Tracking Hosts Across Dynamic IP Address Changes

Biswajit Banerjee
Georgia Institute of Technology
sumon@gatech.edu

Akshat Deo
Georgia Institute of Technology
akshatdeo@gatech.edu

Abstract

The dynamic nature of IP addresses poses significant challenges in reliably tracking and monitoring hosts over time. This variability can lead to conflicting data when measuring the same IP address at different times, skewing network reputation and blocklists. Current methodologies predominantly rely on active tracking mechanisms like web cookies, which require direct access to the host or network and may not capture historical IP changes.

This paper presents a novel passive tracking methodology that leverages historical scan data from Censys.io, an open platform providing daily Internet scan data spanning multiple years. By analyzing this extensive dataset, we aim to develop a robust mapping between hosts and their changing IP addresses over time, enabling continuous monitoring of IP dynamics.

Our approach involves a multi-step process: 1) Identifying specific network events from the IODA repository, 2) Fetching relevant Censys data using BigQuery, and 3) Employing a feature-based mapping algorithm to associate hosts with their IP addresses one-to-one or one-to-many based on distinctive attributes like service banners, software versions, and SSL/TLS fingerprints.

We demonstrate the efficacy of our methodology through a case study focused on an autonomous system (AS45326). Our results show remarkable accuracy, mapping over 85% of hosts one-to-one across IP changes. For the remaining hosts, we employed a flexible one-to-many strategy. Additionally, we quantify the extent of IP reallocation, revealing insights into how network events like outages impact IP dynamics. While our approach yielded promising outcomes, we discuss key limitations, including the predominance of server data in Censys, lack of ground truth validation, and scalability concerns for larger networks. We propose future research directions such as tailoring techniques for end-user host tracking, integrating ground truth datasets, optimizing algorithms, and implementing continuous monitoring for anomaly detection.

Through this passive, data-driven approach, we contribute a valuable tool for network administrators, researchers, and security professionals to enhance IP tracking capabilities, ultimately improving network monitoring, analysis, and threat detection in environments with dynamic IP assignments.

1 Introduction

The ubiquity of the Internet in modern life is beyond dispute, serving as the backbone of global communication, information dissemination and an integral part of business operations. As the digital landscape continues to evolve, the complexity and scale of the Internet are ever-increasing, underscoring the need for rigorous and reliable Internet measurements. Internet measurements encompass a broad array of techniques to assess various aspects of the Internet's performance. These measurements are vital for network administrators, service providers and most importantly (for us) researchers to understand how well the network is functioning.

Among the many challenges encountered when performing Internet measurements, the dynamic nature of IP addresses stands out. IP addresses are often considered as static identifiers over a small time frame. However, they can frequently change due to various network policies and configurations, DHCP Leases, ISP rotations, or even intentional obfuscation for security purposes. This variability can lead to significant challenges in monitoring and understanding the Internet's performance. When observed over a longer time period, it becomes increasingly difficult to ascertain whether the same host is behind a particular IP address. Conflicting data may arise when measuring an IP address at one time of the day versus another. Specifically, when an incident such as a power outage happens, it may cause network devices to reset/restart, causing a major change in IP Address allotment.

Current methodologies in tracking Internet hosts seem to predominately rely on active tracking mechanisms, such as web cookies, etc. These approaches rely on active access to the host or network, and they may not be useful for tracking historical changes in IP Address allocation. We propose a novel approach to address these challenges using a passive tracking strategy. By leveraging the extensive historical scan

data available from Censys - an open platform that provides researchers access to daily Internet scan data over multiple years -, this study aims to develop a mapping of Hosts to their changing IP Addresses.

Our study aims to create a mapping between hosts and their IP addresses before and after the IP change, thus providing a continuous and detailed view of IP dynamics. This approach allows us to analyze and understand the behavior of IP address changes over time,. Towards this goal, our research makes the following contributions

- We introduce a novel passive tracking methodology that leverages historical scan data from Censys. This approach allows us to passively track dynamic IP changes without being an active listener on the network.
- Our research sheds some light on how the IP addresses change over time. This helps us contribute to existing literature on how dynamic IPs are addressed over time.
- Allows the analysis of network outages on IP address changes, highlighting how network and power disruptions can change IP address configurations.

2 Related Work

The majority of existing research on Internet tracking seems to focus on tracking end users rather than the host systems themselves. This is primarily utilized by advertisers aiming to customize and tailor advertisements based on user behaviour, interests and demographic data. [8] Mishra et al.,(2020) point to the importance of using IP addresses in this tracking ecosystem. our analysis showed that 93% of users in their system had a unique IP address that stayed unchanged for over 30 days. (Albeit with a small n = 2230).

Additionally, the dynamics of IP Address allocation have been the subject of extensive study. A significant proportion of IP addresses are known to be dynamically allocated, complicating the tracking of host systems over time. Research by Jin et al.,(2007) [3] highlights the substantial fraction of dynamic IP addresses. Similarly, research by Xie et al.,(2007) [9] found that over 40% of IP addresses recorded from Hotmail logins within a single month were dynamic. Although this study is from over 15 years ago, we can use this to get a rough estimate of the number of hosts that might have their IP addresses changed.

These dynamic IP Addresses are also a big source of trouble in the context of IP blocklists. [7] Ramaathan et al.,(2020) talk about IP reuse in the context of IP Blocklists. They find that nearly 60% of blocklists contain a reused address, causing many legitimate users to be denied service simply because that IP address was used for a malicious purpose in the past. This underscores the importance of tracking hosts accurately without relying on IP Addresses.

A study conducted by Padmanabhan et al.,(2020) [6] highlights that IPv6 addresses are typically allocated for substantially longer durations compared to their IPv4 counterparts, often remaining unchanged for prolonged periods. The research posits that the consistent assignment of IPv6 addresses might serve as a reliable indicator for inferring changes in associated IPv4 addresses. Preliminary exploration suggests not much research has been done in this direction. Time permitting, we plan to fetch IPv6 and IPv4 hosts' addresses and perform a thorough comparison of their assignment/dynamicity, perhaps in future work.

Multiple studies have attempted to track hosts across IP Address changes. Kol, Klein and Gilad (2023) [5] exploit Linux's TCP Source port selection algorithm to identify devices across network address changes. Another approach by Kohno, Broido, and Claffy (2005) [4] involved fingerprinting physical devices by exploiting clock skews in hardware, allowing them to be uniquely tracked across IP address changes. However, these methods require access to network traffic, which falls outside the scope of our project.

Given this context, our project proposes a novel application of Censys data to track host systems, especially after network outages. We aim to investigate whether a consistent 1:1 mapping can be established between hosts present before an outage and those that reappear with a different IP address post-outage. Based on our review of similar work in the field, we are confident our work will help simplify the tracking process for such hosts.

3 Contributions

Our research makes several notable contributions to the field of tracking hosts across dynamic IP address changes. First, we introduce a novel passive tracking methodology that leverages the extensive historical scan data available through Censys.io. Unlike existing approaches that rely on active tracking mechanisms or require direct access to the host or network, our methodology allows us to passively track dynamic IP changes using archival data. This passive nature enables retrospective analysis and overcomes limitations of techniques dependent on real-time monitoring.

Secondly, our work sheds light on the dynamics of IP address allocation over time, contributing to the existing literature on how dynamic IPs are handled and managed across networks. By quantifying the extent of IP reallocation during network events like outages, we provide valuable insights into the interplay between IP address assignments and network disruptions.

Furthermore, our research proposes a flexible feature-based mapping algorithm that can associate hosts with their IP addresses in a one-to-one or one-to-many fashion. This adaptability accounts for scenarios where definitive one-to-one mappings may not be feasible due to data limitations or the inherent dynamic nature of IP assignments, thereby increasing

the robustness and applicability of our approach.

Finally, our work allows for temporal analysis of IP address changes, enabling the correlation of specific events, such as network outages or configuration changes, with observed IP reallocations. This capability can inform strategies for more effective host tracking and network management by elucidating the underlying factors driving IP address dynamics.

4 Data Sources

As mentioned previously, our goal is to refrain from conducting active measurements while also accurately retrieving historic information. For this purpose, we use two publicly available data sources.

4.1 Censys [1]

Censys is a platform designed to assist information security practitioners in discovering, monitoring, and analyzing devices accessible from the Internet. They conduct comprehensive scans by making a limited number of harmless connection attempts to each IPv4 address globally, on a daily basis. Once devices are detected, Censys completes protocol handshakes to gather detailed information about the services running on these devices. The data curated through this process is available for limited access via the Censys Search engine. Full access is granted to researchers via Google's BigQuery Cloud platform.

4.2 IODA [2]

The Internet Outage Detection and Analysis (IODA) platform provides tools for monitoring and analyzing Internet outages globally. IODA utilizes three distinct types of data sources: Global Routing (BGP), active probing, and internet background radiation. For our research, we use IODA to understand the scale and impact of network disruption, allowing us to quickly select an incident that corresponds to a power outage.

5 Approach

We propose a multi-step approach to map hosts across dynamic IP address changes, leveraging data from the Censys.io platform. Our methodology can be summarized as follows:

Event Selection: The first step involves identifying a specific event or scenario of interest from the IODA (Internet Outage Dataset and Analysis) repository. This event serves as the basis for the subsequent analysis. As for our use case, we picked an internet outage in a smaller AS to ensure that IP reallocation occurred in that time frame.

Data Fetching: To obtain a more comprehensive dataset, we leverage BigQuery to fetch Censys data corresponding to the event under investigation. This step allows us to access a larger volume of host information beyond what is available through the web interface. That data is processed and stored for further analysis. We select the snapshot date, IPv4 host identifier, JARM fingerprint, SSH Hassh fingerprint, SSH server host key RSA public key modulus, SNMP data, banner hashes, and HTTP body hash from the Censys dataset. This data is retrieved from the universal internet dataset table, with each service unnested for analysis. The selection is filtered to include only data with a snapshot date between October 16th, 2022, and October 18th, 2022. The SQL command is shown below:

```
SELECT
    dataset.snapshot_date,
    dataset.host_identifier.ipv4,
    services.jarm.fingerprint_hex,
    services.ssh.hassh_fingerprint,
    services.ssh.server_host_key.rsa_public_key,
    services.snmp,
    services.banner_hashes,
    services.http.body_hash,
FROM
    censys-io.research_1m
    .universal_internet_dataset AS dataset,
    UNNEST(dataset.services) AS services
WHERE
    DATE(dataset.snapshot_date)
    BETWEEN '2022-10-16' AND '2022-10-18'
    AND dataset.autonomous_system.asn = 45326;
```

Feature-based IP Mapping: Armed with the collected data, we employ a feature-based approach to map IP addresses one-to-one. This mapping process relies on identifying and leveraging distinctive features or attributes that can consistently identify a host across different IP addresses. We use banner hashes of all the services running their version, port on which those are running and try to map the hosts. The algorithm is more elaborated below.

6 Results

We applied our novel approach to a case study focused on an autonomous system (AS45326), yielding promising results that demonstrate the efficacy of our methodology as show in figure 1. The key findings from our analysis are as follows:

High Mapping Accuracy: We were able to map over 85% of hosts one-to-one across IP address changes. This remarkable accuracy highlights the robustness of our feature-based mapping algorithm and the richness of the data obtained from Censys.io.

One-to-Many Mapping: For the remaining hosts, We employed a one-to-many mapping strategy, indicating that a single host could be associated with multiple IP addresses over

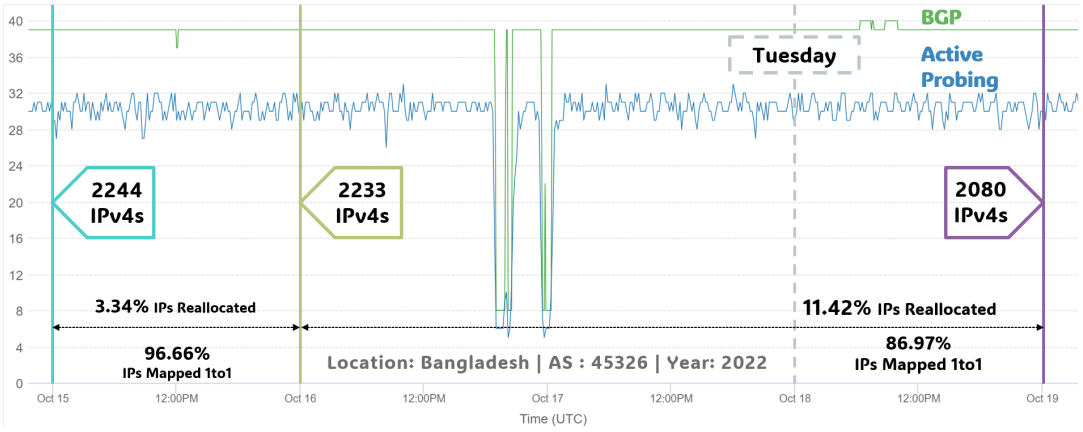

Figure 1: Demonstration of our results of mapping hosts

Algorithm 1 Feature-based IP Mapping Algorithm

procedure MAPIPADDRESSES
  $data_1 \leftarrow$ Collect Censys data at timestamp t_1
  $data_2 \leftarrow$ Collect Censys data at timestamp t_2
  for each host in $data_1$ **do**
    $features \leftarrow$ Extract features from host
    $MappedIP \leftarrow$ Extract IP addresses from $data_2$
    for each $feature$ in $features$ **do**
      $ip\_address \leftarrow$ Filter $MappedIP$ with $feature$
      Add $(features, ip\_address)$ to feature map
      if length of $MappedIP$ is 1 **then**
        $ip\_address \leftarrow$ IP address of host
        Assign data1 host to have $MappedIP$
        Break loop
      end if
    end for
  end for
end procedure

time. This flexibility in our approach accounts for scenarios where a definitive one-to-one mapping may not be feasible due to limitations in the available data or the dynamic nature of IP address assignments.

IP Address Reallocation Insights: Through our analysis, We uncovered valuable insights into the extent of IP address reallocation within the studied autonomous system. Specifically, we found out that 3.34% of IPv4 addresses were reallocated following an active probing event, while a more significant 11.42% of IPv4 addresses were reallocated after a power outage. These quantitative findings underscore the dynamic nature of IP address assignments and the challenges posed to host tracking and monitoring efforts.

Temporal Analysis: Our approach facilitated a temporal analysis of the IP address changes, enabling them to correlate specific events, such as network outages or configuration changes, with the observed IP address reallocations. This capability can prove invaluable in understanding the underlying factors driving IP address dynamics and informing strategies for more effective host tracking and network management.

The results not only demonstrate the practical application of our approach but also highlight its potential to provide quantitative insights into the phenomenon of IP address dynamics. By accurately mapping hosts across IP changes and quantifying the extent of IP address reallocations, We have contributed a valuable tool for network administrators, security researchers, and others involved in network monitoring and analysis

7 Limitaions and Challanges

While our approach has yielded promising results in mapping hosts across dynamic IP addresses, we encountered several limitations and challenges that warrant further discussion and consideration. One of the primary limitations lies in the nature of the data available on Censys.io. As we observed, most of the information appears to be related to server systems

rather than end-user hosts. This limitation may constrain the generalizability of our approach, as the characteristics and behaviors of server systems could differ significantly from those of end-user devices. Addressing this limitation will require tailoring our methodology to account for the unique aspects of end-user host tracking, which may involve incorporating additional data sources or developing specialized techniques.

Another significant challenge we faced was the lack of ground truth data for validation purposes. Without a reliable baseline or reference dataset, quantifying the accuracy of our mapping results objectively becomes a daunting task. In our study, we relied heavily on manual analysis and heuristics to assess the mapping quality, which may introduce subjective biases. Overcoming this challenge will necessitate concerted efforts in collecting and integrating ground truth data, potentially through collaborations with network administrators or by developing controlled experimental setups.

Additionally, our case study focused on a relatively smaller autonomous system (AS45326). While this allowed us to validate our approach and demonstrate its feasibility, the true test of our methodology lies in its scalability and performance when applied to larger and more complex autonomous systems. Expanding our analysis to encompass a diverse range of network environments, with varying sizes and dynamics, is crucial to assess the robustness and broader applicability of our IP mapping technique. Achieving this scalability may require optimizations to our algorithms, parallelization strategies, or leveraging distributed computing resources.

Furthermore, we acknowledge that our approach, while novel, is not without its inherent complexities. The manual data analysis and feature engineering steps can be time-consuming and resource-intensive, potentially limiting the scalability of our solution. Addressing this challenge will require exploring avenues for automating these processes, leveraging machine learning techniques or developing more efficient algorithms to streamline the IP mapping workflow. Finally, our study faced challenges related to querying and processing the large volumes of data from Censys using BigQuery. The sheer size of the data posed computational and storage limitations, and managing these resources effectively was a significant undertaking. Additionally, dealing with empty or missing values in the dataset introduced complexities that required careful handling and imputation strategies.

Despite these limitations and challenges, we remain optimistic about the potential of our approach and its ability to contribute to the development of robust solutions for tracking hosts across dynamic IP address changes. Addressing these challenges through future research and collaboration with the broader community will be crucial in advancing this field and enhancing network monitoring, security, and analysis capabilities.

8 Future Work

Unique Identifier: Expanding on our methods, a Universal Unique Identifier can be created with some matching tolerance threshold that will help map users with some limited accuracy. This will help encode available data points into an ID that can then be assigned and used to track hosts. Additionally, machine learning algorithms can be employed to enhance the accuracy of the matching process. For example, techniques such as clustering or classification algorithms can be applied to learn patterns from the data and improve the matching results.

End-user Host Tracking: Our current approach has primarily focused on server systems, as reflected by the data available on Censys.io. However, to truly address the challenge of mapping hosts across dynamic IPs, future work should explore techniques tailored specifically to end-user devices, which may exhibit different characteristics and behaviors compared to server systems.

Ground Truth Data Collection: A significant limitation of our study is the lack of ground truth data for validation purposes. Future efforts should prioritize the collection and integration of reliable ground truth datasets, potentially through collaborations with network administrators or by developing controlled experimental setups. Having access to ground truth data will enable more accurate assessments of mapping accuracy and facilitate the refinement of our approach.

Scalability and Robustness Testing: While our case study demonstrated the feasibility of our approach on a smaller autonomous system (AS45326), further research is needed to assess its scalability and robustness when applied to larger and more complex network environments. Expanding the analysis to encompass a diverse range of autonomous systems, with varying sizes and dynamics, will provide valuable insights into the broader applicability and potential limitations of our IP mapping technique.

Automation and Efficiency Improvements: Our current methodology involves manual data analysis and feature engineering steps, which can be time-consuming and resource-intensive. Future work should explore avenues for automating these processes, leveraging machine learning techniques or developing more efficient algorithms to streamline the IP mapping workflow. Automating these steps will enhance the scalability and practical utility of our approach.

Continuous Monitoring and Anomaly Detection: To enhance the practical value of our approach, future work could focus on continuously monitoring the number of unique hosts within an autonomous system and creating a graphical representation over time. By establishing a baseline and analyzing deviations from expected patterns, our methodology could be extended to identify potential anomalies or suspicious activities related to IP address dynamics. This capability could prove invaluable for network administrators and security professionals in proactively detecting and responding to

network-related threats or irregularities.

By addressing these future research directions, we aim to contribute to the development of more robust, scalable, and practical solutions to track hosts across dynamic changes in IP addresses, ultimately improving network monitoring, security, and analysis capabilities in various domains.

References

[1] DURUMERIC, Z., ADRIAN, D., MIRIAN, A., BAILEY, M., AND HALDERMAN, J. A. A search engine backed by internet-wide scanning. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security* (New York, NY, USA, 2015), CCS '15, Association for Computing Machinery, p. 542–553.

[2] INTERNET INTELLIGENCE RESEARCH LAB. Internet outage detection and analysis (ioda), 2024. Accessed: 2024-05-03.

[3] JIN, Y., SHARAFUDDIN, E., AND ZHANG, Z. L. Identifying dynamic ip address blocks serendipitously through background scanning traffic. In *Proceedings of the 2007 ACM CoNEXT Conference* (New York, NY, USA, 2007), CoNEXT '07, Association for Computing Machinery.

[4] KOHNO, T., BROIDO, A., AND CLAFFY, K. C. Remote physical device fingerprinting. *IEEE Transactions on Dependable and Secure Computing 2*, 2 (apr 2005), 93–108.

[5] KOL, M., KLEIN, A., AND GILAD, Y. Device tracking via linux's new tcp source port selection algorithm (extended version). *arXiv* (2022).

[6] PADMANABHAN, R., RULA, J. P., RICHTER, P., STROWES, S. D., AND DAINOTTI, A. Dynamips: Analyzing address assignment practices in ipv4 and ipv6. In *The 16th International Conference on emerging Networking EXperiments and Technologies (CoNEXT '20)* (New York, NY, USA, 2020), ACM, pp. 1–16.

[7] RAMANATHAN, S., HOSSAIN, A., MIRKOVIC, J., YU, M., AND AFROZ, S. Quantifying the impact of blocklisting in the age of address reuse. In *Proceedings of the ACM Internet Measurement Conference* (New York, NY, USA, 2020), IMC '20, Association for Computing Machinery, p. 360–369.

[8] VIKAS, M., PIERRE, L., ANTOINE, V., WALTER, R., ROMAIN, R., AND MARTIN, L. Don't count me out: On the relevance of ip address in the tracking ecosystem. In *Proceedings of The Web Conference 2020* (New York, NY, USA, 2020), Association for Computing Machinery, pp. 808–815.

[9] XIE, Y., YU, F., ACHAN, K., GILLUM, E., GOLDSZMIDT, M., AND WOBBER, T. How dynamic are ip addresses? In *Proceedings of the 2007 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications* (New York, NY, USA, 2007), SIGCOMM '07, Association for Computing Machinery, p. 301–312.