Predicting IPv6 Services Across All Ports

David Budziwojski[1] Manda Tran[1] Grant Williams[2] Liz Izhikevich[1]

[1] *University of California, Los Angeles* [2] *Georgia Institute of Technology*

Abstract

Recently, there has been a need to transition from Internet Protocol Version 4 (IPv4) address space to Internet Protocol Version 6 (IPv6) address space, resulting in a need to update the means of scanning to work efficiently in IPv6's increased search space. This work aims to apply probabilistic modeling to IPv6 address space, enabling abilities to analyze security as well as vulnerabilities. We leverage ideas from previous successful works in IPv4: GPS's predictive features that are hypothesized to carry over into IPv6 space. The initial step to begin verifying this suspicion comes in the comparison between Layer 4 Correspondence, which we show in this work is different between both IPv4 and IPv6. Most notably, the variety and density of port correlations unveils parts of the reason of the success of the approach in IPv4, while also showing that the sparsity of IPv6 hints to being a less meaningful predictor with the exception of ports 443 and 80.

1 Introduction

Internet-wide scanning is an important process that maps and characterizes the internet and provides a means to analyze a variety of characteristics (e.g., security, usage, speed, etc.). Until recently, the Internet Protocol Version 4 (IPv4) was the primary case study, resulting in a variety of tools, methodologies, and systems. To throw things more into twist, a significant finding in the IPv4 space while scanning for services was that IANA-assigned ports allocations were not necessarily assignments that were used in practice. That is, there were notable services assigned to various "unusual" ports, which meant that simply following the IANA assignment while looking at services hosted by a given IP would result in overlooking the services offered [3]. Nonetheless, with the development of tools like GPS [4], a probabilistic modeling approach to discovering services, there was an effective and efficient means to do so.

However, while this characterizes both the progression and success in the IPv4 domain, the rapid decrease and eventual run-out of IP addresses called for the adaptation of a new Internet Protocol: Internet Protocol version 6 (IPv6). While there are a variety of characteristics that differentiate the two versions, the most notable is the increase of length from 2^{32} to 2^{128}, increasing the search space by an enormous amount. Although this enables more IP addresses to be assigned, fixing a significant issue of its predecessor, this once again increased the difficulty of the Internet-wide scanning task. With more than 40% of users beginning to adopt IPv6 [2], and this number continuing to grow rapidly each year. Now more than ever, there is a needs to be an approach to efficiently scan in IPv6 that finally tames the large search space and allows larger scale analysis of the domain. Moreover, this issue does not stop there because simply detecting active IP addresses is just the first step.

The search space again grows by a factor of 2^{16} when adding ports and services into the mix. Nonetheless, it has been shown that probabilistic modeling has proven successful in leveraging correlations to detect ports and services despite the "unusual" assignments in IPv4.

In this work, we introduce a preliminary study on IPv6, specifically focusing on setting the groundwork to efficiently scan in IPv6 space by analyzing IPv4 probabilistic methods in the context of IPv4. We focus primarily on a the **Layer 4 Correspondence** (1) and the subtle intricacies present in the comparison from IPv6 to IPv4.

2 GPS: An IPv4 Success Story

The growth of scanning-systems over the last 20 years has been enormous, gradually getting more and more fine-tuned. Most recently, **GPS** took a step back from the recent trend towards utilizing a machine learning approach to predict services, reverting to a simpler (but successful) approach: a probabilistic model-based approach. The system introduces simple conditional probabilities that leverage "intuitive" predictive features to predict services across all IPv4 ports.

The system utilizes a 4 step pipeline:

1. **Building seed data**
2. **Identifying predictive patterns** to create a probability mapping
3. **Predicting the first service**
4. **Finding additional services** using the built probability mapping

When implemented using ZMap, ZGrab, LZR, and BigQuery, the system is able to leverage both effective scanning and computations of predictive patterns. The system was found to discover 92.5% of services with 131 times less bandwidth and 204 times more precision compared to the brute force scanning. Furthermore, when compared to leading scanning systems, GPS out performs them while simultaneously utilizing less bandwidth [4].

2.1 Conditional Probabilities

GPS relies on conditional probabilities that are empirically calculated using the port numbers, autonomous system (AS) numbers, and application fingerprints for the seed data to create the probability mappings used to later predict other services on different IP addresses. Formally, the four conditional probabilities are defined as:

1. Layer 4 Correspondence:

$$P\big(\mathrm{Port}_a \mid \mathrm{Port}_b\big) \tag{1}$$

2. Layers 3 and 4 Correspondence:

$$P\big(\mathrm{Port}_a \mid \mathrm{Port}_b, \mathrm{AS}_b\big) \tag{2}$$

3. Layers 3 and 7 Correspondence:

$$P\big(\mathrm{Port}_a \mid \mathrm{Port}_b, \mathrm{App}_b\big) \tag{3}$$

4. Layers 3, 4, and 7 Correspondence:

$$P\big(\mathrm{Port}_a \mid \mathrm{Port}_b, \mathrm{App}_b, \mathrm{AS}_b\big) \tag{4}$$

3 Methodology

This work focuses on a **predictive comparison** of the Layer 4 Correspondence (1), utilizing both a raw and weighed comparison to uncover underlying structures and biases present in the datasets.

3.1 Data

Datasets for IPv4 and IPv6 were both acquired from LZR and LZRv6, respectively, where the former was scanned over 42 ports while the latter was scanned over 55 ports. The data was stored in Google BigQuery[1], allowing easy use of query-based joins, filtering, and de-duping. The IPv6 dataset was filtered much more substantially compared to IPv4 due to 2606:3dc0::/32, 2a05:bb80::/29 and 2620:0025:6000::/48 being strangely behaving aliases that could be ignored.[2] This resulted in a substantial decrease in number of unique IP addresses in IPv6 to 26 million, which compared to IPv4's 134 million is around 5 times more. Furthermore, due to IPv6's port scans being a superset of IPv4's port scans, the 13 additional ports were removed from the dataset so that a more direct comparison could be done between the two (again decreasing things, but this time for both cases).

3.2 Predictive Comparison

Two forms of comparisons were used on the dataset:

1. Raw Comparison
2. Weighed Comparison

The raw comparison simply investigates characteristics without any transformations and seeks to find apparent trends and biases within both the raw data (no correlations computed) and the computed Layer 4 Correspondence.

The weighed comparison comes about because of a key feature of the computed Layer 4 Correspondence: the sample size used to compute the given probability of a port-to-port relationship. A favorable situation is one where the sample size is large, indicating a sense of reliance; however, that is not always the case. There are many instances were probabilities are close to 1 or 0 simply because there are not many instances of the given relationship. Thus, there is a need to quantify these probabilities by some factor to allow their values to be used in more meaningful ways. This is done by weighing the probabilities by their sample size and then normalizing their values. Formally, the weighed probability is calculated as follows,

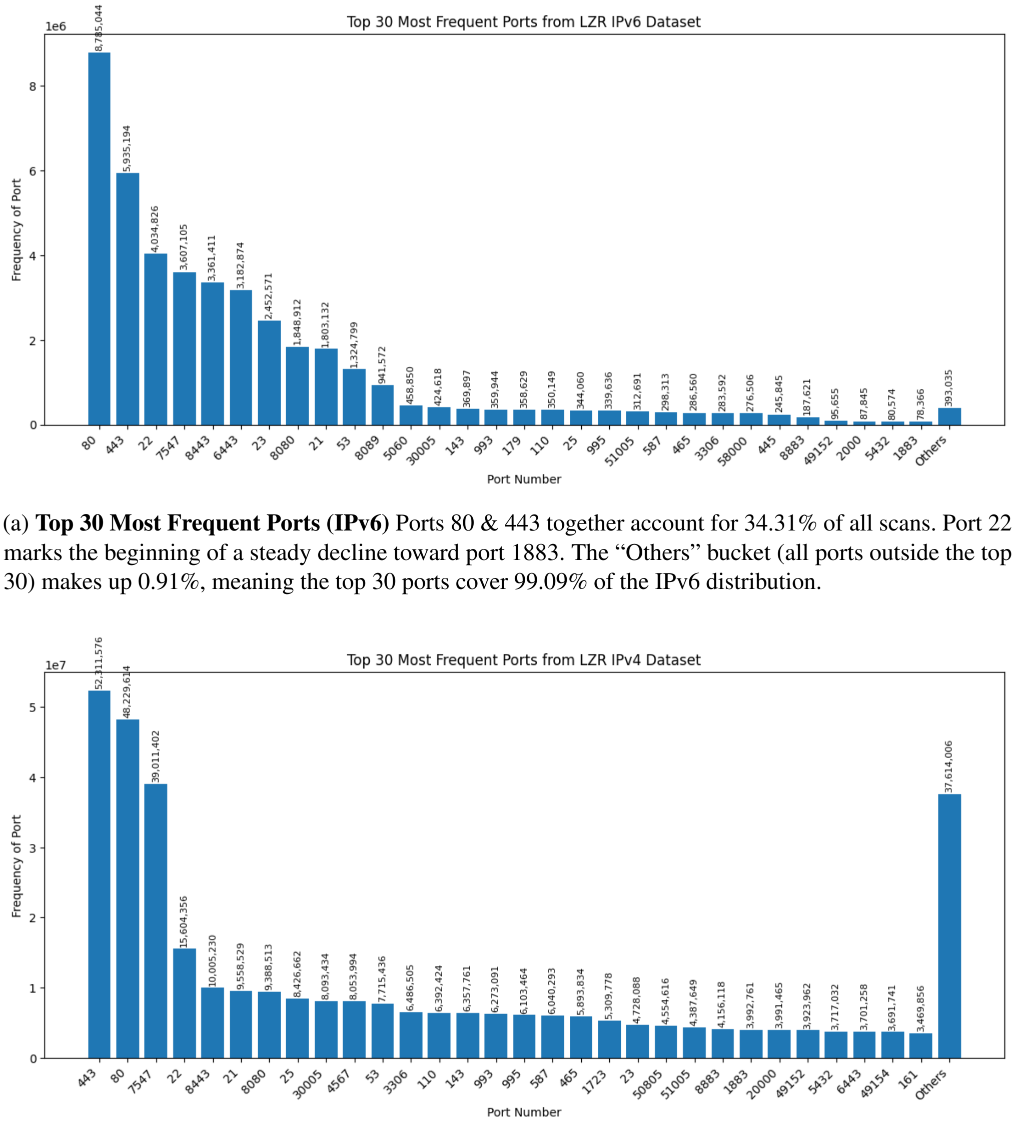
$$P_{weighed}(a|b) = \frac{P(a|b) \cdot log(|b|)}{log|B|} \tag{5}$$

where $|b|$ is the sample size of b and $|B|$ is the max sample size found in the computed correspondence (e.g. the largest sample size in IPv4 or IPv6 respectively).

Now that the values are weighed by their "reliance" and normalized by the largest "reliance" factor, the probabilities

[1] Google BigQuery is a serverless database that parallelizes computations efficiently and was actually used to compute the other probabilities as well

[2] BigQuery was found to take a long time when trying to match prefixes that needed to be filtered, which resulted in the use of 6Sense's AS tool (using pyasn and pytricia).



(a) **Top 30 Most Frequent Ports (IPv6)** Ports 80 & 443 together account for 34.31% of all scans. Port 22 marks the beginning of a steady decline toward port 1883. The “Others” bucket (all ports outside the top 30) makes up 0.91%, meaning the top 30 ports cover 99.09% of the IPv6 distribution.

(b) **Top 30 Most Frequent Ports (IPv4)** Ports 80 & 443 together account for 28.98% of all scans. Port 22 again begins a steady decline toward port 161. The “Others” bucket composes of 10.84% of the scans, illustrating the top 30 ports cover 89.15% of the IPv4 distribution.

Figure 1: Comparison of the Top 30 most-frequently scanned ports in the LZR IPv6 (a) and IPv4 (b) datasets.

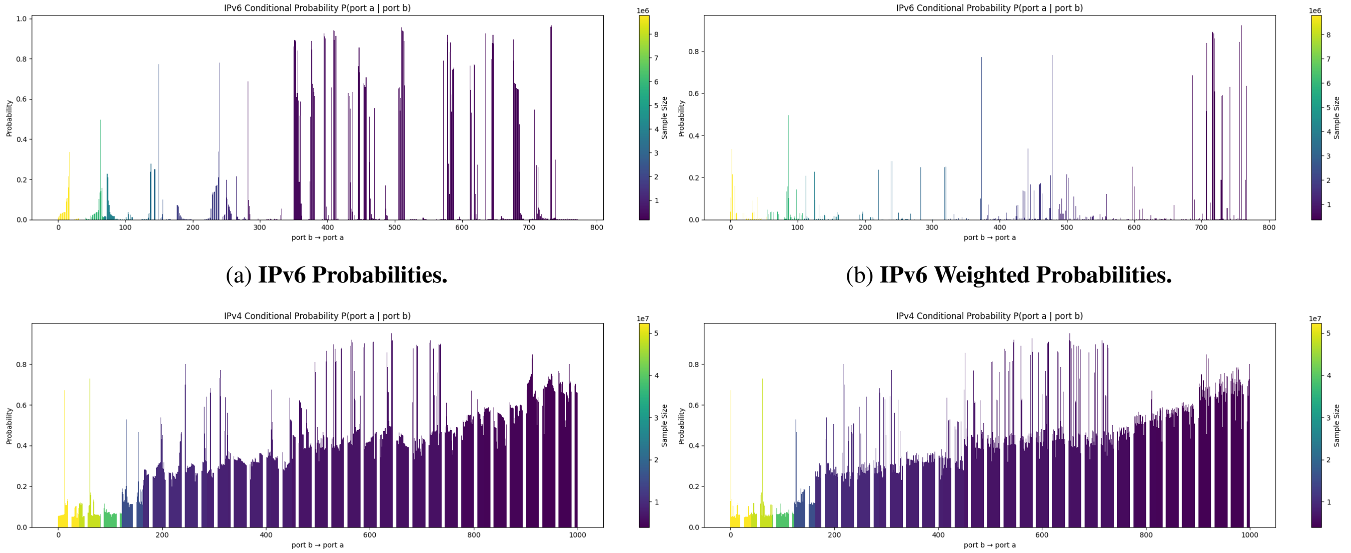
can be take closer to face value. That is, comparisons between IPv4’s and IPv6’s Layer 4 Correspondence are done in a classical statistical approach, directly comparing the conditional probabilities that were successful in GPS on the entire datasets of IPv4 and IPv6 to see if the predictive features used previously will carry over to IPv6 successfully.

4 Comparative Evaluation on IPv4 and IPv6

In this section, we conduct an initial comparison of port distribution between the IPv4 and IPv6 datasets and then continue by identifying notable differences in the Layer 4 Correspondence from both the plain and weighed probabilities.

4.1 Port Distributions

A meaningful comparison between the IPv4 and IPv6 datasets is one that looks at the distribution of the ports across both datasets. In Figures 1a and 1b, we find distributions of ports across both IPv6 and IPv4. The general the leftward distribution stemming from ports 80 and 443 and the rapid tail-off towards the "Others" category is present in both cases. There is a little more of an abrupt drop in IPv4 compared to IPv6



(a) **IPv6 Probabilities.**

(b) **IPv6 Weighted Probabilities.**

(c) **IPv4 Probabilities.**

(d) **IPv4 Weighted Probabilities.**

Figure 2: **Port-probability comparison for both IPv6 (top row) and IPv4 (bottom row)** Each plot is sorted according to the "reliability" of their data (the sample size used to find the probability). While the weighed probabilities do not technically need to be sorted by sample size since the values are scaled, sorting them makes the comparisons easier. (a) IPv6 scan probabilities (b) IPv6 weighted scan probabilities (c) IPv4 scan probabilities (d) IPv4 weighted scan probabilities

which tails off more slowly for the top 8 ports, however, the "Others" category is much larger relatively for IPv4 compared to IPv6. The heaviness towards 443 and 80 being maintained in IPv6 makes sense because they typically map to http and https, which remain some of the most commonly hosted services (and really is not something that is expected to change). Similarly, the fact that the first 10 ports are almost the same set (with the exception of a couple) hints again that the uses (and services being hosted) in IPv6 and IPv4 resemble each other.

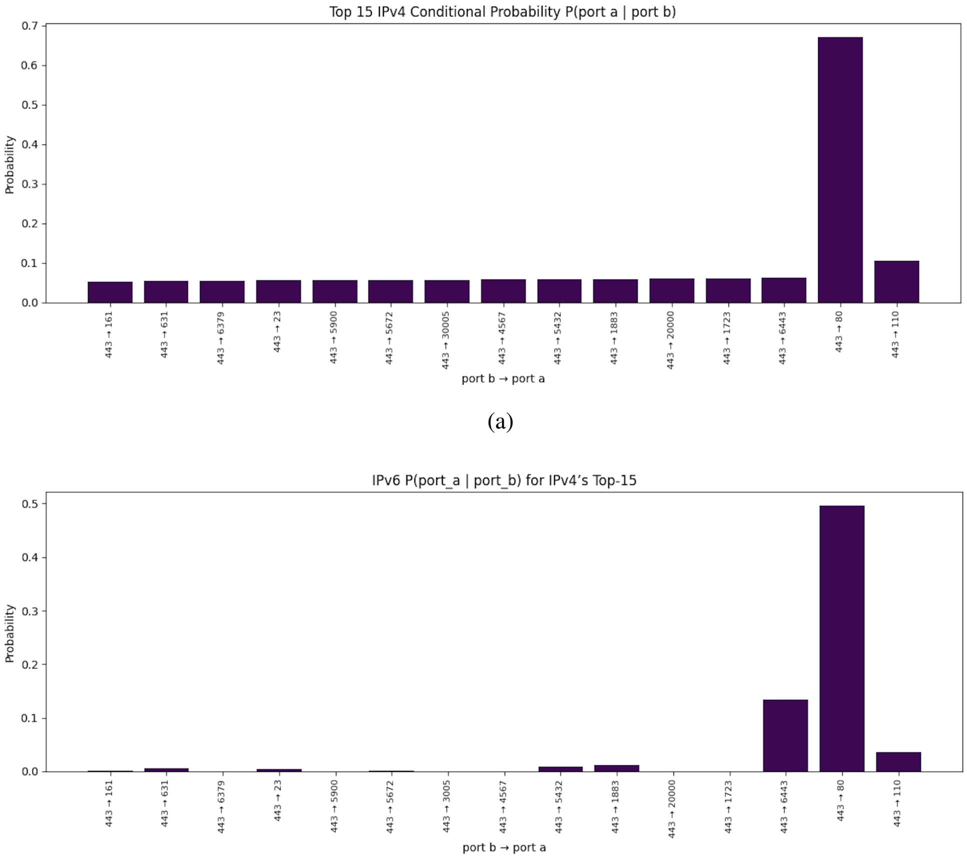
4.2 Probability Comparison

We look at two forms of the probabilities: the plain probabilities and the weighed probabilities. In figure 2, we plot both forms of the probabilities for both IPv6 and IPv4, where all of them are ordered by the sample size (an indicator of reliance) in descending order.

Figures 2a and 2c present the plain probabilities, where we already see a clear difference between IPv4 and IPv6. Firstly, the scale of the sample size between the IPv4 is a full magnitude more than IPv6—which makes sense because the dataset was smaller, but also indicates the sparsity of IPv6 compared to IPv4. Moreover, the gradual growth of the probability from the yellow region to the purple region for IPv4 is much more subtle compared to IPv6, which appears earlier on and more sporadically compared to the linear growth in IPv4. This illustrates that IPv6 has less predictive port-to-port relationships compared to IPv4, which has many more correlations despite some even occurring on less commonly used ports.

Figures 2b and 2d present similar trends seen in the plain probabilities, but with a couple notable changes. The IPv6 Weighed Probabilities have several large spikes disappear due to their respective weight getting overpowered by the normalization of the log of the largest sample size. Furthermore, in the cases of the more "unreliable" spikes that haven't disappeared (e.g. the ones towards the back), the probabilities have somewhat decreased. While on the other hand, the more reliable cases (e.g. the ones closer to the front), the probabilities have practically not changed much at all because their multipliers are closer to 1. The IPv4 Weighed Probabilities are even more similar to their plain probabilities. While we see the same effect of the "unreliable" regions getting weighed less and reliable regions getting weighed by values closer to 1, the fact that the scale of figures 2c and 2d is a magnitude larger than figures 2a and 2b means that in general the "unreliable" case is still somewhat more reliable in many cases found in IPv6. Nonetheless, we see that the weighed plot ends up a bit smoother and linearly increases nicely with some of the more outlying jumps getting settled down.

We see that in the IPv6 case that many of the larger probabilities come from fairly low sample sizes and are characteristics that should not be trusted very much. While IPv4, which inherently comes from a larger sampling set, remains more trusty and indicates more instances of different variety of port-to-port relationships (the reason for less sparsity in the plot), characterizing a clear distinction on the types of port-to-port correlations present in the two datasets.



(a)

(b)

Figure 3: **Comparison of the Top 15 most-frequently scanned ports in the LZR IPv6 (a) and IPv4 (b) datasets** Figure (a) simply is the zoomed part of Figure 2a, illustrating the 15 most popular ports in IPv4 based on sample size. Figure (b) is the same ports illustrated in IPv4 except from IPv6. While the probabilities are not from the upmost top of Figure 2a, the values are very close to that point (port 443 comes after port 80 in terms of sample size for IPv6) and can be considered reliable.

We further investigate the distinction between the types of relationships by zooming in on 2c and examining the Top 15 Ports for IPv4 (Figure 3a) then comparing those same ports for IPv6 (Figure 3b)[3]. We see that in Figure 3a that most probabilities are in the range between 0.04 and 0.08. The exception being $P(80|443)$ which is almost 0.7, and illustrates the common notion of setting up both http and https as services. While the values are fairly low, it has been shown in [4] that low probabilities are not necessarily a bad thing and can still be used quite well for predicting other services. When we plot the same port relationships for IPv6 (port 3005 had to be manually added and set to 0 because no instance occurred in the dataset), all values except for $P(80|443)$ and $P(6443|443)$ are very close to zero (some of which practically do not appear visible on the plot). Compared to IPv4, $P(80|443)$ decreases from just under 0.7 to slightly under 0.5 in IPv6.

Ultimately, the similarities and differences presented in the previous figures indicate a couple unique trends about the type of services hosted. Firstly, we see that the popularity of port 80 and 443 is still quite notable in IPv6 despite it becoming a slightly weaker predictor. Secondly, IPv6 has less instances of port-to-port relationships compared to IPv4. For instance, in the 443 case seen in Figure 3, there are many more non-negligible correlations in IPv4, hinting at the idea of multi-hosting is fairly popular. On the other hand, IPv6 sees the multi-hosting behavior less frequently as even some of the most popular ports do not have strong correspondences with other less popular ports. It is likely that this difference comes from a combination of the fact that IPv6 is newer and still changing, it has different setups, or that the data has some form of sampling bias since the dataset is fairly small.

[3] Since we are looking at the top part of the plots (the reliable part), using the weighed or plain probabilities does no really change anything.

5 Discussion

The work focused primarily on the Layer 4 Correspondence and a little on the port distributions, but there are still the other three correspondences (2), (3), and (4) that were not analyzed. Thus, some future work could be a similar analysis on them, which would provide further insights and would characterize more completely if the probabilistic modeling used in IPv4 carries over fully to IPv6. Additionally, a work that properly scans or psuedo-scans ports to verify the GPS system would also further provide insights about the system in the IPv6 realm. Specifically, comparisons between the usage of probability mappings, seed size, and more could be analyzed more carefully. Furthermore, additional work can be done on discovering which predictive features should be used, specifically because IPv4 and IPv6 appear to be more divergent than initially thought.

Moreover, there has to be some caution taken when looking at the results because IPv6 is still emerging (which likely is reason for the dataset samples being around 5 times less for IPv6 compared to IPv4), meaning that some of the findings could change depending in the manner in which ports (and services) in IPv6 are utilized by future users. Moreover, the attempt to weigh things based on their respective probability dataset (e.g IPv4 was weighed using the max size of a sample from IPv4) means that technically things were not scaled to the exact same point. However, this choice stems from the fact that IPv4 is presently still more widely used, but this decision may need to be changed in the future when both are used by a more similar amount (a threshold that in its own right must be decided).

6 Related Works

Since the dawn of the internet, there has been an necessity to understand the reality of the internet. This gave rise to some of the earliest forms of network measurement utilities and tools like ping, traceroute, etc. However, it quickly became clear, especially with the advent of countless new platforms, applications, and services being hosted online that more rigorous means of measurement were required; which led to the development of many systems that enabled IPv4 to truly be measured and analyzed.

6.1 Current State of IPv4

Developments in IPv4 have come very far and has become a backbone for various other forms of internet measurement-related fields.

ZMap, a stateless packet scanner system, revolutionized the scanning world by enabling the entire IPv4 space to be examined in less than 45 minutes [1] and has become an integral to various other scanning systems. Moreover, then with the advent of**LZR**, a system that significantly increased the effectivity of application-level scanning and is capable of identifying 99% of identifiable services despite possible unexpected assignments [3], another decrease in time required to scan occurred.

6.2 IPv6: A Similar but Different Task

However, when it comes to IPv6, the scanning methods have to take a step back. IPv6 scanning finds itself in a similar state to the beginnings of IPv4 scanning, where a current issue is actually detecting what IPs are actually responsive. Unlike IPv4, which currently (or almost) is fully populated, responsiveness is not an issue. As such, Target Generating Algorithms (TGAs) are a must and currently still is an open investigation.

Nonetheless, **6Sense** provides one of the first complete systems for end-to-end internet-wide scanning in the IPv6 space that enabled a comprehensive introduction to security applications that were only hinted at previously [5]. Moreover, we are beginning to see systems like ZMap and LZR get "refactored" for IPv6, which has proven to be successful (with the addition of some changes to the system). Thus, while IPv6 is in its early stages, the benefit of having IPv4 as a predecessor allows for ideas to be reused and modified to better fit the IPv6 space.

Nonetheless, due to the still recent emergence of the field and the bottleneck of finding responsive IPs, there has been little analysis on services and their distributions across ports in IPv6. While 6Sense touches very briefly on port distributions, there has been no work that compares IPv6 services in a manner of detail seen in [4]. With the growth of scanning ability in IPv6, it is gradually beginning to be possible to learn more subtleties as well as to take the first steps of determining if GPS's approach is translatable to IPv6.

7 Conclusion

We show that despite Layer 4 Correspondence being a crucial feature to successfully predict other active ports in GPS that the correspondence is quite different between IPv4 and IPv6. The former has a higher quantity of correlations (even between less common ports) compared to IPv6 which has a significantly lower quantity. That is, IPv6 is likely to have trouble predicting other ports when relying solely on Layer 4 correspondence, except when looking at very popular ports like 80 and 443 (but even then the sparsity aspect of IPv6 impacts things quite a bit). Additionally, we begin to understand the underlying structure of the success of the predictive features in IPv4, where quantity and sample sizing contributes to the stability of the predictability of port-to-port correspondences.

References

[1] Zakir Durumeric, Eric Wustrow, and J. Alex Halderman. Zmap: Fast internet-wide scanning and its security applications. In *Proceedings of the 22nd USENIX Security Symposium*, Washington, D.C., USA, August 2013.

[2] Google. Ipv6 adoption statistics. `https://www.google.com/intl/en/ipv6/statistics.html`, 2025.

[3] Liz Izhikevich, Renata Teixeira, and Zakir Durumeric. Lzr: Identifying unexpected internet services. In Michael D. Bailey and Rachel Greenstadt, editors, *Proceedings of the 30th USENIX Security Symposium*, pages 3111–3128. USENIX Association, 2021.

[4] Liz Izhikevich, Renata Teixeira, and Zakir Durumeric. Predicting ipv4 services across all ports. In *SIGCOMM '22: Proceedings of the 2022 ACM SIGCOMM Conference*, SIGCOMM '22, Amsterdam, Netherlands, August 2022. Association for Computing Machinery.

[5] Grant Williams, Mert Erdemir, Amanda Hsu, Shraddha Bhat, Abhishek Bhaskar, Frank Li, and Paul Pearce. 6sense: Internet-wide ipv6 scanning and its security applications. In *33rd USENIX Security Symposium (USENIX Security 24)*. USENIX Association, 2024.