Don't count me out: On the relevance of IP addresses in the tracking ecosystem

Vikas Mishra, Pierre Laperdrix, Antoine Vastel, Walter Rudametkin, Romain Rouvoy, Martin Lopatka

Don't Count Me Out: On the Relevance of IP Address in the Tracking Ecosystem

Vikas Mishra
Inria / Univ. Lille
vikas.mishra@inria.fr

Pierre Laperdrix
CNRS / Univ. Lille / Inria
pierre.laperdrix@univ-lille.fr

Antoine Vastel
Univ. Lille / Inria
antoine.vastel@inria.fr

Walter Rudametkin
Univ. Lille / Inria
walter.rudametkin@univ-lille.fr

Romain Rouvoy
Univ. Lille / Inria / IUF
romain.rouvoy@univ-lille.fr

Martin Lopatka
Mozilla
mlopatka@mozilla.com

ABSTRACT

Targeted online advertising has become an inextricable part of the way Web content and applications are monetized. At the beginning, online advertising consisted of simple ad-banners broadly shown to website visitors. Over time, it evolved into a complex ecosystem that tracks and collects a wealth of data to learn user habits and show targeted and personalized ads. To protect users against tracking, several countermeasures have been proposed, ranging from browser extensions that leverage filter lists, to features natively integrated into popular browsers like Firefox and Brave to combat more modern techniques like browser fingerprinting. Nevertheless, few browsers offer protections against IP address-based tracking techniques. Notably, the most popular browsers, Chrome, Firefox, Safari and Edge do not offer any.

In this paper, we study the stability of the public IP addresses a user device uses to communicate with our server. Over time, a same device communicates with our server using a set of distinct IP addresses, but we find that devices reuse some of their previous IP addresses for long periods of time. We call this *IP address retention* and, the duration for which an IP address is retained by a device, is named the *IP address retention period*.

We present an analysis of 34,488 unique public IP addresses collected from 2,230 users over a period of 111 days and we show that IP addresses remain a prime vector for online tracking. 87 % of participants retain at least one IP address for more than a month and 45 % of ISPs in our dataset allow keeping the same IP address for more than 30 days. Furthermore, we also detect the presence of cycles of IP addresses in a user's history and highlight their potential to be abused to infer traits of the user behaviour, as well as mobility traces. Our findings paint a bleak picture of the current state of online tracking at a time where IP addresses are overlooked compared to other techniques like cookies or fingerprinting.

CCS CONCEPTS

• **Security and privacy → Privacy protections**; *Social aspects of security and privacy*.

KEYWORDS

IP address tracking, online privacy

WWW '20, April 20–24, 2020, Taipei, Taiwan
2020. ACM ISBN 978-1-4503-7023-3/20/04.
https://doi.org/10.1145/3366423.3380161

ACM Reference Format:
Vikas Mishra, Pierre Laperdrix, Antoine Vastel, Walter Rudametkin, Romain Rouvoy, and Martin Lopatka. 2020. Don't Count Me Out: On the Relevance of IP Address in the Tracking Ecosystem. In *Proceedings of The Web Conference 2020 (WWW '20), April 20–24, 2020, Taipei, Taiwan.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3366423.3380161

1 INTRODUCTION

Advertising is a popular way to monetize content on the web. While originally, online advertising consisted of simple ad-banners, nowadays the online advertising ecosystem has become more complex. To increase their incomes, advertisers propose advanced features that enable to target specific users based on a wide range of criteria ranging from their demographics to their prior purchases [6].

In order to target users with relevant ads, advertisers use trackers, usually JavaScript files or 1x1 pixel images, to gather data about their browsing routines [20]. To keep track of a user across different browsing sessions and websites, trackers generate a *Unique User IDentifier* (UUID) that is stored on the user's device. Moreover, trackers leverage several storage mechanisms to replicate this identifier so that, if a user deletes one of them, the identifier can be still be retrieved using the others [5]. Although trackers heavily rely on persistent storage APIs proposed by the browsers, such as cookies, local storage or indexedDB, they also divert the function of browser features not focused on persistent storage, such as the E-Tag HTTP header to store identifiers. Finally, advertisers also developed stateless tracking techniques, such as browser fingerprinting, that do not require storing any identifiers on the user's device [16].

To protect against these different forms of tracking, several countermeasures have been proposed, ranging from simple browser extensions based on filter lists and heuristics [3, 11–14] to more complex machine learning-based approaches [15] that leverage features extracted from HTML elements, HTTP requests and JavaScript. Recently, most of the major browser vendors have also started to implement advanced anti-tracking features. In 2017, Safari, through WebKit, integrated an *intelligent tracking protection* [26] to protect against both stateless and stateful tracking. Firefox also added countermeasures specifically designed against browser fingerprinting [8]. Moreover, in 2019, Chrome proposed *Privacy Sandbox* [24], a set of open standards to enhance privacy on the web. Even though all these features may be used as a commercial argument, which is out of the scope of this paper, there seems to be a trend indicating that browser vendors try to protect the privacy of their users.

Nevertheless, besides Opera that natively integrates a VPN to hide the IP address of its users [22], none of the aforementioned browsers protect against IP address based tracking techniques. Thus, we argue that many of these efforts present few benefit if users can be tracked solely because of their IP address. Indeed, while mobile network IP addresses are commonly shared by multiple users because of carrier-grade NAT [18], it is less the case of residential IP addresses. Even though most *Internet Service Providers* (ISP) provide dynamic IP addresses, they still remain the same for long time, as long as the user does not turn off their WiFi router for long enough. Nevertheless, no large-scale longitudinal study has been conducted on the duration for which such IP addresses are retained by users and its implications in terms of privacy.

Our study leverages a dataset of public IP addresses collected over a period of 111 days from 5,443 users. The IP addresses were collected using two browser extensions advertised on the AmIUnique website.[1] Using this dataset, we study the stability of the public IP addresses a user's device uses to communicate with our server. The public IP addresses we obtain could be those that are directly assigned to the users' devices or, more commonly, the users' devices are behind a gateway, such as a residential router, in which case, our server obtains the IP addresses of the routers. Over time, a same device communicates with our server using a set of distinct IP addresses, but we find that devices reuse some of their previous IP addresses for long periods. We call this *IP address retention* and, the duration for which an IP address is retained by a device, is known as the *IP address retention period*, or simply retention period. In many cases, a device's list of retained IP addresses show repetitive patterns. In its simplest form, this may be a home-work-home routine. We define these patterns as *IP address cycles*.

We first study the retention period of each IP address, how it varies from country to country, the presence of short-lived and long-lived IP addresses, and when a device uses a new, previously unknown, IP address, we test to see its similarity to previous IP addresses by removing the last—least-significant—octet. Then, we derive cycles of long-lived IP addresses for each user, which shows potential to discriminate and track users. Our intuition is that, given the portable nature of personal devices, IP cycles reflect human behavior and can be used as a proxy to infer other information, like user routines. Our evaluation also shows that even simple metrics, such as the Jaccard similarity between sets of IP addresses, provide unique and stable information that could be used by trackers, and could be also be used for respawning cookies.

In summary, this paper reports on the following results:

(1) 87 % of users have at least one IP address that was retained for more than 30 days,
(2) Among the 10 countries that contributed the most IP addresses to our dataset, the Netherlands shows the highest average IP address retention period of 36.96 days,
(3) 93 % of users have a distinct set of long-lived IP addresses, that is, the set of IP addresses with a retention period at least 30 days proves to be highly discriminating,
(4) 20 % of users have at least one cycle of long-lived IP addresses that lasts for more than 30 days.

[1] https://amiunique.org/

The remainder of this paper is organized as follows. Section 2 describes the background and related Work. Section 3 details our input dataset and how we cleaned it. Section 4 goes over our methodology and the analysis of our dataset. Section 5 discusses our results and some of the privacy implications. Section 6 concludes the paper.

2 BACKGROUND & RELATED WORK

An *Internet Protocol* (IP) address is a numerical label assigned to each device connected to a computer network that uses IP for communication. IP serves two main functions: host or network interface identification and location addressing. The IP address space is managed by *Internet Assigned Numbers Authority* (IANA) and by 5 regional registries for different parts of the world. These registries assign blocks of addresses to *Internet Service Providers* (ISP) who further assigns an IP address to each device connected on its network. These IP addresses can be *static* or *dynamic*, depending on the usage.

In general, static IP addresses are used to host services and are more expensive. Most residential IP addresses are *dynamic*, allowing the ISP to optimize how the address space is used. They are assigned by the ISP using the *Dynamic Host Configuration Protocol* (DHCP), and each address is given a *lease* with an expiry period. If the lease is not renewed before the expiry period, the address is released back to the DHCP server and can then be assigned to a different device. However, if the lease is renewed the device retains the same IP address. Depending on policies and configuration, this lease can be renewed an indefinite amount of times. In practice, if our WiFi set-top-boxes remain connected and the ISP's policies allow it, contrary to being *dynamic*, the same IP address may be retained for a long duration.

Dynamic IP addresses constitute a significant portion of assigned addresses. In 2007, Xie *et al.* [28] observed that more than 40 % of IP addresses collected from Hotmail user logins in a month were dynamic. They also studied the volatility of dynamic IP addresses and they showed that over 30 % of dynamic IP addresses were shared by more than one user within 1 to 3 days. Despite the very large dataset, the IP address is only collected when a user logs in, and important information, like the ID of the device or any IP changes between two user logins, is lost. As we propose in this paper, we believe that better understanding the stability of IP addresses requires a finer-grained dataset that can associate IP addresses to specific devices, and retrieve IP address changes rapidly.

Maier *et al.* [19] studied the dominant characteristics of residential broadband traffic in 2009. On a DSL network, they found that 50 % of IP addresses were assigned to residential routers at least twice in 24 hours, whereas 1–5% of IP addresses were reassigned almost 10 times a day. In our study, we look at the public IP addresses from which devices connect to the Internet, not at any specific ISP. We observe that some IP addresses are used by devices for only a few hours and then they change. Our study confirms that devices access the Internet across a large amount of short-lived IP addresses, but we also found that devices have some very long-lived IP addresses that they reuse over time, with induces important privacy risks, such as online tracking.

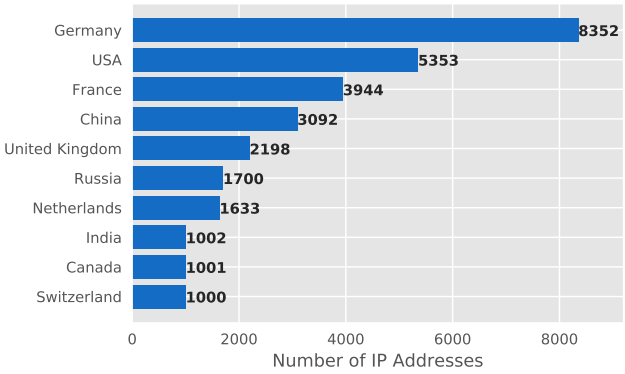

Figure 1: Top 10 countries contributing to our dataset.

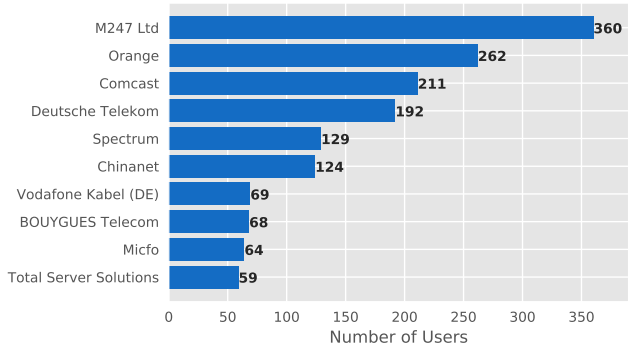

Figure 2: Top 10 ISPs contributing to our dataset.

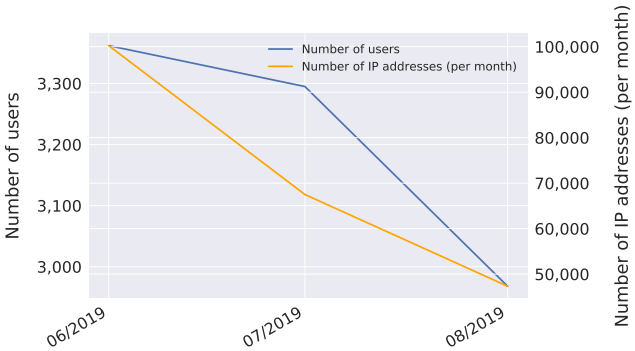

Figure 3: Evolution of the number of active extensions and IPs collected per month.

Although online tracking has been a focus of much recent research [9, 10, 25], studies on the relevance of IP addresses for tracking are still lacking. Yen *et al.* [29] were able to correctly identify 80 % of the users who deleted their cookies purely based on the IP address and the user-agent of the browser. Sebastian *et al.* [30] showed that the IP addresses along with browsing history of a user can be used effectively for cross-device tracking. However, their dataset is quite small with only 126 users and 1,994 IP addresses. Their objective was to study cross-device tracking, but they do not study the stability of IP addresses, how they are assigned, how portable devices, like laptops and cellphones, cycle through IP addresses when moving across networks.

We show the existence of IP address cycles, that is, IP addresses that the devices use to connect to the Internet over long periods of time. We argue that this is a proxy to the user's behavior, inherent in the way modern, portable, personal devices are used, and can reveal information on the user daily routine. De Montjoye *et al.* [7], in a study in 2013, showed that human mobility traces are highly unique. They showed that 95 % of individuals from a dataset of 1.5 million individuals can be uniquely identified if their location is specified hourly. The results of this study highlight the impact of learning an individual daily routine.

In this paper, we focus on studying *i)* the stability of IP addresses, including variations among countries, *ii)* the uniqueness of a set of IP addresses assigned to a user device, and finally, *iii)* we show the presence of cycles of IP addresses allocated to a user device, which has privacy implications, as suggested by De Montjoye *et al.* [7].

3 INPUT DATASET

The dataset we study contains 41,566 unique IP addresses from 5,443 browser instances and has been collected between June 6, 2019 and September 25, 2019 using the AmIUnique browser extensions for Chrome and Firefox.[2] AmIUnique [2] is a website that aims to study browser fingerprinting. To better understand stability of fingerprints over time, AmIUnique also provides two browser extensions, available for Chromium-based and Gecko-based browsers. Our dataset focuses on desktop devices, as these extensions are not supported by mobile browsers. Users are fully aware that these extensions collect data about their fingerprint, the timestamp and the IP address, since it is the main purpose of the extension and is clearly stated when the users install it. Every four hours, the extension sends a browser fingerprint and IP address to the AmIUnique servers. The AmIUnique initiative, including the online website and the associated extensions, obtained the agreement from the IRB to collect such information.

We used the Maxmind GeoLite2 database[3] to gather geolocation data on the collected IP addresses. Our dataset of 41,566 IP addresses originates from 131 countries, with the majority of them coming from either Europe or North America. Figure 1 reports on the Top 10 countries contributing IP addresses to our dataset and Germany is the highest represented one with 8,352 distinct IP addresses.

We also gather the ISP (*Internet Service Provider*) associated with the IP addresses in our dataset from the website *Who is my ISP?*[4] for this purpose. Our dataset of 41,566 IP addresses are assigned and owned by 3,087 ISPs. Figure 2 reports on the Top 10 ISPs contributing users to our dataset—*i.e.*, those ISPs which are used by the most number of users in our dataset.

Figure 3 summarizes the evolution of the number of active users for our extensions, together with the number of collected IPs over time. It is important to note that the x-axis reports browser instances, and not users: a user can have the extension installed in multiple browsers and we do not try to link those two installations to a single user. Furthermore, the extension assigns a unique identifier at install time and, if users reinstall the extension, a unique identifier is generated again and we consider it as a new instance.

[2] Available from https://amiunique.org/timeline

[3] https://dev.maxmind.com/geoip/geoip2/geolite2/

[4] https://www.whoismyisp.org

Nevertheless, for simplification purposes, we use the term *user* interchangeably with *browser instance* in the rest of the paper.

We can observe from the graph that, even though the number of extension users remains somewhat stable over the course of three months, the number of IP addresses collected per month decreases. This is because we get a stable number of new users every month, which are present for only a few days, thus, we remove these users from our study for better quality of results.

Dataset cleaning. We remove users that contribute IP addresses for less than 60 days from our dataset, as they do not contribute to our objective of studying stability of IP address over long durations. After removing these users from the dataset, we are left with 2,230 users and 34,488 distinct IP addresses in the dataset.

4 DATA ANALYSIS

In this section, we introduce our observations from the analysis of the input dataset described in Section 3. We first study the stability of IP addresses assigned to a user, then analyze the uniqueness of a set of IP addresses assigned to a user, which enables the discrimination of users based on their sets of IP addresses. Finally, we report on the presence of cycles in the user IP addresses and present its impact on privacy and tracking.

4.1 Stability & Uniqueness of IP Addresses

4.1.1 Stability. We specifically want to answer the following question: *How long does a device retain the same IP address on a given network?* Devices can obtain different IP addresses when connected to different networks (*e.g.*, home, office, leisure networks). When reconnecting to a previous network, more often than not, the device is assigned its previously used IP address. We call this duration the *retention period*.

The IP addresses collected in our dataset are public IP addresses from desktop and laptop devices, and it is entirely possible that the user physically moves to a different location and connects to a different network. Hence, if we observe a change of IP address for a user at two timestamps, it does not necessarily mean that the IP address for the user has changed on *a given network*. Thus, we simply calculate the retention period of a specific IP address irrespective of the fact that the user had a different IP address somewhere in the middle. We compute the retention period of each IP address for every user in the dataset by finding the time between the first and the last time the IP address appeared in our dataset for that user.

Retention period of an IP address. The average retention period of an IP address is 9.3 days, whereas 760 IP addresses (~2 %) were retained for more than 100 days. ~70 % of users (1569) had at least 1 IP address that was retained for more than 2 months. Figure 4 depicts a CDF plot for the retention period of IP addresses in days. From this graph, one can observe that the green line describes the overall retention period of IP addresses. We can see that most IP addresses are retained for a very short periods of time, for example, 90 % of IP addresses are retained for less than 10 days. However, when we plot the average retention of IP addresses per-user (red line), we observe that IP addresses are retained, on average, for longer periods. This is because, although the majority of IP addresses are short-lived, users also have very

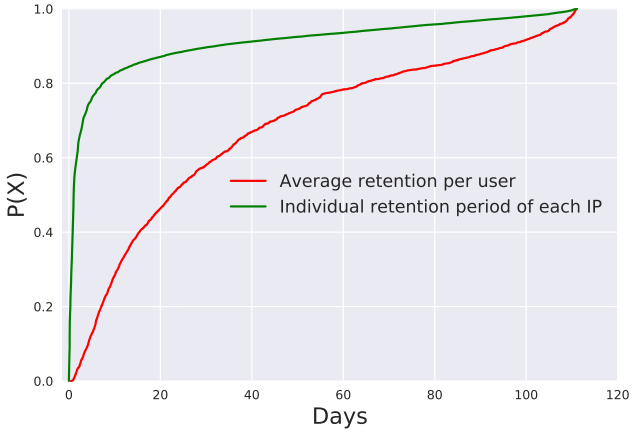

Figure 4: Retention period of IP addresses. The green line shows that, individually, most IP addresses are very short lived (90% are retained for less than 10 days), while the red line shows that, on average, users retain their IP addresses much longer (53% of users have an average retention of 20 days or more).

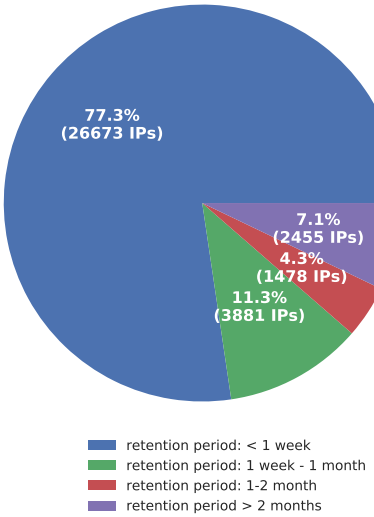

Figure 5: Number of IP addresses with retention periods ranging from less than a week to more than 2 months.

long-lived IP addresses. This makes it apparent that the users have both short-lived and long-lived IP addresses. Intuitively, when users are traveling or connected to a WiFi hot-spot, these IP addresses are retained for a short period of time, whereas the IP addresses at home and work are retained for longer durations.

Short-lived and long-lived IP addresses. To further understand the distribution of such short- and long-lived IP addresses for each user, we compute the number of IP addresses retained by a user for less than 1 week, 1 week, 1 month and 2 months in our dataset. 135 users (6 %) have at least one IP address that lasted for the entire duration of the dataset (111 days). Figure 5 depicts a pie chart of the number of IP addresses retained for different durations.

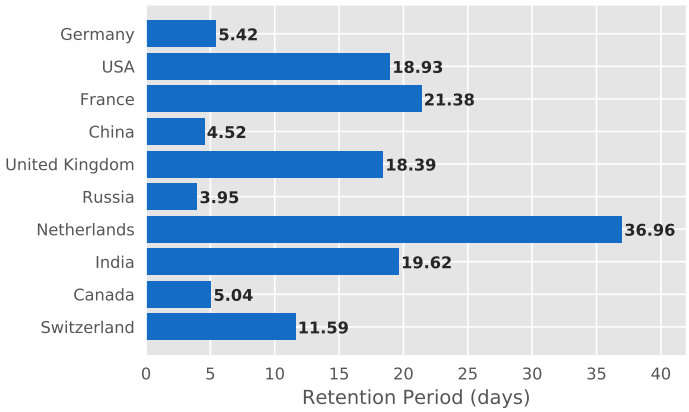

Figure 6: Average retention period for the Top 10 contributing countries. Countries ordered from highest number of IP addresses in our dataset (top), to lowest (bottom).

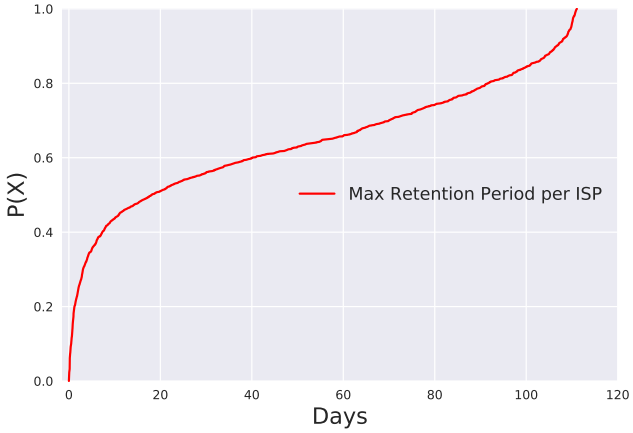

Figure 7: Maximum Retention period per ISP

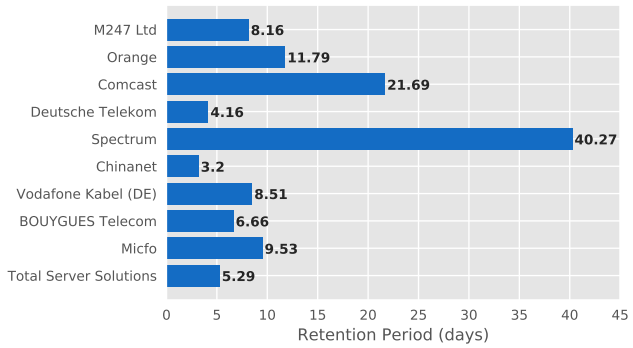

Figure 8: Average Retention period for the Top 10 contributing ISPs. ISPs ordered from highest number of users in our dataset (top), to lowest (bottom).

About 77 % of all the distinct IP addresses in our dataset have a retention period of less than a week while 7.1 % of IP addresses are retained for more than 2 months. Furthermore, only 22 users in our dataset do not have any IP address retained for more than a week, whereas 99 % (2, 208) of users have at least one IP address retained more than a week, ~87 % (1, 948) users have at least one IP that was retained for at least one month, and 70 % (1, 569) users have at least one IP address that was retained for more than 2 months.

Impact of country on retention period. We also study how the retention period varies per country. We gather country information using Maxmind's geolite database and compute the average retention period of each IP address per country. Figure 6 shows the average retention periods of the Top 10 countries. There are 31 countries with an average retention period greater than 10 days, with Turkey having the maximum retention period of about 97 days and Indonesia having the minimum with about 4 hours. We believe the reason for differences in retention period among these countries is due to the difference in the lease time set by DHCP servers from ISPs in different countries. However, we could not verify this hypothesis as it is challenging to gather lease times of specific ISPs given that their IP address configurations are often considered sensitive and are not published.

Impact of ISPs on the retention period. As explained in Section 2, the retention period of an IP address depends on the lease time and renewal policy used by the DHCP servers of an *Internet Service Provider* (ISP). Padmanabhan *et al.* [23] reported that some ISPs renew IP addresses after a fixed duration, ranging from one day to two weeks, and even more. Here, we want to see if we can observe similar behaviors in our dataset. By querying the `whoismyisp.org` website, we identified 3,087 distinct ISPs and we plotted, in Figure 7, the maximum observable retention period for each of them. We consider the maximum period we were able to identify in our dataset, per ISP, as the lower bounds of the worst case scenario in terms of IP retention for that ISP. This is because we collected data over a limited amount of time and this might not reveal the complete behavior of IP assignments; the longest retention periods may indeed be higher. We can see from the figure that about 45 % of ISPs in our dataset have at least 1 IP address that was retained for at least 30 days. Moreover, there are 392 ISPs for which the maximum retention period is more than 100 days. This shows that a significant portion of ISPs all over the world suffer from the issue of assigning long-lived IP addresses, making these users susceptible to IP-based tracking. Furthermore, in Figure 8, we present the average retention periods of the top 10 ISPs contributing the highest number of users in our dataset. In conclusion, we observe significant differences between ISPs. It remains to be seen if these results generalize at a larger scale with a more representative numbers of users and ISPs per country.

Amount of changes in IP addresses. Finally, we also observed several cases wherein, after a change in IP address, the new IP address is very similar to the previously assigned address. IPv4 addresses are made of only 4 octets, and we observed several cases where only the last octet of the IP address changed. For instance, assume a user had an IP address `89.158.242.220` that changed to `89.158.242.120`, while on the same network. This occurs often because ISPs generally assign IPs from a relatively small range. To measure the prevalence of such changes in our dataset, we calculate the average retention period of each user

by ignoring the last octet of every IP address in our dataset. Thus, IP `89.158.242.120` simply becomes `89.158.242` for further calculations of retention period. The average retention period more than doubles from 9.3 days to 19.51 days after ignoring the last octet of the IP address. Thus, showing that even in cases when IP addresses change, they often do not change completely, and retain much of their previous information.

Our analyses and observations show that users have both short- and long-lived IP addresses. This can intuitively be explained by the fact that IP addresses are a proxy to user behavior, as users travel, they connect to new networks and obtain short-lived IP addresses. However, other networks are more prevalent in the user's routine, for example, their home or work networks, or the WiFi hotspot of their favorite places. Moreover, whenever IP addresses do change, much of the time the changes involve the least-significant octet of the address. This shows potential privacy issues for users in regards to tracking.

4.1.2 Uniqueness. Our observations from the above subsection gives us the key insight that users retain IP addresses for long durations and reuse a given IP address multiple times over a period of weeks or months in some cases. The fact that IP addresses are retained and reused for a long period of time alone cannot be exploited to track people over time. Multiple users can share the same IP address behind a NAT router. For instance, in our dataset, there are 1,046 users who share 1 IP with another user. However, we observed in our dataset that there are multiple IP addresses that are reused and retained for a long duration by a user. The chances that two users connect to the same network at multiple locations and share multiple IP addresses is quite low. Thus, set of IP addresses retained by a user over a long duration can be considered to be unique, as well as stable, for a long duration.

To evaluate this hypothesis, we compute a pair-wise comparison between the set of long-lived (greater than 30 days) IP addresses for each user in our dataset and compute the Jaccard similarity to verify if a set of IP addresses can be used for discriminating different users. We use the following formula to calculate the Jaccard similarity between 2 users A and B with IP_A and IP_B being the sets of IP addresses used by the corresponding users.

$$J(A, B) = \frac{IP_A \cap IP_B}{IP_A \cup IP_B}$$

This results in us performing 2,485,335 comparisons for 2,230 users in our dataset. From this pair-wise comparison, we find that the Jaccard similarity is zero for 99.97 % of pairs of users in our dataset. We further observed that ~93 % of users have a unique set of long-lived IP addresses. Thus, showing the high uniqueness offered by sets of IP addresses assigned to a user. Furthermore, we also observed a perfect Jaccard similarity of 100 % for 44 pairs of users. After manually analyzing the browser fingerprints of these users, we came to the conclusion that these users are actually the same, but are marked as different. We believe the reason for this is that these people use our extension in multiple browsers, thus, having multiple extension IDs.

4.2 Presence of Cycles

With our dataset, we have observed that the list of distinct IP addresses is enough to identify, and potentially track, our user base over time. We also observed that there are users in our dataset who exhibit connectivity patterns in their IP addresses, with one IP in the morning and the other at night, often repeated over long durations. We define these repeated patterns as *cycles*.

In Section 4.1.1, we showed that there are both short-lived and long-lived IP addresses for each user. In this section, we only look at those cycles, which are made of IP addresses with retention periods greater than a month, since they are more interesting in terms of tracking a user's routine. We observed 443 users in our dataset who have at least 1 cycle, repeating every week (frequency of 1 week) and lasting more than a month. The maximum duration of a cycle detected in our dataset is 105.5 days. These cycles can consist of various numbers of IP addresses. 443 users have a cycle of 2 IP addresses, 78 users have a cycle of 3 IP addresses. The maximum size of cycle, in terms of number of distinct IP addresses, observed in the dataset is 4 IP addresses, which was observed for 7 users.

4.2.1 Impact on Privacy. Our hypothesis is that the IP address cycles present in our database can be attributed to a simple cause: the user's behavior and daily routines. As we have shown in Section 4.1.1, IP addresses are stable over time. If we observe the same IP address multiple times for a given user over long durations, and this IP address is interleaved by other IP addresses, this likely means that the user reconnected through the same network, from the same physical location as before. In essence, the user came back to the same network. These cycles then present two major privacy problems, as we describe next.

Insight into the user routine. First, it can give an adversary incredible insight into someone's weekly or daily routine. If we observe the same set of addresses during the day and a different one during the night, we can deduce when the user is at work or at home. From our dataset of IP addresses, we observe that 5, 633 of them are always present in either day or night. Furthermore, we also detected 58 users in our dataset, which have IP cycles that resemble mobility traces. Their IP address changes every morning, remains the same until the evening, and the same cycle continues the next day. We also saw differences for these users over weekends, where the IP address that was observed between morning and evening was not present for these users on Saturdays and Sundays. It is intuitive to think that the reason behind this could be that these IP addresses are assigned to the user's device at work. If one is able to acquire fine-grain data, identifying cycles can even help predict where the user will be next. In 2015, Libert *et al.* crawled the Alexa Top 1M websites and found that 78.07 % of websites initiated third-party HTTP requests to a Google-owned domain [17]. Such a massive presence on the web gives Google a lot of power in acquiring a very fine-grained set of IP addresses that could be used to identify one's routine with precision.

Tracking through IP address cycles. The second problem is that the identification of cycles can ease user reidentification. In the case a user deletes her cookies or uses a different device, identifying

a cycle that has been seen in the past can help in linking two distinct profiles to the same user.

In the end, the stability of individual IP addresses combined with the uniqueness of sets of IP addresses involved in the cycles creates an important privacy threat. While software bugs can be fixed with patches or redesigns, human behavior is not so simple. These cycles are inherently linked to peoples' behavior, as exemplified by the typical home, work, home daily routine. Using VPNs or the Tor network can mitigate the effects of IP address tracking, at the cost of changing the user's quality of experience, while including the inherent added risks of trusting a potentially unknown server (*e.g.*, TOR exit node snooping [27], VPN server). However, we believe that further research should be done to study these aspects of cycles and we argue that the problem should be fixed at its core, through smaller retention periods and higher diversity of IP addresses.

5 DISCUSSION

While the current tracking ecosystem exploits numerous ways to track users online, particularly with the use of UUIDs in persistent storage (*e.g.*, cookies), not enough attention has been given to tracking using IP addresses. In this study, we looked at the stability of public IP address and found that long-lived addresses are stable and can enable long-term tracking of users. IP addresses are retained, and reused, and can form cycles that could potentially allow inferring information about a user behavior or routine. These findings are comparable to what can be seen with mobility traces. In a study published in 2013, De Montjoye *et al.* [7] showed that human mobility traces are highly unique. They found that 95 % of individuals from a dataset of 1.5 million individuals can be uniquely identified if their location is specified hourly. Their results highlighted the impact of learning an individual's daily routine and the privacy violations that go with it. Our intuition is that IP addresses of portable, personal devices can be used as a proxy to a user's mobility traces.

To mitigate the privacy implications of IP tracking, we can look at the advantages and disadvantages of different solutions. In the field of geolocation, methods like the one proposed by Andrés *et al.* [4] where noise is added to the traces of a user can increase privacy without impacting the quality of the desired service. However, for IP addresses, one cannot simply change the IP address and modify bits of it. The packets will never reach their destination and the quality of service will simply drop to zero.

One solution may include using VPNs, but this also presents some shortcomings. For a VPN to fully work against IP tracking, a user always has to use it. VPNs introduce additional latency, most are for-profit services, and in locations such as the work environment, user's might not be allowed to use them or may find themselves unable to access the local services. Some services can also detect the use of VPNs, adding information to enable tracking, and even block them, such as Netflix who attempts to geo-lock their content. Finally, users must also trust the VPN service since they are ideally positioned for snooping the users' activities, as has been shown in existing work [21].

Another solution would be to use the Tor Browser [1] that funnels all network packets through the Tor network. This solution works against IP tracking, but the user has serious drawbacks. Tor exit node IP addresses are public, meaning any service can know if the client is connecting through the Tor network, and many services already block Tor. The user also faces a degraded experience on the web, as the Tor browser limits some functionality in order to improve privacy and security. And finally, the Tor network increases latency by a considerable amount, which may, or may not be, acceptable to the user.

Finally, in 2016, Padmanabhan *et al.* [23] measured how often some IP addresses change depending on the ISP or the country. They found that some ISPs in Germany and in France dynamically change residential IP addresses every day. Moreover, they observed that IP address changes typically span prefixes and most users are not reassigned addresses that are very close numerically to their old ones. This study proves that dynamic changes are possible on a daily basis without causing drastic service interruptions or routing inflation. In the end, we argue that the simplest solution that can benefit all users is to work closely with ISPs to reassign new addresses more frequently and ensure, as best as possible, that the address changes are difficult to link. This would mitigate most of the long-term impacts of IP tracking and prevent trackers from building IP address cycles that can give insight into users' lives.

Threats to Validity Our dataset is inevitably biased as we collected data from desktop users who visited the AmIUnique website and installed our extension. Our findings are intended to be challenged with a larger population and data coming from mobile devices. Moreover, we do not have concrete details on the renewal policies of different ISPs that would confirm and put into perspective the retention periods we observed in our study. Yet, our results show that IP-based tracking is still very much alive. With both long-lived IP addresses and the presence of cycles, most users in our dataset can be tracked just because of the network environments they connect to.

6 CONCLUSION

In this paper, we studied IP addresses to assess their impact in the current tracking ecosystem. With the help of 2,230 participants, we collected IP addresses over 111 days and analyzed both their stability and uniqueness. Out of 34,488 distinct IP addresses we collected, we found that 11.3 % of them were present for more than a month. Many endpoints from which our users connected to the Web remained stable throughout the 111 day period and, as such, renders users vulnerable to long-term tracking based solely on their IP address. For about 20 % of users, we also identified cycles of long-lived IP addresses, which have the potential to provide insights into users' lives, and also allows easier re-identification in the case of cookie removal. We found 58 users with a very specific cycle which resembles a typical home, work, home daily routine, including weekends off.

Our analysis shows that 93 % of users in our dataset had a unique set of long-lived IP addresses (minimum retention period of 30 days) that remain stable over time, presenting a bleak picture of the state of online tracking. Despite much effort to fight stateless and stateful tracking mechanisms, such as cookies, cookie syncing or browser fingerprinting, the role of IP addresses in tracking remains significant and, in regards to recent large-scale research studies, an

overlooked threat to online privacy. Based on these preliminary findings, we believe more research is required to get a better picture of the risks involved in IP address tracking.

ACKNOWLEDGEMENTS

This work is partially supported by ANR-Bottlenet under grant number ANR-15-CE25-0013 and partially funded by the ANR FP-Locker project under grant number ANR-19-CE39-00201.

REFERENCES

[1] 2019. Tor Browser - Tor Project Official website. https://www.torproject.org/projects/torbrowser.html.
[2] Accessed on 2019-10-04. *AmIUnique: Platform to collect browser fingerprints.* https://amiunique.org
[3] AdBlock. 2018. AdBlock. https://getadblock.com/
[4] Miguel E. Andrés, Nicolás E. Bordenabe, Konstantinos Chatzikokolakis, and Catuscia Palamidessi. 2013. Geo-indistinguishability: Differential Privacy for Location-based Systems. In *Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security (CCS '13)*. ACM, New York, NY, USA, 901–914. https://doi.org/10.1145/2508859.2516735
[5] Mika D Ayenson, Dietrich James Wambach, Ashkan Soltani, Nathan Good, and Chris Jay Hoofnagle. 2011. Flash cookies and privacy II: Now with HTML5 and ETag respawning. *Available at SSRN 1898390* (2011).
[6] Facebook business. 2019. Help your ads find the people who will love your business. https://www.facebook.com/business/ads/ad-targeting
[7] Yves-Alexandre De Montjoye, César A Hidalgo, Michel Verleysen, and Vincent D Blondel. 2013. Unique in the crowd: The privacy bounds of human mobility. *Scientific reports* 3 (2013), 1376.
[8] Arthur Edelstein. 2019. Protections Against Fingerprinting and Cryptocurrency Mining Available in Firefox Nightly and Beta. https://blog.mozilla.org/futurereleases/2019/04/09/protections-against-fingerprinting-and-cryptocurrency-mining-available-in-firefox-nightly-and-beta/
[9] Steven Englehardt and Arvind Narayanan. 2016. Online tracking: A 1-million-site measurement and analysis. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*. ACM, 1388–1401.
[10] Steven Englehardt, Dillon Reisman, Christian Eubank, Peter Zimmerman, Jonathan Mayer, Arvind Narayanan, and Edward W Felten. 2015. Cookies that give you away: The surveillance implications of web tracking. In *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 289–299.
[11] Electronic Frontier Foundation. 2018. Privacy Badger. https://www.eff.org/fr/node/99095
[12] Cliqz International GmbH. 2018. Ghostery. https://www.ghostery.com
[13] Eyeo GmbH. 2018. Adblock Plus. https://adblockplus.org/
[14] Raymond Hill. 2018. uBlock Origin - An efficient blocker for Chromium and Firefox. Fast and lean. https://github.com/gorhill/uBlock
[15] Umar Iqbal, Peter Snyder, Shitong Zhu, Benjamin Livshits, Zhiyun Qian, and Zubair Shafiq. 2020. ADGRAPH: A Graph-Based Approach to Ad and Tracker Blocking. *IEEE Security and Privacy* (2020).
[16] Pierre Laperdrix, Walter Rudametkin, and Benoit Baudry. 2016. Beauty and the beast: Diverting modern web browsers to build unique browser fingerprints. In *2016 IEEE Symposium on Security and Privacy (SP)*. IEEE, 878–894.
[17] Timothy Libert. 2015. Exposing the Invisible Web: An Analysis of Third-Party HTTP Requests on 1 Million Websites. *International Journal of Communication* 9, 0 (2015). https://ijoc.org/index.php/ijoc/article/view/3646
[18] Ioana Livadariu, Karyn Benson, Ahmed Elmokashfi, Amogh Dhamdhere, and Alberto Dainotti. 2018. Inferring carrier-grade NAT deployment in the wild. In *IEEE INFOCOM 2018-IEEE Conference on Computer Communications*. IEEE, 2249–2257.
[19] Gregor Maier, Anja Feldmann, Vern Paxson, and Mark Allman. 2009. On dominant characteristics of residential broadband internet traffic. In *Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement*. ACM, 90–102.
[20] Jonathan R Mayer and John C Mitchell. 2012. Third-party web tracking: Policy and technology. In *2012 IEEE symposium on security and privacy*. IEEE, 413–427.
[21] Xianghang Mi, Ying Liu, Xuan Feng, Xiaojing Liao, Baojun Liu, XiaoFeng Wang, Feng Qian, Zhou Li, Sumayah Alrwais, and Limin Sun. 2019. Resident Evil: Understanding residential ip proxy as a dark service. In *Resident Evil: Understanding Residential IP Proxy as a Dark Service*. IEEE, 0.
[22] Opera. 2019. Free VPN | Browser with built-in VPN | Download | Opera. https://www.opera.com/computer/features/free-vpn
[23] Ramakrishna Padmanabhan, Amogh Dhamdhere, Emile Aben, kc claffy, and Neil Spring. 2016. Reasons Dynamic Addresses Change. In *Proceedings of the 2016 Internet Measurement Conference (IMC '16)*. ACM, New York, NY, USA, 183–198. https://doi.org/10.1145/2987443.2987461
[24] Justin Schuh. 2019. Building a more private web. https://www.blog.google/products/chrome/building-a-more-private-web/
[25] Antoine Vastel, Pierre Laperdrix, Walter Rudametkin, and Romain Rouvoy. 2018. FP-STALKER: Tracking Browser Fingerprint Evolutions. In *IEEE S&P 2018-39th IEEE Symposium on Security and Privacy*. IEEE, 1–14.
[26] John Wilander. 2017. Intelligent Tracking Prevention | WebKit. https://webkit.org/blog/7675/intelligent-tracking-prevention/
[27] Philipp Winter, Richard Köwer, Martin Mulazzani, Markus Huber, Sebastian Schrittwieser, Stefan Lindskog, and Edgar Weippl. 2014. Spoiled onions: Exposing malicious Tor exit relays. In *International Symposium on Privacy Enhancing Technologies Symposium*. Springer, 304–331.
[28] Yinglian Xie, Fang Yu, Kannan Achan, Eliot Gillum, Moises Goldszmidt, and Ted Wobber. 2007. How dynamic are IP addresses?. In *ACM SIGCOMM Computer Communication Review*, Vol. 37. ACM, 301–312.
[29] Ting-Fang Yen, Yinglian Xie, Fang Yu, Roger Peng Yu, and Martin Abadi. 2012. Host Fingerprinting and Tracking on the Web: Privacy and Security Implications.. In *NDSS*, Vol. 62. Citeseer, 66.
[30] Sebastian Zimmeck, Jie S Li, Hyungtae Kim, Steven M Bellovin, and Tony Jebara. 2017. A privacy analysis of cross-device tracking. In *26th USENIX Security Symposium (USENIX Security 17)*. 1391–1408.