Share or Not: Investigating the Presence of Large-Scale Address Sharing in the Internet

Sebastian Zander, David Murray
School of Engineering and Information Technology
Murdoch University
{s.zander, d.murray}@murdoch.edu.au

***Abstract*—Network Address Translation (NAT) allows multiple devices with private addresses to share one public address. NAT was mainly confined to home gateways, but with the exhaustion of the IPv4 address space, large-scale NATs have been deployed. Other technologies causing large-scale address sharing are on the rise as well (e.g. VPNs). Large-scale address sharing is problematic, since it limits the number of concurrent TCP connections and severely limits geolocation and geoblocking. We investigate the presence of large-scale address sharing in the Internet, including how frequently it occurs, in which types of organisations it occurs, where it occurs geographically, how many users share addresses, and whether its presence is linked to IPv4 address shortage. Our results show that there are thousands of addresses with significant large-scale sharing with up to a few thousand users sharing a single address. Most of this sharing occurs within ISPs, many of which are located in countries with IPv4 address shortage, indicating that large-scale NATs may be a consequence of IPv4 shortages.**

***Index Terms*—IP Address Sharing, Network Address Translation (NAT), Carrier Grade NAT, Internet Measurement**

I. Introduction

In the early days of the Internet, each connected device had a globally unique IP address. However, with the looming shortage of IPv4 addresses in the 1990s, Network Address Translation (NAT) was invented to slow down the allocation of IPv4 addresses. NAT allows multiple devices with private addresses to share one public IP address. Traditionally, NAT was largely confined to home gateways. However, with the advent of a completely exhausted IPv4 address space [1], some organisations have deployed large scale NATs or carrier-grade NATs (CGNATs). Furthermore, other technologies that cause address sharing have also become more prominent in recent years, such as Virtual Private Networks (VPNs) or anonymisation networks (e.g. Tor network [2]).

Address sharing poses a number of problems. Firstly, if protocols above the network layer encode IP addresses, the gateway or tunnel endpoint must also translate these addresses in higher-layer protocols. For example, the FTP protocol requires NAT helpers to do the translation of IP addresses. Secondly, to initiate a connection to a server or peer behind NAT, port forwarding must be set up. In some cases, protocols have been redesigned to solve this problem, such as FTP passive mode [3]. In other cases NAT helper protocols, such as Universal Plug and Play (UPnP), are needed to communicate through NAT. Address translation also poses a challenge for newer or more complex protocols. The Stream Control Transfer Protocol (SCTP) provides a stronger checksum, multi-homing and chunk multiplexing but its operation through NAT requires complex workarounds, which are also fragile [4]. In other cases NAT may cause problems for remote workers connecting through IPSec VPNs [5].

Large-scale address sharing resulting from CGNATs causes further problems. With many different homes/users mapped to one external address, the number of TCP sessions per user is limited, IP-geolocation is less accurate or even useless, and there are issues for law enforcement. A CGNAT also introduces a single point of failure which may impact reliability [6]. Furthermore, services, which use IP bans to prevent spam or brute force attacks are presented with an ultimatum: block the offending IP address, including the many legitimate users who share the same IP, or find other mechanisms to identify and block only the offending user. A single misbehaving user can block a service to a wide range of users.

Large scale address sharing is clearly problematic, but the current extent of large-scale address sharing is unknown. While some recent work on mechanisms to detect NAT exist [7], [8], to our best knowledge the only previous work that measured the usage of NAT on the Internet dates back to the early 2000s [9], [10]. Our study fills this gap and provides some answers with respect to large-scale address sharing:

- How much of it is present in the Internet?
- How large is it, i.e. how many users share addresses?
- In which geographical regions is it common?
- What organisations / industry sectors make use of it?
- Is it correlated with IPv4 shortage or IPv6 adoption?
- Are there any trends, i.e. is it on the rise or not?

Based on anonymous user session logs from a large Internet gaming platform with more than 40 million users world-wide (Valve's Steam), we investigate the presence of large-scale address sharing. The logs allow the identification of different users behind a single IP address at any point in time and thus to detect large-scale address sharing. While our paper is based on this particular dataset, our approach is more general and could be used with other session logs. We also developed a technique to map IP addresses to organisations, allowing us to not only investigate sharing based on IP addresses, but also based on organisations to whom these addresses belong.

The key findings of our study are that there are thousands of addresses with significant large-scale sharing with up to

DOI 10.1109/LCN.2017.85

a few thousand users sharing a single address. Most sharing occurs within ISPs, many of which are located in countries with IPv4 address shortages, indicating that large-scale NATs may be a consequence of IPv4 shortages. This results of this research may be particularly useful to inform research activities investigating the IPv6 transition, IPv4 address markets, or those needing to quantify the extent of NAT traversal issues.

Section II reviews technologies that lead to address sharing and previous work on measuring address sharing in the Internet, mainly focussed on detecting NAT. Section III describes our dataset and explains how we detect address sharing. Section IV presents the results of our analysis. Section V discusses limitations and future work. Section VI concludes the paper.

II. Related Work

We briefly discuss technologies that cause large-scale address sharing and then review previous studies on NAT detection techniques and measurement of NAT use in the Internet.

A. Address sharing

Address sharing is caused by several different technologies: NATs, VPNs, anonymisation networks, and other proxies.

While IPv6 is the long-term solution, using only IPv6 is infeasible in the short to medium term with many Internet services currently inaccessible via IPv6. CGNATs will be used to provide IPv4 access even when an IPv6 address is available. The following CGNAT technologies are already standardised and deployed: NAT444, NAT A+P, DS-Lite and NAT64. Even the most progressive of these transition technologies will result in large IPv4 address sharing.

Internet access is frequently restricted by organisations to prevent access to non-work sites or by nation states to censor content or impose sanctions. Also, many content providers use geoblocking to enforce price discrimination in different locales, and in some countries network monitoring is routine. VPNs are commonly used to circumvent restrictions, monitoring, or geoblocking and also manifest as large-scale address sharing since users share tunnel endpoints.

Anonymisation networks, such as Tor [2], allow anonymous access to Internet services. A service can only observe that traffic is coming from an exit node of the anonymisation network but cannot identify its true origin. As there are more users than exit nodes, anonymisation networks also cause large-scale address sharing.

Other types of services also manifest as large-scale address sharing. One example, specific to our dataset, is the use of Amazon EC2 infrastructure for cloud gaming [11]. Users can run high-end games on Amazon EC2 and stream the game to low-end clients, such as laptops. Since Amazon EC2 uses address sharing between different virtual machines, we observe large-scale address sharing for Amazon in our dataset.

B. Prior measurement of address reuse

Detecting and measuring the presence of address sharing has several uses. Detecting the number of hosts sharing a single IP address may provide a better estimate of the number of hosts on the Internet [10]. The uptake and use of NAT is also an important Internet statistic [9]. Several approaches have been used to detect NATs.

Armitage [9] set up Quake 3 game servers in different locations and observed players from around the world requesting game information from the server – Quake 3 clients query every online server for game information [12]. By default the Quake 3 client uses UDP source port 27960; any client creating a connection using a different source port was assumed to be transiting through a NAT. Armitage concluded that, in 2002, NAT was on approximately 17–25% of public/private internet access boundaries. The study underestimates the true number of NATs, as the approach cannot detect port-preserving NATs with sole Quake 3 players behind them.

Bellovin [10] noted that many operating systems (OSs) used the "Identification" field in the IP header (IP ID) as a simple sequential counter. Without NAT there is a string of consecutive IP IDs from a single IP address. With multiple machines behind a NAT there are multiple streams of IP IDs in different parts of the 16-bit number range, since the IP IDs of different machines are not synchronised. Suitable processing makes it possible to determine the number of machines behind the NAT. Unfortunately, the IP ID is not sequential for all OSs and thus there is a degree of inaccuracy to this method. Furthermore, this approach requires packet traces making it infeasible for an Internet-wide study.

Kohno et al. [13] proposed an approach for the remote fingerprinting of physical devices. The technique exploits minute deviations in the hardware of devices: clock skews. Clock skews can be measured remotely based on obtaining samples of devices' clocks. Kohno's technique can be used to count the number of hosts behind NATs, even if the hosts use random or constant IP IDs to counter the approach proposed in [10]. However, the fact that the method requires series of clock samples (per packet data) means this approach is practically infeasible to use for an Internet-wide study.

Maier et al. [7] collected packet data from more than 20,000 DSL lines. They analysed IP Time-to-live (TTL) values and HTTP user-agent strings. The initial TTL varies with different OSs and can potentially detect NATs with different OSs behind them. The HTTP user-agent string is sent by a user's web browser to the web server. It provides details of the computer accessing the website, such as OS, browser name and version. By combining the TTL and user-agent string, Maier's approach can detect different computers behind a NAT, although it may be inaccurate for office environments, where the OS and web browser is standardised. Similar to [10], this method requires packet traces.

Komarek [8] describes a technique based on the statistical analysis of HTTP logs and user-agent strings. Unlike previous studies, their goal was to determine whether a host is connected directly or through NAT. This approach is best used to detect unauthorised NATs, installed by users to extend the network or provide access to unauthorised devices. Komarek's approach requires data collection from the server or access to

packet payload. As packet payloads are increasingly encrypted, access to the browser's user-agent string will be impossible in many scenarios and limit the analysis to server-side logs.

Previous approaches focused on NAT and may, apart from [9] and [8], require packet data, complicating large-scale studies. Armitage's approach using Quake 3 is outdated and newer game clients only query a regional subset of servers, impeding Internet-wide measurement.

Our study is broader than prior work and investigates address sharing in general, including VPNs and proxies. For content providers, VPNs and proxies cause all the problems of NATs with a number of additional problems, such as the inability to link a user to a country. Secondly, our study is global, since it is based on a dataset from a popular international gaming platform. Thirdly, our study does not rely on a specific protocol or application features that will be obsolete in the near future. Even if the gaming platform would loose popularity, our approach could be applied to data of other services that track user sessions.

III. Methodology

In this section we describe our datasets, the preprocessing of the datasets, how we identify large-scale address sharing, how we map shared-addresses to organisations and how we map organisations to industry types.

A. Datasets and preprocessing

Our data is based on session data from Steam [14], one of the largest gaming platforms in the Internet. Each dataset contains a list of (anonymised unique player ID, IPv4 address[1], login time, logout time) tuples. When a user logs into the PC, the Steam client will automatically start and perform a login. When a users logs out or shuts down the PC, the Steam client will perform a logout. If the Steam client is restarted due to an update or a client looses its Internet connection, this will also trigger login and logout events. Effectively, users have Steam sessions while they are logged into their computers and not just while playing.

The session data allows us to identify the times a particular player was logged into the system and from which public IP address the player was logged in. In the common case of a player behind a NAT router, this is the IP address of the NAT router. Note that when a Steam client does not terminate gracefully, the server will time out the session relatively quickly (in 60 seconds or less). Due to the way the data is collected on Steam, some preprocessing is required to remove part of the records from which we cannot determine the duration of a session.

Several datasets, each spanning a whole week, from 2015, 2016 and 2017 have been obtained. For 2016, two datasets from adjacent months provide insight into short term 'noise' (useful for delineating noise and long-term trends). The details of the datasets *after* preprocessing are shown in Table I.

[1]Steam does not support IPv6 yet.

TABLE I
Collected datasets (after preprocessing).

Dataset	D1	D2	D3	D4
Time period	1/10/2015–7/10/2015	1/4/2016–7/04/2016	1/5/2016–7/5/2016	20/2/2017–27/2/2017
Sessions	385.5 M	394.2 M	389.0 M	503.6 M
Unique IPs	53.2 M	58.5 M	57.8 M	67.3 M
Unique /24	3.29 M	3.47 M	3.46 M	3.62 M
Unique users	34.5 M	39.9 M	39.7 M	48.0 M

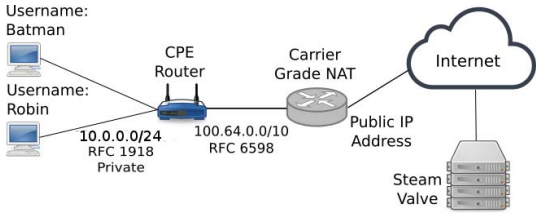

Fig. 1. Address sharing detection based on logins to Valve's Steam gaming platform.

B. Detecting NATs

The mechanism to detect address sharing is simple. On a single PC, there can only be one Steam client logged in at any point in time. Hence, if multiple different users are logged in simultaneously from the same IP address (logged by Steam), then these users shared one IP address. A simplified topology of the approach is shown in Figure 1. Based on the session data we then compute the following metrics:

- Maximum number of concurrent users: We count the maximum number of unique users that were observed behind the same IP address concurrently, which provides a lower bound on how many users may have shared the single IP address.
- Total number of users: We count the total number of unique anonymised user IDs seen for each shared IP address over the whole period each dataset covers.

Our metrics may include false positives, i.e. the number of users we observed may be higher than the actual number of users, due to Steam clients that terminated ungracefully (with recorded session end times after users effectively logged out). However, given that ungraceful terminations are the exception rather than the norm (based on Steam data) and the relatively short session timeout of 60 seconds or less, we think this error is very small. Also, this error is most likely dwarfed by the effect of not observing all Internet users but only the users that use Steam.

For all addresses identified as shared addresses we also identify the responsible organisation, the Regional Internet Registrar (RIR) that allocated the IP space, and the country of the organisation using whois and GeoIP data. Due to limitations of our whois database we only have whois data for approximately 99% of IP addresses. We also use reverse

DNS lookups to obtain the domain name for each shared IP (if a name exists). For performance reasons we only performed one reverse DNS lookup per /24 subnet and assumed that the domain name is the same for all IP addresses in the same /24.[2] Based on this information we aggregate data for all shared IPs of one organisation as described in Section III-D.

C. Country mapping

To determine the country of origin of an IP address, each address was mapped to a country using the free GeoIP GeoLite2 database [15]. We do not use the country code from the whois data since 1) some whois entries use the EU country code and cannot be associated to a single country and 2) coverage for GeoIP was better. Previous research demonstrated that IP geolocation on country-level is fairly reliable [16], [17]. However, to assess the consistency and accuracy of the data, we also compared the GeoIP country codes with whois data country codes. In 99% of the cases both country codes are consistent.

We also map anonymised user IDs to countries based on the IP addresses from which the users were logged in. For most users the country code is static, but for 1% of users their sessions originated from different country codes. In the latter case, we map the user to one of the country codes randomly. Further analysis of user mobility is left for future work.

D. Organisation mapping

One of our goals is to identify the organisations with the largest address sharing. To do that we need to merge entries for different IPs if they belong to the same organisation. The merging is based on domain names obtained from reverse DNS (rDNS) lookups on the IP addresses as well as on the whois data for the Autonomous System (AS) each IP belongs to. Our algorithm uses the following rules:

- Two entries are classified as belonging to the same organisation if (1) the last two parts of the rDNS domain name are identical, (2) the second last part is at least four characters long (to prevent gov.au and com.au from matching) and (3) the last part is not equal to "arpa" (to prevent default entries from matching).
- Two entries are classified as belonging to the same organisation if the combined string of AS name, AS description, AS address and domain part of the AS contact email address pass a similarity test.[3]
- If the AS numbers of two entries are the same, these are not automatically classified as belonging to the same organisation, since they could be two smaller organisations (potentially with private AS numbers) behind one public AS (e.g. ISP). However, if the two AS number are the same, the probably that these two entries belong to the one organisation is higher and to account for that the threshold used for the similarity test is lowered.

[2]This may not always be true, but since all RIRs allocate minimum address block sizes of /24 or larger, all addresses in one /24 network likely belong to the same organisation (although some may be leased to customers).

[3]Note that we only use the first 25 characters of the AS description and the first 15 characters of the AS address. This is to increase performance and to prevent that a very long AS address dominates the string matching.

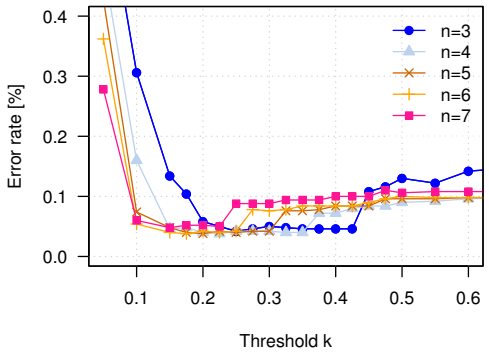

Fig. 2. Dimensioning of parameters k and n for string similarity test.

As string similarity test we use n-gram string comparison [18]. Two strings pass the test if the similarity value exceeds a threshold k given an n-gram size of n. To determine how to dimension k and n we created a dataset of 500 randomly sampled entries from the top 50,000 IP addresses with the largest sharing (from D1). We then classified and merged the entries by hand to create ground truth. Next, we ran our aggregation algorithm varying k and n and compared the result for each (k, n) against the ground truth. In each test, each IP address that is not merged with the correct set counts as error. Figure 2 plots the error in percent for varying k and n (zoomed in on the region of the minimum error). We varied k in the range 0.05–0.95 and n in the range 1–9, but only show part of the range in the plot ($n \in [1, 2]$ gives very bad results for smaller k as expected, since $n = 3$ is the default setting).

From Figure 2 it is clear that there is a sweet spot for $k = [0.15, 0.25]$ and $n = [4, 6]$ where the error rate is lowest (4%). The fact that this sweet spot is relatively large means our approach is not overly sensitive to small changes of k and n. Hence, we selected the parameters at the centre of this area ($k = 0.2$ and $n = 5$) as parameters for the aggregation. In case of equal AS numbers, we set $k = 0.1$ (we found this worked best in initial tests). We also manually inspected the aggregation for the top 1000 IP addresses with the most simultaneous users behind them for D1. The top-1000 IP entries are aggregated into 152 organisations. Of the aggregations all but one looked correct (error of approx. 0.1%).

During the aggregation we compute the set of Steam users for the aggregated shared IP addresses and from that we get the number of users behind shared IPs of the organisation. We also count the total number of shared IPs per organisation.

E. Organisation type classification

The types of organisations, performing large-scale sharing, are also of interest. Based on the whois data, we used text analysis to classify organisations into Military, Education, Government, Company, ISP and Cloud/Hosting. An Unknown class captures organisations the classifier cannot classify. The

keywords that were used, based on the whois data, were carefully tuned and were slowly expanded from a couple of hundred obvious entries, such as ‘apple.com’ or ‘at&t’, to the final size of 1470 keywords.

The whois data uses an English language character set, but our keywords also include many non-English words. For example, words such as ‘ecole’ and ‘universitaet’ matched against Education while ‘bundesministerium’ and ‘ministerstwa’ matched against Government. Probable misspellings such as ‘ministery’ were included and all text was converted to lower case for comparison.

The program which analysed the dataset wrote entries receiving no matches to a separate file. This file was then fed into text analysis software which ranked the most common 1, 2, 3, 4 and 5 word combinations. This allowed us to add strings such as ‘Cosmopolitan of Las Vegas’ where the individual words in isolation would have been too generic to be meaningful. To further improve the accuracy of our classification, we wanted to reduce the number of entries matching multiple categories. Every entry that matched in two or more categories was written to a separate file. Text analysis of this file identified keywords, such as ‘.com’ and ‘communication’, that were too generic.

By iterating between removing generic words, such as ‘.com’ and ‘gmbh’, and inserting more specific words, such as ‘ford.com’ or ‘htp gmbh’, the accuracy of our classification was improved. Throughout the process, we were mindful that misclassification was more harmful than failure to classify (Unknown). A final validation was performed by manually checking two hundred random samples from the 1,000,000 IP addresses with most users behind them (for D1). Only a few (2%) ambiguities were noted. Approximately 20% of the whois entries could not be classified by our algorithm.

F. Geographical coverage

Our datasets cover users and IPs from over 220 country codes. For a number of very small countries (many island states or independent territories) the number of users is very low, but for about 135 countries (varies slightly with dataset) we have at least 1000 active users. The top countries with the most users are: USA, Russia, China, Germany, Great Britain, Philippines, Brazil, France, Canada, Poland, Turkey, Indonesia and Australia (quite consistent across all datasets).

Figure 3 shows the number of users and IP addresses per RIR for D4 (graphs for D1–D3 are similar qualitatively). Nearly half of the users are from Europe (RIPE NCC), but the number of users observed from Asia (APNIC), North America (ARIN) and South America (LACNIC) are also sizable – about 12 million, 9 million, and 4 million users respectively. The number of users for Africa (AfriNIC) is about 250,000.

IV. Results

First, we look at how many IP addresses are shared and how many users share the top-shared addresses. Next, we investigate the fraction of organisations with sharing and for which types of organisations we see large-scale sharing. Then we look at geographical differences and differences over time.

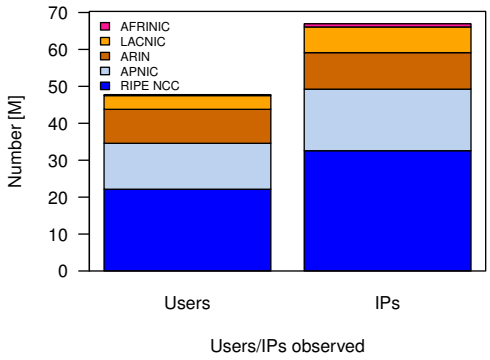

Fig. 3. Number of users and IP addresses per RIR.

Finally, we investigate whether sharing is correlated with IPv4 shortage or IPv6 uptake.

A. Largest address sharing

Figure 4 and 5 show the largest 5,000 and 50,000 cases of address sharing that occurred in the four datasets. The largest, top 100, cases of address sharing each consisted of hundreds to thousand of simultaneous users. Unless these NATs, gateways or proxies were set up for the sole purpose of gaming through Steam, they may have thousands or tens of thousands of actual users. While the number of users decreases relatively quickly, there are tens of thousands of IP addresses that were shared by at least 10 users. All the distributions have long tails and the total number of addresses with some sharing (at least two users) is 1–1.1 million depending on the dataset.

The distributions for the different datasets are broadly similar, but there are some notable differences which relate to notable plateaus present in all four datasets. An investigation into the owner of the related IP addresses reveals that they are almost exclusively Amazon EC2 IP addresses. As shown in Figure 5, D4 contains over 10,000 cases where 40 simultaneous steam users shared a single IP address.

We wanted to know whether this was a single user group that had cycled through numerous addresses over the week or whether these were all separate users. We found that there were many users who logged in to multiple IP addresses over the week but that many users appeared only once. It is difficult to determine the exact reason for this but a few scenarios are plausible. One possibility is that Amazon EC2 instances were used for VPN services, perhaps to circumvent state based restrictions. Another possibility is that users ran games on Amazon EC2 and streamed them to their PCs [11].

B. Organisations with large-scale address sharing

Section IV-A demonstrates some pronounced differences between organisations in terms of the number of IP addresses shared and this may cause a bias when comparing different countries, e.g. US companies have generally more IPv4 addresses so they may have larger pools of shared addresses with smaller share ratios whereas in countries with fewer IPv4 addresses the pools may be smaller but have higher share

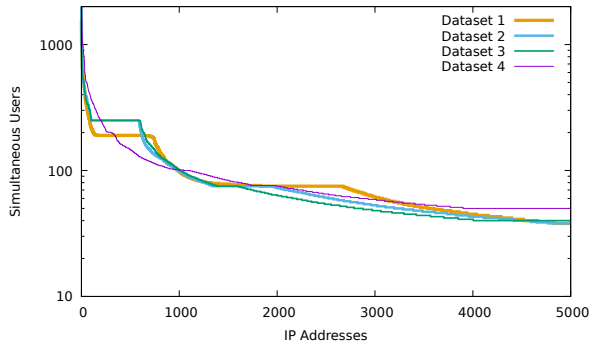

Fig. 4. Top 5,000 IP addresses with the largest sharing (log y-axis).

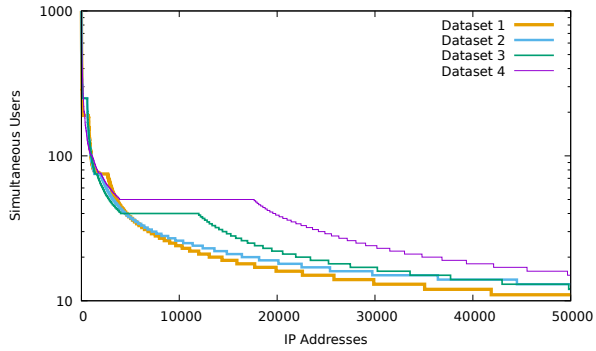

Fig. 5. Top 50,000 IP addresses with the largest sharing (log y-axis).

ratios. Also, the distribution of shared IP addresses has a very long tail due to lots of small NATs, such as home NATs.

To mitigate the effect of IP allocation size we now look at organisations. Table II shows the mean and coefficient of variation (Cvar) of the proportion (both as percent) of organisations with sharing across different countries for the different datasets (D1–D4). Each row shows the mean and coefficient of variation for organisations with sharing where at least t users shared at least one IP address.

With increasing t the mean reduces (as expected) but at the same time the coefficient of variation increases. When considering all organisations with sharing, the fraction of organisations with sharing is generally high and while there are differences between different countries, they are relatively small. We think this is because this case includes smaller NATs, such as NATs in homes, student dorms etc. which exist in all Telcos/ISPs in all countries. However, when only considering organisations with large-scale address sharing there is a lot more variation between countries (as indicated by the coefficient of variation). We also found that there is a larger jump of the coefficient around $t = 10$. Hence, in the rest of the analysis we use $t = 10$ to focus on large sharing.

Since we only have four data points over time and there is considerable noise (as indicated when comparing D2 and D3), we need to be cautious when investigating the trend over time.

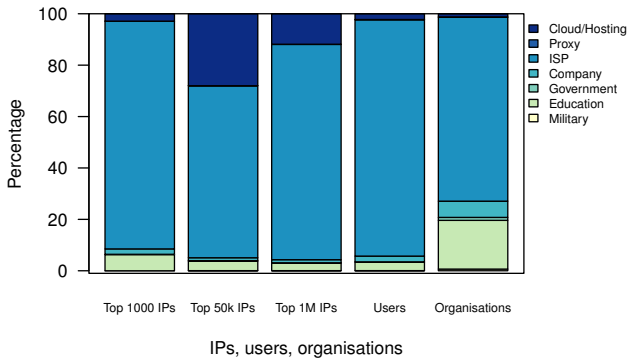

Fig. 6. Organisation types with larger address sharing (10 or more users).

Nevertheless, the mean for D4 is consistently higher than the mean for D1 for each t, indicating an upward trend.

Figure 6 shows the organisation type distribution for IP addresses and organisations with large-scale address sharing for D4 (graphs for other datasets are similar). The left-most three bars show the distributions for the top 1000, 10,000 and 1 million IP addresses with the highest sharing for organisations with large-scale sharing (at least 10 users). The following two bars show the distributions based on the number of users and organisations, again only those with large-scale sharing. To have more resolution for the different types each bar in the figure is normalised on the organisations we can classify. For the top IP addresses we cannot classify 20–25% of addresses, for users the percentage that cannot be classified is negligible and for organisations we cannot classify 50% of organisations.

Overall, the largest identifiable fraction is ISPs in every case. As Steam is a gaming platform, it is intuitive that the majority of address sharing is from home users connected to ISPs, with relatively little from military, government and companies. ISPs also have large numbers of addresses assigned to home users, but in terms of addresses there is also a sizable fraction of Cloud/Hosting providers. While Education is small in terms of IP addresses or users, it is sizable as percentage of organisations. This is plausible as there is often gaming in education networks, e.g. student dorms are often part of University networks, but possibly exaggerated due to the fact that our classifier is better in identifying these than some other types of organisations (often they are in .edu). As stated in the previous section, Amazon EC2 was responsible for a significant percentage of the largest NATs and it dominates the fraction of cloud organisations.

While we tried to detect VPNs or privacy proxies in the data set, many of these services are not identifiable from the whois data as anonymity is the primary goal of these services. We suspect that they are a larger contributor to address sharing than Figure 6 suggests.

C. Sharing depending on time and location

Figure 7 shows the fraction of IP addresses with large-scale sharing (at least 10 users), the fraction of unique users behind

TABLE II
AVERAGE PROPORTION OF ORGANISATIONS WITH SHARING AND VARIANCE ACROSS DIFFERENT COUNTRIES.

Min. Users t	Mean D1	Cvar D1	Mean D2	Cvar D2	Mean D3	Cvar D3	Mean D4	Cvar D4
2	59.6	25.7	63.6	16.3	64.8	30.1	62.7	32.7
5	43.1	46.0	47.4	40.6	48.2	41.4	50.9	38.2
10	21.3	61.5	25.9	55.7	29.7	59.9	33.0	60.3
30	7.3	88.6	8.6	92.5	12.3	83.5	14.4	94.9

these shared addresses[4] and the fraction of organisations for which we observe large-scale sharing for the different RIRs. For IPs the fraction is highest in Asia (APNIC), followed by South America (LACNIC) and Europe (RIPE NCC). Africa (AfriNIC) and North America (ARIN) show the lowest percentage. The picture is similar for users, although in this case Africa has a relatively higher fraction. For organisations the picture is more even, but the fraction is still the highest for South America and Europe. The three regions with the highest fractions also experience the highest shortage of IPv4 address space [19], which could mean that NAT deployment due to IPv4 address shortage is a major factor (this is consistent with recent work indicating the presence of large NATs in parts of the Internet [20]). A longitudinal comparison suggests a trend showing that the fraction of organisations with large scale address sharing is increasing. Furthermore, in the two largest regions with IPv4 shortages (Asia and Europe) the fraction of IPs shared has also consistently increased.

Figure 8 shows the 14 countries with the highest (Top-14) and lowest (Bottom-14) fraction of organisations with large-scale sharing for D4 considering only countries which, based on our data, have at least 50 organisations. This creates a bias towards excluding South American and African countries, but data for small countries is very noisy and hard to interpret.

The figure shows that the Top-14 is dominated by countries from Eastern Europe or Asia, such as Russia and China, which have relatively small address IPv4 address spaces relative to the number of citizens (note that India is in the 16th spot from the top). It also includes Denmark, which has a very small IPv6 uptake compared to most Northern European countries [21]. The Bottom-14 countries are largely wealthy developed countries, which have relatively large IPv4 allocations, such as the USA and Germany. Romania and Bulgaria stand out in the Bottom-14. However, unbeknownst to many people Romania has a relatively large IPv6 uptake [21] and thus potentially needs fewer NATs.

D. Sharing vs. IPv4 shortage and IPv6 uptake

This section investigates whether the fraction of organisations with address sharing is correlated to the shortage of IPv4 addresses or the uptake of IPv6 using D4. Again, we only consider countries with at least 10 organisations. As measure for correlation we use Spearman's rank correlation (ρ) and we use hypothesis testing to test for statistical significance.

Figure 9 shows a scatter plot of the percentage of organisations with address sharing versus the IPv6 capability of each country. To provide a measurement of IPv6 capability we use the IPv6-capable metric from APNIC [21], i.e. percentage of hosts that are IPv6-capable, based on data from February 15, 2017 (results for APNIC's IPv6-preferred metric are very similar and not shown here). The plot shows that there is a slight tendency for countries with higher sharing percentage to have lower IPv6 capability. Spearman's $\rho = -0.11$ but there is no significant correlation at 95% significance level ($p = 0.27$).

Figure 10 shows a scatter plot of the percentage of organisations with address sharing versus the number of allocated IPv4 addresses per citizen for each country. The number of allocated IPv4 addresses is reported by the RIRs at the end of 2016. As the IP address space is almost completely allocated, more recent numbers are only subject to minor changes. The plot shows that there is a tendency for countries with higher sharing percentage to have fewer IP addresses per citizen. Spearman's $\rho = -0.28$ and this correlation is significant at 95% significance level ($p = 0.006$).

We also analysed the percentage of organisations with address sharing versus the number of used IPv4 addresses of each country, i.e. the number of addresses estimated to be used at the end of 2014 using multiple IPv4 datasets and capture-recapture methods [22]. The plot, not shown for space reasons, indicates that there is a tendency for countries with higher sharing percentage to have a higher percentage of used IPv4 addresses. Spearman's $\rho = 0.17$, but this correlation is not significant at 95% significance level ($p = 0.1$).

V. LIMITATIONS AND FUTURE WORK

Our study uses the data of one Internet service, but it is a very large service with nearly world-wide coverage. Ideally, we could combine our data with data from other services, but currently we do not have data for other large services – such data is hard to obtain due to privacy concerns. As Steam does not work over mobile devices, such as phones and tablets, the dataset and the resulting analysis is likely to be PC-centric. Our results may include a smaller proportion of mobile broadband Internet connections than might normally be observed in the real world.

Determining whether a shared address was caused by NAT or a VPN would be valuable, but is infeasible with the current

[4]We compute this as count of the total number of unique user IDs behind shared IPs in all organisations for each RIR divided by the total number of unique user IDs for each RIR. Since most IP address changes are confined to one organisation, the approach of counting the number of unique users behind shared IPs for each organisation prevents double counting of users.

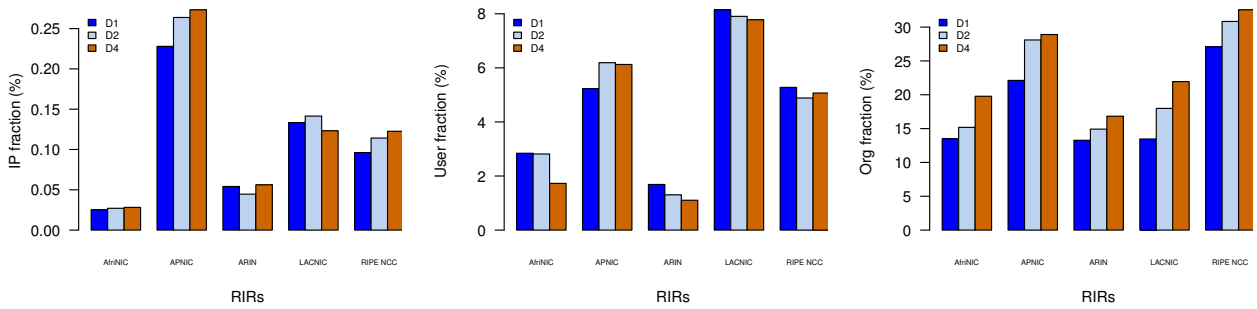

Fig. 7. Fraction of IP addresses (left), users (middle) and organisations (right) with large scale sharing depending on RIR.

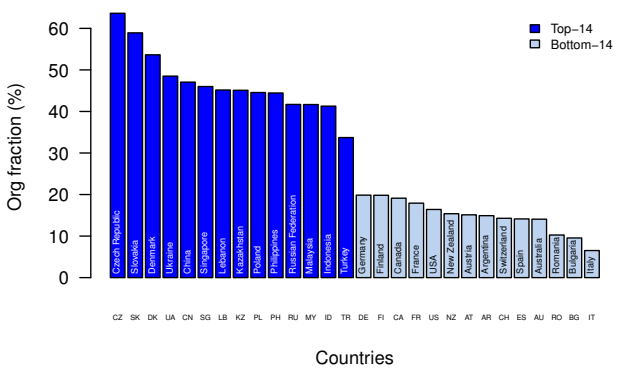

Fig. 8. Countries with highest and lowest fractions of organisations with IP sharing.

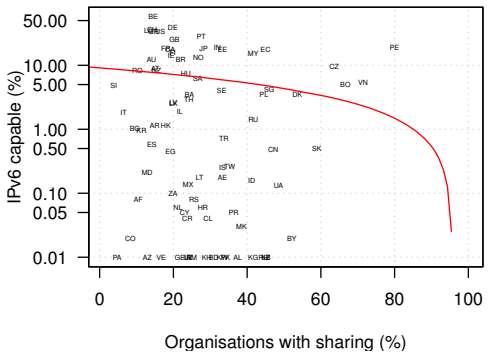

Fig. 9. Address sharing vs IPv6 capable per country.

data. We performed rDNS lookups of the IP addresses with observed sharing to investigate if the DNS names reveal information allowing a classification. For example, a name "nat.provider.com" provides an indication that the sharing is due to NAT. We created rules manually to detect NATs, VPNs and proxies, but our keyword classifier is only able to classify 3–5% of the shared IPs. Of the classified IPs the vast majority (96%) are NATs while the rest are VPNs/proxies.

Future work could investigate combining session data from Internet services with network traffic data, active measure-

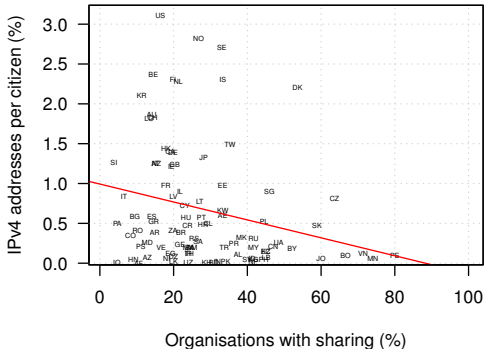

Fig. 10. Address sharing vs number of IP addresses per citizen per country.

ments or third-party data sources. For example, it may be feasible to detect VPNs based on a lower Maximum Segment Size (MSS) or identify VPNs/proxies by consulting databases, such as Maxmind [23]. Another avenue for future work is to estimate the *total* number of IPs with sharing by applying privacy-preserving capture-recapture methods [22].

VI. Conclusions

Large-scale address sharing, due to NAT, VPNs or proxies, is problematic since it limits the number of concurrent TCP connections for users and also severely limits geolocation and geoblocking. With most IPv6 transition technologies using a form of NAT, the problem will grow even with IPv6 adoption in the future. We analysed data from a large online gaming platform to investigate the presence of large-scale address sharing in the Internet.

The results show that there are thousands of IP addresses that are shared by ten or more users and a few hundred IP addresses are shared by hundreds or even thousands of users. Given that our dataset is limited to one particular application, the real number of users sharing these addresses is likely much larger. Cloud providers account for a large proportion of the most extreme cases, but when looking across all cases of sharing, ISPs and Telecommunications companies are usually responsible.

Asia, South America and Europe have the largest fractions of shared addresses, the largest fractions of users behind shared

addresses and also the largest fractions of organisations with large-scale sharing. For organisations these differences are reduced. The fraction of small scale-address sharing shows some consistency between countries. However, when looking at large-scale sharing, 10 users or more, there are much more profound differences between different countries. We also find that large-scale address sharing is significantly negatively correlated with the number of IPs per citizen, indicating that sharing is more common in countries with smaller number of IPs per citizen and thus NATs are the likely reason for address sharing.

Acknowledgements

We thank Valve Corporation, in particular Alfred Reynolds, for providing the datasets for this research. We also thank LACNIC for providing bulk access to their whois database.

References

[1] G. Huston, "IPv4 Address Report," 2016, http://www.potaroo.net/tools/ipv4/index.html [retrieved 22/4/2017].

[2] Tor Project, https://www.torproject.org/ [retrieved 22/4/2017].

[3] S. Bellovin, "Firewall-Friendly FTP," RFC 1579, 1994.

[4] J. B. David A. Hayes and G. Armitage, "Issues with Network Address Translation for SCTP," *ACM SIGCOMM Computer Communication Review (CCR)*, vol. 39, no. 1, pp. 24–33, Jan 2009.

[5] B. Aboba and W. Dixon, "IPsec-Network Address Translation (NAT) Compatibility Requirements," RFC 3715, 2004.

[6] I. C. Communications, "MC/159 Report on the Implications of Carrier Grade Network Address Translators," OFcom Whitepaper, 2013.

[7] G. Maier, F. Schneider, and A. Feldmann, "NAT Usage in Residential Broadband Networks," in *Proceedings of the 12th International Conference on Passive and Active Measurement*, 2011, pp. 32–41.

[8] T. Komarek, "Passive NAT Detection using HTTP Logs," M.S. thesis, Czech Technical University in Prague, 2015.

[9] G. J. Armitage, "Inferring the Extent of Network Address Port Translation at Public/Private Internet Boundaries," CAIA Technical Report 020712A, 2002.

[10] S. M. Bellovin, "A Technique for Counting NATted Hosts," in *Proceedings of the 2Nd ACM SIGCOMM Workshop on Internet Measurment*. ACM, 2002, pp. 267–272.

[11] L. Land, "Revised and much faster, run your own high-end cloud gaming service on EC2," https://lg.io/2015/07/05/revised-and-much-faster-run-your-own-highend-cloud-gaming-service/-on-ec2.html [retrieved 22/4/2017].

[12] S. Zander, D. Kennedy, G. Armitage, "Dissecting Server-Discovery Traffic Patterns Generated By Multiplayer First Person Shooter Games," in *NetGames 2005*, October 2005.

[13] T. Kohno, A. Broido, kc claffy, "Remote Physical Device Fingerprinting," in *Proceedings of IEEE Symposium on Security and Privacy*, May 2005, pp. 211–225.

[14] Valve, "Steam," http://store.steampowered.com/, 2017.

[15] MAXMIND, "GeoLite2 Free Downloadable Databases," http://dev.maxmind.com/geoip/geoip2/geolite2/ [retrieved 23/5/2017].

[16] I. Poese, S. Uhlig, M. A. Kaafar, B. Donnet, B. Gueye, "IP Geolocation Databases: Unreliable?" *SIGCOMM Comput. Commun. Rev.*, vol. 41, no. 2, pp. 53–56, Apr. 2011.

[17] Y. Shavitt, N. Zilberman, "A Geolocation Databases Study," *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 10, pp. 2044–2056, December 2011.

[18] W. B. Cavnar, J. M. Trenkle, "N-Gram-Based Text Categorization," in *Proceedings of 3rd Annual Symposium on Document Analysis and Information Retrieval (SDAIR)*, 1994, pp. 161–175.

[19] S. Zander, L. L. H. Andrew, G. Armitage, "Capturing Ghosts: Predicting the Used IPv4 Space by Inferring Unobserved Addresses," in *Internet Measurement Conference (IMC)*, November 2014.

[20] P. Richter, G. Smaragdakis, D. Plonka, A. Berger, "Beyond Counting: New Perspectives on the Active IPv4 Address Space," in *ACM Internet Measurement Conference (IMC)*, 2016.

[21] APNIC, "IPv6 Capable Rate by Country," http://stats.labs.apnic.net/ipv6 [retrieved 31/03/2017].

[22] S. Zander, L. Andrew, G. Armitage, "Collaborative and Privacy-preserving Estimation of IP Address Space Utilisation," *Elsevier Computer Networks*, vol. 119, pp. 56–70, 2017.

[23] Maxmind, "minFraud Fraud Detection Services - Tools for online fraud prevention," https://www.maxmind.com/en/home [retrieved 22/4/2017].