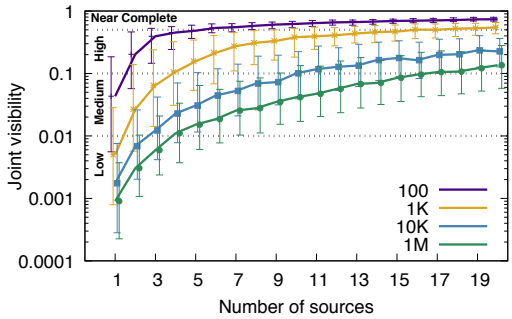
Visibility weakly correlates with content popularity. Next we investigate if visibility of our targets by AS-sized monitors depends on the popularity of an AS's content, and find significant but weak correlation. We measure content popularity by using the Alexa's top 1M list. We look up addresses associated with domains from the Alexa list, and map each address into an AS. The rank of an AS will be the lowest rank (i.e., the highest popularity) that any of its addresses has. We then run a Spearman correlation test between the rank and the AS's visibility of our targets. There is negative correlation for all three targets, which is significant ($p < 2.2 \cdot e^{16}$) but weak: $-0.23 \leq r \leq -0.19$. We believe that this low correlation occurs because local popularity of content (as measured by visits from our target) does not match global popularity (as measured by Alexa). For example, Akamai ranks as number 1 by *USC* visibility, but its hosted content (large retailers, Fox news, Hulu) has an Alexa rank >256.

Popularity's influence on measurement. We next explore how many monitors would be needed, and how popular they need to be to achieve a certain visibility goal. Since visibility is a property both of a monitor and its relationship with a target, we cannot provide a general estimate, but we can calculate specific estimates for our Known Networks as targets.

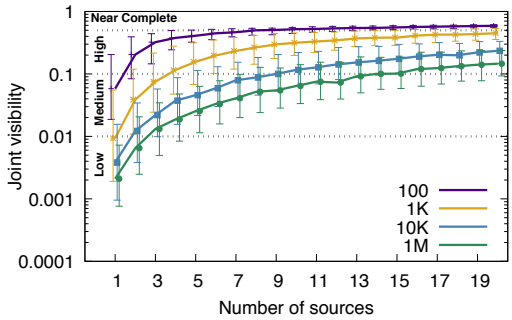
To obtain these estimates we observe monitors as belonging to four popularity classes by their Alexa rank: top 100, top 1K, top 10K and top 1M. Fig. 3 shows median and error bars for address visibility in our Known Networks, where each data point is obtained by 1,000 random draws of 2–20 monitors from a given popularity class. We see that popularity has a very strong influence on the number of monitors that are required. Given a visibility of at least 10% of a Known Network's addresses as a goal, each decrease in monitor popularity by a factor of 10 roughly doubles the number of monitors: 2 top-100 monitors, 3–4 top-1K monitors, 7–10 top-10K monitors or 12–16 top-1M monitors are all roughly equivalent in visibility. In addition, the variance increases greatly as monitor popularity declines, as measured by larger error quartiles shown with error bars. We see similar trends across all three Known Networks. These examples show the importance of having popular monitors, and having multiple monitors, for completeness.



(a) USC



(b) CSU



(c) UGA

Fig. 3: Visibility of addresses given a number of sources drawn from a certain Alexa rank.

V. Bipartite Traffic and Interest Sparsity

We next show that most addresses in our Known Networks are *clients* or *client-servers*. This leads to mostly *bipartite* traffic patterns, with server-heavy monitors seeing our targets much better than client-heavy monitors. We then use all of our datasets to explore what our client-heavy (universities) and server-heavy (our CDN) monitors see of the world. This confirms bipartite structure: client-heavy networks see server-heavy networks well and vice versa, but neither has good visibility within its own type.

V-A What is Seen in Our Known Networks

Looking at address labels in our Known Networks (targets) there are many clients (55–71% of addresses), with a smaller number of servers (4–14%), client-servers (10–22%) and responders (1–13%). Looking at the compositions of /24 blocks, addresses that are clients and servers are spread over many blocks. This diversity in location makes it likely that all classes

of remote monitors will see many blocks, even if they see only a few addresses in each block.

V-B Who Sees Us

We next investigate *who* sees our targets well and *why*. The top ten AS monitors by address visibility each see around 60–70% of each target and belong to: (1) academic networks hosting caches for Google, Netflix and Akamai (confirmed via reverse DNS), (2) large content providers, hosting companies and CDNs, e.g., SoftLayer, Akamai, Google, (3) a handful of aggressive scanners. Further, most monitors in high and moderate visibility regions ($V_{m,t} > 1\%$) see our targets because these targets send client payload-flows to the monitor network, i.e. due to user interest. Content providers among the top ten AS monitors see clients and client-servers well, but they miss servers and responders—entire classes of addresses that may have unique features and behaviors. Scanners exist among the top ten AS monitors only for *USC* as our other two targets filter aggressive scanners. The scanners see not just *USC* clients and client-servers, but also servers and responders. While their visibility is smaller than that of content providers number-wise (<60%), it is more diverse.

To generalize this relationship between visibility and the nature of the monitor network (server-heavy or client-heavy) we formed an unbiased set of block-monitors with different powers. For each Known Network (*USC*, *CSU* and *UGA*) we randomly select 300 block-monitors, 100 from each of the low, medium and high visibility range. This results in 900 blocks (875 unique blocks) forming the *Random Far Monitors* set. We then manually label the AS of each Random Far Monitor as client- or server-heavy, using the approach from §III-C, resulting in 313 labeled ASes. Next, we calculate the representation of server-heavy and client-heavy Random Far Monitors in monitor groups that have high, medium and low visibility into each of our targets. All three targets are seen in similar ways. Client-heavy monitors make 80–90% of the low-visibility monitors, but only 10–20% of medium- and high-visibility monitors. The rest of the monitors are server-heavy. This analysis generalizes our observations that mostly servers have good visibility of our client-heavy targets.

V-C What We See of Others

Our analysis of who sees us (§V-B) shows that visibility of *client-heavy targets* is bipartite. To get a more diverse set of targets, we next investigate what liveness information can be gleaned by our three *client-heavy* and two *server-heavy* monitors about different targets around the Internet. This analysis is regular (forward, not inverted) analysis of microscopic Internet-wide liveness as estimated by our monitors.

Methodology. We normalize our datasets for this study in two ways. First, we retain the four datasets that are collected at the same time, omitting the later *UGA* dataset. Second, we remove spoofed blocks using statistical filtering (§III-B). The original datasets see many non-routable blocks with apparent traffic (16% of *USC* and 0.04% in *CSU*), but after filtering we are left with a negligible 0.1% (*USC*) and 0.000004% (*CSU*), showing spoofing removal is successful.

We include an active source in our investigation, because recent studies of Internet liveness [8], [23], [20] used a combination of active and passive to achieve macroscopic completeness. We focus on understanding how well active sources achieve microscopic completeness, and compare them with our passive sources.

Our active source is the union of all ISI Internet censuses (called here *it60all*). Further, the CSU dataset contains significant number of active probes as CSU was hosting of one of our probers for *it60all*. We thus break the *CSU* dataset into an active source, which includes all ICMP traffic (CSU_I), and a passive source, which includes TCP and UDP traffic (CSU_{UT}).

Tab. II shows the number of total ASes, blocks and addresses seen by each of our sources. Our set of targets contains ASes that host Random Far Monitor blocks from §V-B. This includes 313 ASes—43 server-heavy and 170 client-heavy ASes. We call this set *Random Far Targets* and study how much of each target is seen by our monitors. Since we do not have ground truth about live blocks or addresses in any of these targets we use joint observations from our six datasets as the ground truth. We take the union of all blocks seen as block-ground-truth for a target. Again, because of how our data is anonymized, we cannot use the union of addresses for address-ground-truth. Instead we use the lower-bound estimates (maximum seen by any one source) of the number of addresses per block, and sum these over all blocks in a target, as described in §III-B.

Macroscopic Visibility. Several prior studies have reported that the combination of passive and active sources increase coverage, with each contributing unique addresses and blocks. For completeness, we compare our macroscopic visibility with two prior studies that list specific contributions of active and passive sources, shown in Tab. III.

The contribution of passive is much lower in our study (11% blocks, 7% addr.) than in Richter et al. [20] (20% blocks, 40% addr.), even though both include a CDN source. The difference in discovery occurs because their passive observation is longer than ours (16 weeks vs. our 1), allowing for discovery of more dynamic addresses (25% [20]). Our shorter passive collection approaches *instantaneous* liveness (for example, from a one-shot census), while their longer observation accumulates more addresses, reflecting *cumulative* liveness.

To get a deeper understanding of unique contributions, we show the visibility into the total of our Random Far Targets (broken into client-heavy and server-heavy) by our sources in the *Client* and *Server* columns in Tab. II.

Microscopic Visibility. Because macroscopic measures may be biased by ASes with large address/block occupancy, we also show cumulative distributions of the fraction of ground-truth addresses and blocks that our sources see in each client- and server-heavy target in Fig. 4, i.e., microscopic visibility. Assuming that near-complete microscopic visibility is the desired goal (as defined in §III-A), we report percentages of client- and server-heavy targets for which a source achieves this goal in the *Near-complete* column in Tab. II.

We see that *active sources are powerful, especially in*

	Source	Total ASes	Total blk	Total addrs	Client blk	Client addrs	Server blk	Server addrs	Near-complete vis. srv-blk	srv-addrs	cl-blk	cl-addrs
passive	*CDNloc*	20 k	2.2 M	202 M	52%	40%	16%	3%	14%	0%	30%	16%
	CDNglob	41 k	46 M	573 M	84%	77%	55%	13%	47%	6%	87%	66%
	CSU_{UT}	28 k	1 M	2 M	18%	0.2%	32%	1.7%	44%	2%	5%	0%
	USC	30 k	1.8 M	13 M	33%	1.3%	51%	9.6%	56%	2%	19%	0%
	passive	42 k	4.8 M	614 M	86%	86%	78%	20%				
active	$Census$	44 k	5 M	486 M	92%	72%	83%	92%	90%	94%	95%	75%
	CSU_I	38 k	4 M	421 M	76%	65%	62%	76%	82%	74%	83%	60%
	active	45 k	5 M	510 M	93%	75%	83%	93%				
pass & act.		45 k	5.7 M	741 M								

TABLE II: Contributions of sources, in blocks (blk) and addresses (addrs).

TABLE III: Comparison of numbers of blocks (and addresses) found by recent evaluations of Internet liveness.

source	unique blocks (addrs) passive		active	
CAIDA [7]	10%		19%	
Large CDN [20]	20%	(40%)	10%	(10%)
us	**11%**	**(7%)**	**15%**	**(6%)**

TABLE IV: Recommendations for source selection to get good coverage of desired target.

Target	Recommended Sources
Server block	active or passive client-heavy
Server addr.	active
Client block	active or passive server-heavy
Client addr.	active and passive server-heavy
All blocks	active or passive cl+srv-heavy
All addr	active and passive srv-heavy

finding servers. Active sources (Census and CSU_I, red lines in Fig. 4) have near-complete *macroscopic* visibility into all four categories of targets (65–92% in Tab. II), and CDFs that place many blocks near 100% visibility. However, their *microscopic* visibility into client-heavy addresses is lower than client-heavy blocks, server-heavy blocks, and server-heavy addresses (75% near-complete, vs. 90-95%; see Fig. 4(d) vs. others in Fig. 4).

We next show that *passive CDN observations underrepresent server-heavy networks*. Not surprisingly, our CDN source (a server-heavy observer) sees client blocks and addresses well (Figures 4(b) and 4(d)), and global observation is much better than local (compare the solid and dotted lines). *CDNglob* has good visibility of client-heavy addresses and blocks (macroscopic 77–84%, microscopic 66–87% near-complete), but its visibility into server-heavy blocks is often lower than that of university networks (yellow solid line above blue in Fig. 4(a), macroscopic 55%, microscopic 47% near-complete), and its visibility into server-heavy addresses is very low (macroscopic 13%, microscopic 6% near-complete).

Finally, *passive data from client-heavy networks provides limited visibility of addresses and client-heavy blocks*. Our university networks (client-heavy observers) have good visibility of server-heavy blocks (macroscopic 32–51%, microscopic 44-56% near-complete) and lower visibility of client-heavy blocks (macroscopic 18–33%, microscopic 5–19% near-complete). Their visibility of server-heavy addresses is very low (macroscopic 1.7–9.6%, microscopic 2% near-complete) and that of client-heavy addresses is even lower (macroscopic 0.2–1.3%, microscopic $<$1% near-complete).

Overall, 99% of server-heavy monitors outperform client monitors when observing client addresses, and 82% of client-heavy monitors outperform server-heavy monitors when observing server addresses, confirming strong bipartite visibility.

V-D Implications for Measurement Studies

An important implication of our findings is that measurement studies using *passive data from only one source (clients or servers) will systematically miss parts of the Internet*—they will have poor microscopic visibility into entire classes of networks.

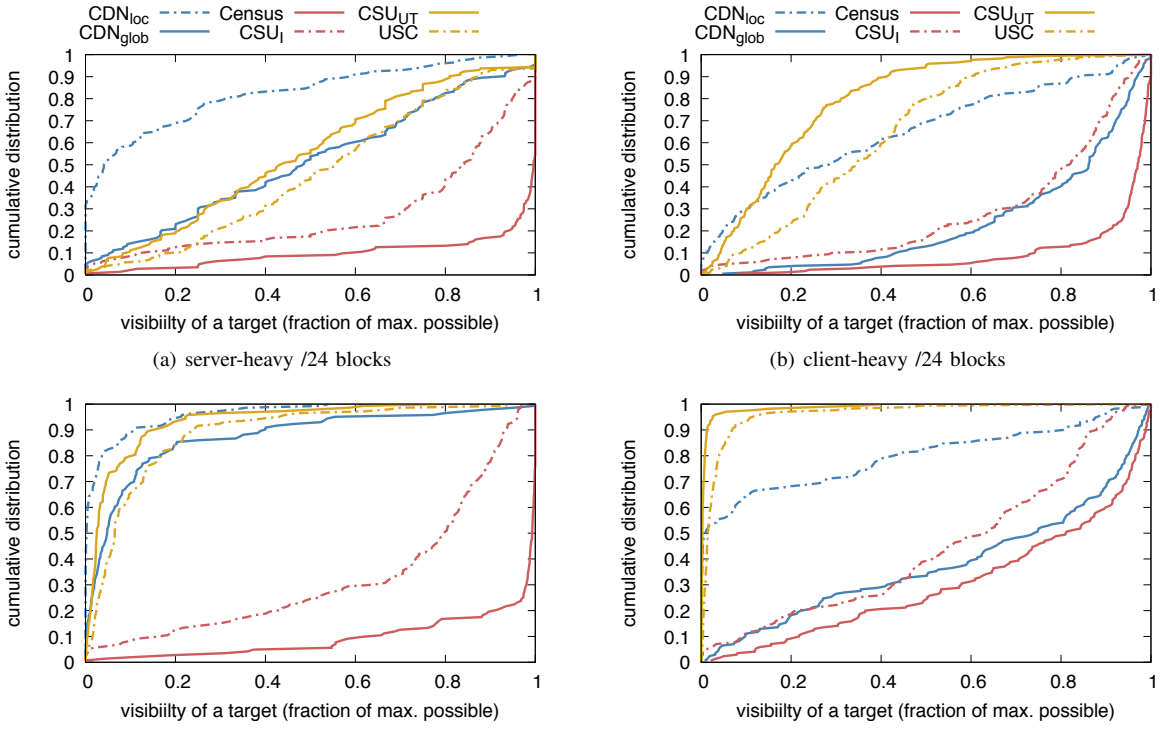
Tab. IV suggests selection methods for sources that will provide near-complete visibility of a desired measurement target (client- or server-heavy), summarizing §V-C. The key factor is a researcher must use *either* active probing, *or* passive data of the opposite type than the target, for good visibility of blocks. Further passive and active discover complementary addresses in all targets. Active probing outperforms passive observation for 99% of server addresses, and passive observation at popular servers outperforms active probing for almost 100% of client addresses. Visibility of server-heavy addresses thus requires active probing, and visibility of client-heavy addresses requires passive observation at popular servers.

Our findings can also help interpret prior measurement studies. For example, works on traffic policing [11] and DHCP churn [20] are based on data from large CDNs. While they hold macroscopically, neither is completely "Internet-wide", since we show CDNs have poor visibility into server-heavy networks. While it is unlikely that servers are policed or change DHCP, generalization requires care.

VI. Temporal and Coverage Sparsity

Although ultimately visibility is driven by interest, observation duration (temporal sparsity) and sampling (one kind of coverage sparsity) may also affect visibility. In this section we seek to quantify these effects. Due to space limitations, we report findings only for USC and only for addresses, but we see similar results for our other two targets and for blocks [16].

Temporal Sparsity reflects the importance of listening "long enough". Prior work demonstrates visibility increases logarithmically with time, with 70–90% of the addresses being discovered within 3-days [2], [7], but they quantify duration effects only on *their* monitors. Inverted analysis allows study of how temporal sparsity impacts different types of monitors.



(a) server-heavy /24 blocks

(b) client-heavy /24 blocks

(c) server-heavy addresses

(d) client-heavy addresses

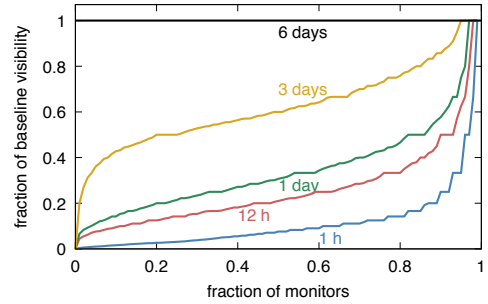
Fig. 4: Visibility of /24 blocks and addresses in server- and client-heavy targets, in each of our sources.

We look at how visibility changes as we vary the duration of passive observation and compare this with a monitor's visibility of each of our targets. We consider only monitors that see our given target in at least ten one-hours periods over the full trace.

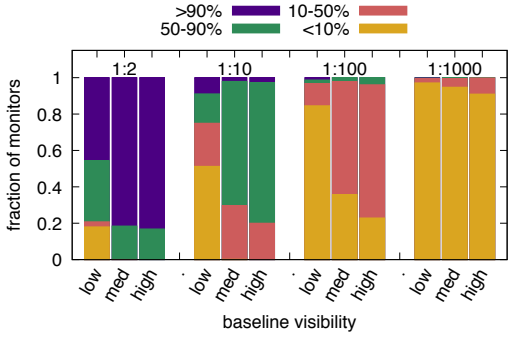
Fig. 5(a) shows the distribution of visibility across monitors as a function of time, compared to the baseline of evaluation over the full duration. 54% of monitors have seen at least 70% of their full visibility within three days. The lowest visibility monitors, reach their full coverage quickly—5% of monitors only need one day and for these monitors longer duration does not improve coverage. For higher visibility monitors, their coverage exhibits logarithmic growth, with longer observations bringing reduced benefit. Those monitors that have high visibility all reach 60–80% of their full visibility within a day, and reach more than 90% of full visibility within three days. Very low-visibility and medium to high-visibility monitors converge within days to at least 70% of their visibility. Low-visibility monitors experience linear growth in their visibility and need longer observations to converge.

Coverage Sparsity includes down-sampling of monitored traffic and view-point omissions such as missing traffic on specific links. The effects of view-point omissions are highly site-specific, but down-sampling during measurement is common and generalizable, so here we consider sampling effects.

To investigate down-sampling, we artificially discard packets from flows with a given probability. If all packets are discarded, we remove that flow. We then perform inverted analysis on the remaining flows, and compare visibility on non-sampled vs. sampled flows.



(a) Address visibility after one hour, twelve hours, one day and three days as a fraction of 6-day visibility. Block visibility is similar.



(b) Reduction in visibility of addresses when packets are sampled.

Fig. 5: Effects of duration and down-sampling.

Fig. 5(b) shows the percentage of monitors in several ranges of remaining visibility—more than 90%, 50-90%, 10-50% and <10% of the baseline visibility over non-sampled data. Monitors are grouped into low, medium and high visibility groups based on their baseline visibility. Each group of bars shows a different sampling rate: 1-in-2, -10, -100, or -1,000.

We find that resiliency of monitors to sampling depends on their baseline visibility. High- and medium-visibility monitors are barely affected by 1-in-2 sampling and at 1-in-10 sampling, most achieve well above 50% of their baseline visibility. Even at 1-in-100 sampling rate, these monitors achieve 10–50% of their baseline visibility. On the other hand, low-visibility monitors are severely affected by sampling. At just 1-in-2 sampling, 60% of low-visibility monitors lose half of their baseline visibility. At 1-in-10, half of the low-visibility monitors retain just 10% or less of their baseline visibility. We also find block visibility is much more robust to sampling than address visibility.

VII. Related Work

Our work is motivated by studies of network services [2] and address liveness [8], [7], [23] with passive sources.

Early work compared passive and active techniques for discovering services in a campus network [2] and showed that popular servers are discovered quickly, that scanners help discover many otherwise inactive addresses, and that continuous estimation of liveness is necessary due to dynamic addressing.

Dainotti et al. were the first to apply passive discovery to study Internet-wide liveness, complementing and expanding on active probing [8], [7]. They recognize the importance of filtering spoofing, and the importance of multiple data sources. Similarly, Richter et al [20] use passive observations from a large CDN and active probing to study dynamic addressing in the Internet. Our work builds on these prior works to explore the root causes in the visibility provided by different monitors and the role of clients and servers.

Zander et al. apply the capture-recapture framework from biology to extend prior passive and active estimates of Internet liveness [23]. They validate their approach on six chosen networks and use many data sources, but do not explore reasons why their sources provide different information, while we do.

VIII. Conclusion

This paper investigated what passive observers can learn about address liveness and why. We proposed *inverted analysis* to study many virtual passive monitors using a small number of real monitors at edge networks. We also identified interest sparsity as a key factor for limitations of passive sources.

Our key result is that the *type* and *popularity* of passive observers matter. While prior studies often gathered as many sources as possible [7], [23], [20], in §V-D we summarize our guidance to selecting sources and our understanding of their limitations for microscopic observations of different populations.

Acknowledgments: The authors would like to thank Christos Papadopoulos (Colorado State U.), Roberto Perdisci (U. Georgia), Yuri Pradkin (USC), and our industrial partners for their help with data collection and processing.

John Heidemann's work is partially sponsored by the U.S. DHS S&T Dir., Cyber Security Div., via SPAWAR Systems Center Pac. (Contract No. N66001-13-C-3001), via BAA 11-01-RIKA and AFRL (agreements FA8750-12-2-0344 and FA8750-15-2-0224). The U.S. Government is authorized to make reprints for Governmental purposes notwithstanding any copyright. The views in this paper are of the authors and not DHS or U.S. Government.

References

[1] L. Alt, R. Beverly, and A. Dainotti. Uncovering network tarpits with Degreaser. In *Proc. of ACM Annual Computer Security Applications Conference*, page xxx, New Orleans, Louisana, USA, Dec. 2014. USENIX.

[2] G. Bartlett, J. Heidemann, and C. Papadopoulos. Understanding passive and active service discovery. In *Proc. of ACM IMC*, pages 57–70. ACM, Oct. 2007.

[3] Bloomberg. Example profile page. http://www.bloomberg.com/research/Stocks/private/snapshot.asp?privcapId=183695812.

[4] P. Borgnat, G. Dewaele, K. Fukuda, P. Abry, and K. Cho. Seven years and one day: Sketching the evolution of internet traffic. In *INFOCOM 2009, IEEE*, pages 711–719. IEEE, 2009.

[5] R. Bush, J. Hiebert, O. Maennel, M. Roughan, and S. Uhlig. Testing the reachability of (new) address space. In *Proc. of ACM Workshop on Internet Nework Management*, pages 236–241, Kyoto, Japan, Aug. 2007. ACM.

[6] CAIDA. The CAIDA anonymized internet traces 2011 dataset. http://www.caida.org/data/passive/passive_2011_dataset.xml.

[7] A. Dainotti, K. Benson, A. King, kc claffy, E. Glatz, X. Dimitropoulos, P. Richter, A. Finamore, and A. C. Snoeren. Lost in space: Improving inference of ipv4 address space utilization. *CoRR*, abs/1410.6858, 2014.

[8] A. Dainotti, K. Benson, A. King, kc claffy, M. Kallitsis, E. Glatz, and X. Dimitropoulos. Estimating Internet address space usage through passive measurements. *ACM Computer Communication Review*, 44(1):42–49, Jan. 2014.

[9] Z. Durumeric, E. Wustrow, and J. A. Halderman. ZMap: Fast internet-wide scanning and its security applications. In *Proc. of USENIX Security*, 2013.

[10] X. Fan and J. Heidemann. Selecting representative IP addresses for Internet topology studies. In *Proc. of ACM IMC*, 2010.

[11] T. Flach, P. Papageorge, A. Terzis, L. Pedrosa, Y. Cheng, T. Karim, E. Katz-Bassett, and R. Govindan. An internet-wide analysis of traffic policing. In *Proc. of ACM SIGCOMM*, 2016.

[12] J. Heidemann, Y. Pradkin, R. Govindan, C. Papadopoulos, G. Bartlett, and J. Bannister. Census and survey of the visible Internet. In *Proc. of ACM IMC*, 2008.

[13] IANA. Service Name and Transport Protocol Port Number Registry.

[14] M. Liljenstam, Y. Yuan, B. Premore, and D. Nicol. A mixed abstraction level simulation model of large-scale internet worm infestations. In *Proc. of IEEE MASCOTS*, 2002.

[15] J. Mirkovic, G. Bartlett, J. Heidemann, H. Shi, and X. Deng. Do you see me now? sparsity in passive observations of address liveness (extended). Technical Report ISI-TR-2016-710, USC/Information Sciences Institute, July 2016.

[16] J. Mirkovic, G. Bartlett, J. Heidemann, H. Shi, and X. Deng. Do you see me now? sparsity in passive observations of address liveness (extended). Technical Report ISI-TR-710, Information Sciences Institute, University of Southern California, July 2016.

[17] D. Moore, C. Shannon, D. J. Brown, G. M. Voelker, and S. Savage. Inferring internet denial-of-service activity. *ACM Transactions on Computer Systems (TOCS)*, 24(2):115–139, 2006.

[18] PeeringDB. PeeringDB. https://www.peeringdb.com/.

[19] L. Quan, J. Heidemann, and Y. Pradkin. Trinocular: Understanding Internet reliability through adaptive probing. In *Proc. of ACM SIGCOMM*, 2013.

[20] P. Richter, G. Smaragdakis, D. Plonka, and A. Berger. Beyond counting: New perspectives on the active IPv4 address space. Technical Report arXiv:1606.00360v1, arXiv, June 2016.

[21] USC/LANDER Project. Address space scan. PREDICT ID USC-LANDER/internet_address_census_it60w(it60c,it60j)-20140619, June 2014.

[22] J. Xu, J. Fan, M. H. Ammar, and S. B. Moon. Prefix-preserving IP address anonymization: Measurement-based security evaluation and a new cryptography-based scheme. In *Proc. of IEEE ICNP*, 2002.

[23] S. Zander, L. L. Andrew, and G. Armitage. Capturing ghosts: predicting the used ipv4 space by inferring unobserved addresses. In *Proceedings of the ACM IMC*, 2014.