Visualization of Network Traffic for Network Analysis and Intrusion Detection

Anthony Orlowski

Introduction

To enable their organizations business goals, enterprise information technology (IT) departments build, secure, and operate computer networks. To enable global collaboration and business they extend these networks to the Internet exposing their organizations to cyber-attacks. IT departments uses special purpose network traffic monitors and tools to collect and analyze computer network traffic to plan, build, forecast, and protect their organizations' computing resources and data. By analyzing computer network traffic, IT analysts provide valuable insight into these efforts, however the volume, variety, velocity, and veracity of the traffic complicates its analysis in its raw log form. Network engineers, IT architects, cyber security analysts, and IT management struggle to generate actionable insights from millions of network traffic logs. In this paper, we discuss the visual analysis of network traffic including prototype chord diagrams and treemaps that help analysts generate insights in an interactive exploration of network traffic data flows. These insights can be applied to multiple IT department decisions.

Computer network traffic is machine generated data that is encapsulated in layers of network protocols to enable communication between computers attached to a network. When viewed from the perspective of the TCP/IP model these layers include the application layer, transport layer, network layer, and the physical layer. To analyze communication patterns and flows in a network we can identify unique communication sessions using what is commonly known as the five-tuple which consists of the source IP address (network layer), source port (transport layer), destination IP, destination port, and transport protocol (transport layer). The five tuple identifies a bi-directional flow of communication between a client and a server. Using a timestamp and header flags from transport protocols—including session establishment and teardown identifiers—we can distinguish unique sessions over time. By dropping the application data portion of network traffic, or using a smaller set of fields computed from its content—known as deep packet inspection—we can reduce the volume of data required for analysis, but still answer many IT department insight needs.

Table 1. Data Analysis Reporting

Feature	Unique Values	Sample Values
Ipv4 Src Addr	224,568	101.132.106.252
Ipv4 Src Addr (group)	10	172.31.64.X, INTERNET
Ipv4 Dst Addr	33,614	172.31.64.5
Ipv4 Dst Addr (group)	9	172.31.65.X, INTERNET
L4 Src Port	65386	50515,
L4 Dst Port	36309	23, 22, 80, 53
IP Protocol Types	6	1, 2, 6, 17, 47, 58
Attack Types	15	Bot, Benign, FTP Bruteforce

1 Insight Needs

IT Departments use network traffic analysis, specifically network session logs, to plan, build, forecast, and protect an organization's computer network. There are many unique stakeholders residing in the organizations' IT departments.

1.1 Stakeholder Analysis

Many distinct stakeholders use network traffic analysis to generate insight: network engineers plan for network capacity, monitor the network for failures, and verify network configuration; IT architects ensure business critical applications and data are accessible and protected, cybersecurity analysts identify and investigate network incidents, and IT management creates policy to meet organizational security and service-level-agreement goals. These objectives include overlapping needs for identifying distributions of network traffic in a network, trends of network traffic volume and variety over time, ordered, ranked, and sorted analysis of network traffic types, and relationships between communicating machines. Our chord diagram is apt at providing relationships and comparisons of network traffic between individual machines and subnets. Analysts use different filters to select traffic that is required for their insight needs.

2 Data Acquisition

In real-world operation, IT departments use a variety of network taps, storage servers, and network traffic analysers such as CICFlowmeter [1], or Zeek [2] to generate network session data from raw network traffic. As organizational network traffic contains sensitive and proprietary data, it is difficult to obtain a realistic dataset for academic analysis. For this paper, we analyze traffic from a simulated network built by the Canadian Institute for Cybersecurity [3]. With a focus on building a standard dataset for network traffic anomaly detection, the CIC built a mock enterprise network consisting of six subnets including 420 user machines, 30 servers, a simulated Internet, and an attack network comprising of 50 machines. Network traffic was collected at each machine in the form of raw network capture (PCAP), and run through a statistical profiling tool called CICFlowmeter. The traffic is a combination of simulated benign network traffic, and seven categories of cyber-attacks. However, the parsed network session data they provide does not include the complete five-tuple, preventing the flow analysis we seek. Fortunately, a group from the University of Queensland generated a new set of logs from that same study, named NF-CSE-CIC-IDS2018, that include these fields which we use for our analysis [4].

2.1 Description of Data

The data is approximately 3.14 GB in size, and consists of 18,893,708 network sessions consisting of 43 features. Attacks account for 2,258,141 (11.95%) sessions and the remaining 16,635,567 (88.05%) are benign. The records contain 43 features. For comparison, these logs were generated from raw network capture in excess of 220 GBs. Besides the five tuple and timestamp that we are primarily interested in for this paper, there are a variety of statistical measures of the network sessions that describe size, length, and protocol operation of the network sessions. This data set is also labelled to identify if it was part of an attack. See Table 1 for a description of the data fields we focused our analysis on.

3 Analysis Methods

Our analysis is primarily a network study that investigates the communication relationships between machines. However, due to the large number of machines and types of network traffic we are required to aggregate and filter the network traffic to simplify visualizations to contain only the necessary data needed to meet user's insight needs.

3.1 Analysis Methods

We first present the analyst a macro view of network communication traffic. We first group source and destination IPs into a set of 10

distinct groups. These groups reflect the subnet and departmental boundaries in the simulated organization. For example, the CIC labelled the five user subnets R&D, Management, Technician, Operations, IT, and implemented a separate server room subnet. We add additional groups to distinguish a DNS server that is an outlier in number of received network sessions, the Internet, the non-routable IP address 0.0.0.0, and a group of unidentified internal machines that receive a negligible amount of traffic. Next, we count the number of network sessions between each source group and destination group.

Next we provide a prototype interactive analysis by allowing the user to zoom in on a subnet or machine from the current view, or filter traffic from the current view based on a feature or time. In our example, we focus on filtering network traffic specifically within the sever subnet and the aggregated Internet group. We compute the number of network sessions between each machine in the server network and the aggregated group.

4 Visualizations

Next, we present two prototype visualizations designed to provide analysts an understanding of the network communication relationships in the network and provide a volumetric depiction of the number of network sessions by grouped type.

4.1.1 Chord Diagrams

The first prototype, demonstrated by Fig. 1 and Fig. 2., is a chord diagram. In [5] we identified similar methods to visualize network flows and anomalies in industrial networks. We however, focus on enterprise networks. We generated the chord diagrams with the assistance of a Tableau Prep workbook template by Marc Reid [6] and a Tableau workbook template by Luke Stanke [7] and improved on by Marc Reid [6]. These templates did most of the visualization heavy lifting including node layout, chord sizes and the brushed and linked bar charts. We created a separate Tableau Prep workbook to compute the aggregate network sessions for the input and then modified the templates to work with our data source.

In the diagram, we utilize chords to represent the number of network sessions from a source to a destination. The width of the of the chord represents the number of network sessions, and as chords split off the output node, they maintain their proportion of the width based on the number of network sessions to that destination. Text labels identify the nodes.

The bars in the FROM Dimension bar chart are the total number of network sessions by source. They further provide a brushed-and-linked view of the chord diagram and the TO Dimension bar chart. Selecting a node will highlight its chords in the chord diagram, and update the TO dimension bar chart with that node's network sessions by destination. We plan to further filter traffic by features, such as destination port or a time window. For example, in Fig. 2. we have filtered to only the communication between the server network machines, and an aggregated internet node. These user filters will interactively update the diagrams.

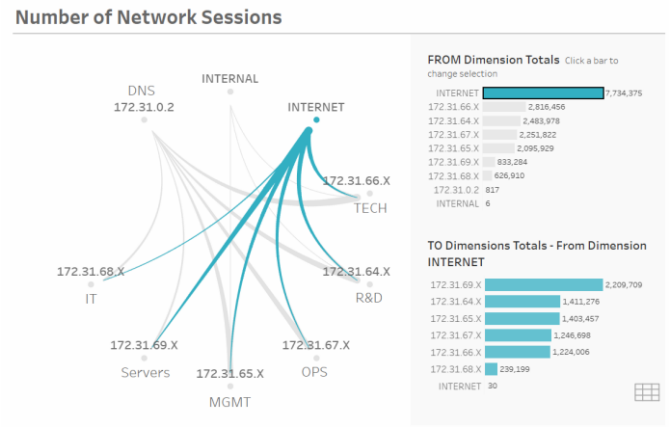

Fig. 1. Chord diagram for all traffic by source and destination.

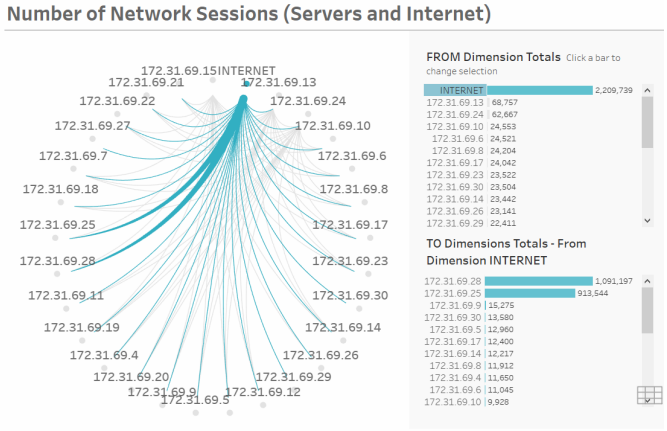

Fig. 2. Chord diagram within the server subnet and between the internet group.

4.1.2 Treemaps

The second set of prototypes are interactive tree-maps. In Fig 3. We group the network sessions by source group, destination group, and destination port. Text labels in the largest boxes reflect these data fields, while a tooltip hover on boxes reveals these data fields as well as the number of network sessions. The area of the boxes also represents the number of network sessions. Color and position are used to group the boxes by destination group. The user is able to interactively update the treemap by selecting and excluding boxes they are not interested in. The treemap is then recomputed and displayed. While not implemented in the prototype, the user will be able to easily filter the network sessions displayed. In Fig 4. we have displayed only those network sessions that were sourced from the internet, and removed large groups of expected network traffic such as web traffic to internal websites. This reflects how a cyber security analysts may

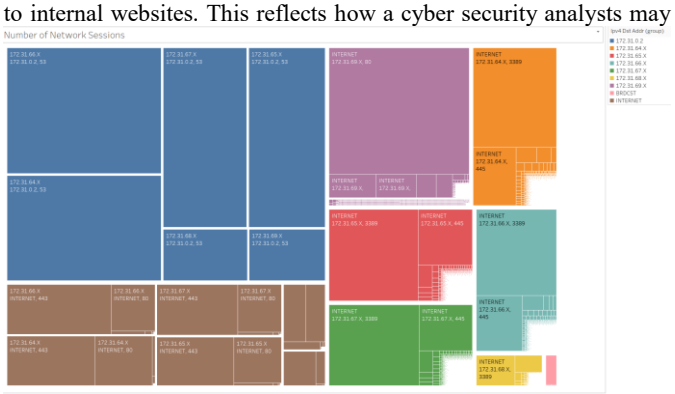

Fig. 3. Treemap of all network sessions by source, destination and destination port.

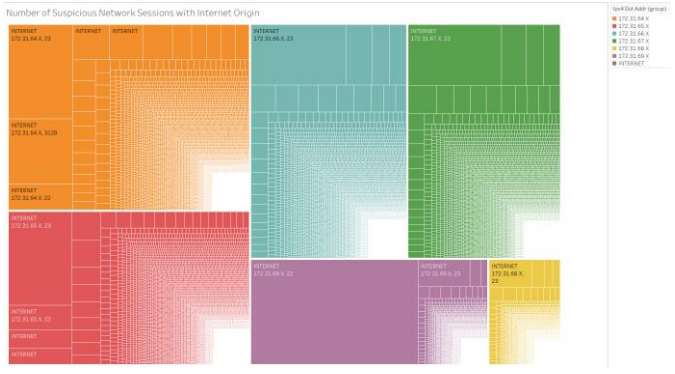

Fig. 4. Treemap of all internet initiated network sessions by destination and destination port.

prune expected traffic to identify a treemap of suspicious network traffic to investigate.

5. Interpretation of Results

By demonstration we show how these visualizations can provide successful insight to IT departments. One stakeholder that benefits greatly is cyber security analysts. For example, we can identity suspicious traffic based on source, destination, and traffic filters. The chord diagram further provides powerful comparison insight that allows you to compare traffic flows to similar nodes which increases an analyst's ability to identify unusual behavior. As demonstrated by our pivot from Fig 1. To Fig 2, we have identified two servers in the server subnet that receive most of the traffic sourced from the internet. Looking at the CIC data labels we can verify that these two machines, 192.168.69.25 and 192.168.69.28, are indeed victims of denial of service (DOS) and brute force attacks that generate many network sessions.

Another pair of interesting hosts are 192.168.69.13 and 192.168.69.24, as we see they communicate to all other servers in their subnet with a small number of network sessions. This could indicate port scanning or network enumeration attack, which we verify is indeed a part of the infiltration attack carried out by the CIC researchers on these machines.

In the treemap in fig. 4 we have filtered out internet sourced traffic that we expected to see. For example, traffic inbound to web servers on port 80 or 443, which is commonly used for HTTP web servers. While this filter misses some attacks, such as the SQL injection and web attacks that exist in the data set, it does quickly highlight concerning SSH and telnet access from the internet directly to a variety of subnets. Indeed, the largest purple box in fig 4. includes an SSH bruteforce attack that was carried out against the server subnet. The other concerning terminal traffic, while concerning in a real network, is likely an artifact of the simulated environment, highlighting limitations of simulated network traffic datasets.

We further see in fig 4. many small boxes indicating many combinations of source IP, destination IP, and destination port. In normal traffic profiles, destination ports of servers are generally a set of well-known application ports. Here we see a far greater number than we expect to see which could indicate a comprehensive port scan, or issues with our network session parsing tool.

Another cause for concern is the internet traffic destined to subnets other than the server subnet. As these are user workstation subnets, we do not expect internet hosts to initiate communication with these machines. While these are not a part of the attack data set, we see that there is still room for improvement of this simulated network's routing and firewall rules to more closely reflect modern enterprise networks. Likewise, there is no traffic between the user subnets and the server subnet, indicating this network is lacking intranet data flows typical to an organization.

Network engineers and IT architects can use these visualizations to conduct capacity planning. For example, in both visualizations we see the DNS server is a single host that is the recipient of most of the internally sourced traffic, this may identify an opportunity to introduce additional recursive resolvers lower in the network architecture to free up vital network bandwidth in the network core. While we only looked at the number of network sessions, we could easily aggregate session bytes, session length, or number of session packets to provide the specific insight users are looking for. Users can also use these visualizations to quickly verify network configurations. For example, they can quickly question "Why does our external firewall permit direct session initiation to user workstations?"

IT management is further interested in verifying that their critical business applications and data are protected and available. These network visualizations provide an easy way to comprehend if network firewall rules are appropriate, and provide a way to monitor active network failures. For example, "why is our external website not receiving any traffic this hour, did it crash?"

Beyond demonstration by example, we can validate these visualizations by having cyber security analysts filter and highlight suspicious network traffic. As this is a labeled dataset, we can measure how accurate they are at identifying network attacks with just the raw logs with an appropriate log analysis tool, and then compare this to their accuracy when we further provide them our visualization tools.

6. Challenges and Opportunities

There are many challenges and opportunities that arose while generating these visualization prototypes.

For example, we manually conducted the analysis to generate the aggregate data for each prototype using Tableau Prep. This analysis needs to be automated to enable the envisioned interactive pivoting and filtering, enable additional analyses on other network session features such as size and length, and to interface with streaming network data.

The size of the data set used, at 3GB initially, and 1.3GB after removing unneeded fields, was reaching usability limits for Tableau on a modern laptop computer. Investigations into larger computing resources or a tool more apt for a large data set is needed.

Additional mapping of raw port numbers to protocol names and host IP addresses to hostnames will further enable analysts to identify suspicious traffic.

Further opportunities exist to turn this dataset and visualization into a game designed to train cyber security analysts to conduct network traffic analysis. The labeled data gives way to score them on their ability to identify malicious patterns they highlight in the visualizations.

As identified by Sec 5. we see there are many patterns in this simulated enterprise environment that do not accurately reflect real world networks. This highlights the need for more realistic network traffic data sets for network analysis and intrusion detection studies.

Acknowledgments

The author wishes to thank Marc Reid and Luke Stanke for their valuable tutorials and templates on how to generate a chord diagram in Tableau, and Katy Borner, Andreas Bueckle, and the teaching staff of ENGR-E 583 "information visualization" Spring 2021 for their valuable visualization insights and tutorials.

References

[1] Lashkari, Arash Habibi. CICFlowmeter. June. 2020. [Online]. Available: https://github.com/ahlashkari/CICFlowMeter

[2] Zeek. February. 2020. [Online] Available https://zeek.org/

[3] CIC IDS 2018Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron Spectroscopy Studies on Magneto-Optical Media and Plastic Substrate Interface," *IEEE Trans. Magnetics,* vol. 2, pp. 740-741, Aug. 1987. (IEEE Transactions)

[4] M. Sarhan, S. Layeghy, N. Moustafa, and M. Portmann, Netflow datasets for machine learning-based network intrusion detection systems, 2020. arXiv:2011.09144 [cs.NI]

[5] Iturbe, Mikel & Garitano, Iñaki & Zurutuza, Urko & Uribeetxeberria, Roberto. (2016). Visualizing Network Flows and Related Anomalies in Industrial Networks using Chord Diagrams and Whitelisting. 99-106. 10.5220/0005670000990106.

[6] Reid, Mark, Creating a Chord Diagram with Tableau Prep and Desktop. [Online]. Available https://datavis.blog/2020/07/02/creating-chord-diagram-in-tableau/

[7] Stanke, Luke, Tutorial Chord Diagram. February. 2019. [Online] Available https://www.tessellationtech.io/tutorial-chord-diagram/

Appendix: Figures Repeated for Visibility

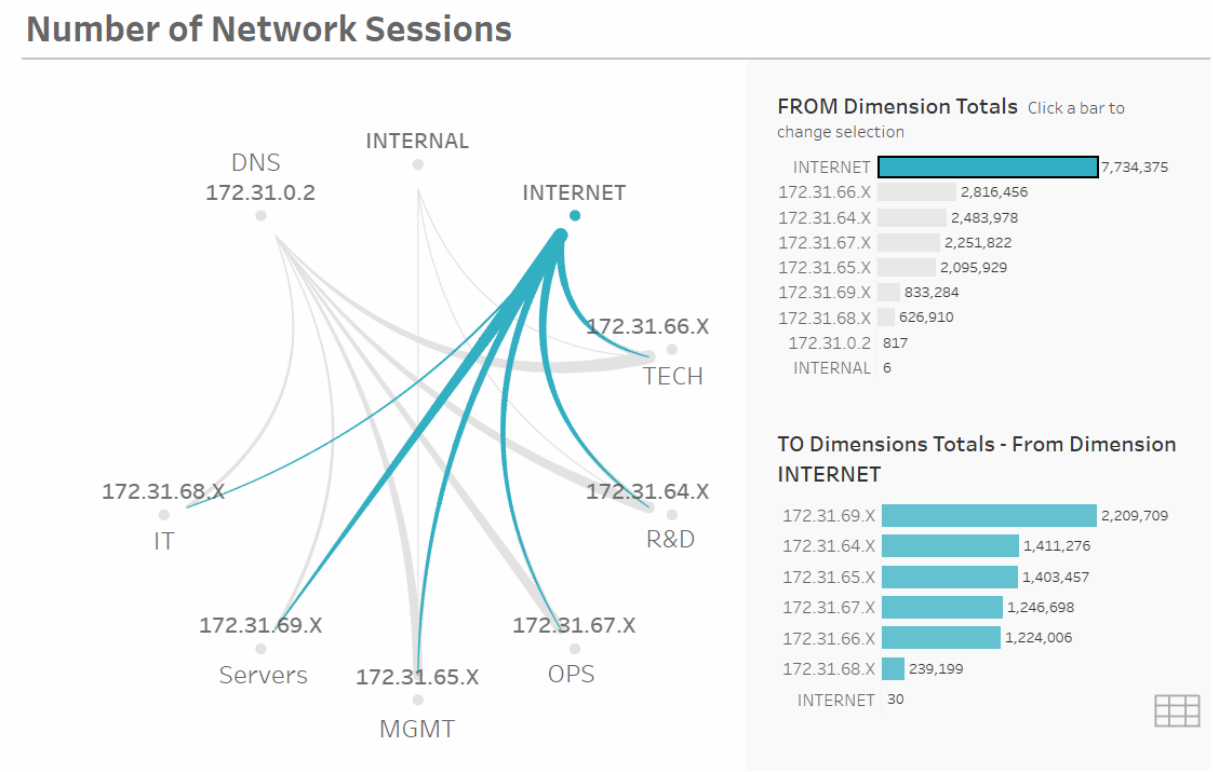

Fig. 1. Chord diagram for all traffic by source and destination.

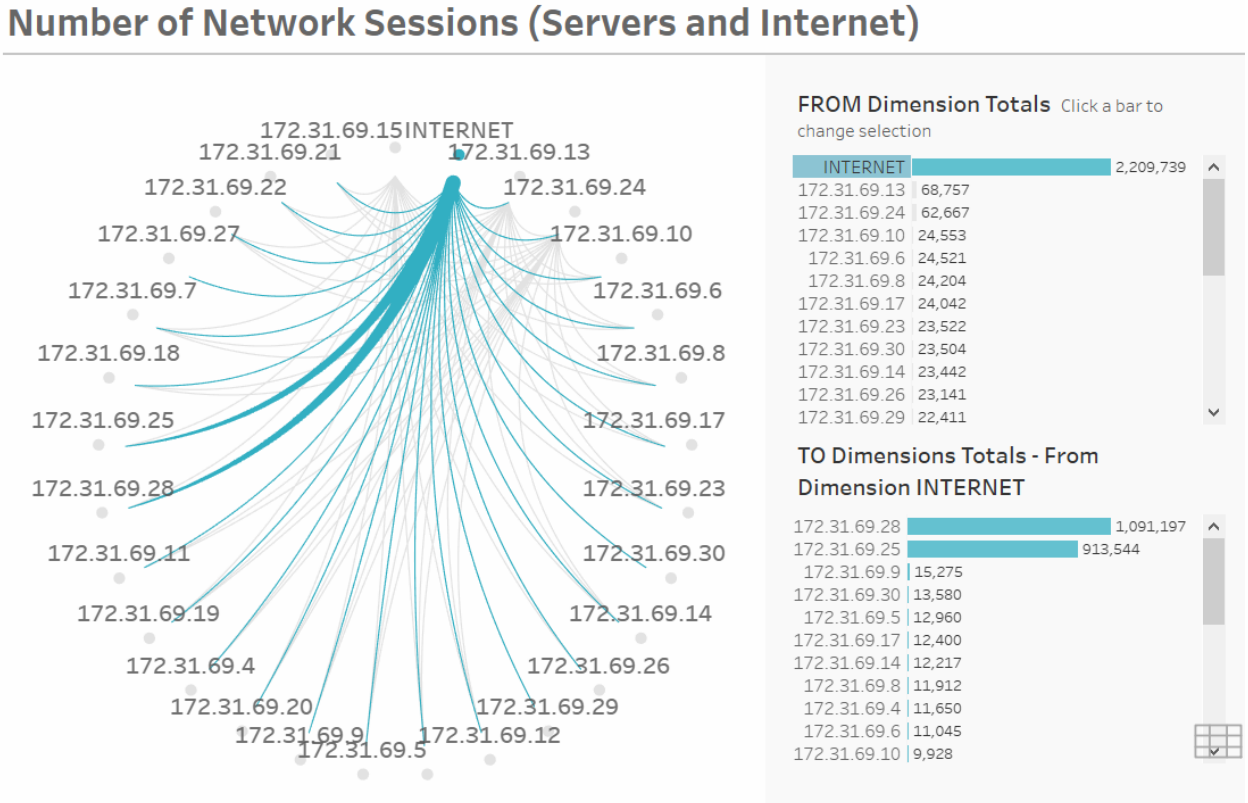

Fig. 2. Chord diagram within the server subnet and between the internet group.

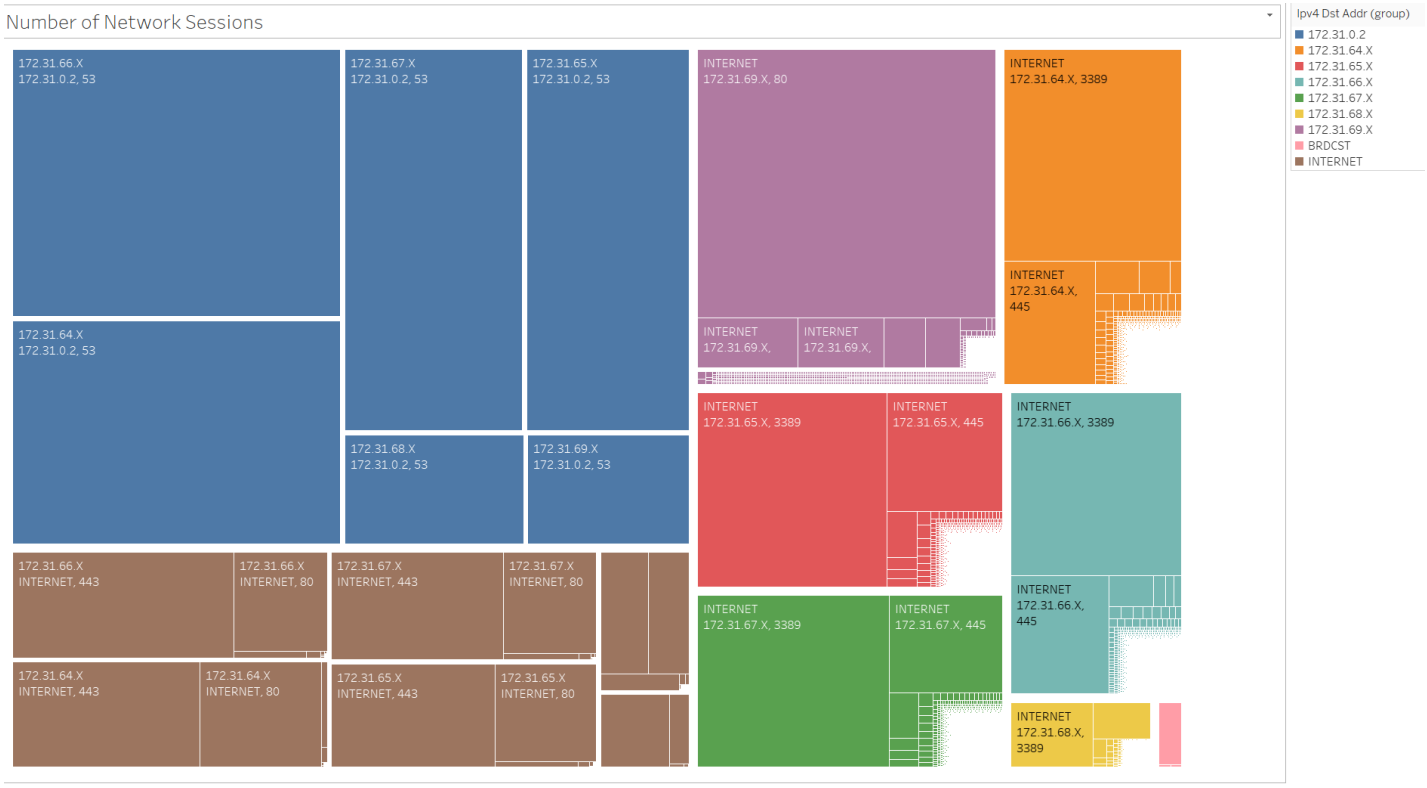

Fig. 3. Treemap of all network sessions by source, destination and destination port.

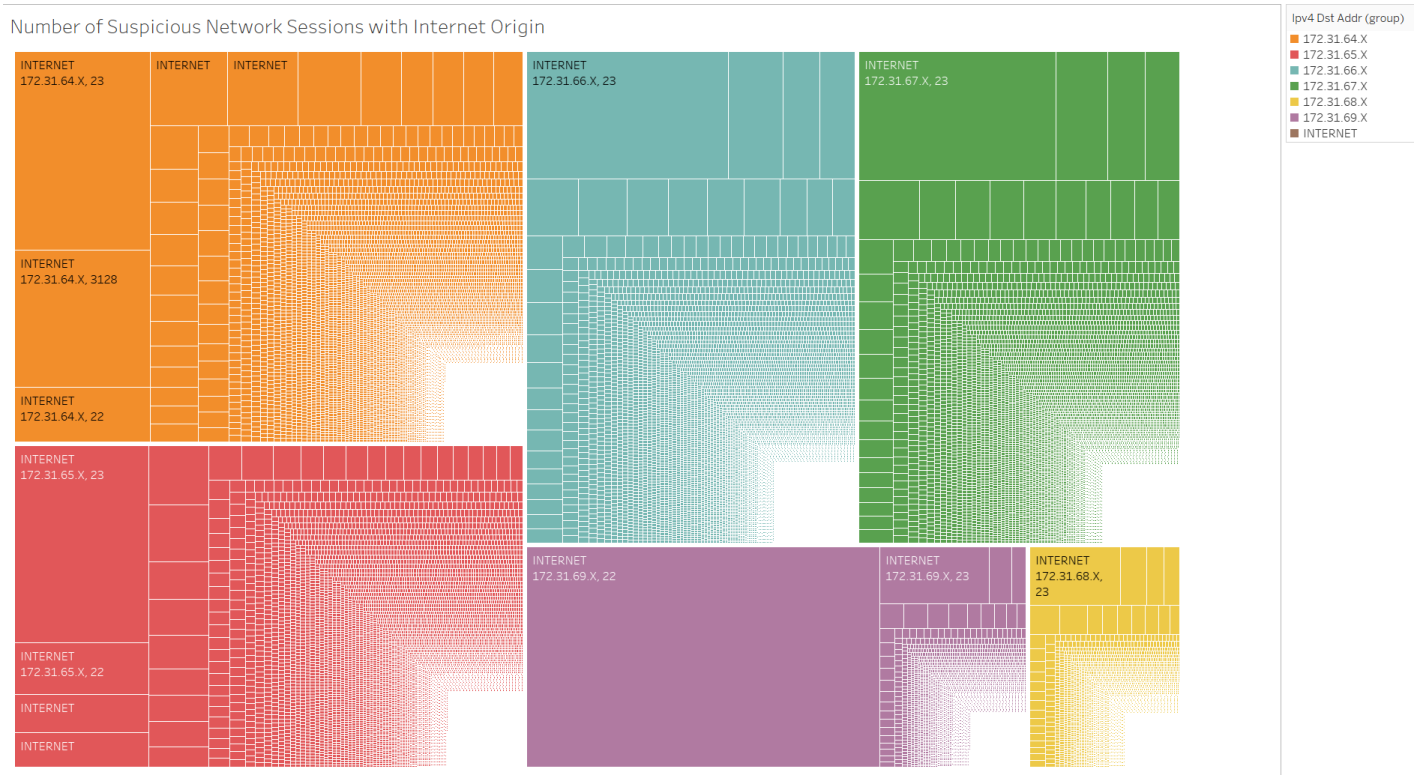

Fig. 4. Treemap of all internet initiated network sessions by destination and destination port.