Lost in Space: Improving Inference of IPv4 Address Space Utilization

Alberto Dainotti*, Karyn Benson*, Alistair King*, Bradley Huffaker*,
Eduard Glatz†, Xenofontas Dimitropoulos‡,
Philipp Richter§, Alessandro Finamore¶, Alex C. Snoeren*
*UC San Diego, †ETH Zurich, ‡Forth, §TU Berlin, ¶Politecnico di Torino

Abstract

One challenge in understanding the evolution of Internet infrastructure is the lack of systematic mechanisms for monitoring the extent to which allocated IP addresses are actually used. In this paper we try to advance the science of inferring IPv4 address space utilization by proposing a novel taxonomy and analyzing and correlating results obtained through different types of measurements. We have previously studied an approach based on passive measurements that can reveal used portions of the address space unseen by active approaches. In this paper, we study such passive approaches in detail, extending our methodology to new types of vantage points, identifying traffic components that most significantly contribute to discovering used IPv4 network blocks. We then combine the results we obtained through passive measurements together with data from active measurement studies, as well as measurements from BGP and additional datasets available to researchers. Through the analysis of this large collection of heterogeneous datasets, we substantially improve the state of the art in terms of: (i) understanding the challenges and opportunities in using passive and active techniques to study address utilization; and (ii) knowledge of the utilization of the IPv4 space.

I. Introduction

In September 2015 the American Registry for Internet Numbers (ARIN) expects to exhaust its IPv4 address space, making it the fourth RIR unable to allocate new IP addresses. This historical event has been anticipated for decades, accompanied by intense debates over address management policy, IPv6 transition, and IPv4 address markets [31]–[33]. However, only one project (Heidemann et al. [29]) presently measures which allocated addresses are actually being visibily used, where *used* is defined as “directly responding to an ICMP echo request”. Unfortunately, measurement campaigns based on Internet-wide active probing can only illuminate a portion of the used address space, because of (i) operational filtering of scanning or (ii) potential violation of acceptable usage policies, triggering either complaints or blacklisting of the measurement infrastructure. Recently, Dainotti et al. [20] proposed an approach based on passive measurements, which is complementary to [29] and promises significant improvement when surveying Internet address usage at /24 address-block (/24 blocks, in the following) granularity. Passive measurements may also compensate for active approaches' inability to scale for use in a future IPv6 census [54].

Building on Heidemann's landmark work and on the novel concepts introduced in [20], our goal in this study is to improve the science of Internet address usage inference in a systematic way. We contribute to this field from different angles:

- *Taxonomic*. We propose a taxonomy of address space utilization that pertains to the whole address space and we introduce metrics to analyze the results of census studies.
- *Methodological*. We extend the passive-measurement approach presented in [20] to vantage points and network measurements of different type. In total we consider: (i) full packet traces from a large darknet; (ii) NetFlow logs from a national academic network; (iii) sampled packet traces from one of the largest IXPs worldwide; (iv) traffic classification logs from residential customers of a European ISP. Thanks to the availability of these diverse data we scrutinize the general applicability and limitations of this approach. We analyze how inferences of active address blocks can be influenced by characteristics specific to traffic observation vantage points, such as traffic composition, size of the monitored address space, and duration and time of the measurement. We find that all the four types of VPs are reasonably robust to variations in these characteristics and we provide insights to guide researchers in replicating our methodology on other VPs.
- *Knowledge and implications*. We combine seven passive and active measurement datasets to perform the first extended IPv4 Census using our taxonomy. We compare our results to the state of the art represented by the ISI census [29] and obtain an increase of 15.6% over ISI. In this process, we also learn novel insights about the views obtained through active and passive measurements (e.g., we identify special categories of address blocks that do not seem to generate traffic on the public Internet, unless solicited) which can inspire additional work in surveying address space utilization [9].

  We then analyze the results of our census, which estimates that *only 37% of the usable IPv4 space is used*, and that 3.4M assigned /24 blocks are not even visible in the global

BGP routing system. We analyze how unused space is distributed across RIRs, countries, continents, and ASes and we infer that only 9.5% of the legacy /24 blocks are *used* and that most *unused* address blocks are in the U.S.

Finally, we discuss how scientific studies of Internet-related phenomena might change if they used this extended dataset instead of other related data sets to estimate the address space of ASes or countries. As an example, we show the impact on CAIDA AS Rank [49].

Section II and Section III describe related work and the datasets we use in our study. Section IV introduces our new taxonomy for IPv4 address space utilization and provides a first insight in our findings. Section V extends and provides a detailed evaluation of our passive traffic methodology. Section VI combines passive and active measurement approaches and examines their different contributions. Section VII characterizes the utilization of the address space and the potential impact of using our dataset (shared through the PREDICT repository [4]) in other research studies. Section VIII offers promising directions for applicability and extension of this work.

II. Related work

Huston [26], [31]–[33] has provided a wealth of statistics and projections related to allocated and routed IPv4 address space, although he does not attempt to discern if allocated or routed addresses are actually *used* (for any definition). Allocated and routed addresses are also studied by Meng *et al.* [52], who found that most IPv4 prefixes allocated between 1997 and 2004 appeared in the global routing system within 75 days.

With respect to measurement to evaluate actual address usage, USC's long-standing effort [15], [29] periodically probes the entire IPv4 space with ICMP echo requests. Probing every routed IPv4 address over ~30 days, repeated multiple times between 2005 and 2007, they observed only 3.6% of allocated addresses responding [29]. In developing their methodology, they compared ICMP and TCP probing to passive traffic observation of USC addresses on USC's own campus network, finding 14% more USC IP addresses visible to ICMP than to TCP, and 28% more USC IP addresses visible to passive traffic observation than to either ICMP or TCP active probing. But each method observed some IP addresses missed by other methods. Also, Bartlett *et al.* [13] found that passive traffic observation and active probing complemented each other for the purpose of discovering active network services on campus. In this work, we also find that active and passive methods are able to observe different subsets of addresses (Section VI), but unlike [29], we use our passive monitors to infer usage about the entire Internet instead of only hosts internal to a network we monitor.

However, passive measurements introduce their own challenges, most notably the presence of traffic using spoofed source IP addresses, which can badly pollute estimates if not removed. In [20], we introduced a methodology validated on two sources of traffic data available to us in 2012. In this work, we extend this approach to two additional types of data sources – the most challenging of which is sampled traffic captured at an IXP – and we then examine how resulting inferences can be influenced by characteristics specific to observation vantage points, such as traffic composition, size of the monitored address space, and duration and time of the measurement.

Others have also explored the use of passive data to estimate specific usage characteristics of IPv4 addresses. Zander *et al.* [63] estimated the *number* of used IPv4 addresses by applying a capture-recapture method for estimating population sizes on active and passive measurement logs of IP addresses collected from sources such as web servers and spam blacklists. This work is largely complementary to ours, since it does not focus on improving active and passive methodologies to collect census data and understand their complementarity, but rather proposes an approach to estimate the *size* of the used space that such methodologies fail to observe.

Durumeric *et al.* [23] explored the system challenges of active Internet-wide scanning in developing Zmap, a scanner that probes the entire IPv4 address space in under 45 minutes from a single machine. Accelerated scanning was also a goal of an Internet Census illegally (and anonymously) performed in 2012 from a botnet [30], although their methods were neither well-documented nor validated [5]. Finally, Cai *et al.* [15] explore (and undertake several) potential applications of clustering active probes to infer address usage, including understanding how efficiently individual address blocks are used, assessing the prevalence of dynamic address management, and distinguishing low-bitrate from broadband edge links.

III. Datasets

Table I summarizes the datasets we use, which include 4 types of passive traffic traces (from a darknet, an academic ISP, an IXP, and a residential ISP), 3 types of active measurements, BGP data, IPv4 address allocation data, and derived data about geolocations and ASes. They were collected between July and October 2013.

Dataset	Source type	Data format	Period
UCSD-NT [2]	Traffic: Darknet	full pkt traces	July 23 to August 25, 2013
SWITCH [59]	Traffic: Live Academic Net.	Netflow logs	July 23 to August 25, 2013
IXP [8]	Traffic: IXP	sFlow packet samples	July 8 to July 28, August 12 to September 8, 2013
R-ISP [25]	Traffic: Residential ISP	Tstat [24] logs	July 1 to September 31, 2013
ISI [41]	Active Probing: ICMP ping	logs	July 23 to August 25, 2013
HTTP [28]	Active Probing: HTTP GET	logs	October 29, 2013
ARK-TTL [34]	Active Probing: traceroute	logs	July to September, 2013
BGP [6], [57]	BGP announcements	RIBs	July to September, 2013
Available Blocks [27]	IANA/RIRs	IP ranges	October 1, 2013
NetAcuity Edge [22]	IP Geolocation	IP ranges	July 2013
prefix2AS [16]	BGP announcements	prefix to ASN	July 2013

TABLE I: We infer used /24 blocks from passively collected traffic (UCSD-NT, SWITCH, IXP, R-ISP) and active probing (ISI, HTTP, ARK-TTL). The remaining datasets are used to infer both usable and routed prefixes, or label prefixes according to geolocation and AS.

Passive Data-plane Measurements. We apply our passive methodology for inferring used /24 blocks to the following four vantage points (VP), each of which retains traffic data in different formats and thus requires different approaches to filtering for use in a census (Section V). **SWITCH**: We collected unsampled NetFlow records from all the border routers of SWITCH, a national academic backbone network serving 46 single-homed universities and research institutes in Switzerland [59]. The monitored address range of SWITCH contains 2.2 million IP addresses, which correspond to a continuous block slightly larger than a /11. **R-ISP**: We collected per-flow logs from a vantage point monitoring traffic of about 25,000 residential ADSL customers of a major European ISP [25]. The VP is instrumented to run Tstat, an open source passive traffic flow analyser [24] that stores transport-level statistics of bidirectional flows, and uses internal network knowledge to label flows as inbound or outbound. **UCSD-NT**: We collected full packet traces from the /8 network telescope operated at the University of California San Diego [2]. **IXP**: Our fourth VP is one of the largest IXPs in the world, which is located in Europe, interconnects $O(100)$ networks, and exchanges more than 400 PB monthly [8]. We have access to randomly sampled (1 out of 16K) packets, capturing the first 128 bytes of each sampled Ethernet frame exchanged via the public switching infrastructure of this IXP. A sample includes full Ethernet, network- and transport-layer headers, along with a few payload bytes.

Active Measurements. **ISI**: We used the ISI Internet Census dataset *it55w-20130723* [41], obtained by probing the routed IPv4 address space with ICMP echo requests[1] and retaining

[1] We did not use reverse DNS PTR scans of the IPv4 space for the same reasons articulated in [29], namely that many active IP addresses lack DNS mappings, and many unused IP addresses still have (obsolete) DNS mappings.

only those probes that received an ICMP echo reply from an address that matched the one probed (as recommended [42]). Note that the ISI Census experiment was designed to report at a /32 (host) rather than /24 (subnet) granularity, but we apply the resulting data set to a /24 granularity analysis. **HTTP**: We extracted IP addresses from logs of Project Sonar's HTTP (TCP port 80) scan of the entire IPv4 address space on October 29, 2013 [28]. For each /24 block, we stored how many IP addresses responded to an HTTP GET query from the scan. **Ark-TTL**: We processed ICMP traceroutes performed by CAIDA's Archipelago to each /24 in the routed IPv4 address space between July and September 2013 [34]. Specifically, we extracted the ICMP Time Exceeded replies sent by hops along the traceroute path.

Address Allocation and BGP Data. We analyzed BGP announcements captured by all collectors (24 collectors peering with 184 peers) of the Routeviews [6] and RIPE RIS [57] projects. For each collector we took all routing tables (dumped every 2 hours by Routeviews and 8 hours by RIPE RIS) and built per-day statistics for each peer. For each /24 block, we computed the maximum number of peers that saw it reachable at any time within the full observation period of 92 days. To determine which address blocks are available for assignment, we used a dataset compiled by Geoff Huston [27], which merges the extended delegation files from the 5 RIRs [7], [10], [11], [45], [56] with IANA's published registries [35]–[40].

Mapping to ASes and Countries. To establish a mapping from /24 block to ASN, we merged all CAIDA's Routeviews Prefix to AS [16] mappings files for July 2013. For each /24 in the IPv4 address space, we identified the set of overlapping prefixes and chose the most specific. We found 116k /24s (out of more than 10M) that mapped to multiple ASNs (due to multi-origin ASes and AS sets), which we omitted from our per-AS computations (Sections VI and VII). We geolocated each /24 block using Digital Element's NetAcuity Edge [22] database from 6 July 2013. For each /24, we identified the unique set of country codes to which overlapping blocks map. We found 27k /24s (out of more than 14M) that map to multiple countries, which we excluded from the geographic visualization in Section VII.

IV. A taxonomy of Internet address space utilization

How to break down the whole IPv4 address space based on what use is made of it? Of the unrouted space, which is assigned vs. available?

We propose a taxonomy of the IPv4 address space according to the tree in Figure 1, where

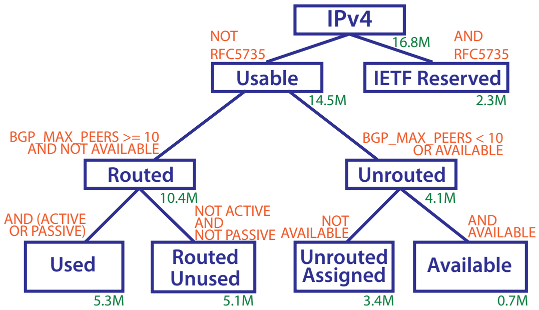


Fig. 1: IPv4 address space taxonomy. Nodes are annotated with the estimated /24 population of each category (Section VII) and the filter applied to arrive at the estimate (Sections IV through VI).

blue labels set the terminology that we use throughout the paper and red annotations summarize the classification criteria. While this taxonomy is generally applicable, in this paper we analyze the IPv4 address space with /24 block granularity. There is no universal IP address segment boundary space (due to sub-netting and varying size of administrative domains), but using a /24 granularity mitigates the effects of dynamic but temporary IP address assignment (e.g., DHCP), as well as having an intuitive relationship with both routing operations and address allocation policy.

All address blocks dedicated to special use (multicast, private networks, etc.) are *IETF reserved* and are covered by RFC5735 [51] (≈2.3M /24 blocks). To classify the remainder into *routed* and *unrouted*, we must distinguish legitimately routed address blocks from those that appear in BGP announcements due to router misconfigurations. We consider a /24 block as *routed* only if covered by a prefix visible by at least 10 BGP peers. RIPE recommends this threshold [62], which we believe is reasonable since it removed from BGP measurements 99.93% of the /24 blocks we previously determined were reserved by IETF or *available* (defined in the next paragraph) and thus could not be legitimately routed via BGP.

Of the 4.1M *unrouted* /24 blocks, we classified as *available* any /24 block (≈.7M) falling in address ranges marked in Geoff Huston's dataset (Section III) as either "available" (i.e., allocated to an RIR but not yet assigned to an LIR or organization) or "ianapool" (i.e., IANA has not allocated it to an RIR) [27]. This data does not have LIR granularity, thus we considered any block allocated to an LIR as assigned (i.e., not available). The remainder – the *unrouted assigned* category – is made of 3.4M /24 blocks that are assigned to organizations (many of whom announce other IPv4 address space) and yet are not routed. In other words, we find that

≈53 /8's worth of address space are not used for the purpose of global BGP reachability.

Our filtering yields 10.4M *routed* /24 blocks that we further classify as *used* or *unused*. We define a /24 block as *used* if at least one of its IP addresses is assigned to a machine that will exchange packets on the public Internet with such address in the IP header. In Sections V and VI, we discuss the inference methodologies – based on both active and passive measurements – that we use for this purpose. Figure 2, provides an overview of our final results according to our taxonomy and breaking the space by RIR and legacy allocations. This visualization succinctly represents "where" in the allocation system, and how, large portions of address space appear unutilized.

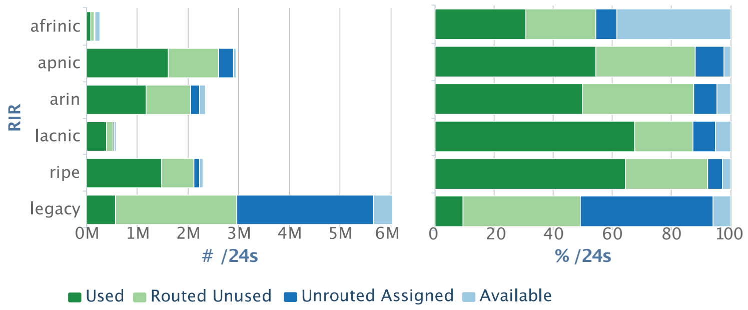


Fig. 2: Our final inferences classified by RIR-allocated (and legacy) address space. We identify legacy addresses per /8 [36], but include some /8s that are presently administered by RIRs. Only 9.5% of the legacy addresses are used.

V. Analysis of Passive Traffic

Is the approach of passive measurement for inferring address space utilization generally applicable? How does it depend on different network types, trace types, and other parameters?

We first extend the method of [20], which used data from a darknet and an academic ISP, to work with the fundamentally different types of traffic collected at a residential ISP and an IXP, showing how to filter out spoofed traffic in different trace types (Section V-A). Second, we evaluate the impact on our inferences of varying aspects of the vantage points: traffic composition, size of monitored address space, duration and time of measurement (Section V-B).

Vantage	Original Traffic			After Applying Heuristics		
Point	/24 blocks	Unrouted	Dark	/24 blocks	Unrouted	Dark
UCSD-NT	10,884,504	1,284,219 (31.6%)	D-SWITCH: 4,553 (90.9%)	3,152,067	2,123 (0.05%)	D-SWITCH: 2 (0.04%)
SWITCH	4,679,233	35,585 (0.69%)	UCSD-NT: 429 (0.68%)	3,599,558	178 (0.004%)	UCSD-NT: 0 (0.00%)
R-ISP	5,233,871	344,188 (8.5%)	UCSD-NT: 7,287 (11.6%)	3,797,544	271 (0.006%)	UCSD-NT: 0 (0.00%)
IXP	14,461,947	4,068,232 (78.5%)	UCSD-NT: 62,838 (100%)	3,091,021	376 (0.009%)	UCSD-NT: 3 (0.004%)

TABLE II: Applying our heuristics to remove spoofed traffic reduces the number of unrouted and dark (i.e., likely spoofed) /24 blocks at all VPs. For each VP, we report the absolute number and percentage of all /24 blocks that are unrouted. For the dark category (4th and 7th column), we use the /24 blocks of SWITCH that did not generate bidirectional flows (D-SWITCH) to evaluate UCSD-NT, and the addresses monitored by UCSD-NT to evaluate all other VPs.

A. Removing spoofed traffic

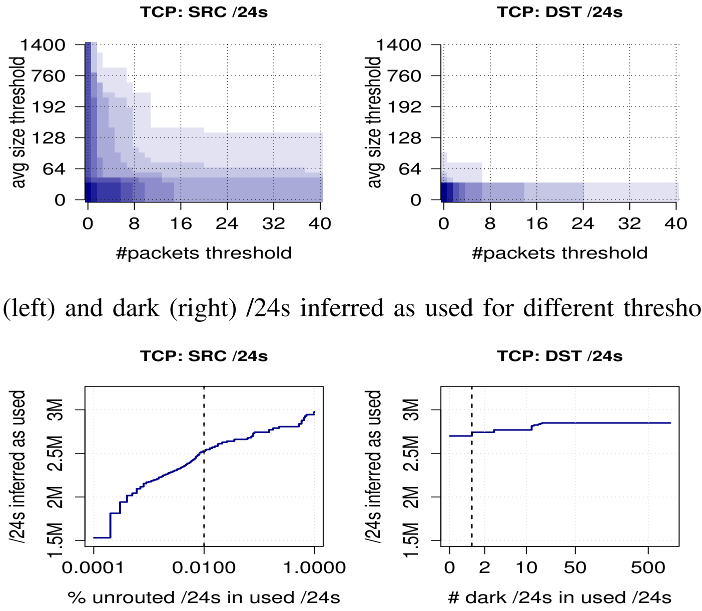
The main challenge in curating traffic data for use in a census is to remove spoofed traffic from the data sets, since it can severely distort estimates of address utilization. Since the R-ISP data retains bidirectional flow information and is guaranteed to see both directions of every flow, filtering out spoofed traffic is easy. For the IXP, the sampled data collection and the frequently asymmetric traffic flow (i.e., only one direction of a flow may traverse the IXP) mean that we cannot use the obvious and most reliable technique to infer spoofed traffic (i.e., failed TCP flow completion, variants of which we use for R-ISP and SWITCH data). Indeed, the IXP data sees only one packet for the vast majority of flows. The IXP data also introduces a new challenge: filtering out packets with potentially unused destination addresses (e.g., scanning packets).

Although each VP's data set requires its own technique, we tune and validate each technique using the same assumption: packets appearing to originate from [or destined to] *unrouted* blocks are likely spoofed [or scanning] packets. As an additional source of validation, we compare our results against other network blocks that we know to be unused. Specifically: (i) at the SWITCH, R-ISP, and IXP VPs we use the dark /24 blocks in the UCSD-NT address space [2] (62,838 /24 blocks); (ii) at the UCSD-NT VP, we use the /24 blocks from SWITCH that we infer to be dark because they did not generate a single bidirectional flow in the whole observation period (5,003 /24 blocks). We use these data only with UCSD-NT because their observation periods exactly match. Table II shows the numbers of /24s found by each VP before and after applying our heuristics.

[2] Some addresses within this "darknet" are actually used and their traffic is not collected.

1) IXP (large IXP): For the IXP, we consider only TCP traffic and discarded TCP packets with the SYN flag set, which reduced the number of observed /24s from 14.4M to 5.7M /24s. We then used a heuristic that tries to filter out /24s observed only due to spoofing noise. This heuristic is based on thresholds to the number of packets and average packet size from and to a given /24 block. These thresholds impose a trade-off between false positives and false negatives: if we set the packet count threshold high enough, we are more likely to filter out /24s that contain only IP addresses being used in spoofed source address packets, but we will also lose many legitimately used /24 blocks (especially since we only have 1:16K sampled packet data in to begin with). The average packet size threshold complements the packet count threshold by increasing the likelihood of retaining /24s that are actually exchanging TCP payload.

The left plot in Figure 3a shows how we can use packets appearing from unrouted IP addresses (i.e., likely spoofed) to tune our heuristic: the darker the color the higher the number of unrouted blocks that we infer as used based on source addresses of the sampled packets (log scale).



(a) Unrouted (left) and dark (right) /24s inferred as used for different threshold combinations.

(b) Trade off between introduced error (unrouted /24s, dark /24s) and the number of /24s inferred as used.

Fig. 3: IXP: Threshold selection for inference of used /24s.

Interestingly, we see almost no packets in the IXP data set destined to unrouted /24 blocks, perhaps because there are no default routes advertised across BGP peering (vs. transit) sessions at the IXP, so only explicitly routed addresses will be observed as destinations. We can still use dark but routed destination addresses as indicators of scanning traffic; the right plot in Figure 3a shows the number of dark /24 blocks inferred as used when considering the destination addresses of packets. The average packet size is highly efficient at removing scanning traffic.

To find an appropriate combination of thresholds, Figure 3b plots ROC-like curves that show the number of /24 blocks inferred as used (y axis) against the percentage of unrouted source addresses (left diagram) and number of dark destination addresses (right diagram) /24 blocks. For a given requirement (e.g., "less than 0.1% unrouted"), we find the combination of thresholds (minimum number of packets and minimum average packet size) that results in the largest set of used /24 blocks.

To keep the error in our inference low, we use very conservative thresholds (shown as dashed vertical lines in Figure 3b) and select used /24s from the SRC and DST addresses independently. The resulting numbers are depicted in Table II. Our antispoofing approach is efficient, reducing the number of unrouted and dark /24s dramatically, even for sampled traffic. Nevertheless, we point out that the number of used /24s directly depends on the thresholds applied and that the false-negative rate increases with more conservative thresholds. Hence, we likely miss used /24s with our threshold selection.

We found similar behavior with UDP (as TCP) but we needed to set thresholds more conservatively, particularly for average packet size. We did not include UDP-based inferences in our final dataset, since the additional gain in terms of /24s was not significant.

2) R-ISP (residential ADSL ISP): Unlike the other traffic data sources, the R-ISP's use of Tstat automatically removes essentially all TCP spoofed traffic, since to be logged a TCP flow must complete the 3-way handshake. For UDP traffic, our approach is to extract only bidirectional flows initiated locally with at least 1 packet with payload transmitted in both directions. We consider both source and destination addresses from the selected TCP and UDP flows. Table II confirms the accuracy of our approach.

3) UCSD-NT: (a large darknet): In [20] we looked deeply into several *spoofing events* to derive filters that would allow us to filter such events from darknet traffic in general. Two phenomena that we found to be indicators of a spoofing event were: (i) spikes in the numbers of both unrouted and overall /24 blocks per hour, and (ii) traffic using the same ports and protocols with a high fraction of unrouted source /24 blocks. We developed general filters (properties of the traffic that always indicate spoofing), and filters specific to individual events. Many types

Spoofed Traffic Filter	Total /24s	Unrouted /24s
TTL> 200 and not ICMP	10,588,879	1,278,027
Least signif. byte src addr 0	45,382	7
Least signif. byte src addr 255	444,346	6,691
Non-traditional Protocol	56,502	2,209
Same Src. and Dst. Addr.	96	0
No TCP Flags	3,449	638
UDP Without Payload	545	114
All Specific Filters	10,587,049	1,280,826

TABLE III: Types of spoofed traffic observed and removed at UCSD-NT. Total and unrouted /24s seen in each traffic type. All non-general filters are grouped as "All Specific Filters".

of spoofing captured by our generic filters in our 2012 study [20] were also present in 2013 (see [20] for details on methodology and filters). In addition, we added two general filters: TCP packets with no flags set and UDP packets without payload. Table III reports the number of /24 blocks matching each filter.

After applying our filters, we observe more than 3 million /24 blocks. Table II shows that our filtering heuristics reduce traffic appearing to originate from unrouted or dark networks to around 0.05% (compared to 31.6% and 90.9% unrouted and dark blocks, respectively, before filtering).

4) SWITCH (academic network): To filter spoofed traffic, we use the same heuristic we introduced in [20], which extracts from Netflow records bidirectional TCP flows with at least 5 packets and 80 bytes per packet on average and we use both source and destination addresses. We performed a sensitivity analysis on these thresholds in [20], and found that they diminish the probability that the remote IP address is spoofed. Using this heuristic leads us to infer as used only 0.004% and 0% of the unrouted and the UCSD-NT /24 blocks, respectively (Table II).

B. Effect of vantage points characteristics: traffic, network address segment, time, duration

After filtering spoofed traffic, we analyze the impact of four characteristics specific to a given vantage point on the number of /24s observed: type of traffic, size of address space monitored, and duration or specific time of monitoring. We find that all four VPs are reasonably robust to variations in these characteristics, i.e, we observe a substantial fraction of address space at all VPs or when observing from smaller fractions of the address spaces (where we could test that), and each VP saw a consistent number of /24 blocks over a two-year period.

1) Influential Traffic Components: How do traffic characteristics specific to a VP influence its contribution to the inferences?

R-ISP Traffic Class	/24 Blocks	Unique	Volume
P2P[a]	3,172,439 (91.2%)	610,438	34.1%
Teredo	914,533 (26.3%)	1,467	1.4%
VoIP (RTP,RTCP)	892,488 (25.7%)	3,619	0.5%
HTTP/HTTPS	234,586 (6.8%)	20,274	57.7%
Other[b]	196,503 (5.7%)	62,406	1.9%
Unknown[c]	2,691,300 (77.4%)	115,869	4.5%

[a] eMule, ED2K, KAD, BitTorrent, PPLive, SopCast, TVAnts, and PPStream

[b] DNS, POP3, SMTP, IMAP4, XMPP, MSN, RTMP, SSH

[c] Flows unmatched by the classification engines.

TABLE IV: At the R-ISP VP, P2P traffic contributes almost 3.2M /24 blocks, including 610K unique. HTTP/HTTPS is a smaller component, despite accounting for 57.7% of the volume.

Darknet Traffic Class	/24 Blocks		Unique
BitTorrent	2,210,257	(70.2%)	321,474
Encrypted[a]	1,349,578	(42.8%)	34,290
UDP Qihoo 360 bug	1,343,911	(42.7%)	115,951
Other P2P (eDonkey,QQLive)	834,657	(26.5%)	5,361
Encapsulated IPv6 (Teredo,6to4)	745,092	(23.7%)	11,322
Conficker	604,877	(19.2%)	61,836
Backscatter	388,095	(12.3%)	53,277
Scanning (non-Conficker)[b]	194,649	(6.2%)	4,269
Other	2,038,150	(64.7%)	143,066

[a] Packets where entropy(payload)$\approx\log_2$ len(payload).

[b] Meeting Bro's definition of a scanner: sent same protocol/port packets to at least 25 destinations in 5 minutes [61].

TABLE V: At UCSD-NT, BitTorrent traffic contributes the most /24 blocks, instead of activities traditionally observed in darknets (scanning, Conficker, backscatter).

Characterizing traffic at our VPs assists with two objectives: (i) highlighting how the VP contributes to the census; and (ii) ensuring that traffic components specific to a VP do not skew our findings or make them not generally applicable. That is, to legitimately use passive traffic data for a census, we need to convince ourselves that a given VP is not observing a special set of /24 blocks. For objective (i), SWITCH's popular services attract users from many /24 blocks, while R-ISP and UCSD-NT contribute many /24 blocks as the result of P2P traffic. However, without these traffic components, each VP would still see at least 69.9% of the /24 blocks it would normally observe, implying that traffic composition at a particular VP does not skew our results (i.e., objective ii). We could not analyze traffic composition from the IXP due to the sampled packet capture.

SWITCH. SWITCH hosts many popular services that attract end users to the monitored address space, including: a website hosting medical information (exchanging traffic with hosts in 1.8M /24 blocks), a SourceForge mirror, PlanetLab nodes, university web pages, and mail servers. The top 100 most popular IP addresses (i.e., the top services) in SWITCH each observe over 70K /24 blocks, and collectively contribute 91.2% of the /24 blocks observed at this VP.

Compared to the UCSD-NT and R-ISP vantage points, SWITCH's value as a VP depends more on popular IP addresses. If SWITCH did not host its top 1000 most popular IP addresses, it would observe only 69.9% of the /24 blocks it otherwise observes, compared to 89.7% and 97.5% at R-ISP and UCSD-NT respectively. This finding can be easily explained: all three VPs see a large fraction of traffic from client hosts, but the prevalence of P2P traffic in R-ISP and UCSD-NT makes the monitored IP addresses more "interchangeable", whereas the top services at SWITCH tend to attract varying client populations.

R-ISP. Table IV aggregates the Tstat-identified traffic categories observed at R-ISP into five traffic components accounting for 97% of /24 blocks observed at the ISP. While HTTP and HTTPS account for 57.7% of the traffic volume, they contribute only 6.8% of the /24 blocks observed at the VP. Instead, the largest source of /24 blocks comes from client-to-client communication (e.g., P2P and VoIP). P2P is a key contributor, as 610k /24 blocks are only observable through P2P traffic.

UCSD-NT. Surprisingly, P2P also plays a key role at the UCSD-NT VP, where we observe 2.2M /24 blocks (357k unique) from traffic with a BitTorrent payload (see Table V), probably caused by index poisoning attacks [46]. Qihoo 360 updates using a P2P network [1] and a byte-order bug in the software results in traffic from sources in over 1.3M /24 blocks, 40% of which geolocate to China. To a lesser extent, networks with end users are exposed through malware-infected hosts (e.g., Conficker and scanning). Alternatively, the backscatter traffic (a result of spoofed DoS attacks) reveals networks likely hosting services. In [14], we present a thorough analysis of (unspoofed) traffic reaching large darknets.

2) Impact of Vantage Point Size: We analyze vantage point size (the number of IP addresses monitored) to determine the extent to which our results depend on access to large datasets. Unfortunately, the analysis of vantage point size is not straightforward due to the non-uniform nature of the monitored address space. Notwithstanding the extraordinary popularity of some IP addresses, as well as non-uniform assignment of hosts within an address subnet, we found an interesting correlation: for each vantage point, the median number of /24 blocks observed is roughly proportional to the log of the number of monitored IP addresses. Consistent with this observation, the utility of a monitored IP address declines as the size of the vantage point increases. While our results benefit large datasets, halving or doubling the size of our vantage points is unlikely to have a substantial impact on the number of /24 blocks we infer as used.

3) Impact of Time: How does the duration or time of collection affect the inference of which /24s are used?

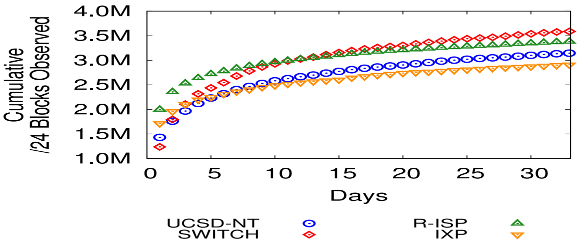


Fig. 4: The cumulative number of /24 blocks observed grows logarithmically at each vantage point.

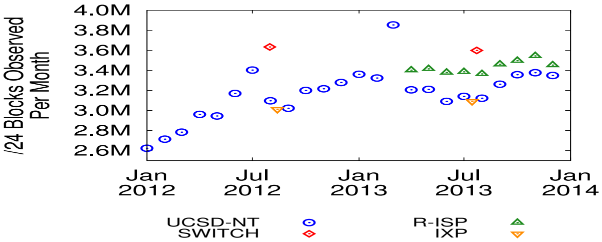


Fig. 5: In our data, taken over two years, every VP observed at least 2.6M /24 blocks per month. The fluctations in UCSD-NT data are the result of changes in IBR composition.

Figure 4 shows the logarithmic but varied growth of the number of /24 blocks collected over time for our four VPs. SWITCH, which initially captures the fewest /24 blocks has the fastest growth rate; while the R-ISP and IXP VPs capture more /24 blocks initially but they grow more slowly. Other factors that can influence inferences are strong changes in traffic composition, e.g., flash events. Our traffic data sets all had low (max 2%) standard deviation in the number of /24 blocks observed per week, with no abnormal events observed.

However, when observing measurements from a broader time frame, we found evidence of flash events and changes in traffic. For example, in August 2012 (the year preceding our datasets), SWITCH web sites hosting content about shark protection experienced a sharp increase in visits (and thus observed /24 blocks); the Discovery Channel's Shark Week aired that month. Figure 5 shows per-month sample measurements using our methodology over a period of two years. The SWITCH and IXP VPs oberved a similar number of /24 blocks approximately one year prior to our census. R-ISP consistently observed between 3.4M and 3.6M /24 blocks for nine consecutive months. At UCSD-NT, changes in IBR composition resulted in a corresponding increase in visible /24 blocks. Specifically, (i) in July 2012, there was an increase in BitTorrent traffic; (ii) in March 2013, there was a large increase in the darknet's backscatter category, possibly related to the DDoS attacks on Spamhaus [55]. Such events may increase the number of /24 blocks inferred as used, but our technique does not appear to significantly depend on one-off events: in our data, every VP observed at least 2.6M /24 blocks per month.

VI. Combining Active and Passive

In this section, we first combine our seven datasets obtained from active and passive measurements to break down the *routed* node in Figure 1 into *used* and *routed unused* categories (we filtered all the datasets used in this section to include only /24 blocks marked as *routed* according to Section IV). We then compare our results to the state of the art represented by the ISI census (Section VI-B).

A. Active vs Passive

What are the respective strengths and limitations of active and passive measurements? Are passive measurements from multiple VPs useful?

The top half of Table VI shows the number of /24 blocks discovered by each active approach and their unique contribution. The large number of /24 blocks found by ISI and HTTP, and their distinct contributions within the set of active measurements, are unsurprising because we know that ICMP and TCP port 80 probing are among the most effective active probing methods that capture different but overlapping populations [29], [50]. More surprising is the 40k additional /24 blocks that we obtain from the Ark dataset; we speculate that routers may be sending TTL exceeded packets using a different source address from what they use in ICMP echo responses.

The bottom half of Table VI compares the contribution of our passive measurements. The merged results from our four passive VPs do not entirely cover the set observed by active measurements, missing about 840k /24 blocks. However, the same data includes 470k /24 blocks not observed through active measurements, demonstrating the value of combining active and passive datasets.

Each passive vantage point offers a unique contribution, shown in the third and fourth columns of Table VI, suggesting that these measurements are not exhaustive and that using more vantage

Dataset	# /24s	# Unique /24s within same family	# Unique /24s among active + passive
Active			
ISI	4,589,213	1,319,283	398,334
HTTP	3,161,064	189,831	76,189
Ark-TTL	1,627,363	40,284	24,533
All Active	4,837,056		
Passive			
SWITCH	3,599,380	147,220	54,905
UCSD-NT	3,149,944	61,443	24,134
R-ISP	3,797,273	176,721	59,278
IXP	3,090,645	195,328	55,155
All Passive	4,468,096		
Total	5,306,935		

TABLE VI: Each data set used to infer address space utilization offers a unique contribution. Unrouted /24 blocks are not represented here. The third column is the number of /24s observed in the data set that were not also observed in the (top) other active data sets or (bottom) other passive data sets; the fourth column is the number of /24s observed that were not observed in any other data set. The final total is the number of /24s we infer as *used* (lower left node of tree in Figure 1).

points would improve the coverage. In particular, when we examine the portion of the address space observed exclusively by passive approaches (470k /24 blocks, not shown in the table), we find that only 17% of it was visible by all four vantage points, while ≈ 41% came from the sum of each unique contribution (4th column in Table VI).

Since 3 out of 4 vantage points are in Europe, we test for the possibility of geographical bias in the passive measurements. Table VII shows the percent increase of /24 blocks discovered by merged passive+active data vs. active measurements alone. The larger increase in European coverage vs. other continents (middle column) is consistent with a slight bias from the European vantage points, but on a per-continent basis the marginal increase spreads more evenly across continents (right column, noting that the lower three continents have so much less address space that any increase will be relatively large in percentage terms.)

We also explored why a significant portion of space is discovered only by the active measurements in our data sets. On one hand, there are limitations in passive approaches, some of which will be subject of future work: (i) some of our heuristics to remove spoofed traffic may be too conservative and remove much legitimate traffic; (ii) for IXP and SWITCH, we included only TCP traffic which could have limited our view; curating UDP and other traffic would probably

	% of newly discovered /24 blocks	per-continent % increase of /24 blocks
Europe	32.44%	11.11%
North America	26.54%	9.08%
Asia	25.31%	7.64%
South America	8.56%	10.85%
Africa	4.65%	30.18%
Oceania	4.33%	29.24%

TABLE VII: Absence of significant geographical bias in passive vs active measurements: of the number of /24 blocks discovered by passive approaches and not seen by active ones, a slightly larger portion is geolocated to Europe (where 3 of our 4 passive VP are). But on a per-continent basis (right colum), the increase is more even across continents (Southern continents have little address space so any increase will be relatively large in percentage terms.)

# VPs	# ISI-special /24s	# single-IP /24s without ISI-special
0	94,266	58,132
1	13,057	19,414
2	9,674	19,115
3	4,959	27,185
4	2,465	13,091

TABLE VIII: Most /24 blocks with only a single IP address ending in .0, .1, .255 are not observed by any of our passive measurements (first row and middle column). In contrast, if a /24 had only a single responding address ending in another octet, it was more likely to be observed sending traffic (3rd column). We conclude that most /24s represented in the middle column likely do not send traffic to the public Internet.

improve coverage; (iii) as discussed, adding more VPs would also bring an improvement. On the other hand, our results also reveal fractions of IPv4 space that does not seem to spontaneously generate traffic on the public Internet, since visible only by active measurements and showing special properties. For example, we found that most /24 blocks from the ISI dataset with a single responding address whose last octet was 0, 1, or 255 (*isi_special* column in Table VIII) were not observed in our passive measurements. Table VIII shows the distribution of the number of passive vantage points that saw such /24 blocks (2nd column), as well as all /24s in the ISI data that had only a single non-special responding IP address (3rd column). The progression from /24 blocks observed by 1 to 4 VPs shows a rapid decay for *isi_special* blocks (middle column), while there is almost no trend for /24s in the right column. We conclude that most of the /24s represented in the middle column likely do not send traffic to the public Internet. This finding poses the question of wether such addresses – even if matching our definition of *used* – are actually utilized for the purpose of global reachability, which suggests to extend our taxonomy in the future by defining *used* subcategories to provide additional insight.

We manually investigated other cases of network blocks only visible to active probing, identifying special cases that suggest that they are not used on the public Internet, including clusters of /24 blocks apparently used as internal CDNs by large service providers. In a study led by the Naval Postgraduate School [9], we also identified network *tarpits* (a form of defensive cyber-deception, whereby a single host or appliance can masquerade as many fake hosts on a network and slow network scanners) as large as /16, polluting Internet census data. We plan a thorough investigation of all these behaviors (and their taxonomization) as future work.

The last row of Table VI shows the final number (5.3M) of /24 blocks we infer as *used* combining our 7 active and passive datasets (leftmost leaf in Figure 1). We subtract this from the total amount of BGP-routed space (10.4M) to arrive at an estimate of **5.1M *routed unused* /24 blocks, an impressive quantity of unused but BGP-reachable IPv4 space**. A similar number of routed but unused /24 blocks is also corroborated by Zander *et al*. [63], who used a capture-recapture methodology to estimate the number of /24 blocks missed by passive VPs.

B. Coverage

What is the improvement of our combined approach to infer utilization in the routed space with respect to the state of the art (ISI census)?

We consider the ISI Census [29] to be the state of the art in inferring address space utilization within the routed space. Since there is no global ground truth available about which routed space is actually utilized, we present our results in terms of additional IPv4 space coverage we obtain when combining our 7 datasets (which include ISI). We define coverage at three different levels: (i) the percentage of routed /24 blocks inferred as used (*global coverage*); (ii) the percentage of ASes announcing the /24 blocks inferred as used out of the ASes that announce at least one BGP prefix (44,628 ASes) (*AS-level coverage*); (iii) for each AS, the percentage of routed /24 blocks inferred as used (*intra-AS coverage*). AS-level coverage is the only case in which we expect the upper bound to approximate ground truth (i.e., it is reasonable to assume that an AS announcing prefixes on BGP uses at least one /24 block).

We found 718k previously undiscovered used /24 blocks (difference between last and 1st row of Table VI), bringing global coverage from 44% to 51%. Figure 6 shows that adding just a single dataset can greatly improve the global coverage. As we include our additional datasets, there is considerable amount of overlap. If we were to include additional measurements of used address space, the actual number of /24 blocks would be highly dependent on the quality and diversity of the datasets. However, if we consider the logarithmic trend suggested by our observations, increasing the number of additional datasets from 6 to 12 would result in approximately 200k more /24 blocks.

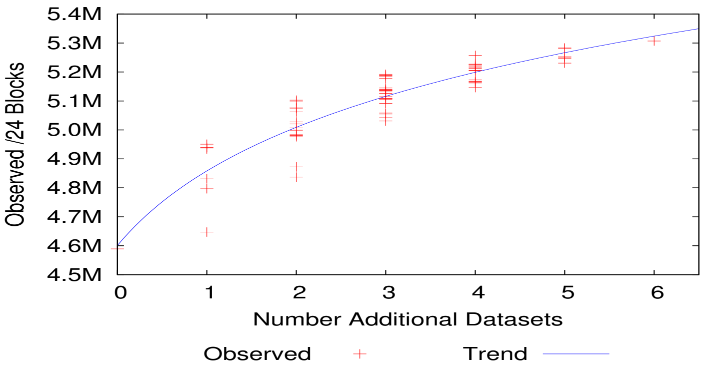


Fig. 6: We improve the global coverage of ISI (0 additional datasets) by considering ISI + any number of our datasets. The actual number of /24 blocks observed are shown in red, and the seemingly logarithmic trend is shown in blue.

Our AS-level coverage is 98.9% versus 94.9% found by ISI. We manually analyzed whois and BGP data for the 489 ASes for which we did not infer a single used /24 block. We found that 37 ASes associated with U.S. military organizations accounted for 79% of the (17,080) /24 blocks advertised by these 489 unobserved ASes. We suspect such networks do not transmit ICMP, TCP or UDP traffic over the public Internet (but they may be tunneling traffic using, e.g., IPSEC, which we did not capture in our passive measurements.) The vast majority of the remaining ASes (399 out of 452) announce 10 or fewer /24 blocks.

Figure 7 shows the intra-AS coverage obtained with our combined approach as a function of results obtained by ISI (the graph is sorted by increasing ISI intra-AS coverage, with bins of 2%). For example, in the first bin, the bar from the median to the upper quartile shows intra-AS coverage between 23% and 100%. The graph shows visible increments across the whole x axis (decreasing as ISI intra-AS coverage approaches 100%). This result shows that even for ASes which responded to ISI's pings ($x! = 0$), our additional datasets reveal new /24 blocks (i.e., ASes do not exhibit a uniform behavior across their used subnets with respect to ICMP echo requests). In most of the bins, for half of the ASes (i.e., two bottom quartiles) we obtain a few percentage point increase. The two upper quartiles show more significant increments, e.g., up to $x = 20$, for ASes in the upper quartile we see about 20% more /24 blocks (at least). The

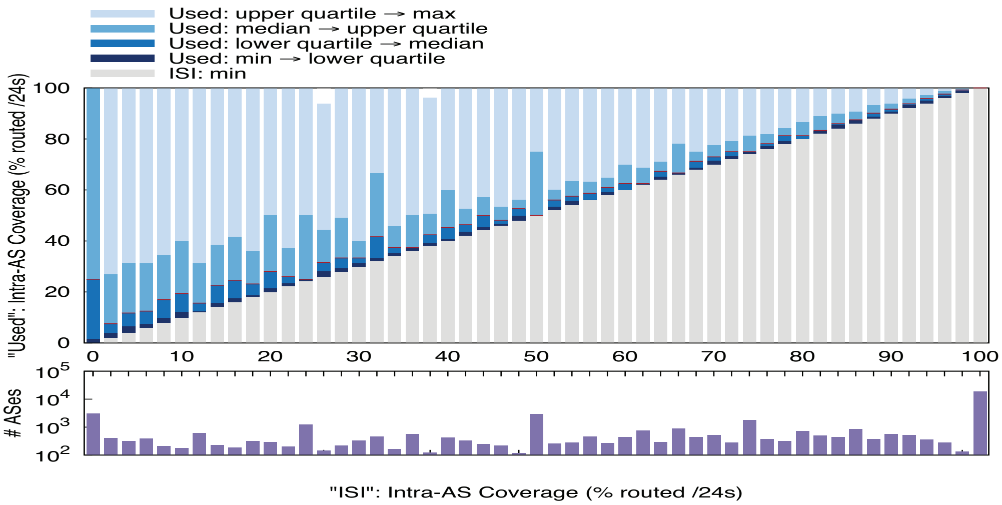


Fig. 7: Comparing the intra-AS coverage of our combined approach ("Used") against ISI's. The graph is sorted by increasing intra-AS coverage in ISI's data, with bins of 2%. The bottom graph shows the number of ASes per bin. In the top graph, the bottom grey bar represents the intra-AS coverage obtained by ISI for ASes in the bin, whereas the remaining 4 (colored) bars refer to the intra-AS coverage obtained by our combined approach (which includes ISI data). Each of these 4 bars represents a quartile of the ASes in the bin. For each bar, its bottom and top show on the y axis, respectively, the lower and upper bound of the coverage we obtain for ASes in that quartile (e.g., in the first bin, the bar from the median to the upper quartile shows intra-AS coverage between 23% and 100%).

first bin shows different behavior, with at least 25% of ASes covered entirely by our method (although most of these ASes announce only one /24).

SWITCH is the only AS for which we can derive better reference data (rather than simply using the 100% upper bound): from 23 July to 25 August 2013, all 9,271 /24 blocks within SWITCH were announced in BGP, but only 49% of these blocks generated bidirectional flows. Assuming these are the only used /24 blocks in SWITCH, we should not infer an intra-AS coverage above 49% for this AS (instead of considering 100% of the routed /24 blocks according to our definition of upper bound). ISI's infers 20.9% intra-AS coverage for this AS; our combined approach (without data from the SWITCH VP) reaches 33.1%. Still almost 16% of the blocks of the AS (which are used) are not discovered by our approach, showing space for further improvement. However, for all other ASes we would include the SWITCH VP in our analysis (thus using 4 VPs instead of 3), potentially resulting in a higher intra-AS coverage.

VII. IPv4 Census 2013

How is (un)used space distributed across RIRs, ASes, countries and continents? Which ASes or countries make the worst use of the space they have been assigned? Would previous scientific *studies of Internet-related phenomena change if they used this dataset instead of other related data sets?*

Finally, we examine IPv4 address space utilization from the perspective of our inferences. We emphasize that our inferences do not provide complete coverage of the used IPv4 address space, but it is the first dataset which includes ASes and network blocks that do not reply to ICMP probing. In addition, the estimated number of missing addresses is small [63]. All our data is from approximately the same period (from July 2013 through Oct 2013). We assume that usage of the address space does not change significantly within a period of 4 months.

Figure 8 illustrates[3] a Hilbert map of IPv4 address space utilization based on our results, taxonomized in Figure 1. The *IETF reserved* space accounts for 2.3M address blocks, or 13.7% of the entire IPv4 address space (grey). The remaining usable 14.5M address blocks consist of 5.3M (37%) *used* (light blue), 5.1M (35%) *routed unused* (dark blue), 3.4M (23%) *unrouted assigned* (purple), and 0.7M (5%) *available* (black). It is striking that most of the usable address space is actually unused. An enormous amount of IPv4 address space is assigned to organizations that do not even announce it on the Internet (i.e., there is no need to perform inference through additional active/passive measurements to sketch this phenomenon). In addition, since we verified that several of these organizations announce on BGP other address blocks they have been assigned, such number also suggests that our inference of large unused routed space is realistic.

A. View by allocation, geographic area, and AS

Figure 2 classifies IPv4 addresses by their RIR region, or as *legacy* addresses if they were allocated before the RIR system began. Legacy addresses were allocated by the central Internet Registry prior to the RIRs primarily to military organizations and large corporations such as IBM, AT&T, and Apple. Some of this space is now administered by individual RIRs. We use the IANA IPv4 address space registry [43], which marks legacy space and its designation at a /8 granularity. The figure shows that 42% of the usable address blocks are legacy; these blocks are more lightly utilized (9.5% of the legacy) and include more *unrouted assigned* (45% of the legacy) addresses than the RIRs (56% and 7.7% of the RIR address blocks, respectively). Interestingly, the combined set of legacy *routed unused* and *unrouted assigned* addresses is similar in size

[3] Full resolution of this image and other visualizations from this work are available at [17].

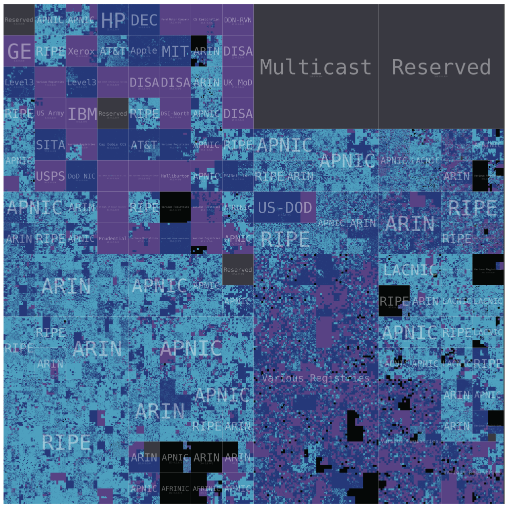


Fig. 8: Hilbert map visualization showing the utilization of the address space according to our taxonomy. The IPv4 address space is rendered in two dimensions using a space-filling continuous fractal Hilbert curve of order 12 [53], [58]. Each pixel in the full-resolution image [17] represents a /24 block; *light blue* indicates used, *dark blue* routed unused, *purple* unrouted assigned, *black* unassigned, and *grey* reserved by RFC blocks.

(5.1M /24s) to the entire *used* address space (5.3M /24s)! ARIN, RIPE, APNIC, and LACNIC have 50%, 65%, 54% and 68% of their address blocks *used*, respectively, in contrast to AFRINIC which has fewer of their blocks *used* (31%) and many more *available* (38%) address blocks than other RIRs (6.7% of other RIR addresses are *available*).

Table IX(a) lists the top-5 continents and countries in *routed unused* and *unrouted assigned* /24s. 52.2% of the *routed unused* space and 72% of *unrouted assigned* space is in North America, primarily in the U.S., where most legacy allocations were made. Asia follows, with China owning 8.79% and 5.7% of the global *routed unused* and *unrouted assigned* space, respectively, and then Europe. Other continents (South America, Oceania, and Africa) have between 0.93% and 2.13% of the global *routed unused* and *unrouted assigned* space. Figure 9 visually illustrates the per-country ratio of *assigned unused* (the sum of *routed unused* and *unrouted assigned*) over *assigned* (that is, *usable* minus *available*) space, suggesting which regions are using space most and least efficiently. The U.S. is red in this map due to a few very large allocations, while some African countries are red because they use a very small fraction of their (also small) assigned space.

Figure 10 compares address space assigned to countries to per-country population [19] and Gross Domestic Product (GDP - we used "purchasing power parity" from CIA's World Factbook [18]). We observe notable disparities between used /24s and population. For example,

Top Continents			
By Routed Unused /24s		By Unrouted Assigned /24s	
North America	52.2%	North America	72.0%
Asia	22.3%	Asia	13.1%
Europe	19.7%	Europe	12.1%
South America	2.13%	Oceania	0.97%
Oceania	1.92%	Africa	0.93%

Top Countries			
By Routed Unused /24s		By Unrouted Assigned /24s	
USA	49.8%	USA	67.5%
China	8.79%	China	5.70%
Japan	6.22%	United Kingdom	5.39%
Germany	4.85%	Japan	4.21%
South Korea	2.72%	Canada	3.73%

(a) Top continents and countries in unused and unrouted assigned /24s.

Top ASes in unused /24s	
AS Name & Number	Routed Unused /24s (%)
DoD NIC (721)	190k (3.82%)
Level 3 (3356)	157k (3.16%)
HP (71)	126k (2.54%)
China Telecom (4134)	106k (2.13%)
UUNET (701)	105k (2.12%)

(b) Top ASes in routed unused /24s

TABLE IX: Top continents, countries, AS names, and AS numbers in unused and unrouted assigned /24s. North America and USA have a large fraction of the assigned, but unused or unrouted address space.

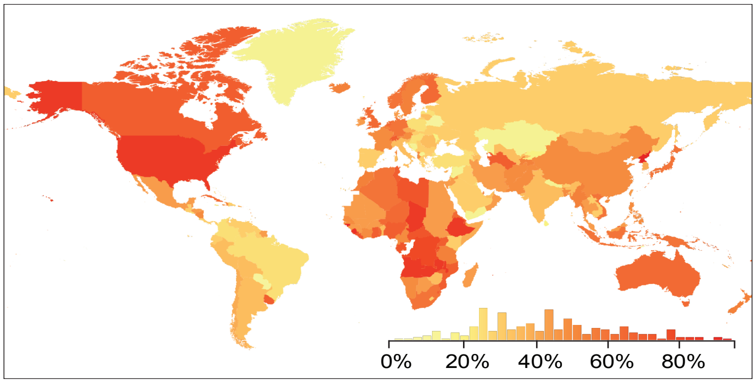


Fig. 9: Per-country percentage of unused space (*routed unused* + *unrouted assigned*) out of the assigned. The U.S. is red in this map due to a few very large allocations heavily unutilized, while some African countries are red because they use a very small fraction of their (also small) assigned space.

USA has 25% of the used /24s, but only 4.44% of the population. In contrast, African countries have only 1.8% of the used /24s, but 16% of the world population. The per-country used /24s correlate much better with the distribution of GDP (0.960 correlation), than with population (0.517 correlation), suggesting that economic inequalities could explain the differences in the used /24s. We can also observe disparities in the distribution of used and unused addresses: due to legacy allocations USA holds 49.8% of the *routed unused* and 67.5% *unrouted assigned* space, but 25% of the used space. The distribution of the *unrouted assigned* space is more uneven than of the *routed unused* space.

Table IX(b) lists the top ASes by *routed unused* /24s (we do not have per-AS data for *unrouted assigned* space). The top ASes are the Department of Defense (DoD) Network Information Center

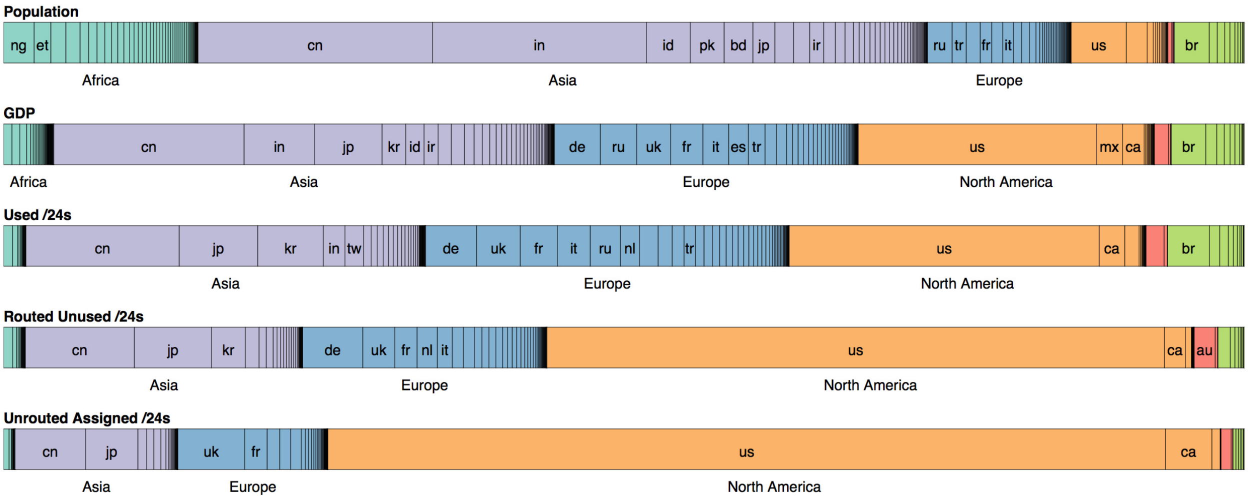


Fig. 10: Comparison of address space assigned to countries with per-country population and GDP. The width of a country (and continent) represents its relative size within a dataset. E.g., the top bar shows the percentage that each country contributes to the global population, with China (*cn*) contributing the most (1.36B, 18.9%). The correlation between datasets can be observed by comparing bars. We observe that there is not a strong correlation between population (top bar) and number of *used* /24 blocks of a country; in large part due to high usage by the USA. There is however, a strong correlation between the GDP (2nd from top) and number of used /24 blocks of a country (3rd bar). Not only does the USA dominate /24 block usage, it also represents a significant portion of both the *routed unused* and *unrouted assigned* bars, with 49.8% and 67.5% respectively. An interactive version of this visualization is available at [17].

(NIC), followed by Level 3, HP, China Telekom, and finally UUNET.

B. *Implications for other studies*

This type of census dataset also has implications for a range of scientific research of the Internet, most notably projects that incorporate routed address space metrics into estimates of the size, degree, type, or maliciousness of ASes [21], [44], [48], [49], [60]. More accurate metrics of address space usage could also potentially improve the accuracy of analysis of (or prediction of likely future) address blocks transfers in the grey market [47]. Figure 11 shows the overestimation error one would make by using a canonical *BGP-routed address space* metric to reflect how much address space an AS is actually observably using, for five types of network providers of various sizes. Median overestimation error generally increases with the size of the AS, perhaps due to large ASes under-utilizing their allocations. Large Enterprise ASes (>1k /24s) result in the most dramatic overestimation, with a median overestimation error of 96%. Figure 12 shows the overestimation error when using the same (*BGP-routed address space*) to

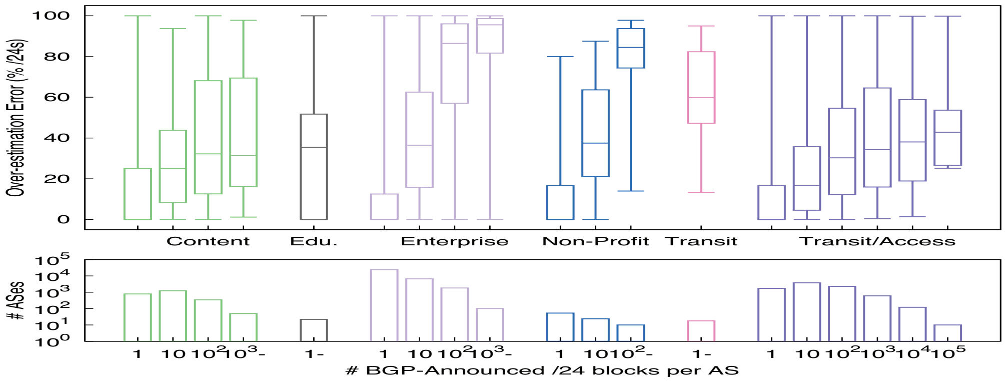


Fig. 11: Overestimation error (top graph) when using *routed address space* instead of our census as a rough metric for AS size. ASes are grouped according to the classification scheme proposed by Dhamdhere et al. [21] and sorted by number of routed /24 blocks (the x label indicates the minimum value in the bin). The bottom graph shows the number of ASes per bin. Median overestimation error generally increases with the size of the AS, perhaps due to large ASes under-utilizing their allocations.

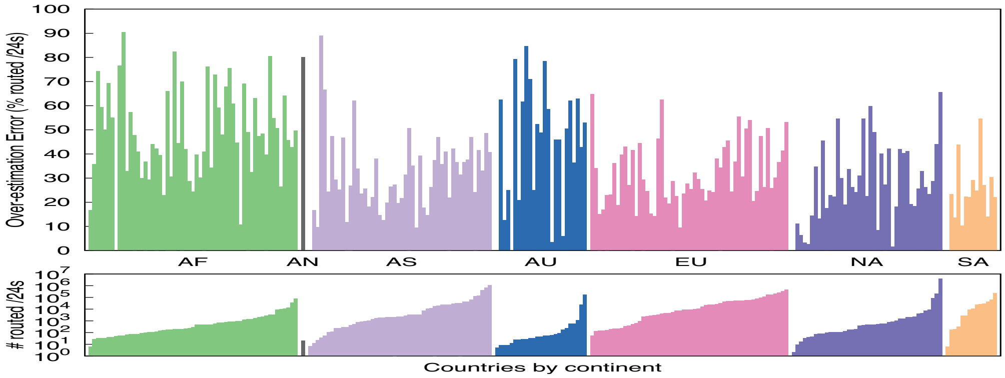


Fig. 12: Overestimation error when using *routed address space* instead of our census as a rough metric for a country's footprint of activity on the Internet. Countries are grouped by continent and sorted by number of routed /24 blocks (y value on bottom graph). The top graph shows the overestimation error for each country. As is also evident in Figure 9, activity in African countries would be significantly overestimated using *routed address space*. Most importantly, there is no significant correlation between the the number of per-country routed /24s and the resulting overestimation error.

reflect each country's Internet footprint. Both figures also show that there is no simple formula to translate between routed address space and actually used address space – the difference varies widely by AS and country, independently from the number of routed /24s.

Finally, as a concrete example of how switching from BGP data to our census dataset can impact research results, we show the case of CAIDA AS Rank [49]. CAIDA uses publicly available BGP data to infer business relationships among ASes and provides a ranking of ASes based on a measure of their role in the global Internet routing system. Ideally, this would be done by counting the amount of traffic transited across an individual AS, but organizations consider this information proprietary. Instead, CAIDA uses as a proxy the size of an AS customer cone, which is the number of /24 blocks that the AS can reach via its customers, i.e., by (recursively) crossing only customer links. To capture more complex peering relationships than those inferred from the simple provider/customer/peer model, CAIDA uses a slightly more restrictive definition of AS customer cone, which only includes blocks from the set of prefixes that the AS is observed announcing to its peers or providers [49].

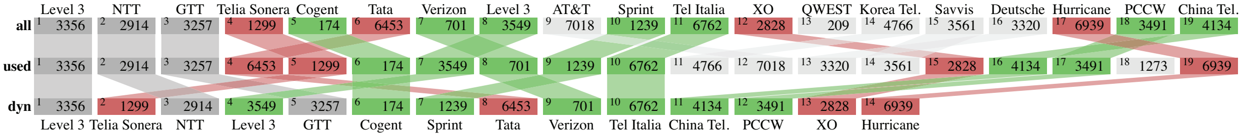


Fig. 13: ASes ranked by: the number of /24 blocks in their customer cone (**all**), *used* /24 blocks in their customer cone (**used**), and Dyn's transit addresses (**dyn**). AS color is dependant on if **used** or **all** rank was closer to **dyn**. *Green* means **used** was closer to **dyn**'s ranking. *Red* means **all** was closer to **dyn**'s ranking. *Dark grey* means they are at the same distance from **dyn**'s ranking. *Grey* means **dyn** provided no ranking. **Used** ranked 7 ASes closer to their **dyn** ranking, while **all** had only 4.

Recomputing the AS customer cones by filtering out /24 blocks that are *routed unused* (basically we consider only *used* blocks), we observe changes in the AS ranking order. Interestingly, we find (Figure 13) that using our census dataset, the resulting ranking makes CAIDA AS rank closer to Dyn's *IP Transit Intelligence AS ranking*, which apparently is based on transit traffic measured or inferred with their proprietary techniques [3]. We discuss this comparison in further detail in a CAIDA blog post [12].

VIII. Future Directions

In addition to the applications of census measurements that have been well articulated by [15], there are many possible future directions for this work. To enhance our methodology, we would like to further improve our ability to infer spoofed traffic and validate such inferences, perhaps by responding to darknet traffic. We would also like to investigate the use of UDP or other protocol traffic at R-ISP and IXP vantage points, and analyze in more detail what addresses are less visible to traffic measurement e.g., internal CDNs or quiet networks. As always, additional vantage points and ground truth information from operators would help improve the integrity of the method.

For a periodic global Internet census that tracks changes over time, we imagine a hybrid approach that first infers active IP address blocks based on passive measurements from one or more (live or dark) traffic vantage points, then probes only those address blocks that cannot be confidently inferred as in use. This approach could dramatically improve coverage over state of the art methods, while minimizing measurement overhead and potential irritation of network operators with aggressive firewalls. When performing a periodic census, our proposed taxonomy and definitions of coverage will help to quantify and track changes in space utilization over time. Finally, the unscalability of active scanning to the IPv6 address space was one motivation to explore our hybrid apporach, but we do not know how well distributed passive traffic observation alone could effectively support a future IPv6 census.

References

[1] 360 Total Security Software License and Service Agreement. www.360safe.com/totalsecurity/en/licence.html.

[2] UCSD Network Telescope. http://www.caida.org/projects/network_telescope/.

[3] Dyn IP Transit Intelligence. http://dyn.com/ip-transit-intelligence/trial/, 2015.

[4] PREDICT: Protected Repository for the Defense of Infrastructure Against Cyber Threats. http://predict.org, 2015.

[5] A. Dainotti, A. King. CAIDA Blog: Carna botnet scans confirmed. http://blog.caida.org/best_available_data/2013/05/13/carna-botnet-scans/.

[6] Advanced Network Technology Center, University of Oregon. Route Views Project. http://www.routeviews.org/.

[7] AFRINIC. AFRINIC delegation file. ftp://ftp.afrinic.net/pub/stats/afrinic/delegated-afrinic-extended-latest.

[8] B. Ager, N. Chatzis, A. Feldmann, N. Sarrar, S. Uhlig, and W. Willinger. Anatomy of a large european ixp. In *Proceedings of ACM SIGCOMM 2012*, SIGCOMM '12, pages 163–174, New York, NY, USA, 2012. ACM.

[9] L. Alt, R. Beverly, and A. Dainotti. Uncovering network tarpits with degreaser. In *Proceedings of the 30th Annual Computer Security Applications Conference*, ACSAC '14, pages 156–165, New York, NY, USA, 2014. ACM.

[10] APNIC. APNIC delegation file. http://ftp.apnic.net/stats/apnic/delegated-apnic-latest.

[11] ARIN. ARIN delegation file. http://ftp.arin.net/pub/stats/arin/delegated-arin-extended-latest.

[12] B. Huffaker. CAIDA Blog: Whats in a Ranking? comparing Dyns Bakers Dozen and CAIDAs AS Rank. http://blog.caida.org/best_available_data/2015/07/02/whats-in-a-ranking-comparing-dyns-bakers-dozen-and-caidas-as-rank/.

[13] G. Bartlett, J. Heidemann, and C. Papadopoulos. Understanding passive and active service discovery. In *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement*, IMC '07, pages 57–70, New York, NY, USA, 2007. ACM.

[14] K. Benson, A. Dainotti, k. claffy, A. Snoeren, and M. Kallitsis. Revisiting Internet Background Radiation for Opportunistic Network Analysis. In *Proceedings of the 2015 Conference on Internet Measurement Conference*, 2015.

[15] X. Cai and J. Heidemann. Understanding block-level address usage in the visible internet. In *Proceedings of the ACM SIGCOMM 2010 Conference*, SIGCOMM '10, pages 99–110, New York, NY, USA, 2010. ACM.

[16] CAIDA. Routeviews prefix to AS mappings dataset for IPv4 and IPv6. http://www.caida.org/data/routing/routeviews-prefix2as.xml.

[17] CAIDA. Supplemental data: Lost in Space: Improving Inference of IPv4 Address Space Utilization. http://www.caida.org/publications/papers/2014/lost_in_space/supplemental/, 2014.

[18] CIA. The world factbook: GDP (purchasing power parity). https://www.cia.gov/library/publications/the-world-factbook/rankorder/2001rank.html.

[19] CIA. The world factbook: Population. https://www.cia.gov/library/publications/the-world-factbook/rankorder/2119rank.html.

[20] A. Dainotti, K. Benson, A. King, k. claffy, M. Kallitsis, E. Glatz, and X. Dimitropoulos. Estimating Internet address space usage through passive measurements. In *ACM SIGCOMM Computer Communication Review (CCR)*, 2014.

[21] A. Dhamdhere and C. Dovrolis. Twelve Years in the Evolution of the Internet Ecosystem. *IEEE/ACM Transactions on Networking*, Oct. 2011.

[22] Digital Element. Netacuity. http://www.digital-element.net/ip_intelligence/ip_intelligence.html.

[23] Z. Durumeric, E. Wustrow, and J. A. Halderman. ZMap: Fast Internet-wide scanning and its security applications. In *Proceedings of the 22nd USENIX Security Symposium*, 2013.

[24] A. Finamore, M. Mellia, M. Meo, M. M. Munafø', and D. Rossi. Experiences of internet traffic monitoring with tstat. *IEEE Network*, 25(3), 2011.

[25] V. Gehlen, A. Finamore, M. Mellia, and M. M. Munafò. Uncovering the big players of the web. In *TMA*, TMA'12, pages 15–28, Berlin, Heidelberg, 2012. Springer-Verlag.

[26] Geoff Huston. IPv4 Address Report. http://www.potaroo.net/tools/ipv4/.

[27] Geoff Huston. Delegated address space: extended report, 2013. http://bgp.potaroo.net/stats/nro/archive/delegated-nro-extended-20131001.

[28] H.D. Moore. Project Sonar), 2008. https://community.rapid7.com/community/infosec/sonar/blog.

[29] J. Heidemann, Y. Pradkin, R. Govindan, C. Papadopoulos, G. Bartlett, and J. Bannister. Census and survey of the visible Internet. In *8th ACM SIGCOMM conference on Internet measurement*, IMC '08, 2008.

[30] J. Horchert and C. Stöcker. Mapping the internet: A hacker's secret internet census. Spiegel Online, March 2013.

[31] G. Huston. Ipv4: How long do we have? *The Internet Protocol Journal*, 6(4):2008–2010, 2003.

[32] G. Huston. Ipv4 address depletion and transition to ipv6. *The Internet Protocol Journal*, 9(10):18–28, 2007.

[33] G. Huston. The changing foundation of the internet: confronting ipv4 address exhaustion. *The Internet Protocol Journal*, 11(3):19–36, 2008.

[34] Y. Hyun, B. Huffaker, D. Andersen, M. Luckie, and K. C. Claffy. The IPv4 Routed /24 Topology Dataset, 2014. http://www.caida.org/data/active/ipv4_routed_24_topology_dataset.xml.

[35] IANA. Autonomous system (AS) numbers. http://www.iana.org/assignments/as-numbers/as-numbers.xml.

[36] IANA. Ipv4 address space registry. https://www.iana.org/assignments/ipv4-address-space/.

[37] IANA. IPv4 recovered address space. http://www.iana.org/assignments/ipv4-recovered-address-space/ipv4-recovered-address-space.xml.

[38] IANA. IPv4 special-purpose address registry. http://www.iana.org/assignments/iana-ipv4-special-registry/iana-ipv4-special-registry.xml.

[39] IANA. IPv6 global unicast address assignments. http://www.iana.org/assignments/ipv6-unicast-address-assignments/ipv6-unicast-address-assignments.xhtml.

[40] IANA. IPv6 special-purpose address registry. http://www.iana.org/assignments/iana-ipv6-special-registry/iana-ipv6-special-registry.xhtml.

[41] Information of Sciences Institute, University of Southern California. LANDER project:Internet address census it49c-20120731. http://www.isi.edu/ant/traces/internet_address_census_it55w-20130723.README.txt, 2012.

[42] Information of Sciences Institute, USC. Internet Address Survey Binary Format. http://www.isi.edu/ant/traces/topology/address_surveys/binformat_description.html, 2012.

[43] Internet Assigned Numbers Authority (IANA). IPv4 Address Space Registry. http://www.iana.org/assignments/ipv4-address-space/ipv4-address-space.xhtml.

[44] M. Konte and N. Feamster. Re-wiring activity of malicious networks. In *Proceedings of the 13th International Conference on Passive and Active Measurement*, PAM'12, pages 116–125, Berlin, Heidelberg, 2012. Springer-Verlag.

[45] LACNIC. LACNIC delegation file. ftp://ftp.lacnic.net/pub/stats/lacnic/delegated-lacnic-latest.

[46] J. Liang, N. Naoumov, and K. W. Ross. The index poisoning attack in p2p file sharing systems. In *INFOCOM*. IEEE, 2006.

[47] I. Livadariu, A. Elmokashfi, A. Dhamdhere, and K. Claffy. A first look at ipv4 transfer markets. In D. Papagiannaki and V. Misra, editors, *CoNEXT 2013*, pages 7–12. ACM SIGCOMM, December 2013.

[48] A. Lodhi, N. Larson, A. Dhamdhere, C. Dovrolis, and k. claffy. Using PeeringDB to Understand the Peering Ecosystem. *ACM SIGCOMM Computer Communication Review (CCR)*, 44(2):21–27, Apr 2014.

[49] M. Luckie, B. Huffaker, A. Dhamdhere, V. Giotsas, and k claffy. AS relationships, customer cones, and validation. In *IMC*, Oct. 2013.

[50] M. Luckie, Y. Hyun, and B. Huffaker. Traceroute Probe Method and Forward IP Path Inference. In *Internet Measurement Conference (IMC)*, pages 311–324, Vouliagmeni, Greece, Oct 2008. Internet Measurement Conference (IMC).

[51] M. Cotton and Leo Vegoda. RFC 5735: Special Use IPv4 Addresses, 2010. https://tools.ietf.org/html/rfc5735.

[52] X. Meng, Z. Xu, B. Zhang, G. Huston, S. Lu, and L. Zhang. Ipv4 address allocation and the bgp routing table evolution. *SIGCOMM Comput. Commun. Rev.*, 35(1):71–80, Jan. 2005.

[53] R. Munroe. xkcd: MAP of the INTERNET 2006. http://blog.xkcd.com/2006/12/11/the-map-of-the-internet/.

[54] D. Plonka and A. Berger. Temporal and Spatial Classification of Active IPv6 Addresses. In *Proceedings of the 2015 Conference on Internet Measurement Conference*, 2015.

[55] M. Prince. The DDoS That Almost Broke the Internet. blog.cloudflare.com/the-ddos-that-almost-broke-the-internet, March 2013.

[56] RIPE NCC. RIPE NCC delegation file. ftp://ftp.ripe.net/ripe/stats/delegated-ripencc-latest.

[57] RIPE NCC. Routing Information Service (RIS), 2008. http://www.ripe.net/ris/.

[58] A. N. Shannon V. Spires. Exhaustive search system and method using space-filling curves. Patent, 10 2003. US 6636847.

[59] SWITCH. Swiss Tele Communication System for Higher Education. http://www.switch.ch/.

[60] H. Tangmunarunkit, J. Doyle, R. Govindan, W. Willinger, S. Jamin, and S. Shenker. Does as size determine degree in as topology? *SIGCOMM Comput. Commun. Rev.*, 31(5):7–8, Oct. 2001.

[61] The Bro Project. TCP Scan detection. bro.icir.org/sphinx/scripts/policy/misc/scan.html, 2014.

[62] R. Wilhelm. RIPE Labs Blog: How to define address space as 'routed'? https://labs.ripe.net/Members/wilhelm/content-how-define-address-space-routed.

[63] S. Zander, L. L. Andrew, and G. Armitage. Capturing Ghosts: Predicting the Used IPv4 Space by Inferring Unobserved Addresses. In *Proceedings of the 2014 Conference on Internet Measurement Conference*, 2014.