Visualizing the Structure of Network Traffic Features Across the IP Address Space

Eugene Tan

University of Oregon
eugenet@uoregon.edu

June 2022

1 Introduction

With the growing scale and complexity of the internet, there is an increased need to understand and manage the portions of the internet space that are able to be controlled. From the growth of the internet and increased integration of technology there exists an abundance of network traffic data that can be collected, understanding and identifying useful features found within network traffic data across the IP address space could provide insight and discovery into trends and patterns of IP addresses. For the purpose of identification and observation of patterns and spatial trends of IP addresses, visualization is an important feature. Some visualization techniques often applied for visualizing network traffic data such as Hilbert Curves are especially useful for the purpose of visualizing the IP address space. However, prior studies on the visualization of network traffic data have various limitations such as exploring limited traffic features, visualizing various prefix lengths, and restrictive visualization of data.

Therefore, the goal of this thesis is to explore how visualization using a Hilbert curve can be used to examine network traffic features and their correlation with the structure of IP addresses in order to expose fractal trends or patterns within network traffic data. To approach this study, network traffic data is captured and processed to reveal per-IP-address high-level network traffic features which are subsequently used within visualization for study. With this resultant network data, a Hilbert Curve is used to produce visualizations of the structure of the IP address space with various features. We apply this approach to three different real-world questions with large traffic datasets provides additional insight into the structure of network traffic and the efficacy of using visualization to study network traffic data. The developed software used within this thesis for visualization is publicly available for use within future examinations of network traffic data[17].

2 Background

2.1 Introduction to Project

With the large amounts of network traffic data that can be collected for data driven analysis, visualization techniques are appealing due to their ability to capture larger trends and expose interesting characteristics and patterns. The network traffic data examined within this thesis are per-IP-address traffic features computed over a sequence of packets received from a particular IP source address and the set of all IP source addresses and their prefixes. Therefore, further references to network traffic data in this thesis refer to this particular grouping of network data. By capturing these larger trends and patterns, observations can be made regarding the structure and data attributes of network traffic data found within visualization. However, with many existing visualization techniques applied within existing works, there exist various limitations within the produced visualization such as differing features examined, examination of network traffic at a singular granularity, and inability to study the relationship between traffic features. From the ideas presented within this thesis, a new visualization method is introduced using a Hilbert curve and mapping features to color to demonstrate how through visualization it is possible to examine network traffic data and discover fractal patterns and trends within the data.

2.2 Introduction to the Hilbert Curve

An important contextual concept of the thesis is the use of a Hilbert curve[16] for the visualization of network traffic data. A Hilbert curve is a fractal structure often applied in visualization techniques of IP addresses due to its mapping capability. The Hilbert curve is capable of mapping one-dimensional data sets into two-dimensional data representations[2]. Therefore, within the context of network traffic IP addresses which can be observed as a one-dimensional data set, the Hilbert curve is able to map IP addresses to a two-dimensional figure representation. This behavior is one of the primary properties that make using a Hilbert curve appealing for the purpose of visualization. An additional property of the Hilbert curve which lends itself to be useful for the purpose of analysis is its data locality preserved through mapping. Data which is close within the one-dimensional data set retain this locality when mapped to a two-dimensional plane, making this behavior an ideal feature for mapping IP addresses. IP addresses which share a similar prefix will be mapped closer together within the curve by this property. From this set of behaviors, the use of a Hilbert curve in the context of visualization is increasingly appealing.

As a fractal structure, the Hilbert curve has a parameter of order of the curve[16][2]. This order of the curve correlates to the dimensions of the resulting two-dimensional space. Larger orders of the curve indicate larger powers of two that become the x and y-axis lengths within the two-dimensional space. To ensure, data is represented correctly, the order of the curve must be large enough

to encapsulate all data points to ensure all data points have a corresponding coordinate. To capture this idea, this interaction is exampled in Figure 1 which illustrates Hilbert curves with orders 2 through 5.

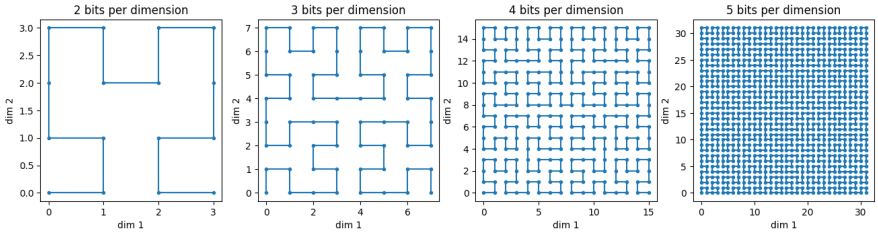


Figure 1: Hilbert Curves of order 2, 3, 4, and 5 respectively.

From Figure 1, it can be observed that increasing orders result in larger coordinate fields. Leveraging this parameter of order, Hilbert curve spaces can grow large to encapsulate all IP addresses to be visualized. As a result, from the properties of the Hilbert curve as well as its possible applications to visualization of network data, the Hilbert curve is used to introduce new methods of visualizing network traffic data.

2.3 Introduction to Color

In order to produce meaningful visualizations, interpretation of these visualizations is important. Interpretation is largely guided by how characteristics or attributes found within the visualization can be described visually. The use of color within the visualization method presented serves as a method of discerning characteristics of the data as well as distinguish differences observed within network traffic data. Understanding how colors are constructed and the factors that affect the creation of these colors is an important feature within this thesis often not explored within many existing visualizations of the IP address space.

Colors observed are composed of hue, saturation, and lightness (HSL) factors. Each of these factors affects how a color is represented, as each color factor influences how color is constructed. These HSL values can be directly translated into red, green, blue values (RGB) that can be used as colors within the visualization. With HSL containing three color factors, affixing lightness as a factor enables a two-dimensional examination of hue and saturation. Therefore, leveraging the color factors of both hue and saturation provides a visual attribute to network traffic features. Using this two-dimensional color analysis, color can be applied to network traffic feature visualization as a method of indicating relative network traffic values. From the concept of color introduced, color is used within the visualization method as an important aspect of both visualization and study.

3 Prior Work

Prior studies have relied on various approaches to capture and characterise different aspects of network traffic including application-specific scheme[8], simulation[1], geo-informed[13], or other type of visualization.

With many existing visualization techniques on visualizing IP addresses as a two-dimensional coordinate graphic using Hilbert curves, multiple methods of approaching network traffic data examination are identified. A prior work which provided valuable context and information regarding the application of Hilbert curves was "Towards Geolocation of Millions of IP Addresses" [4]. This work focuses on the mapping of IP addresses to their corresponding geographic location, using the Hilbert curve to map IP addresses into a two-dimensional plane with geolocation feature data as color. With this prior work[4], the use of the Hilbert curve's property of spatial locality allows for the clustering and coloring of IP addresses based upon their geolocation features of latitude and longitude. This is identified through the visual mapping of IP addresses to the world map [10]. While mapping of IP addresses and ideas of feature coloring resemble the methodology proposed within this thesis, there exist a few limitations regarding the use of this existing work. The prior work[4] does not examine network traffic features and instead focuses on the geolocation features associated with an IP address. Demonstrated through the use of longitude and latitude values as hue and lightness to show a relationship between their location features [4]. Another limitation was the inability to visualize IP addresses at varying levels of granularity. Examination of IP addresses focused on the visualization of the entire IP address space, rather than specific prefixes that retain the same level of granularity or level of detail. While offering insight into the mapping of IP addresses, limitations within this work prevent the investigation of new network traffic features, as the feature attributes studied differ in both feature and color mapping method, and visualization of network traffic at varying levels of granularity.

The differing approach to color mapping is important in the context of visualization, as within "Towards Geolocation of Millions of IP Addresses" [4] features of latitude and longitude are mapped to lightness and hue. Introducing features as attributes of color enables visual examination and study of feature value and feature relationships. In this thesis, the methodology regarding feature color differs compared to the geolocation mapping ideas discussed within the paper by leveraging new color factors in visualization[4]. However, a key idea introduced within the paper[4] was the inclusion of a legend graphic that demonstrates the color spectrum produced by their feature values. This legend graphic[4] that accompanied the Hilbert curve graphic produced offered valuable insight into the feature relationship between the two features of longitude and latititude by correlating color to geolocation. Applying this idea of feature relationships through color is important and of great interest, as it offers visual clarification of observed pattern or trends within the curve and insight into the structural distribution of features.

An additional work that coincides with the existing paper[4] discussed above

is the USC ISI ANT project's[10] census of the entire internet[5] using a Hilbert curve. This study offered interesting insight regarding how a Hilbert curve could be used to produce the structural distribution of IP addresses across the entire IP address space. The use of color and labels within the visualization offer additional information regarding the geographic features mapped for each observed IP address. An interesting utility within this project's examination of the IP address space was its study of the visualization of different prefix lengths[10]. By using multiple prefixes of increasing lengths, it can be seen how the granularity of the entire IP address space changes to reflect the change in prefix length[5]. This prior work is important as it alludes to investigating chunks of the IP address space in greater detail to expose more detail regarding the underlying structure within the Hilbert curve. From this census, increasing the prefix length provides insight into the entire address space by increasing the detail, however there exist a few limitations surrounding this visualization. A primary limitation of this work is that the visual examination of a chunk at increasing prefix lengths is unable to visualize the mapped IP addresses to an increased viewing scale. As a result, consecutive images are difficult to discern the structure from and observe at closer detail. Additionally, the visualization only examines geographic features and does not study network traffic features.

Another tool examined that is well known for mapping IP addresses using a Hilbert curve is 'IPv4-Heatmap' [18], a tool used for mapping a given set of IP addresses onto the IP address space. The tool allows visualization of a set of observed IP addresses onto a two-dimensional coordinate plane with one-to-one address to pixel mapping. Additionally, the tool provides small visual adjustment parameters for the resulting visualization such as generation of labels, axis values, and visual labeling of chunks[18]. However, some limitations of the tool's visualization such as the examination of density as its only feature as well as its incompatibility for variable prefix lengths result in missing features and an inability to examine differing prefix structures.

From examining past prior works, there exist various limitations in how visualization is approached. In the existing works examined, visualization is unable to capture network traffic features or their relationships due to the feature set examined. Additionally, within existing visualization techniques it is impossible to examine the structure of the Hilbert curve at varying levels of granularity; preventing detailed study of network traffic clusters found within the Hilbert curve. From the limitations of prior visualization techniques, this thesis aims to introduce a new approach to visualization using a Hilbert curve that addresses these limitations through the introduction of new features and applications of color.

4 Approach

Within this section, the approach to visualizing network traffic data is introduced. We identify modules within the visualization method and how these modules serve to address the limitations found within existing visualization tech-

niques. Additionally, the visualization approach demonstrates how the addition of these modules provides valuable insight into the visualization of network traffic data and its interpretation.

4.1 Rationale for Approach

To guide the approach to applying visualization to examine network traffic data, various prior works were examined to gain insight into the use of Hilbert curves as well as how IP addresses can be viewed and examined. However, within these prior works there exist multiple limitations within the visualizations that they produce. Close examination of these limitations reveal that existing visualization techniques do not visualize data beyond a fixed granularity or study limited features and their relationships. Therefore, to address these limitations a new visualization method for the examination of network traffic data capable of bridging existing gaps in functionality is introduced.

4.2 Approach Overview

This new visualization method uses network traffic data found within network traces. To study the network traffic data from large network data sets, this method leverages multiple modules to manage network traffic data and distill information from this data to be represented through visualization. The following Figure 2 illustrates this visualization method and its various modules.

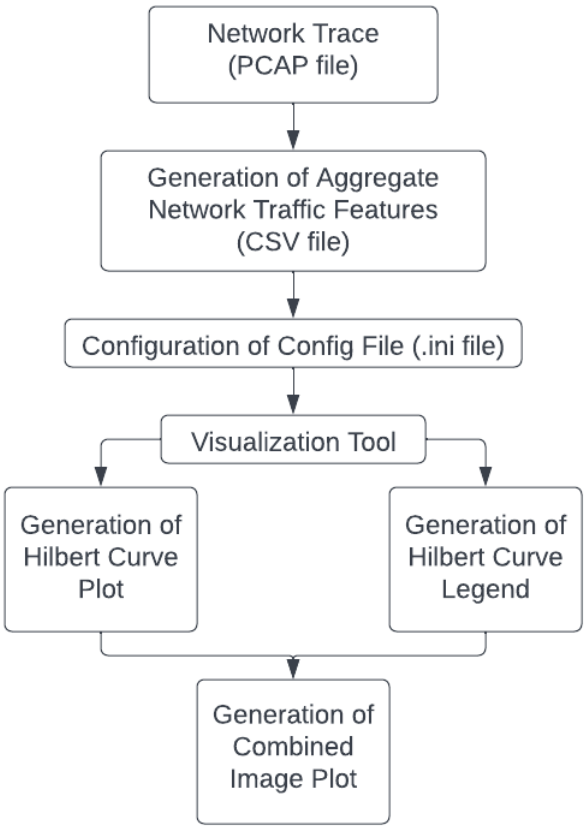


Figure 2: Flowchart of the data handling pipeline that occurs during visualization.

To begin the visualization approach, a network trace is initially used as the data set to be examined. Using a feature generating tool developed by student researchers from ONRG[14] the network trace can be distilled down into a file containing aggregate network traffic features grouped by prefix length and time. With this resulting feature file, examination of network traffic features can be visualized. The use of network traffic features differs significantly from the visualizations produced by existing works, as none of the existing works performed analysis or examination of network traffic features. By including this feature file, configuration of data filtering and visualization can be completed to designate what network traffic is observed and how the resulting visual is depicted.

Subsequently, the visualization of these observed IP addresses is accomplished through a series of data filters applied during visualization to capture the set of IP addresses to be visualized. These IP addresses are subsequently visualized through the use of the Hilbert curve's mapping functionality[2]. IP

addresses with their feature values and associated coordinates are then visualized with a new application of color factors. The color factors applied within this visualization approach differ from existing applications of color seen in prior works[18][4], as this approach leverages factors of hue and saturation to indicate feature relationships and relative feature concentration. The application of color to the visualization is an important concept that adds additional insight into network traffic features and their structural distribution across the IP address space.

An additional module that is used within the visualization tool module which is not found in existing visualization techniques is the capability of producing visualizations at multiple levels of granularity. Leveraging the application of color within this visualization method as well as properties of the Hilbert curve, examinations of specific prefix addresses are able to be carried out. The addition of this visualization utility is a feature that addresses existing visualization limitations and provides valuable insight into structural distributions of IP addresses at differing prefix lengths. Furthermore, the use of this 'zooming' utility is important in the context of network traffic data visualization, as it offers an additional method of study through visualization that has not been seen before.

The result from this visualization method is a two-dimensional coordinate plane of all observed IP addresses with colors derived from the new coloring and visualization methodology introduced. This resulting visualization is able to capture feature relationships as well as structural observations and trends within network traffic data. With this new method of visualization using a Hilbert curve, new examinations of network traffic data can be completed using these new additional modules.

4.3 Coloring of Traffic Features

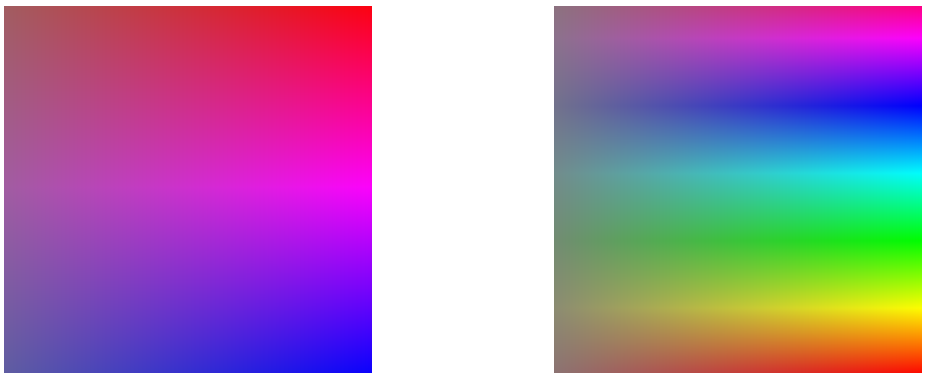
Coloring within the visualization method is an important concept that plays a critical role in the interpretation of feature values as well as the structure of observed IP addresses within the produced visualization. Therefore, careful consideration of how color is applied is required to ensure visibility of data and examination of features.

When initially approaching the concept of coloring, existing works such as the USC ISI ANT project[10] use color to represent limited features. This is seen through their use of latitude and longitude values as the corresponding values for hue and lightness which are color factors[10]. However, their use of color is applied for the examination of geolocation factors, as a result within this visualization method new applications of color are introduced for the examination of network traffic features. To study network traffic features using color, understanding the context behind how color is applied is important to the use of visualization.

Compared to existing works that leverage coloring, using color factors of hue and saturation provided greater visual distinction relative to existing visualization techniques which leverage hue and lightness[10] or a fixed color spectrum. Hue and saturation are seen as two separate factors of color that are represented

separately. Hue represents the spectrum of colors, often visualized in the form of degrees within the color wheel. Colors are often derived from changing hue values but affixed lightness and saturation values. Saturation represents the relative concentration seen of a color, often visualized as a fading of the color based upon saturation percentages. To determine how to best examine network traffic data with hue and saturation, hue values are examined.

Hue values govern the color produced, as changing hue values will alter the color. As a result, careful consideration of how values are mapped to hue value ranges is important in the representation of data. In initial visualizations, similar to the limitations of existing visualization there were difficulties in distinguishing color or differences in feature value. This was largely attributed to how data was interpreted and colored. An observation made regarding how hue affects the efficacy of visualization was short hue ranges introduced confusion regarding the interpretation of values. Conversely, large hue ranges introduced such a wide range of values that it was difficult to clearly identify the correlation between feature value and color within the visualization. The differing results produced by having differing hue ranges are illustrated in Figure 3 by comparing two separate hue ranges side by side.



(a) Hue Range from 0 to 360

(b) Hue Range from 240 to 360

Figure 3: Comparison in the color spectrum produced by differing hue values.

As a result, to provide clarity regarding the feature value corresponding to hue, this visualization method enables flexibility in the range of colors that features can be represented as. This flexibility in coloration and feature representation within the visualization is an important addition that provides clear interpretations of feature data that was not readily seen in many existing visualization techniques which did not utilize color.

Similarly, saturation values directly affect how a color is interpreted. With hue values influencing the range of colors to be seen, saturation affects how each color can be seen by modifying the relative percentage of saturation. Intuitively, colors with high saturation will be visually distinct, whereas colors with lower levels of saturation tend towards the same color. Considering how

saturation affects interpretation of its corresponding feature, within this visualization method saturation is restricted to a range between 25% and 100%. This decision was motivated by the desire to retain clear interpretation of color representation. In saturation values that exist below 25%, feature values that are mapped into saturation are nearly indistinguishable and do not offer valuable insight into any characteristics or attributes that can be found from the visualized IP addresses.

With careful consideration of both hue and saturation, this visualization method applies a new method of coloring that assists in the examination and interpretation of network traffic data. By selecting a hue range that is visually distinct and a saturation range that prevents visual overlap, a visual relationship is formed between features using color. This relationship formed is computed by linearly mapping two features to hue and saturation respectively. Normalizing the feature values within their respective acceptable ranges, the features being examined comprise a complete hue, saturation, and lightness value; as lightness is fixed. Lightness is fixed within this visualization method to reduce the dimensions of analysis to two-dimensions. To compute the saturation values seen within this visualization method, a selected feature for saturation has its values normalized between a value range of 0.25 to 1.00 representing saturation percentage. Similarly, a selected feature for hue value has its values normalized between a hue range of 240 to 360 to produce a distinct color spectrum. As a result, the relationship that can be seen between two features is the color produced by the HSL value created after normalization. To determine this color, the HSL value derived from the hue and saturation value found is translated to a red, green, blue (RGB) value that defines the feature relationship observed within an IP address.

This process of feature coloration is an important addition to the visualization of network traffic features, as it introduces a method for visual examination of network traffic features. Furthermore, this feature coloration is performed for all observed IP addresses using the Hilbert curve to convert IP address integer values to coordinates. The result of applying color in this new method of visualization provides valuable insight into network traffic features and the structural distribution of IP addresses.

4.4 Increasing Granularity of View

A consistent limitation observed within all existing visualization techniques using Hilbert curves was the inability to examine data at differing levels of granularity. Many of the existing visualizations use a single visual granularity that prevents further investigation or examination of a specific region or area of interest within the visualization. Therefore, within this visualization method a new method of visualization is introduced that enables closer examination of specific prefixes and offers valuable insight into the structural distribution of IP addresses at differing granularity.

To study a specific region or cluster within the Hilbert curve, a specific IP prefix must be selected for examination. This specific IP prefix is a parameter

which serves as an additional data filter to filter IP addresses that are not within the specified prefix and also not of the correct prefix length to be studied.The idea of IP prefixes being nested is best shown in Figure 4 which visualizes IP prefixes as a prefix-tree of varying prefix lengths[7].

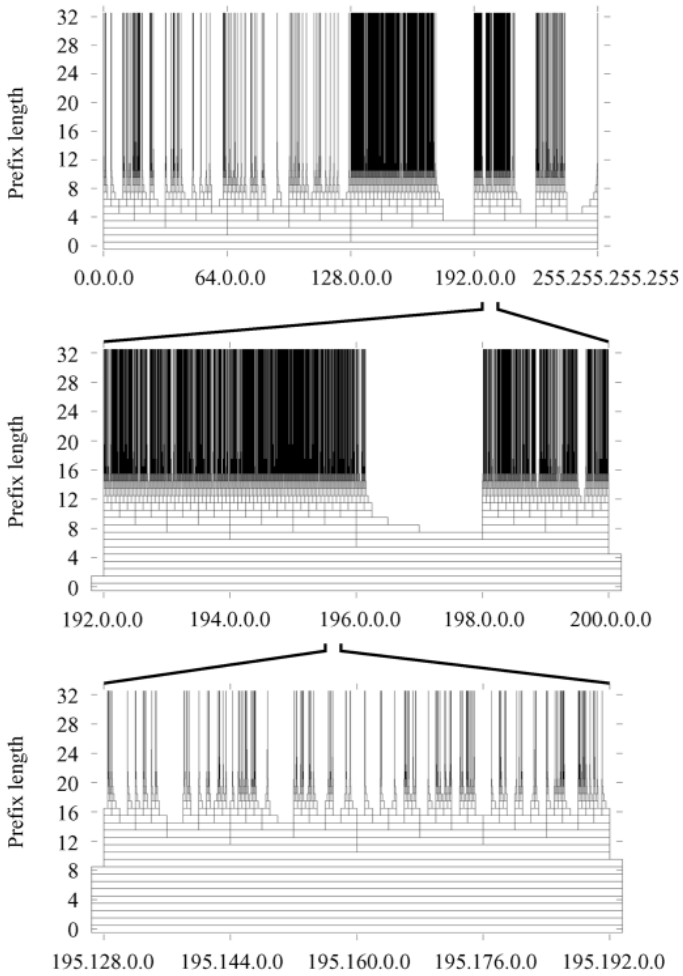


Figure 4: Tree-like structure of IP addresses at varying prefix lengths which represents the membership of IP addresses within the branching structure. Adapted from "Observed Structure of Addresses in IP Traffic" by Eddie Kohler, Jinyang Li, Vern Paxson, and Scott Shenker, Dec 2006, IEEE/ACM Transactions on Networking Observed, 14(6):1207–1218, doi:10.1109/tnet.2006.886288. [7]

By filtering away IP addresses based upon prefix length and sorting these resulting IP addresses based upon membership within a specified prefix length, isolation of a specific prefix is possible. To examine this prefix in greater depth

of detail, this visualization method increases the granularity of view by increasing the prefix length visualized within the Hilbert curve as well as the bit-mask associated with the specified prefix. By increasing the prefix length within the Hilbert curve the visualization increases the structural detail of IP addresses being visualized. And by increasing the corresponding bit-mask the visualization encapsulates only the IP addresses which belong to the specified prefix. The result of this process of increasing the granularity of view is the detailed examination of how IP addresses are structurally distributed within the Hilbert curve and viewed across multiple levels of granularity. Furthermore, this utility for increasing the granularity of view and offering multiple levels of granularity is intuitively useful as it offers multiple perspectives of interpreting the distribution of network traffic data. Having the capability of visualizing the network traffic from multiple levels of granularity enable informed insights regarding how IP addresses are distributed and help expose underlying structural characteristics.

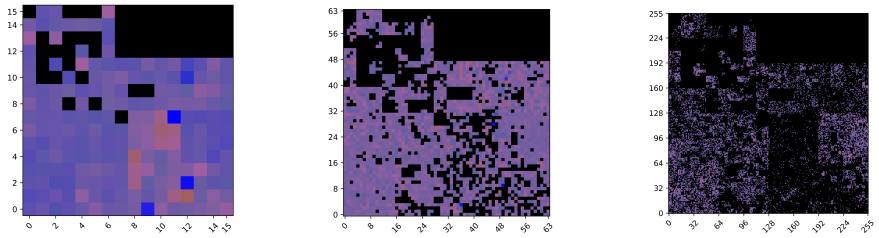
The addition of this module to the visualization method is important as it provides valuable insight into the structural distribution of IP addresses and their features. With existing visualization techniques focusing on a singular granularity for viewing IP addresses with a Hilbert curve, this visualization method enables multiple perspectives and options for viewing the structure of IP addresses within a Hilbert curve.

4.5 Visualization Parameters

An important part of this approach is the use of visualization parameters. These visualization parameters affect how data is filtered as well as how the resulting visualization appears. Primary parameters within the implementation are the prefix length of IP addresses to examine, the specified prefix the visualization should fixate upon, and the visual hue range that is associated with feature coloration. The prefix length parameter is used as a method of data filtration to isolate IP addresses which are of desired prefix length. Similarly the specified prefix parameter also serves as a method of isolating IP addresses which fall into the desired prefix to ensure visualization encapsulates the specified prefix. Lastly, the hue range parameter is an important visual parameter that affects how color is viewed by controlling the spectrum of color that can be used for feature value mappings. These parameters were chosen as the primary visualization parameters because of the impact each parameter has on the creation and interpretation of the resultant visualization. By being able to change the prefix length as well as the prefix block being examined, these parameters enable new visualizations of the structure within prefix blocks at differing levels of detail. Additionally, having a modifiable hue range enables differing methods of interpreting feature data through introducing new color ranges within the visualization. Understanding how each of these three parameters are used and applied within this visualization is important, as the produced visualization is derived from these parameters.

Prefix length indicates what prefix length of IP addresses should be examined. With network traffic feature data being grouped by prefix length, data

filtration is used with prefix length as a parameter to remove prefix lengths that are not examined. The resulting set of IP addresses is a reduced set of data containing IP addresses whose lengths match the specified prefix length. Through the use of this parameter, this visualization method introduces differing scopes of IP address visualization that were not seen previously in prior works such as the IPv4 Heatmap[18] which used a single granularity for visualization. An example of how this parameter affects visualization is illustrated in Figure 5 which depicts the differing prefix lengths of examination within the MAWILab data set[3]:



(a) Prefix Length: 8 (b) Prefix Length: 12 (c) Prefix Length: 16

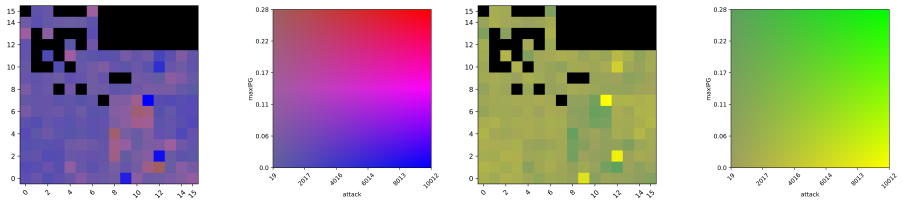
Figure 5: Example of the effects of prefix length on visualization. Visualized plot shows three differing prefix lengths using the maxIPG and density feature from MAWILab[3] datase

From Figure 5, using differing parameter values for prefix length, the structure of IP addresses increase in granularity for increasing prefix lengths. Therefore, prefix length as a parameter for visualization is important as it affects the granularity which data is visualized.

Specified prefix is another parameter that influences the set of IP addresses to be examined. With sets of observed IP addresses, adding a specified prefix on top of the existing prefix length parameter effectively applies an additional level of data filtration. Using a specified prefix and a prefix length for examination, study of IP addresses which are of correct prefix length and reside within the specified prefix is possible. With the specified prefix parameter, discovery and examination of specific clusters or prefixes within the Hilbert curve coordinate structure is possible. Ensuring that all IP addresses that remain after data filtering exist within the specified prefix, enables increasing granularity of view of IP address structures when increasing the parameter of prefix length. As a result, this parameter is another important item within the visualization method that introduces additional possibilities for the examination of network traffic features not seen before in other visualization approaches. The effects of this parameter are best visualized within the evaluation of the visualization method in Case Study: 2.

Hue range affects feature representation due to the color range that feature values are mapped into. Correctly selecting an adequate hue range that of-

fers insight into feature representation while retaining the visual distinction of feature values is important for the ultimate interpretation of the visualization. Hue ranges which are not selected carefully often introduce visual overlap that interfere with the interpretation of feature values resulting in an ineffective visualization. This exampled through Figure 6 which demonstrates the effects of hue range selection.



(a) Hue Value Range: 240 - 360 (b) Hue Value Range: 60 - 120

Figure 6: Example of hue value ranges and the effect on visualization and intepretabilility between two plots using different hue value ranges. Plots are visualized at /8 prefix length with MAWILab data[3]

As evidenced in Figure 6 above, hue ranges can improve or detract from the visualization's ability to convey feature data. Within this visualization approach, a hue range from 240 to 360 is used (blue to red).

4.6 Managing Network Traffic Data

In order to examine network traffic information through visualization, network traffic must be gathered and organized into a readable format for visualization.

Examination of network traffic begins with a network trace containing records of packets that detail the timestamp of each packet, the packet headers associated with each packet, the packet's source and destination addresses, packet length, and other information used within various communication protocols. The network trace is then distilled into high-level aggregate feature values using a feature generating tool developed by ONRG[14] that groups packet records by addresses or address prefixes and computes aggregates for each of these packet groups. This feature generating tool[14] produces a CSV file from the network data trace that has multiple IP prefix lengths and feature values. With multiple IP prefix lengths derived from the network trace, this feature generating tool[14] enables visualization of multiple prefix lengths allowing for examination of the structure of IP addresses at differing levels of detail. Having multiple IP prefixes to examine is also valuable as it may offer additional insight into fractal patterns or behaviors found at different prefix lengths. The feature values produced by the tool[14] are packets to destination (pcktsTo), maximum inter-packet gap (maxIPG), response request difference (respReqDiff), bytes to destination (bytesTo), bytes from destination (bytesFrom), source density, as

well as other additional features. These feature values are valuable as they offer feature-level insight into the network data associated with a given IP address. Therefore, with the CSV file produced by the feature generating tool[14] preliminary data filtering occurs before visualization.

With the configuration of visualization parameters mentioned in the prior section, setting the parameter of prefix length identifies which prefix length should be kept within the feature CSV file. Since the feature generating tool produces multiple prefix lengths, isolating a prefix length for examination requires data filtering to ensure that all observed IP addresses being examined are at the correct prefix length or level of detail specified. The result of this data filtering is the set of observed IP addresses that will be colored according to their feature value and visualized at their corresponding coordinate position found using the Hilbert curve. In order to determine where IP addresses are mapped to within a coordinate plane we must initially compute each IP addresses' integer value. This is required due to the Hilbert curve's mapping utility which maps one-dimensional data sets to two-dimensional spaces[2]. To follow this property, all IP addresses to be visualized are converted into an integer value representation with respect to their prefix depth. The resulting IP addresses' integer representations are then saved within a hashmap along with their aggregate feature values produced by the feature-generating tool[14]. This data mapping is used to reference each IP address and their integer representation. With a Hilbert curve[2] being able to map this integer representation to a coordinate value pair, each IP address is effectively assigned a pixel position. The process of managing and working with network data is visualized in Figure 2

Within Figure 2, we are able to see how data processing is handled and travels in the sequence described. A point of clarification to be made within the flowchart is the visualization tool encapsulates data filtering performed in the initial filtering of aggregate network traffic feature data produced. As a result of this data processing pipeline, the input data required for visualization of network traffic data is produced and prepared. Furthermore, from this data processing feature coloration occurs within each observed IP address for the features being studied enabling the resulting position associated with each observed IP address to take on a color. The final result is a visualization that demonstrates the structural distribution of all IP addresses as well as their associated feature value relationships demonstrated through color.

4.7 Visualization of Hilbert Curve and Legend

While plotting the visualization within a Hilbert curve with feature coloration is valuable, the coloration alone does not provided enough information to interpret the produce Hilbert curve plot with all observed IP addresses and their features. In order for data to be interpreted from the visualization, a legend is introduced along with the factors of color that are used to describe feature relationships. The legend associated with each Hilbert curve plot illustrates the colors produced from differing feature value pairs observed within the Hilbert curve plot. To enable feature value interpretation, feature values are labelled within the leg-

end to associate feature values with specific hues or saturation. This produced legend offers valuable insight into feature relationship representation within the Hilbert curve and allows for examination of how feature values are distributed across the Hilbert curve plot. The addition of these axis labels and markers provided to the legend and Hilbert curve plot is done using matplotlib[6], a library for data processing. The image created with matplotlib[6] demonstrates which feature corresponds to saturation or hue and creates a visual range of feature relationships based upon observed feature values.

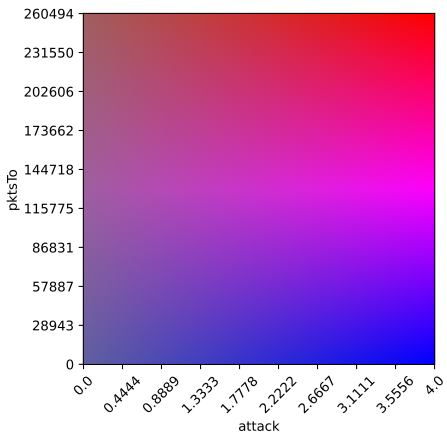


Figure 7: Legend produced for the examination of features: attack and packets to. Values on both axis are scaled by the observed max and minimum feature values. Color range portrayed is from 240 to 360 on the hue value chart.

As the primary output the visualization method produces two separate images: an annotated Hilbert curve with tick marks and labels and an annotated legend image corresponding to the Hilbert curve image with tick marks and feature value labels. To fit the images together so a direct comparison is easier to view, matplotlib[6] is used to plot both images side by side to improve interpretation of data. With the legend providing visual identification for feature relationship values, we can use this color legend to identify the structure of network traffic feature values by observing color distribution within the resulting Hilbert curve plot. To demonstrate this observation, Figure 8 illustrates an example image of the UO_DDoS dataset[12] with 50,000 attacking source addresses.

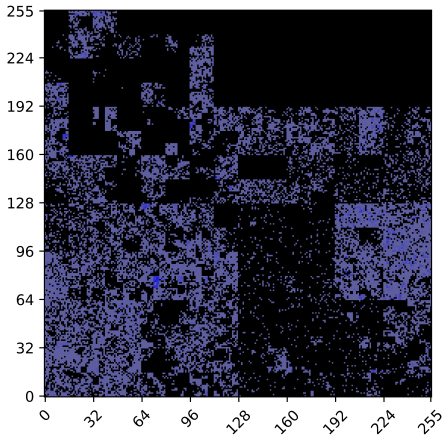


Figure 8: Hilbert curve image at depth /16 using UO_DDoS dataset 50000_dns

From Figure 8 we are able to use the corresponding legend and Hilbert curve image to observe the distribution of IP addresses and their feature relationships through coloring. Through the application of feature coloration as well as how data is handled and visualized using a Hilbert curve, examination of the distribution of IP addresses and feature relationships is visually interpretable. As a result, through visualization, observations and studies surrounding IP addresses, their structures, and their associated traffic features, may offer valuable insight into potential fractal patterns or behaviors that may emerge as a result of visualization.

4.8 Datasets Used

Dataset	Dataset
MAWILab [3]	Date observed: 2019-08-06. 212461 Sources. Packet-level Dataset.
Booters [15]	Date observed: 2013. 29323 Attack Sources. Packet-level Dataset.
UO_DDoS [12]	Date observed:2016-09-08 to 2016-10-31. 500 - 50000 Attack Sources. Packet-level Dataset.

Table 1: Data sets used in the study of structures of network traffic features using the visualization tool to visualize features observed within each data set.

Within Figure 1 are the data sets used within this study. The MAWILab [3] data set is used as a baseline network trace for evaluation within Case Study 1. Secondly, the Booters [15] data set is a combined data set of booter attack traffic collected for use to examine DDoS attacks. Thirdly, the UO_DDoS [12] data set is a combined data set consisting of both the USC ISI Mirai [12] data

set as well as network traces from MAWILab's [3] data set that was designed for the study and examination of DDoS defense systems.

4.9 Performance of Visualization

Dataset	Source Count	Time	Memory Usage
MAWILab [3]	212461	303.77s	53814kB
Booters - Booter 7 [15]	6045	42.60s	6488kB
UO_DDoS - 50000 DNS File [12]	50000	101.17s	16441kB

Table 2: Performance analysis for the data sets used. Time recorded is the entire elapsed time to process and visualize the network data. Memory usage is the memory used by the process for visualization and data processing.

5 Evaluation

In order to evaluate the efficacy of using visualization to examine network traffic data and to address existing limitations found within existing visualization techniques, this visualization method is applied to three real-word case studies. Each case study examines the application of this visualization method and its ability to assist in studying and observing network traffic data.

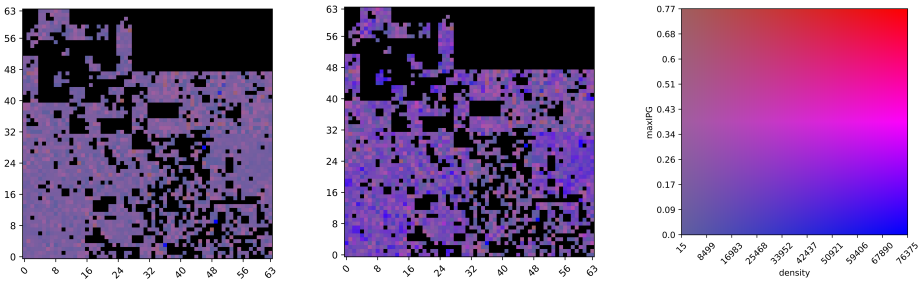
5.1 Case Studies

5.1.1 Case 1: Examining the Effects of Anonymization on IP Addresses

Within this case study we examine whether prefix-preserving anonymization tools such as cryptopan[9] affect the structural properties of observed IP addresses. This guiding question is especially important as researchers often work with anonymized network traces. Using visualization and comparing the structural distribution of IP addresses and feature values may offer insight into how anonymization may affect the structure of IP addresses. By applying visualization with this new method of feature coloration, changes in the structural distribution of IP addresses or their feature representation can be visually captured and observed. To apply the visualization tool to this case study, a comparison is performed between anonymized and non-anonymized IP addresses found within the MAWILab dataset[3] in order to observe any structural differences caused by anonymization.

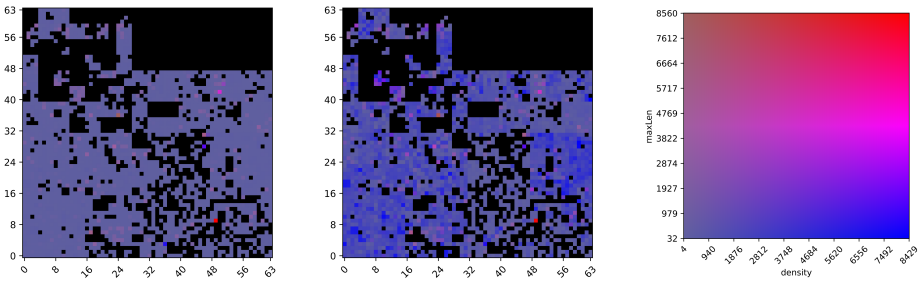
To perform this comparison, an anonymization tool from the USC ISI ANT project[10] called 'dag_scrubber'[11] is used to anonymize the MAWILab dataset[3]. This anonymization tool[11] anonymizes network traces using cryptopan[9] along with a randomized keyfile for randomization. Through this anonymization tool, an anonymized form of the MAWILab dataset[3] is generated. Using both the original and anonymized network trace, high-level aggregate feature values are

produced for both files from the feature-generating tool[14]. These resulting feature files are subsequently both used within the visualization method to visualize both anonymized and non-anonymized network traffic features. This case study offers valuable insight into the effects of anonymization on the structure of IP addresses but also demonstrates effectiveness of using visualization to study network traffic data. Through visualization it is possible to discern clear trends and patterns that emerge within the visualized network traffic data. This is directly demonstrated through the change in feature distribution and color found within Figures 9, 10, and 11.



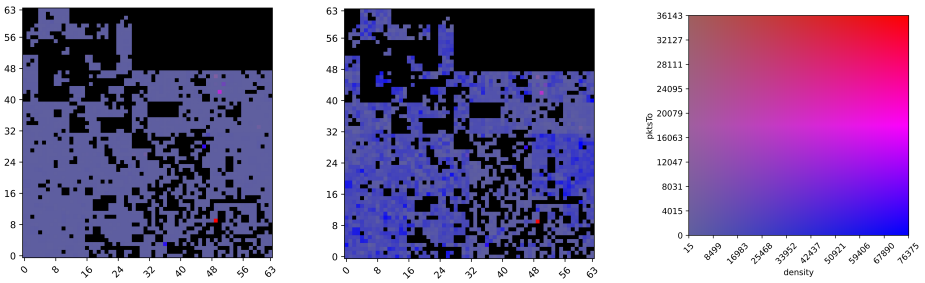
(a) Original PCAP (b) Anonymized PCAP (c) Legend

Figure 9: /12 Comparison of Feature: MaxIPG and Density. Data set used: MAWILab[3] 08/06/2019 trace



(a) Original PCAP (b) Anonymized PCAP (c) Legend

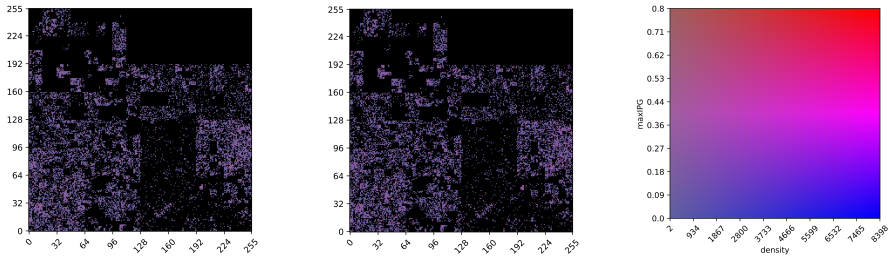
Figure 10: /12 Comparison of Feature: MaxLen and Density. Data set used: MAWILab[3] 08/06/2019 trace



(a) Original PCAP (b) Anonymized PCAP (c) Legend

Figure 11: /12 Comparison of Feature: pktsTo and Density. Data set used: MAWILab[3] 08/06/2019 trace

From Figures 9, 10, and 11 it is evident that there is significant change observed visually between the original and anonymized IP addresses. Through the visualization of both files it can be seen that the original network trace has denser areas of saturation with little color diffusion compared to the anonymized network trace which exhibits a significant difference in feature coloration. With feature coloration being an important addition to this visualization method, it can be observed that the feature relationships observed within the visual comparison changes as a result of anonymization. Further examination reveals that after anonymization the feature values recorded for anonymized network traces for features of source density increase significantly. From the increase in feature value the feature coloration of the visualization directly reflects this development offering visual insight into feature values and their IP addresses. As a result, we can directly see that anonymization plays a large role in the structuring of network traffic within the IP address space, such that areas with little saturation seem to increase in saturation or density after being anonymized. An interesting observation found by applying this visualization method to larger prefix lengths reveals that this visual difference is not readily noticeable within larger prefix lengths. This can be seen in Figure 12 with the /16 image comparison for the maxIPG feature between both files.



(a) Original PCAP (b) Anonymized PCAP (c) Legend

Figure 12: /16 Comparison of Feature: maxIPG and Density. Data set used: MAWILab[3] 08/06/2019 trace

From Figure 12 a visual distinction can be made between the earlier figure comparison and this current figure being examined. The drastic visual difference observed earlier is not readily seen within the figure above. However, examination of feature values found within visualization of larger prefix lengths such as /16 share a similar development in feature values being increased significantly.

Through the application of visualization and the use of the feature coloration to examine network traffic feature data, it can be visibly seen the effect anonymization has on the structure of IP addresses and their traffic features. This is readily seen through the change in feature coloration found between the original and anonymized network trace. The behavior observed requires additional investigation in order to determine why this shift in coloration occurs as well as why similar visual trends are not observed in larger prefix lengths.

5.1.2 Case 2: Identification of Strange Patterns in Booters Data Set.

In case study 2, leveraging the visualization method, visualization is used to examine how the structure within a Hilbert curve changes when examining specific prefixes at differing levels of detail. Using the addition of the zooming utility within the visualization method, it is possible to examine areas or specific prefixes within a Hilbert curve at a greater level of detail or granularity. As a result, by generating increasingly detailed visualizations of network traffic data, it is possible to examine the structure of observed IP addresses at differing levels of detail. This is studied by continuously increasing the prefix length examined for a specified prefix.

To apply this visualization method, the Booters[15] data set is used. The Booters[15] dataset is comprised of booter attack source addresses previously used in studies of DDoS attacks. Applying the zooming utility defined within the visualization method to this dataset will offer valuable insight into how the structural distribution of IP addresses change when examined at different levels of granularity.

The network trace for examination within this case study is Booter 7[15]. This network data was selected due to the unique diagonal clustering distribution[15]

observed when visualized. Therefore, to gain insight into this diagonal distribution observed within the network data, visualization is applied. In order to select a prefix to examine within the diagonal distribution observed for Booter 7[15], IP addresses which demonstrated a high level of saturation or attack source density were selected. Using the Hilbert curve's mapping functionality[2], by identifying a coordinate with high level of feature coloration, mapping this coordinate value back to an IP address provided a specific prefix for examination. The following figures illustrate the application of the zooming utility introduced within the visualization method on the 123.233.58.0/24 prefix at increasing prefix lengths:

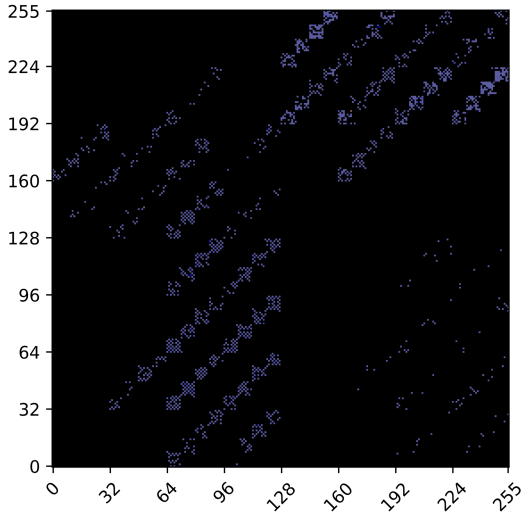
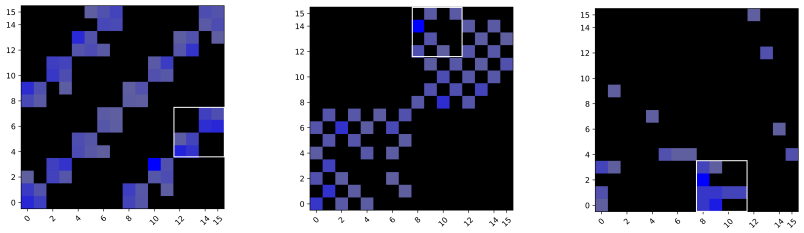
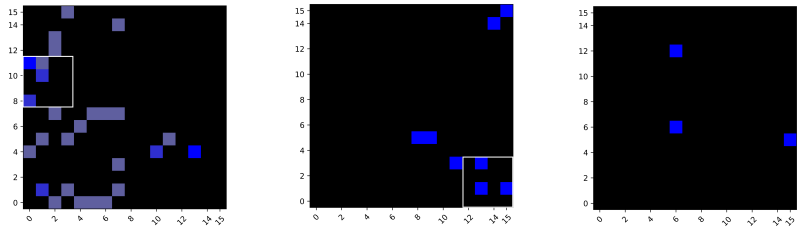


Figure 13: Initial full image of the entire /16 space for Booter 7[15]



(a) /12 (b) /16 (c) /20

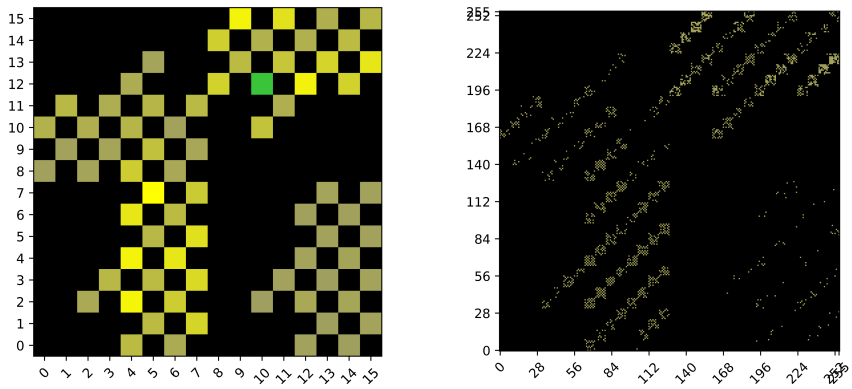


(d) /24 (e) /28 (f) /32

Figure 14: Exploration of 123.233.58.0/24 from Booter 7 within the Booters data set[15]. Feature represented by the blue is attack with increasing saturation indicating more attack sources. Each pixel within the plotted image corresponds to a network address with a length like their label. Data set used: Booters Data Set[15] (1–9)

From the zooming utility within the visualization method, examinations of a specific prefix within the Hilbert curve is illustrated through the use of a white square to denote area of visualization. As can be seen in subsequent images, the area within this white square is visualized at a greater level of detail or granularity enabling observations to be made regarding the structural distribution of IP addresses at differing levels of detail.

This is demonstrated through Figure 13 and Figure 14, as it can be seen within sub-figures 14aand 14b there exists a diagonal distribution of IP addresses, however, subsequent visualizations become increasingly naturally distributed. As shown in sub-figures 14c 14d 14e,and 14f, this diagonal distribution is not visible in the structural distribution visualized. A primary reason as to why this diagonal distribution is not readily visible in following visualizations is due to the effect of increasing granularity. By studying the structure of IP addresses at increasing detail more information is exposed regarding the distribution of IP addresses. When viewing IP addresses at a lower prefix lengths or lower levels of detail, many of these IP addresses tend to merge together resulting in high level observations of structure. This is demonstrated through the Figure 15 which illustrates this observation.



(a) Booters 7 /8 Visualization

(b) Booters 7 /16 Visualization

Figure 15: Comparison in granularity between visualization of Booters 7 data at /8 and /16 representation.

From Figure 15, this change in granularity and level of detail and its effect on the interpretation of the structure of IP addresses demonstrates the importance of the visualization method introduced. By applying the new zooming utility to the Booters 7 data set[15], observations of a specified prefix regarding its underlying structural distribution at increasing prefix lengths can be seen. Without the use of this new zooming utility or visualization, examining and studying the structural distribution of Booter 7[15] IP addresses at a fixed granularity prevents closer examination and discovery of network traffic data. Therefore, through the use of visualization, and application of the visualization method discussed, it is possible to examine changes in the structure of observed IP addresses within network data by focusing and increasing the granularity of visualizations.

5.1.3 Case 3: Illustration of DDoS Indicators Using Feature Coloration

To gain insight into how network traffic features found within network traffic data may offer insight into DDoS attack indications, in case study 3 an examination is conducted regarding feature relationships and their correlation to DDoS attack sources. Through visualization of prefix-level network features and color through feature coloration, valuable insight into whether certain network traffic features exhibit a high degree of association to DDoS source IP addresses can be found.

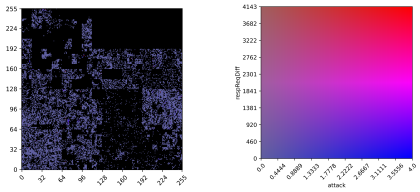
To study the feature relationship between network traffic features and DDoS attack sources, the UO_DDoS dataset[12] is used as the network data of study. This data set[12] contains three differing attack types: ICMP, SYN, and DNS, each grouped into three sub data sets. For this case study, only the DNS DDoS

traffic data is used for visualization.

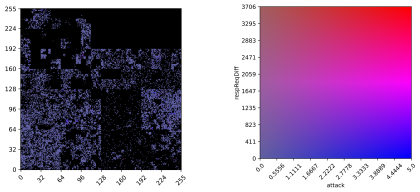
In order to examine if there exists a feature relationship that defines whether a source IP address is likely to be a DDoS source IP address, visualization using feature coloration is applied. With the DNS DDoS data set being the network traffic data of examination, using the visualization method discussed, data is processed and produced by the feature generation tool[14] which is subsequently visualized through the mapping and coloring using a Hilbert curve. The resulting feature coloration defines a relationship between network traffic features such as respReqDiff, pktsFrom, bytesTo, and a few additional traffic features. Visualization using feature coloration is especially important within this study, as the feature relationship defined by color can provide insight into the structural distribution of IP addresses while identifying if there exists a strong correlation between DDoS attack sources and network traffic features.

Through the use of DDoS attack source address density as a feature for examination, the feature coloration and potential correlation between network traffic features and DDoS source addresses can be identified through a series of visualizations. Within each visualization it is possible to examine how the feature relationships change based upon the density of DDoS source addresses and how this feature relationship may provide valuable insight into determining if an IP address is a DDoS source address.

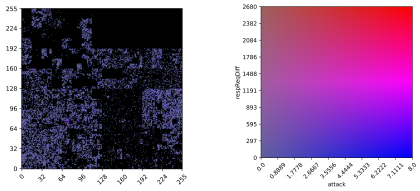
This feature coloration applied through visualization provides insight into identifying specific traffic features that demonstrate high correlation to DDoS attack source addresses. Figures 16 and Figure 17 illustrate this idea through the visualization of network traffic features with increasing counts of DDoS source address density.



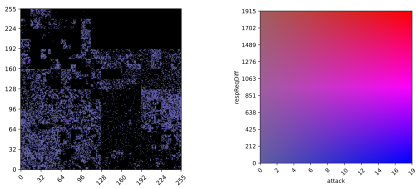
(a) 500 DNS



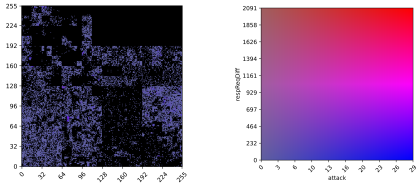
(b) 1000 DNS



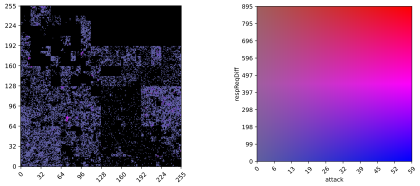
(c) 2000 DNS



(d) 5000 DNS

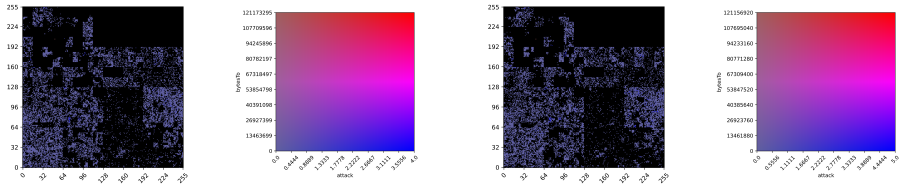


(e) 10000 DNS



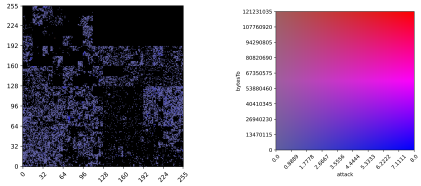
(f) 20000 DNS

Figure 16: Structural /16 visualizations for DNS files 500, 1000, 2000, 5000, 10000, and 20000. Features: Response-request difference and DDoS Source Address Density.

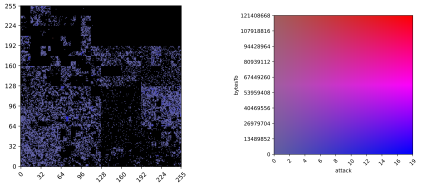


(a) 500 DNS

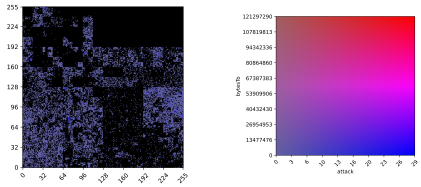
(b) 1000 DNS



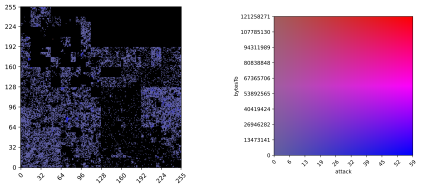
(c) 2000 DNS



(d) 5000 DNS



(e) 10000 DNS



(f) 20000 DNS

Figure 17: Structural /16 visualizations for DNS files with DDoS source address density counts of: 500, 1000, 2000, 5000, 10000, and 20000. Features: BytesTo and DDoS Source Address Density.

From Figures 16 and Figure 17, as the number of DDoS source address density values increase, the saturation found from the subsequent visualization as shown within subfigures (b) through (f) increases significantly. Therefore, using the concept of feature coloration within the visualization, a strong correlation correlates to the presence of high values of hue and saturation as shown through the interpretation of the Hilbert curve plot using a legend. In close examination of the visualizations above which illustrate the feature relationship found between DDoS source address density and network traffic feature respReqDiff, feature correlation can be identified through feature coloration. Within the examination between features of respReqDiff and DDoS source address density, it can be seen that from the feature coloration shown within the produced visualization there is a direct correlation between these two features. This is illustrated through the Figures 18 and Figure 19 demonstrating the feature relationship identified.

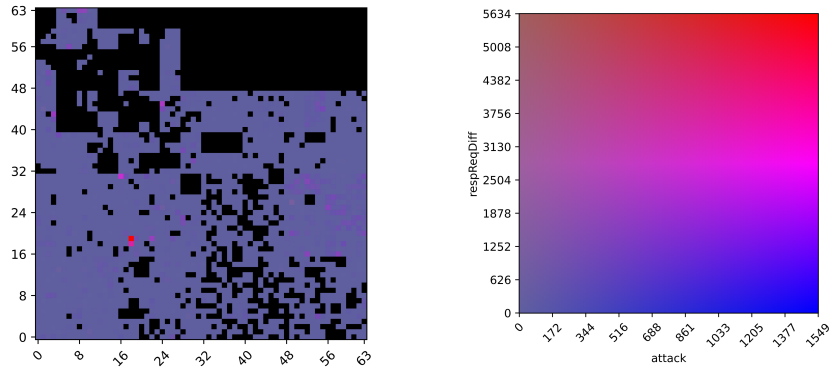


Figure 18: /12 Visualization of UO_DDoS 50000 Source address data using features of respReqDiff and DDoS Source Address density. Coloration produced demonstrates a direct correlation between high levels of respReqDiff and DDoS address density.

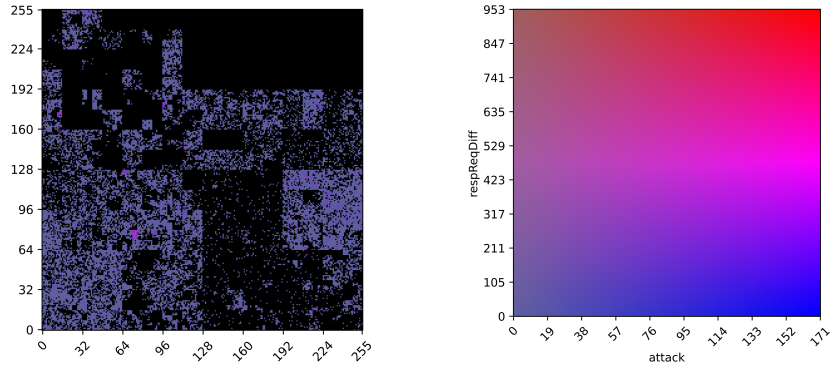


Figure 19: /16 Visualization of UO_DDoS 50000 Source address data using features of respReqDiff and DDoS Source Address density. Coloration produced demonstrates a direct correlation between high levels of respReqDiff and DDoS address density.

This feature relationship between the DDoS source address density and respReqDiff is visible through the intersection of their corresponding feature values found within the associated legend. The high level of correlation is further evidenced through the clustering of red, pink, and light pink pixels visible within both figures which demonstrates the highly coupled relationship exposed using visualization and feature coloration. The behavior observed within Figures 18 and Figure 19 is not found in other network traffic features studied such

as bytesTo visualized above, pktsTo, pktsFrom, and additional other features studied.

As a result, through the use of visualization to study the relationship between features and observed IP addresses, it can be found that the respReqDiff feature within network traffic data shares a high level of correlation to DDoS source address density demonstrated through visual coloration. Therefore, with the application of visualization as a method of studying network traffic, valuable insight can be gained regarding feature relationships found within observed IP addresses.

6 Conclusion and Future Work

6.1 Conclusion

Through the development and approach applied within the visualization method introduced, visualization is able to better examined network traffic data using the addition of feature coloration and zooming utility for discovery. These additions to the visualization method offer valuable additional insight into network traffic features and the structural distribution of network traffic data by introducing additional methods to interpret or visualize data.

Leveraging this visualization method, valuable insight and findings can be found regarding the use of visualization to study network traffic data. Through the series of case studies applying visualization, various findings are discovered which demonstrate the effectiveness of visualization. This is shown through the visible differences found between the visualizations of anonymized network traces, resulting in visibly differing feature representations. Furthermore, using visualization to examine how IP addresses are structurally distributed at different levels of detail and granularity, valuable insight into how the structure of IP addresses changes based upon the level of detail and granularity is found. Additionally, with the visualization method's feature coloration which offers the examination of feature relationships found in observed IP addresses, it is possible to identify relationships between features which may offer insight into the identification of network traffic behavior. From the introduction of a new visualization method, the addition of new visualization techniques that address prior limitations found within existing visualizations, and the application of visualization, visualization proves to be a valuable asset for the examination and study of network traffic data.

6.2 Future Work

Some future considerations that could provide additional insight or explanation to some behaviors observed or discovered are discussed within this section.

An important consideration that might offer additional insight into the examination of network traffic data visualization is the investigation of the root causes observed within these applications. Identifying how these behaviors are

created and how they originate may offer additional insight into the underlying characteristics or behaviors that network traffic data possesses.

Furthermore, within this visualization method the color factor of lightness is fixed to reduce the dimensions of analysis to two. Including lightness as an additional color factor within future examinations of network traffic data may offer additional insight or discovery into feature relationships by enabling a three-dimensional examination of network traffic features.

Another consideration for future work is the improvement of runtime performance for the processing and generation of visualizations, as parallelizing many existing operations within the visualization method taken will offer significant improvements to run-time efficiency. This visualization is implemented using GoLang to process and visualize data.

References

[1] Sandeep Bajaj, Lee Breslau, Deborah Estrin, Kevin Fall, Sally Floyd, Padma Haldar, Mark Handley, Ahmed Helmy, John Heidemann, Polly Huang, et al. Virtual internetwork testbed: Status and research agenda. 1998.

[2] Andrew Brampton. *Hilbert : Go package for mapping values to and from space-filling curves*, 2018. package version: v0.0.0-20181122061418-320f2e35a565. URL: `https://pkg.go.dev/github.com/google/hilbert#NewHilbert`.

[3] Romain Fontugne. Mawilab. URL: `http://www.fukuda-lab.org/mawilab/v1.1/2019/08/06/20190806.html`.

[4] John Heidemann, Zi Hu, and Yuri Pradkin. Towards geolocation of millions of ip addresses. *USC/ISI Technical Report*, ISI-TR-680:1–7, May 2012.

[5] John Heidemann, Yuri Pradkin, Ramesh Govindan, Christos Papadopoulos, and Joseph Bannister. Ant censuses of the internet address space, 2015. URL: `https://ant.isi.edu/address/`.

[6] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007. `doi:10.1109/MCSE.2007.55`.

[7] Eddie Kohler, Jinyang Li, Vern Paxson, and Scott Shenker. Observed structure of addresses in ip traffic. *IEEE/ACM Transactions on Networking*, 14(6):1207–1218, Dec 2006. `doi:10.1109/tnet.2006.886288`.

[8] Ghulam Memon, Reza Rejaie, Yang Guo, and Daniel Stutzbach. Large-scale monitoring of dht traffic. In *IPTPS*, volume 9, pages 1–11, 2009.

[9] Meisam Mohammady, Lingyu Wang, Yuan Hong, Habib Louafi, Makan Pourzandi, and Mourad Debbabi. Preserving both privacy and utility in network trace anonymization. In *Proceedings of the 2018 ACM*

SIGSAC Conference on Computer and Communications Security. ACM, jan 2018. URL: https://doi.org/10.1145%2F3243734.3243809, doi:10.1145/3243734.3243809.

[10] ANT Project. The ant lab. URL: https://ant.isi.edu/index.html.

[11] ANT Project. dag_scrubber. URL: https://ant.isi.edu/software/dag_scrubber/index.html.

[12] ANT Project. Mirai-frgp-scanning-20160908, May 2017. URL: https://ant.isi.edu/datasets/readmes/Mirai-FRGP-scanning-20160908.README.txt.

[13] Amir H Rasti, Nazanin Magharei, Reza Rejaie, and Walter Willinger. Eyeball ases: from geography to connectivity. In *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*, pages 192–198, 2010.

[14] Reza Rejaie and Ramakrishnan Durairajan. URL: https://onrg.gitlab.io/.

[15] Jair Santanna, Roland Rijswijk-Deij, Rick Hofstede, Anna Sperotto, Mark Wierbosch, Lisandro Granville, and Aiko Pras. Booters - an analysis of ddos-as-a-service attacks. In *2015 IFIP/IEEE International Symposium on Integrated Network Management (IM)*, pages 243–251, 06 2015. doi:10.1109/INM.2015.7140298.

[16] Jeffrey Shallit. Hilbert's spacefilling curve described by automatic, regular, and synchronized sequences, Jun 2021. URL: https://arxiv.org/abs/2106.01062.

[17] Eugene Tan and Chris Misa. *Multifractal IP : Visualization of IP addresses using a hilbert curve*, 2022. package version: v0. URL: https://github.com/chris-misa/multifractal-ip.

[18] Duane Wessels and Roy Arends. Ipv4 heatmap: Generate hilbert curve heatmaps of the ipv4 address space. URL: https://github.com/measurement-factory/ipv4-heatmap.