ON THE MULTIFRACTAL STRUCTURE OF OBSERVED INTERNET ADDRESSES

by

MEGAN WALTER

A THESIS

Presented to the Department of Computer and Information Science
and the Robert D. Clark Honors College
in partial fulfillment of the requirements for the degree of
Bachelor of Science

May 2022

An Abstract of the Thesis of

Megan Walter for the degree of Bachelor of Science
in the Department of Computer and Information Science to be taken June 2022

Title: On the Multifractal Structure of Observed Internet Addresses

Approved: *Reza Rejaie, PhD*
Primary Thesis Advisor

As a result of society's increasing dependence on the Internet, we observe a significant increase in Internet attacks and network management issues. However, the growing speed and volume of Internet traffic makes finding portions of traffic responsible for creating problems difficult. Current approaches to classifying connections tend to regard each connection independently of one another. However, the nature of Internet Protocol (IP) addresses points to correlations between addresses located in similar parts of the IP address space. Understanding the structural characteristics of the IP address space could lead to novel ways to create network management algorithms that deal with aggregates of flows.

We examine the structure of observed IP addresses in network traffic collected from border routers at the University of Oregon. Previous work indicates that the characteristics of observed IPv4 address structures are consistent with a multifractal model. We work to solidify the existence of this multifractal structure and provide an initial contribution toward the development of network security and management solutions that aggregate flows by IP address. We use a new method of multifractal analysis using the method of moments to produce an initial characterization of how observed IPv4 addresses relate to one another. We apply this process across traffic samples representing three different timescales, allowing us to look at the temporal dynamics of these multifractal characteristics.

Acknowledgements

Thank you to Dr. Reza Rejaie and Chris Misa for their incredible support and guidance that they provided during the research process for this thesis. This process has been a long and rewarding journey largely thanks to their insight. I would like to thank them as well as Dr. Lindsay Hinkle from the Clark Honors College for serving on my Thesis Committee. I am very grateful to the Honors College for putting me in a position where I have received the support that I have, from mentors and peers alike. The support network that I built in the Honors College and the wider University of Oregon went a long way in helping me persevere throughout all of the challenges that I faced. Finally, I would like to thank my family for their unconditional love and support, and for encouraging me through the pursuit of my degree.

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

The Internet is a world-wide network of computers that interconnects billions of devices so that they can send and receive data from each other. As a society, we have become increasingly dependent on the Internet for many aspects of our daily lives. Accordingly, we can observe a significant increase in Internet attacks and a growing number of network management issues. Considering the many severe consequences that these security issues could lead to, there must be a focus on both detecting and mitigating these events in a timely manner. However, the growing speed and volume of Internet traffic makes finding portions of traffic responsible for creating problems or launching attacks incredibly difficult.

Current approaches to network monitoring tend to regard each individual connection (also known as a flow) independently of all other connections. Despite this, there are many reasons to believe that there are correlations between flows, particularly when considering them in terms of their Internet Protocol (IP) addresses. Understanding the structural characteristics of benign traffic is critical to developing network monitoring techniques as this knowledge can allow for easier detection and mitigation of non-benign connections. For example, a network could implement an IP-whitelisting system to automatically let through connections that are already known to be benign, leaving more resources available to focus on the monitoring of the rest of the traffic passing through the network. However, the structure of general Internet traffic is not currently well understood.

Prior studies in this area of research have shown that the IP addresses of observed Internet traffic exhibit various fractal and multifractal characteristics [Barford et al., Kohler et al.]. That is to say, at multiple spatial scales, IP addresses with shared characteristics irregularly cluster together within the address space. These studies were able to provide mathematical multifractal models to describe this clustering behavior and the structural characteristics of observed traffic. In practice, these models could be applied to simulations of real-world traffic, reducing activity from unwanted sources, and other Internet security tasks [Barford et al., Kohler et al.]. However, these prior works were conducted prior to 2010 and suffer from issues regarding noisy analysis methods and datasets that are considered small and outdated by current standards.

To this end, this thesis tries to address the following key research questions:

- Is the structure of IP addresses observed on a modern campus network still consistent with the multifractal hypothesis set out in prior studies?
- How does the duration of observation (i.e., the duration over which IP addresses are collected) impact the multifractal structure of observed addresses?
- How does the multifractal structure of observed addresses evolve over time?

We want to provide additional evidence supporting the development of network security and management solutions based on the observed Internet address structure by solidifying the existence of this multifractal structure of the observed address space. In addition, we want to contribute an initial analysis of how this structure may evolve depending on different temporal variables.

To accomplish this, we (i) considered and relied on a large set of real-world network traces collected over 9 years at the University of Oregon over three different time scales (24 hours, 60 days, and 8 years) using three different observation durations (30 seconds, 30 minutes, and 24 hours), (ii) developed an efficient program to quickly and accurately extract the spatial address structure of the data over each temporal and spatial scale, (iii) examined the different characteristics of the resulting multifractal structure using a new multifractal analysis method, and (iv) compared this structure across the defined temporal dimensions.

This work differs from prior studies primarily in our multifractal analysis technique, our dataset, and our final observations. In particular, we describe and implement a novel method of multifractal analysis using the method of moments which provides more robust, less noisy results in the context of IP addresses. The technique is suited for the discrete approximation required for IP addresses as opposed to the analysis method used in prior studies which approximates continuous measures on infinite scales. Our complete dataset is far larger in both magnitude of data and timespan over which the data has been collected. Additionally, we perform our analysis on several different samples from our dataset. Finally, from our results, we analyze the effect of the duration of observation on multifractal characteristics and how those characteristics evolve over time.

Through our analysis, we are able to demonstrate that the address structure of observed traffic at the University of Oregon is still consistent with the multifractal hypothesis set out in prior studies. Our results indicate that the duration of observation has a minimal effect on the multifractal characteristics. We also found that the evolution of multifractal characteristics varied depending on which time scale the multifractal analysis was performed on. While the samples along the 60-day time scale showed no visible trends of how multifractal characteristics changed over time, analyzing the 8-year time scale indicated a shallow trend toward more multifractal over the years.

Similarly, the samples along the 24-hour timescale indicated a shallow trend towards more multifractal during times of day when more users tend to be connected to the network.

Chapter 2: Background

We start by presenting a set of terms that are critical for the rest of this thesis as follows:

- A **flow**, or **connection**, is a collection of data packets flowing between two end systems (computers) across the Internet.
- An **IP Address** is an identifier associated with a data packet that indicates the source or destination device of the packet. Each packet has both a source and destination address, as well as a source and destination port number which identifies the specific process/application the packet is associated with. We will be specifically working with version 4 IP addresses (**IPv4)**, which is a 32 binary bit representation of the identifiers.
- An **address prefix** refers to a collection of addresses that share their first *p* bits. The addresses under a prefix are considered adjacent to each other in the address space. In standard notation, a "/24 prefix" signifies that the block of addresses under that prefix share their first 24 bits. As the individual addresses in the block are distinguished by their final 8 bits, there are 2^8 = 256 addresses in the block of a /24 prefix. Similarly, a "/8" prefix block contains addresses that share their first 8 bits. There are 2^{24} = 16,777,216 addresses under a /8 prefix. The concept of an address prefix is further explored in Figure 1.

As the Internet grows, it becomes important to think about aggregates of flows as opposed to individual flows. This is due to a general increase in heavy volumes of Internet traffic and the number of devices connected to the Internet. As these increases occur, the time it takes to run security algorithms also increases and runtime becomes a limitation of performance. Reimagining these algorithms to increase their spatial awareness and deal with an aggregation of flows instead of individual flows combats this issue by reducing the complexity of the program.

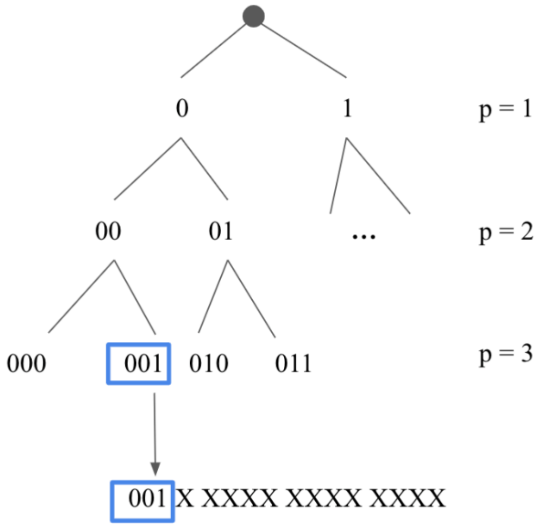

Figure 1: IPv4 Prefix

An IP prefix is a sequence of bits that is shared by a set of addresses and precedes the rest of the address. Each set of addresses with a prefix of length *p* can be considered a subset of the set of addresses with the same prefix up until the *p - 1* bit. In the figure, the addresses with the prefix 001 are also considered to be in the set of addresses with the prefix 0 and 00.

Aggregations defined by IP address prefixes are particularly useful because of the nature of IP address assignment [Kohler et al.]. IP addresses are usually allocated by the Internet Assigned Numbers Authority (IANA) in blocks of adjacent addresses to organizations. The organization will be assigned a block of addresses with a shared address prefix, and every address in that block then exists under the organization's jurisdiction. The individual addresses can then be distributed to sub-organizations or devices. If we partition addresses by their network prefixes, we can generally determine information about relative geographic locations and sub-networks that the addresses belong to [Kohler et al.]. This makes IP addresses a particularly useful criteria of aggregation when trying to classify malicious IP addresses as poorly managed sub-networks can result in a complete infection of hosts within the allocated address block [Barford et al.]. At a high level, this means that we can cluster IP addresses based on their prefixes to determine the chance that an address is a part of malicious traffic based on correlations between addresses.

NetFlow

NetFlow is a utility for traffic measurement in network routers. In order to collect the data, the collecting router sets up a cache to keep track of ongoing flows that pass through it. The router accumulates information about each flow before eventually reporting the data in the form of a record. A record is pushed from the cache to the output file after either (i) a packet header indicates that the flow has ended, (ii) a certain amount of time has passed since the last relevant packet was seen by the router, or (iii) if the cache has run out of space [Estan et al.]. These records contain information about the start and end times of the flow, the source and destination IP addresses and port numbers to indicate the flow's direction, and the number of packets and bytes transferred through the flow.

Fractal and Multifractal Properties

We will define a fractal, or monofractal, to be an infinite pattern that is self-similar at different scales. These sets are too complex to be described by classical geometry, however, such a system can describe a fractal dimension which indicates the complexity of the pattern at different scales [Falconer]. The Cantor Set is a classic example of such a fractal system (Figure 2).

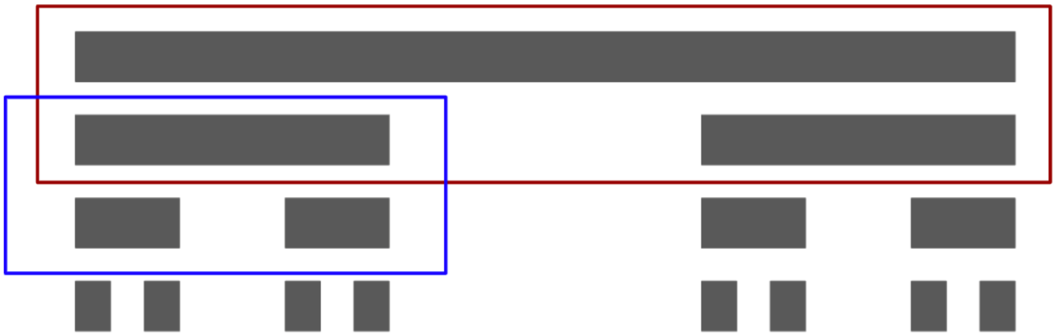

Figure 2: The Cantor Set

The Cantor Set is constructed recursively by removing the middle third segment at each scale. The set is self-similar, with several of the same patterns appearing at different scales throughout the set. The pattern highlighted by the red box is repeated on a smaller scale in the area highlighted by the blue box. The Cantor Set can be described by its fractal dimension $\log_3(2)$.

A multifractal system is a generalization of a monofractal system for which a single fractal dimension is not sufficient to describe the scaling behavior of the pattern. Instead, a multifractal can be described by a spectrum of fractal dimensions, otherwise referred to as the multifractal spectrum of the set [Harte, Salat et al.].

Chapter 3: Prior Work

There have been a few prior studies that explore the multifractal behavior of Internet traffic.

Studies conducted in the mid-2000's used statistical processes to present evidence for the idea that the distribution of addresses in the address space follows multifractal behavior. The multifractal models in these studies were targeted towards realistically simulating traffic data to test security algorithms, improving the performance of IP-based whitelisting systems, and real-time monitoring and detection systems for unwanted traffic [Barford et al., Kohler et al.].

These studies performed Multifractal Spectra statistical analysis through the Histogram Method [Barford et al., Kohler et al.] to demonstrate multifracticality within their test data. While this method has a well-documented history of use, in the context of the address space, it does not give a particularly robust result in the context of IP addresses in the address space. In particular, theoretical multifractal models, such as the Cantor Dust set used in other studies [Kohler et al.] have infinitely many scales whereas the scales of the IP address space are defined by the number of prefix lengths and thus are limited to 32. The discrete number of scales means that it is difficult to fit the data to a multifractal spectrum and that attempting to do so results in a substantial amount of noise.

Another issue with prior studies is that the data sets used are considered too small, outdated, and generally obsolete. For example, Kohler et al. sampled a set of fewer than one million distinct traces that occurred between 1998 and 2001 and Barford et al. sampled just over 10 million addresses that occurred over the course of 7 days in

2004. Furthermore, some characteristics of Internet traffic might have significantly changed over the past decade due to major changes in the popularity of network applications (e.g. peer-to-peer applications [Memon et al.]) and widespread adoption of cloud services [Yeganeh et al]. This raises the question of whether the findings of prior studies are obsolete.

Although these studies provided evidence suggesting multifractal behavior, the age of the studies, the outdated, small datasets, and the noisy analysis methods all contribute to the results of these studies needing additional support.

Chapter 4: Datasets and Data Processing

In this study, we use 8-years worth of unsampled NetFlow data that is collected at the border router of an eyeball Autonomous System [Rasti et al], namely the University of Oregon campus network [Yeganeh et al.]. In our dataset, all IP addresses associated with the University of Oregon have been anonymized to ensure the privacy of related users.

From our collection of NetFlow records, we used the given start and end times of the flows to curate sets of distinct IP addresses that were active during a defined window of time. A flow and its address were considered to be active if the absolute start and end times of the flow indicated the flow was transferring data during any time that intersected the sample window. The process is illustrated further below in Figure 3. This algorithm was built off of a framework provided by the nfdump C library, a popular open-source library for processing NetFlow data [Haag].

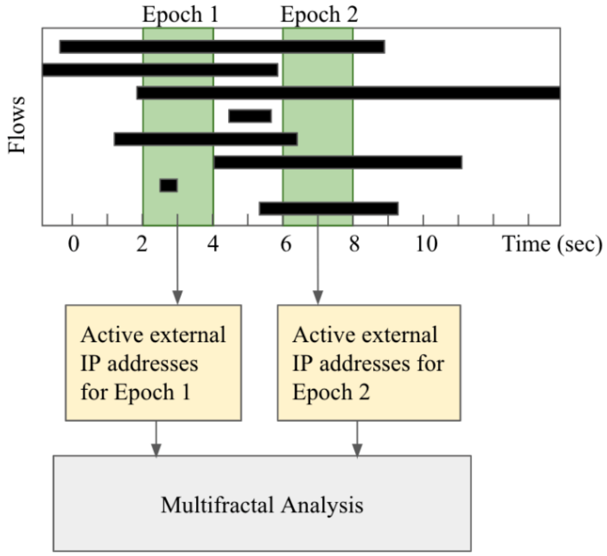
Epoch 1
Epoch 2
Flows
0
2
4
6
8
10
Time (sec)
Active external IP addresses for Epoch 1
Active external IP addresses for Epoch 2
Multifractal Analysis

Figure 3: NetFlow Processing Algorithm

Given a set of flow-level data (represented by black bars), and epochs of collection (indicated by green regions), a flow is considered active during an epoch if the run time of the flow overlaps with the duration of the epoch. In the figure above, any flow that overlaps with a green region is considered active during that epoch. The distinct external IP addresses of any active flows are then extracted into sets and used as an input for multifractal analysis.

Challenges with Using NetFlow Data

The NetFlow data set consists of five-minute bins of NetFlow records. These time bins theoretically contain all flows that passed through the router during that five-minute period. However, in practice, flows are often pushed into time bins several minutes after they were last considered active. This is a result of (i) how records are released from the cache into the time bins and (ii) records only being released from the cache once connections have been marked as terminated. In the case that a flow spans several time bins, it will only be pushed to the final bin. This behavior prevents duplicate records from appearing in several files, however, when determining which flows are active during a certain time period, we need to be able to account for this behavior.

Another challenge that arises is the possibility of flows being split into more than one record. In the case that the NetFlow cache is too full, and subsequently emptied, several flows get cut off and written into records before the rest of their data is collected into separate records.

Addressing Challenges with NetFlow Data

To address the issue of unsorted flow records, we processed more time bins than would seem necessary based on the bin labels. By including more bins, we were able to include a larger portion of flows that were actually active during a time window. However, we were unable to do this for several of the earlier dates in our samples as only a select set of dates in our dataset were archived prior to 2016. As a whole, we did not notice any significant impact of this problem, but we did keep it in mind during our analysis.

The issue of fragmented flows fortunately did not affect our process as we only counted distinct IPv4 addresses. Duplicate appearances of external addresses did not lead to any duplicate appearances of addresses as we only looked at unique addresses.

Sampling Methods

We chose three timescales to take samples over with the intention of comparing the multifractal property of the data across periods of varying lengths (Table 1).

Time Scale	Epoch Length	Sample Size
8-year	24 hours	33
60-day	30 minutes	40
24-hour	30 seconds	40

Table 1: Time Scales and Sample Size

Epoch collection duration and sample size associated with each defined time scale.

8-year Time Scale

The first timescale spans eight years from 2014 to 2022 and takes advantage of the large amount of data that we have access to. From this range, we selected a sample of dates over the years using systematic random sampling. We randomly selected a single date from the first year and then continuously chose dates at an interval of approximately a quarter year from that starting date. This resulted in a sample of 33 days over which we collected the set of IP addresses from the full 24-hour period of the day. One peculiarity to keep in mind with this timescale is that during the earlier years of data collecting, data was not collected every single day. While dates were initially chosen given a regular interval of 93 days, this was modified slightly to be a range of 92-94 as we had to adjust to the dates that were available in our database. The characteristics for each sample in this time scale are detailed below in Figure 4.

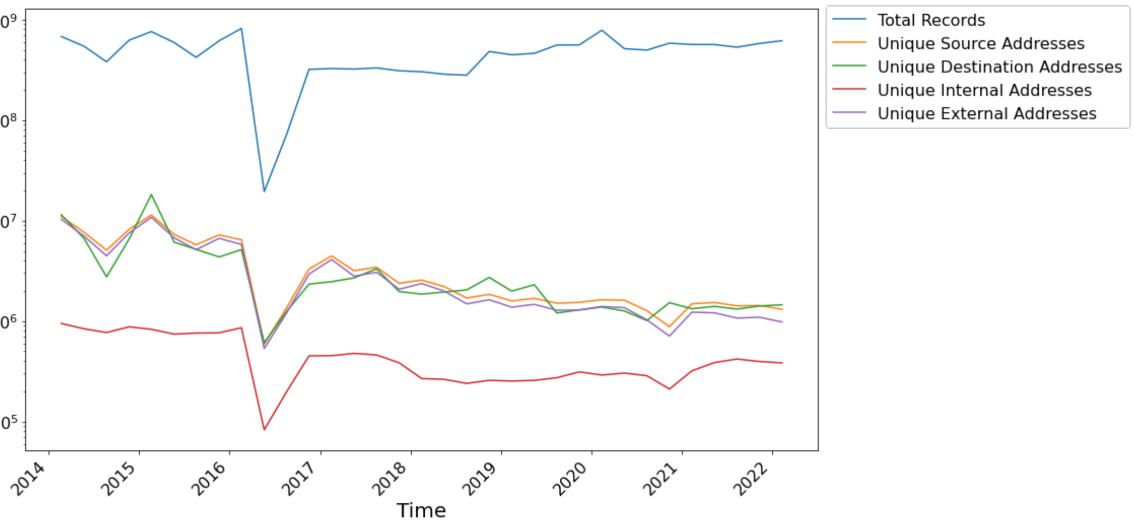

Figure 4: Characteristics of each sample across 8-year time scale

Distribution of different characteristics that define each 24-hour sample over the 8-year time scale. Traces include the total number of records that appeared within the sampling windows over time (Total Records), the number of distinct source and destination addresses in the samples over time (Unique Source Addresses and Unique Destination Addresses), and the number of distinct internal and external source addresses over time (Unique Internal Addresses and Unique External Addresses).

60-day Time Scale

The second timescale spans 60 days in 2019. This timescale provided a more granular look at the dataset but still provided a significant amount of time to consider the multifractal properties over. A period of 60 days can be broken down into 2880 total 30-minute chunks. We randomly selected 40 of these epochs to perform the multifractal analysis over. Specifically, we looked at a 60-day period starting on 2019/05/01 at 00:00:00. This date was chosen due in part to the issues with data collection in the earlier years, but other than that, the decision was quite arbitrary. The characteristics for each sample in this time scale are detailed below in Figure 5.

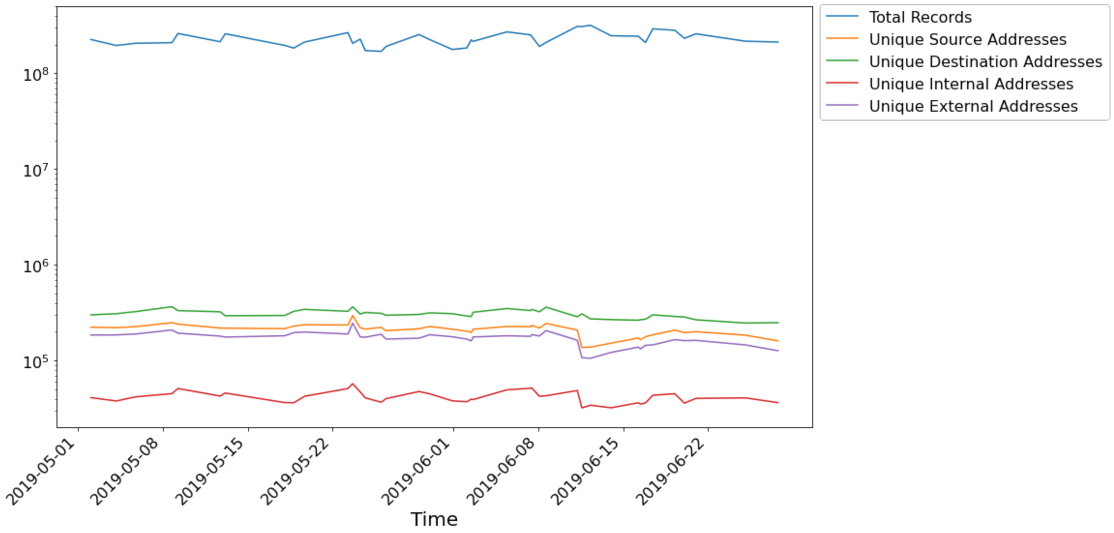

Figure 5: Characteristics of each sample across 60-day time scale

Distribution of different characteristics that define each 30-minute sample over the 60-day time scale from 2019/05/01 to 2019/06/29. Traces include the total number of records that appeared within the sampling windows over time (Total Records), the number of distinct source and destination addresses in the samples over time (Unique Source Addresses and Unique Destination Addresses), and the number of distinct internal and external source addresses over time (Unique Internal Addresses and Unique External Addresses).

24-hour Time Scale

The third timescale takes a far more granular look at the data as it only encompasses 24 hours. In a similar manner to the 60-day timescale, we selected 40 random 30-second epochs from the 24-hour period to perform the analysis on. As a single day has the possibility to be non-representative of any other day in the year, we performed this analysis across 20 random dates from 2019 to combat possible bias. The characteristics for each sample in this time scale for the date 2019/03/11 are detailed below in Figure 6.

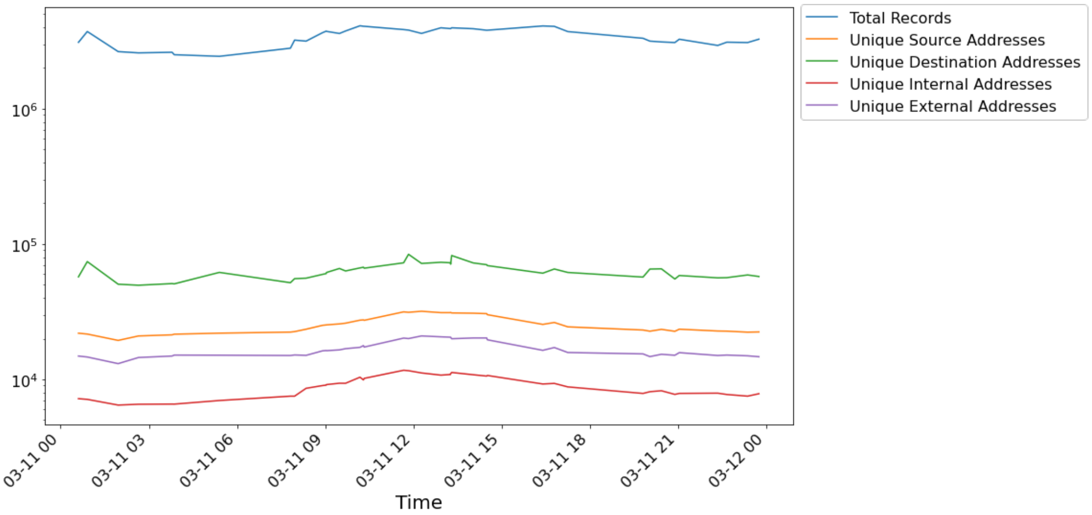

Figure 6: Characteristics of each sample across 24-hour time scale

Distribution of different characteristics that define each 30-second epoch over the 24-hour time scale of the date 2019/03/11. Traces include the total number of records that appeared within the sampling windows over time (Total Records), the number of distinct source and destination addresses in the samples over time (Unique Source Addresses and Unique Destination Addresses), and the number of distinct internal and external source addresses over time (Unique Internal Addresses and Unique External Addresses).

Performance Issues

One issue that arose was the magnitude of the data set, leading to issues in converting the data. Each five-minute NetFlow file contains millions of records and keeping track of 32-bit IP addresses is a memory-intensive task when done at a large scale. This meant the algorithm used hundreds of gigabytes of computer memory when processing a 24-hour period of time which affected the 8-year time scale in particular. To circumvent this issue, we modified the algorithm to only keep track of address accumulators for a limited window of time, using a circular buffer to keep track of addresses, and writing the IPs to an output file once a threshold of flow end-times had been passed (Figure 7).

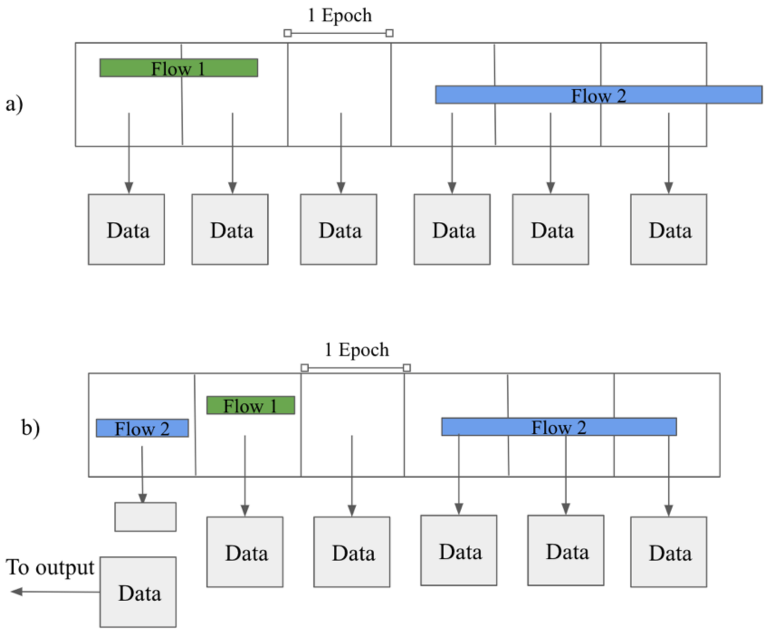

Figure 7: Circular windowed method of collecting data

The NetFlow processor keeps a limited number of epochs in memory. If an encountered flow's end time exceeds the number of epochs kept in memory (a), the least recent epoch's data is written to output and a new epoch takes its place (b).

Using this method reduced the memory usage from over 150GB to 7GB for a 24-hour period. Decreasing the memory usage also led to a decrease in runtime from 2 hours to 15 minutes.

Error across Samples

Unfortunately, our collection method did lead to some degree of data loss. In order to keep track of the flows that were disregarded, the information contained in the flows was written to a separate output file. Doing this allowed us to take a second pass over the data that was not included in the time series later so as to have a more complete view of the data set and giving us the chance to include more data into our analysis. We also quantified the error caused by this data loss by taking the number of flows that were counted and dividing by the total number of flows that occurred during the active time window. We found the data loss to be negligible given the scale of the dataset. The average data loss per sample for each timescale is detailed below in Table 2.

	8-year Time Scale	60-day Time Scale	24-hour Time Scale
Average Percentage of Data Lost	0.459%	1.159%	7.522%

Table 2: Average percentage of data lost across samples for each time scale

The data loss is calculated by counting the percentage of flows that would be considered active during the collection period but were not included in the set of active flows during that period.

Chapter 5: Analysis

Methodology

For the sake of analyzing these sets of IP addresses, we developed a new multifractal analysis method using a statistical technique called the "method of moments" and multifractal formalism to demonstrate that the distribution of IPv4 addresses in the address space exhibit multifractal behavior [Riedi]. We use the method of moments to quantify fractal-like behavior and use multifractal formalism to ensure that we come to the same conclusion that the multifractal spectrum approach would lead to.

Multifractal formalism is a mathematical concept that effectively links the geometric results of the multifractal spectra method used in prior work and the statistical description of the data based on sample moments and their scaling properties. However, the geometric description given by the multifractal spectrum cannot be reliably estimated in practice using real world data. We instead use the indirect statistical approach to arrive at the same conclusion: whether or not the data exhibits multifractal behavior [Riedi].

What differentiates our Method of Moments method from the Multifractal Spectra method used in prior work is that it is better suited toward a dataset with a discrete number of scales. The Multifractal Spectrum method assumes an infinite number of spatial scales. As we have defined the spatial scales over the IP address space to be the different possible prefix lengths, the statistical description of the data provided by the Method of Moments method is easier to compute, and more accurate to the real-world data.

There are four steps to the implementation of this Method of Moments process:

1. Defining **multiresolution quantities** by iterating through prefix lengths.
2. Defining a **structure function** and using the multiresolution quantities as the input.
3. Define a **partition function**, which estimates the slopes of the traces in the structure function.
4. Using the principles of multifractal formalism to demonstrate the **presence of a multifractal structure** using the plots from the partition function.

Multiresolution Quantities

These quantities are a series of values that represent the distribution of IP addresses with a shared prefix throughout the address space for different spatial resolutions, in this case, prefix lengths. Our multiresolution quantities can be described by the expression $\mu(i, l)$ where the output is the number of active external IPs in the i-th subnet with a prefix length of l for $1 \leq l \leq 32$ and $1 \leq i \leq 2^l$.

Structure Function

The structure function $Z(l, q) = \sum_i \mu(i, l)^q$ returns the q-th sample moment for the multiresolution quantities at different prefix lengths. This function relates the multiresolution quantities to each other to give a sense of the characteristics of the distribution of addresses throughout the address space.

Partition Function

The partition function $\tau(q)$ estimates the slopes of the logarithmic plot of the structure function for each moment using the method of least squares. This function is an approximation of a transformation of the multifractal spectrum of the data.

Presence of Multifractal Characteristics

In this context, multifractal formalism asserts that the partition function is the Legendre transform of the fractal dimensions over the data and vice versa. This implies that real-world data is consistent with multifractal scaling given the sufficient condition that the partition function is non-linear. The non-linearity signifying multiple fractal dimensions as opposed to a single one as would be seen in monofractal scaling behavior [Olsen, Riedi]. The partition function deviating significantly from a straight line provides evidence for multifractal scaling behavior.

Non-linearity was estimated by the R-Squared value, also known as the coefficient of determination, applied to the partition function. In statistics, this value represents how well a model fits the given data. In this case, it provides a measure for how well a simple linear model fits the partition function. The coefficient ranges between 0 and 1 with values closer to 0 signifying a higher degree of non-linearity and values approaching 1 meaning the partition function is close to being linear.

Single sample over all prefix lengths

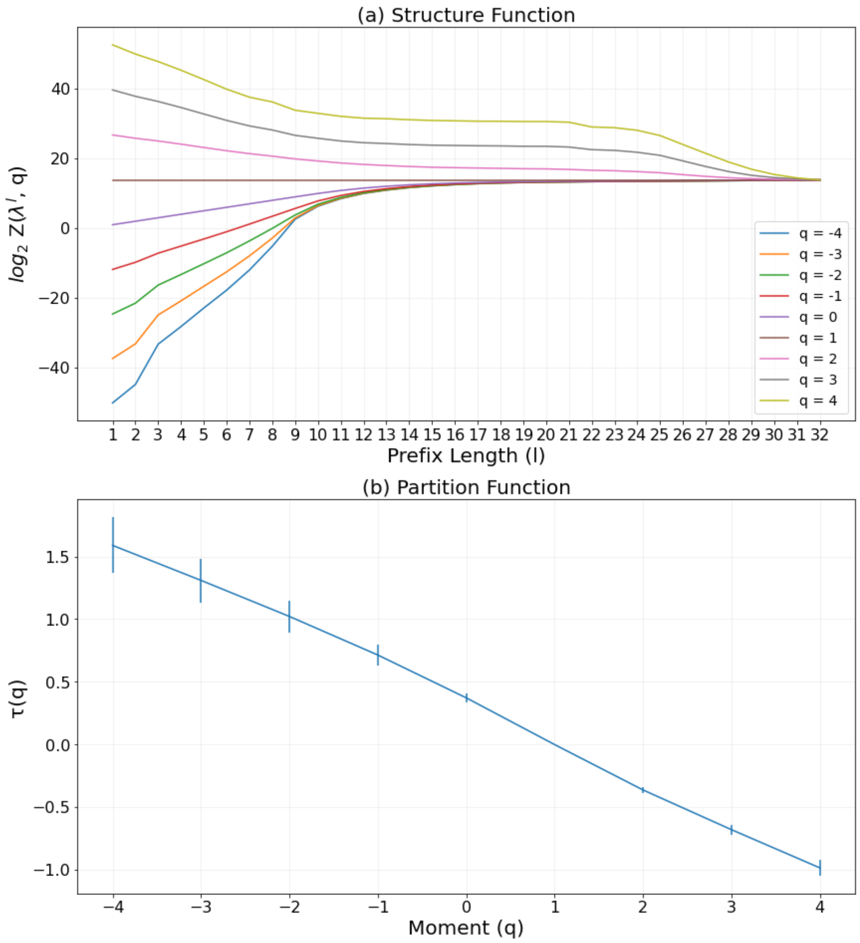

Figure 8: Structure function and partition function of 30-second epoch over all prefix lengths

The structure and partition functions of a single 30-second epoch from the 24-hour timescale (2019/01/31) over all possible prefix lengths ($1 \leq l \leq 32$). Error bars in the partition function plot represent the least-square error of the slope estimate for each moment q. The structure function can be separated into distinct ranges of prefix lengths (2-12, 12-24, and 24-32), in which the slopes of each moment trace all visibly change.

Graphing the structure and partition functions across a range of prefix lengths from 1 to 32 reveals distinct ranges of address prefix lengths that each display a different type of behavior (Figure 8). The graph of the structure function can be separated into distinct regions along its x-axis, each of which display different behaviors in terms of how the slope changes over the ranges of prefix lengths. This phenomenon is consistent across all three timescales. When calculating a linear slope for each moment over all the prefixes to graph the partition function, the distinction between regions is lost. That is to say, the changes in slope along the structure function appear to indicate differences in the multifractal characteristics of the data within different ranges of prefix lengths. As such, generalizing the partition function over the full range of prefix lengths results in a loss of specificity. Given the different multifractal characteristics, we will narrow our focus to a singular range of prefix lengths from 2-12 to gain an understanding of the multifractal behavior within that one range without noise from the rest of the spatial scales.

Comparison of single samples across timescales

Having limited the range of prefix lengths, we turn to our question of whether or not a modern dataset demonstrates evidence for multifractal behavior. Figure 6 shows the plots of the structure and partition functions for a 30-second epoch from the 24-hour timescale. Restricting the range of prefix lengths resulted in a change of shape of the partition function. We can observe a slight elbow in the graph of the partition function as the moments switch from nonpositive to positive. This slight bend, this break in linearity, indicates that there is some evidence for the presence of a multifractal structure within the range of prefix lengths from 2-12.

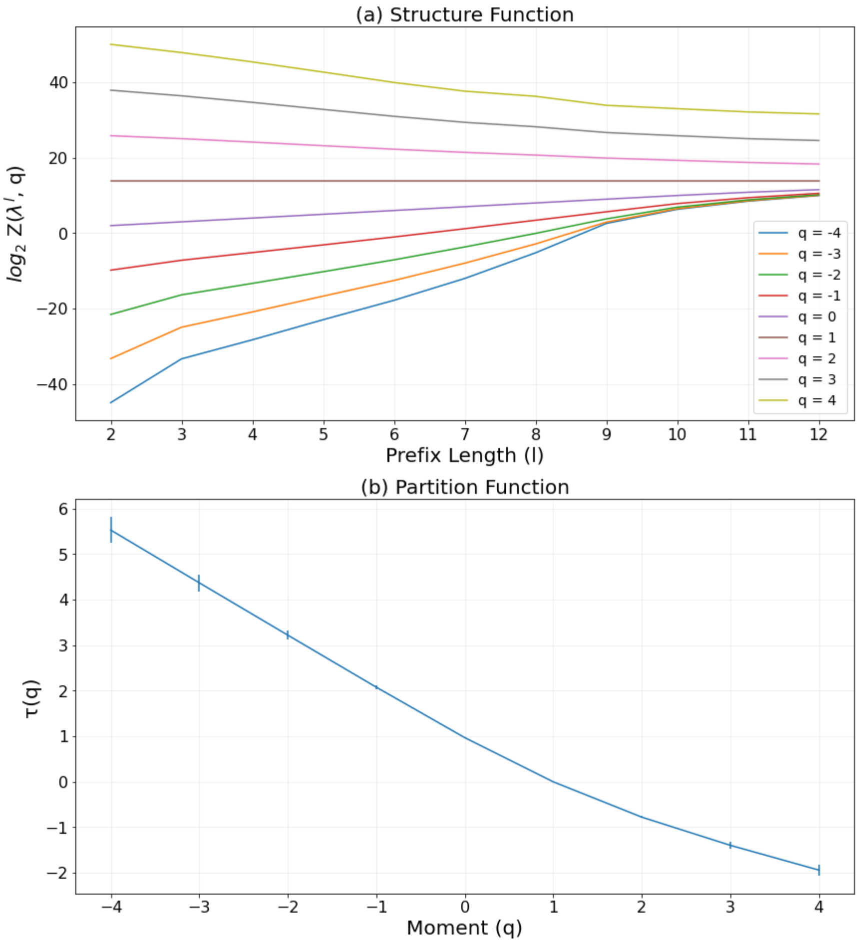

Figure 9: Structure and partition functions over limited range of prefix lengths for 30 second epoch

The structure and partition functions of a single 30-second epoch from the 24-hour timescale (2019/05/24). Limiting the range of prefix lengths has revealed a different shape of partition function which indicates evidence for multifractal behavior.

Comparing the results of an epoch from the 24-hour timescale (Figure 9) to results from the other time scales in Figures 10 and 11, We notice a similar shape in partition functions. As the moments transition from non-positive to positive, we notice the same elbow across all of the different timescales. All three of these samples were taken in late May of 2019 to reduce the number of variables affecting the results. As such, we can use these three samples to compare how the collection duration of addresses affects the multifractal characteristics of the data.

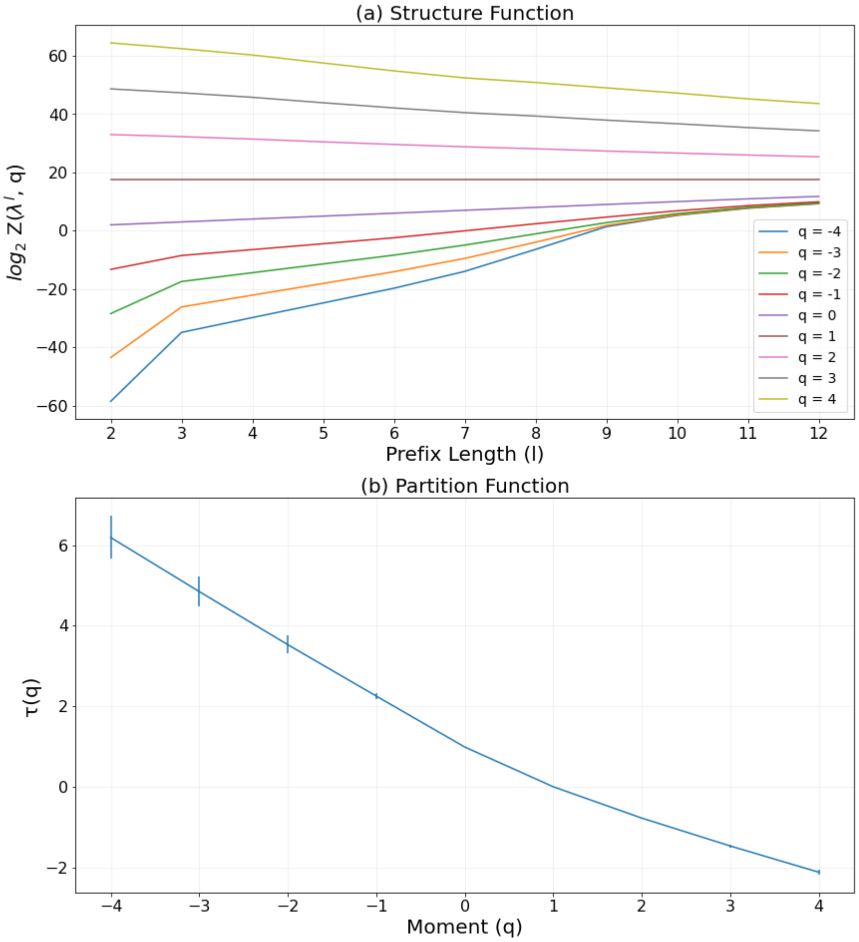

Figure 10: Structure and partition functions over limited range of prefix lengths for 30-minute epoch

The structure and partition functions of a single 30-minute epoch from the 60-day timescale (2019/05/24).

The differences between the samples with different epoch durations appear to be minimal; the shape of the partition function remains the same across all three durations of collection. While the scale of $\tau(q)$ does increase as the collection duration increases–ranging from -2-6 for the 30-second epoch and -2-8 for the 24-hour epoch–we are mainly interested in the linearity of the partition function rather than the actual values.

So, the consistency in shape of the partition function across all three epoch durations suggests that duration of collection does not have a significant effect on the multifractal characteristics of observed IPv4 addresses.

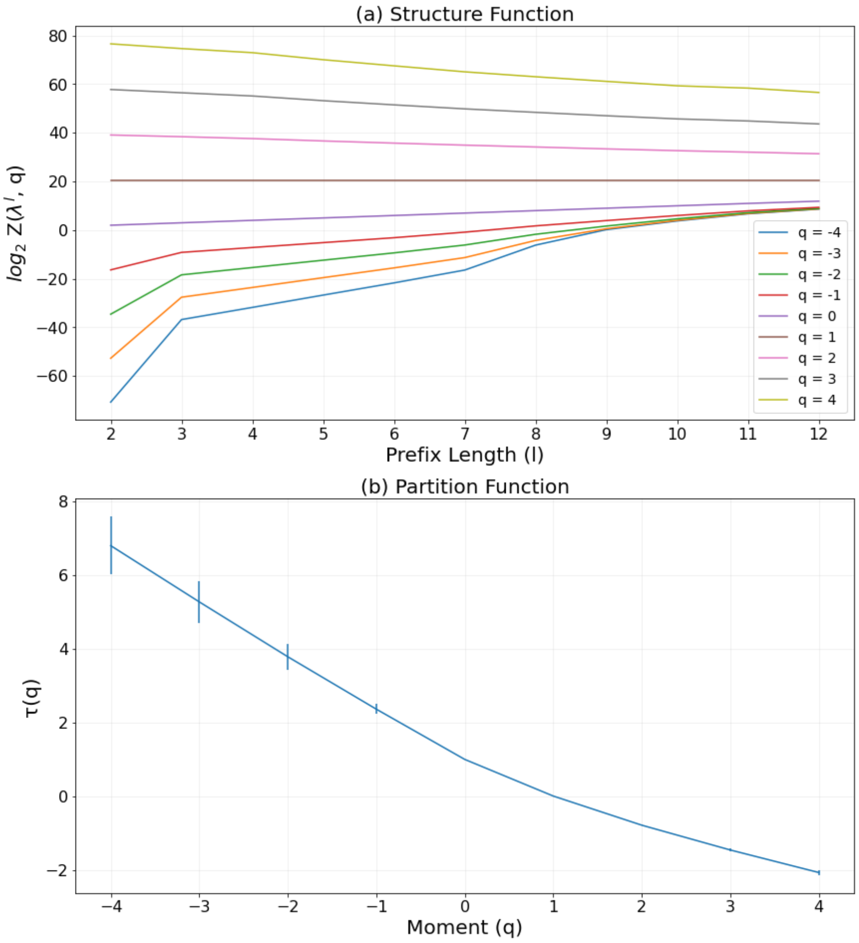

Figure 11: Structure and partition functions over limited range of prefix lengths for 24-hour epoch

The structure and partition functions of a 24-hour epoch from the 8-year timescale (2019/05/16).

24-hour time scale sample dynamic

Given the evidence for multifractal behavior in the sample snapshots that we examined from 2019, we next turned to our question of how this behavior evolves over time. Figure 9 displays how the R-squared value of the least-squares approximation of the slope of the partition function evolved across a short-term 24-hour timescale.

In general, the R-squared values across all time scales were within a margin of 0.1 of 1, indicating a strong tendency towards linearity. This is to be expected considering that in its standard use, the R-squared value is used to determine how well data fits a least-squares approximation. The partition function has a limited number of points that follow a downward trend, leading to a least-squares approximation that fits the limited data points well. In the future, a different metric may be more useful to illustrate the linearity of the partition function, however, the R-Squared value can still inform us about changes in the linearity of the partition function across a time scale.

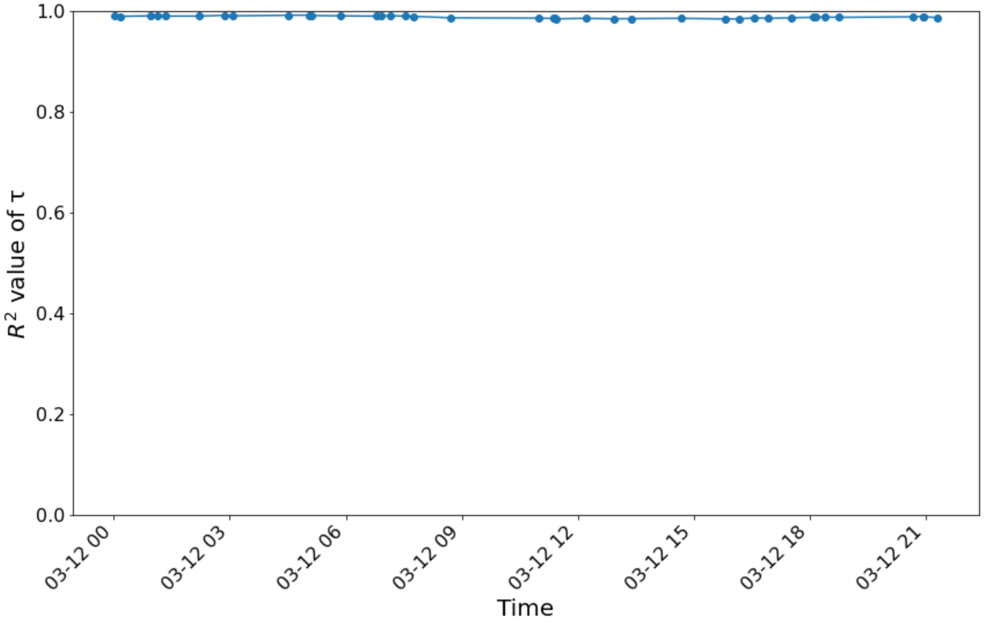

Figure 12: R-Squared values over time for 24-hour timescale

The R-Squared value over a series of 40 random 30-second epochs from 2019/03/12.

Analyzing how the R-squared value of the partition function would evolve over a short-term period, we found that the R-squared values of epochs during the times of day that more users would be connected to the network were lower than those during which the network would be considered less busy. This phenomenon is illustrated in Figure 12 with a slight dip that begins around 9:00 AM and ends around 6:00 PM. This trend indicates a stronger argument for multifractal behavior during the periods of the day when more devices are connected to the network.

60-day time scale sample dynamic

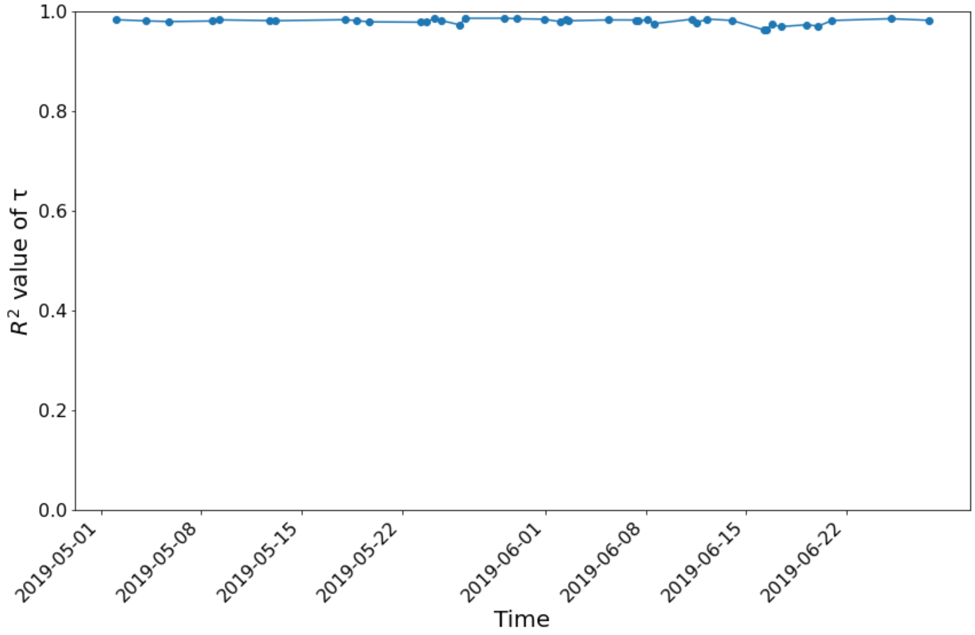

Figure 13: R-Squared value over time for 60-day timescale

The R-Squared value over a series of 40 random 30-minute epochs between 2019/05/01 and 2019/06/29.

The R-Squared values of the partition function over the 60-day timescale (displayed in Figure 13) demonstrated no clear trend. However, several lower R-squared values appeared throughout the 60-day period. While there is no obvious reason for why these changes occur, it is still interesting to note that the strength of the multifractal properties of the observed traffic is subject to fluctuation over time.

8-year time scale sample dynamic

R-Squared values along the 8-year time scale demonstrated a shallow trend towards lower R-Squared values as the years passed. This observation indicates a trend of the network displaying more multifractal behavior over time. This could possibly relate to the results from the short-term 24-hour timescale; as the network has grown busier, the addresses appear to closer fit a multifractal model. We also notice some possible seasonality based on the academic year beginning in 2017 with the R-Squared values decreasing during busier times of the year (Figure 14). In terms of our research question, how the multifractal properties of the real-world data evolve over time, the 8-year sample indicates that over the long-term, the behavior of the University of Oregon network tends to move in the direction of more multifractal over time.

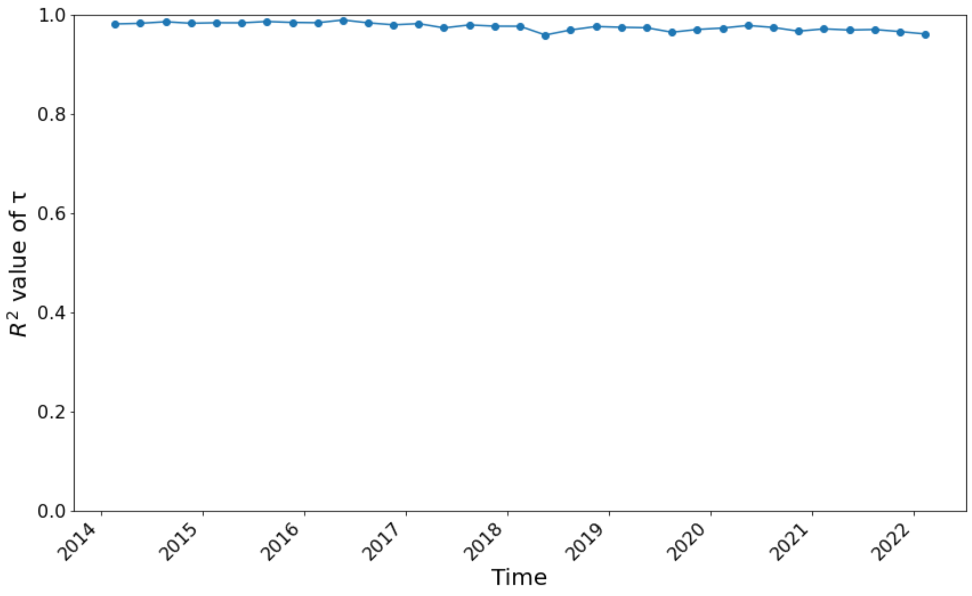

Figure 14: R-Squared value over time for 8-year timescale

The R-Squared value over a series of 33 24-hour epochs between 2014/02/20 and 2022/02/10.

Chapter 6: Conclusion

Altogether, the results of this work demonstrate evidence that the real-world IP addresses collected from a modern network exhibit multifractal behavior across a specific range of IP prefix lengths. Additionally, we found that the duration of collection of data had no obvious effect on the multifractal behavior of the network. Beyond that, the manner in which the multifractal behaviors evolved over short-term and long-term period of time indicate a correlation between the network being busier and stronger multifractal behavior. While this relationship may not be direct, this is an encouraging sign as a busier network creates an environment in which network management solutions are more critical.

Through our research, we have provided a new technique of multifractal analysis of IPv4 addresses that provides more robust, less noisy results than the Multifractal Spectra method used in prior work. While there are still some aspects that could be improved–for example, the use of R-Squared values as a metric of linearity–using the method of moments has provided a more robust set of evidence for the same claims.

Hopefully, this work contributes a meaningful basis to rethink current Internet algorithms to become more spatially aware of how flows relate to one another. Using a multifractal model to represent active addresses can be used to simulate real-world traffic and contribute to the goals that prior works set such as improving whitelisting techniques and improving real-time network monitoring systems.

Future Exploration

One aspect of this study that would be valuable to further explore would be to narrow the scope of the multifractal analysis to specific types of flows. For example, using this method of multifractal analysis on flows that have been classified as malicious as opposed to flows that have been classified as benign. Getting a sense of how active addresses relate to one another beyond a general understanding would help further push spatially aware solutions to network management and security issues.

Another direction to consider would be to understand the relationship between busier times of the network and an increase in multifractal behavior. Understanding the external variables that may be affecting variation in the network's behavior will be valuable information when developing new network monitoring techniques.

Bibliography

Barford, Paul, Rob Nowak, Rebecca Willett, and Vinod Yegneswaran. "Toward a Model for Source Address of Internet Background Radiation.". In *Proceedings of Passive and Active Measurement Conference* (PAM) (pp. 181-190) 2006.

Estan, Cristian, Ken Keys, David Moore, and George Varghese."Building a Better NetFlow." . In *Proceedings of the Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications* (pp. 245–256). Association for Computing Machinery, 2004.

Falconer, Kenneth. Fractal Geometry: Mathematical Foundations and Applications. 2nd ed. Nashville, TN: John Wiley & Sons, 2004.

Haag, Peter (2016) nfdump [Source Code]. https://github.com/phaag/nfdump

Harte, David. Multifractals: Theory and Applications. United States: CRC Press, 2001.

Kohler, Eddie, Jinyang Li, Vern Paxson, and Scott Shenker. "Observed structure of addresses in IP traffic". IEEE/ACM Transactions on Networking 14, no.6 (pp. 1207–1218) 2006.

Memon, Ghulam, Reza Rejaie, Yang Guo, and Daniel Stutzbach. “Large-scale monitoring of DHT traffic.” In *Proceedings of IPTPS* (pp. 1-11) 2009.

Olsen, Lars. “A Multifractal Formalism.” Advances in Mathematics 116 (pp. 82-196) 1995.

Rasti, Amir, Nazanin Magharei, Reza Rejaie, and Walter Willinger. “Eyeball ASes: from geography to connectivity” In *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement* (pp. 192-198) 2010.

Riedi, Rudolf. (1999). An introduction to multifractals.

Salat, Hadrien, Roberto Murcio, and Elsa Arcaute. "Multifractal methodology". Physica A: Statistical Mechanics and its Applications 473, 2016.

Yeganeh, Bahador, Reza Rejaie, and Walter Willinger. “A view from the edge: A stub-as perspective of traffic localization and its implications.” In *Network Traffic Measurement and Analysis Conference* (TMA) (pp. 1-9) 2017.

Yeganeh, Bahador, Ramakrishnan Durairajan, Reza Rejaie and Walter Willinger. “How cloud traffic goes hiding: A study of Amazon's peering fabric.” In *Proceedings of the Internet Measurement Conference* (pp. 202-216) 2019.