2014-06

Temporal comparisons of Internet topology

Wright, Jamar E.

Monterey, California: Naval Postgraduate School

NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

TEMPORAL COMPARISONS OF INTERNET TOPOLOGY

by

Jamar E. Wright

June 2014

Thesis Advisor:	Ralucca Gera
Second Reader:	Robert Beverly

Approved for public release; distribution is unlimited

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704–0188

1. REPORT DATE *(DD–MM–YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From — To)*
13–6–2014	Master's Thesis	2012-07-01—2014-06-20

4. TITLE AND SUBTITLE

TEMPORAL COMPARISONS OF INTERNET TOPOLOGY

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Jamar E. Wright

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Naval Postgraduate School
Monterey, CA 93943

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

United States Army

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited

13. SUPPLEMENTARY NOTES

The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.IRB Protocol Number: XXXX

14. ABSTRACT

Network science and its many applications provide insight into several genres, including biological, neural, logistical and technical problems. The study of complex networks extends to the Internet as well, merging graph theoretical concepts with those of computer science in an effort to perform Internet topology measurements, ultimately contributing to inferred Internet mapping. In this research, we examine whether the time of day is a factor when measuring Internet topology. In doing so, we employ graph measures, statistical measures, and complex network measures to compare graphs inferred from probes of the Internet via network monitors. Using comparisons of these measures, we did not find indication that time was a factor for the seven probing cycles examined in this study.

15. SUBJECT TERMS

Internet Topology, Graph Similarity, Symmetric difference, CAIDA, Temporal Comparison, Complex Network

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			19b. TELEPHONE NUMBER *(include area code)*
Unclassified	Unclassified	Unclassified	UU	75	

THIS PAGE INTENTIONALLY LEFT BLANK

Approved for public release; distribution is unlimited

TEMPORAL COMPARISONS OF INTERNET TOPOLOGY

Jamar E. Wright, Major, United States Army
B.S., United States Military Academy, 2004

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN APPLIED MATHEMATICS

from the

NAVAL POSTGRADUATE SCHOOL
June 2014

Author: Jamar E. Wright

Approved by: Ralucca Gera
Thesis Advisor

Robert Beverly
Second Reader

Carlos Borges
Chair, Department of Applied Mathematics

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

Network science and its many applications provide insight into several genres, including biological, neural, logistical and technical problems. The study of complex networks extends to the Internet as well, merging graph theoretical concepts with those of computer science in an effort to perform Internet topology measurements, ultimately contributing to inferred Internet mapping. In this research, we examine whether the time of day is a factor when measuring Internet topology. In doing so, we employ graph measures, statistical measures, and complex network measures to compare graphs inferred from probes of the Internet via network monitors. Using comparisons of these measures, we did not find indication that time was a factor for the seven probing cycles examined in this study.

THIS PAGE INTENTIONALLY LEFT BLANK

Table of Contents

List of Figures

THIS PAGE INTENTIONALLY LEFT BLANK

List of Tables

THIS PAGE INTENTIONALLY LEFT BLANK

List of Acronyms and Abbreviations

Ark	Archipelago
ARPANet	Advanced Research Projects Agency Network
AS	Autonomous System
ASN	Autonomous System Number
CAIDA	Cooperative Association of Internet Data Analysis
CDN	Content Delivery Network
CI	Confidence Interval
DoS	denial of service
GMT	Greenwich Mean Time
IANA	Internet Assigned Numbers Authority
IP	Internet Protocol
IPv4	Internet Protocol version 4
IPv6	Internet Protocol version 6
IQR	Interquartile Range
ISC	Interface Set Cover
ISP	Internet Service Provider
IXP	Internet Exchange Point
NTC	Network Topology Capture
OSI	Open Systems Interconnection
PII	Personal Identifiable Information
RIR	Regional Internet Registry
RTT	Round-trip time
TTL	Time To Live
WWW	World Wide Web

THIS PAGE INTENTIONALLY LEFT BLANK

Executive Summary

Network science and its many applications provide insight into several genres, including biological, neural, logistical and technical problems. The study of complex networks extends to the Internet as well, merging graph theoretical concepts with those of computer science in an effort to perform Internet topology measurements, ultimately contributing to inferred Internet mapping. In this research, we examine whether the time of day is a factor when measuring Internet topology. Our study employed graph measures, statistical measures, and complex network measures to compare graphs inferred from probes of the Internet via network monitors. The graph measures of vertex and edge count played a significant role in determining our outcome; however, the use of graph measures alone is not sufficient. While the statistical measures allowed for quantitative comparisons of each hourly partition, the small sample size of seven probing cycles was not enough to employ more robust statistical analysis. Complex network measures revealed small differences between the inferred graphs. Using comparisons of these measures, we did not find indication that time was a factor for the seven probing cycles examined in this study.

THIS PAGE INTENTIONALLY LEFT BLANK

Acknowledgments

I would like to thank my advisor, Dr. Ralucca Gera for her patience, insight, and genuine passion for education. I would like to recognize Dr. Robert Beverly, whose enthusiasm and efforts were instrumental to this research. The Naval Postgraduate School Math department also deserves acknowledgment for incredible support given to my family.

Most importantly, I would like to recognize my family, who helped me through the most difficult time in my life and facilitated the completion of this work. Thank you Marisol, for your unwavering support from inception to culmination. Last and certainly not least, I would like to dedicate this work to my father, through whom I learned and saw an example of how to live by doing the best I can, the best I know how. I will never forget his invaluable support and counsel. I love you.

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 1:
Introduction

Over the last half century, the Internet evolved from an additional form of communication for government and educational institutions at its inception to an alternative forum for knowledge and information sharing as well as a major domain of national security, serving as an attack vector for nation states and international hackers against economic and energy infrastructures. As the Internet infiltrates further into human culture, it can be the foundation of political and technological revolutions. With the proliferation of mobile computing across the globe, the Internet serves as a medium of interaction without geographical bounds. The logical absence of these physical limitations on the Internet increases the difficulty of "mapping" the Internet, developing a topology for this continuously growing complex network.

Mathematically, the topology of the complex network known as the Internet relates to graph theory concepts introduced by Leonhard Euler in 1735. Using these concepts, we study topology by modeling the Internet's logical connections as a graph, $G(V,E)$, where $V(G)$ is the vertex set and $E(G)$ is the edge set. The vertices contained in the vertex set $V(G)$ represent the interfaces on routers that logically connect the Internet, while the edges contained in the edge set $E(G)$ represent the logical interconnections of those interfaces.

1.1 Why Measure the Internet?

In its initial stages, the design of the Internet involved the use of the existing telephone infrastructure to route packets of data through a decentralized network [1]. The growth of the Internet from these intended purposes as introduced by the Advanced Research Projects Agency Network (ARPANet) to its current expansion did not occur without challenges and innovation. Internet Service Providers (ISPs) deliver the Internet and World Wide Web (WWW) to residential homes and businesses across the globe. Larger Internet Exchange Points (IXPs) like AT&T and Verizon act as the interconnect between ISPs, Autonomous Systems (ASes), and content providers like Netflix and Amazon. Efficiency is also a key goal in the design of the Internet, one that led to Content Delivery Networks (CDNs) which improve the performance and reliability of Internet interactions [2].

Given the relationship between customer satisfaction and efficiency, Internet measurements can influence economic decisions including those to establish relationships between ASes and IXP

that will maximize customer satisfaction and growth. From an information technology perspective, topology measurements can facilitate planning decisions on the location of resources for network optimization. With information security in mind, topology measurements could reflect the potential impact of vulnerability exploits and responses to mitigate their impact.

1.2 Why Is the Internet Hard To Measure?

Some of the Internet's properties that improve its efficiency and proliferation also present challenges when attempting measurements, including its scale, vastness, and its continuous changes over time [3].

Network operators with a focus on security may design networks with defense in depth to limit information available through network scans and reconnaissance. As the Internet serves as an interconnect for these networks, the implementation of security policies can limit the number of interfaces discovered during topology measurements. The Internet's large size can multiply those limitations as security policies vary. As the number of devices increases, the consistent growth of the Internet compounds these challenges. There are also economic and intellectual factors that contribute as well. Commercial enterprises maximize network security in order to secure intellectual property, customers' Personal Identifiable Information (PII), and financial information.

1.3 Research Questions

In [4], Lee sought to measure the extent of changes in interconnectivity for a large and complex network, the Internet. Our study expands on his efforts by researching the following questions:

- Given the Internet's continuous changes, is time of the day a factor when probing the Internet for measurement?
- Are there existing measurements to depict the significance of the time of day when probing the Internet for measurement?

1.4 Thesis Contribution

Primitive graph measures provide a basis for determining similarity between graphs with minimal granularity. In an effort to refine this basis, we consider statistical measures including summary statistics, confidence intervals, and boxplots to determine if Internet measurements taken at various times of the day are similar. Furthermore, by comparing vertex counts, edge counts, vertex symmetric difference, and edge symmetric difference of various graphs, we can

improve the certainty with which we say inferred graphs of Internet topology are similar. We will model and measure the Internet as inferred by the results of traceroutes contained in the CAIDA data. Our methods include analysis of graph theoretical measures as well as complex network and statistical measures that will quantify the similarity of the inferred graphical representations of the Internet.

1.5 Organization of Thesis

We organize our investigation of the research questions as follows:

- Chapter 1 introduces the motivation for this research.
- Chapter 2 provides an overview of the Internet.
- Chapter 3 introduces the theoretical background for the research.
- Chapter 4 details the data and methodology used in this research.
- Chapter 5 contains the results of graph measures, statistical measures, and complex network measures.
- Chapter 6 summarizes the results and discusses possible areas of expansion, including future work.

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 2:

The Internet

The data analyzed in this work are a result of measurements collected on the Internet. Here, we provide the reader a familiarization with Internet topology. In Figure 2.1, we show a graphical representation of routers and links between routers on the Internet circa 1998 [5].

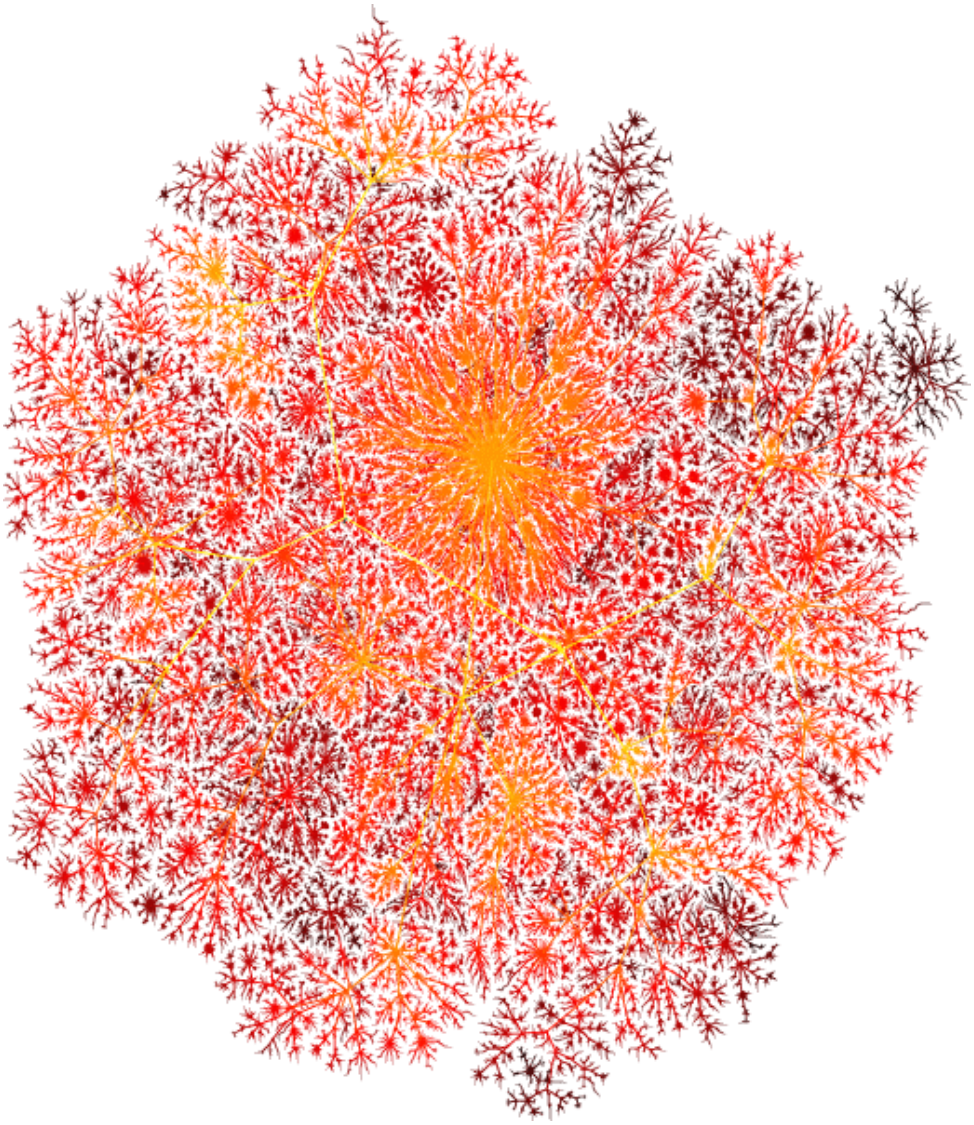


Figure 2.1: Graphical Representation of routers and links on the Internet, circa 1998, from [5].

2.1 Overview of the Internet

The study of complex networks has several applications in applied science, including biological and neural networks or transportation and utility networks. In our research, we consider yet another application, communication networks, particularly the Internet, which is a vast and constantly evolving complex network. The nodes comprising the physical and logical construction of the Internet typically follow the Open Systems Interconnection (OSI) model, which standardizes and models the function of communication networks through the use of seven layers, each with several protocols. The network layer, the third layer, plays a key role in the routing of traffic between two nodes within or through a communications network. This routing occurs through Internet Protocol (IP) addresses assigned to each device on the network. A diagram of the OSI model is in Figure 2.2.

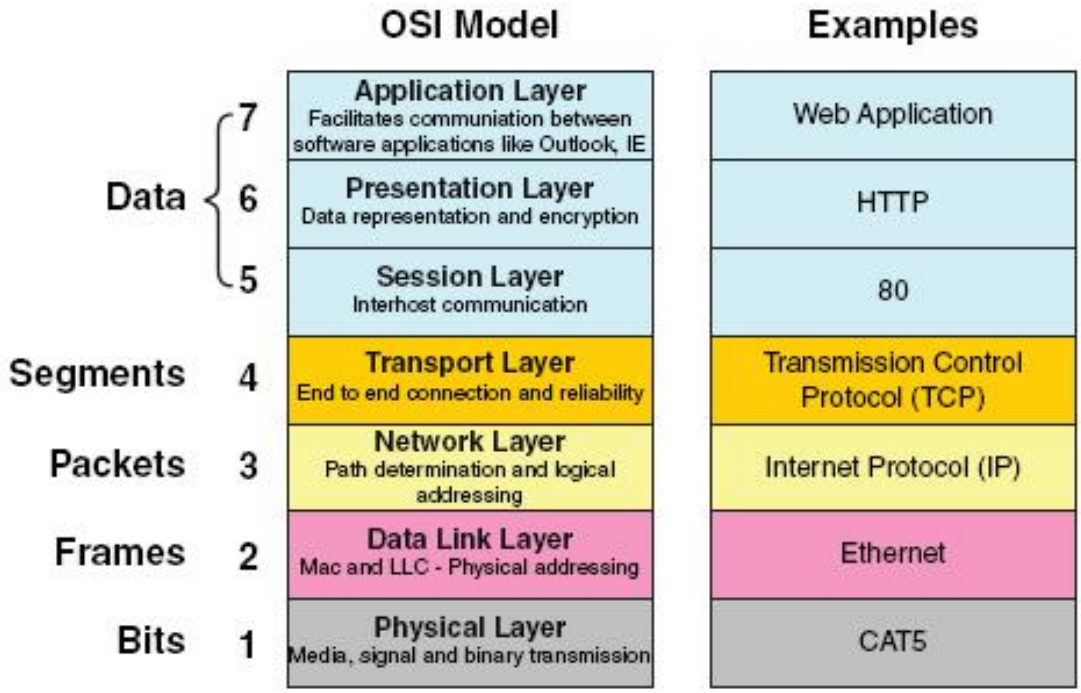


Figure 2.2: Open Systems Interconnect Model, from [6].

The complexity of the Internet is a consequence of the interconnections between several Autonomous Systems (ASes), consisting of a set of devices under a single technical and admin-

istrative control [7]. Examples of ASes include large corporations, university campuses, and Internet Service Providers (ISPs). Examples of ISPs include AT&T, Verizon, Sprint, and Century Link. Each AS receives a unique 32-bit Autonomous System Number (ASN), assigned by the Internet Assigned Numbers Authority (IANA). ASes connect to one another either through a shared ISP or through an IXP, which connects larger ASes and ISPs. Routers, which route traffic across or between networks, do so via routing tables. Using routing protocols, routers at the boundaries of an AS (e.g., R11, R12, R21 and R31 in Figure 2.3) may use protocols that facilitate the sharing of routing tables containing routes to destination IP addresses. In addition to routing traffic between networks, routers can also facilitate internal subnetting, creating multiple networks within some AS, which could allow for more efficient use of network resources by separating network traffic within and outside the AS. An example of an Internet diagram is shown in Figure 2.3 [4]. Here, the routers at the boundary, referred to earlier, serve as the backbone of the Internet, connecting the three ASes (denoted by their ASNs).

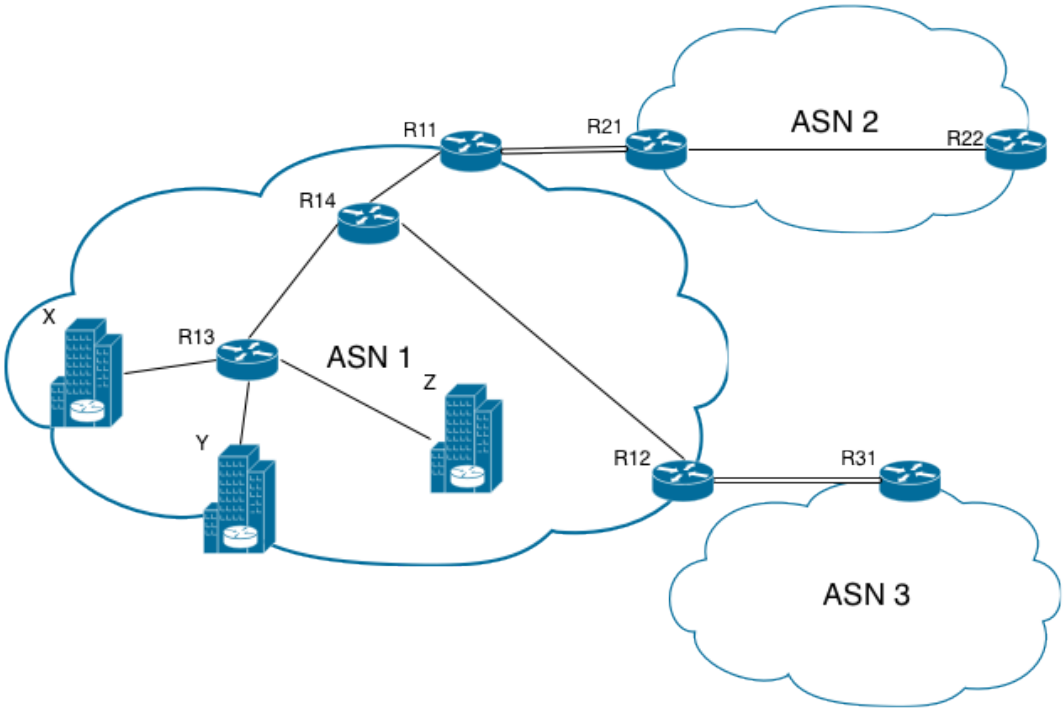


Figure 2.3: Example of an Internet Diagram, from [4].

2.2 Internet Topology

Internet topology involves efforts to map the topological structure of the Internet. The difficulty in mapping the Internet lies not only in its size and complexity, but also the continuous changes in its size and structure over time. With these challenges in mind, attempts to map Internet topology may occur at several levels of the Internet.

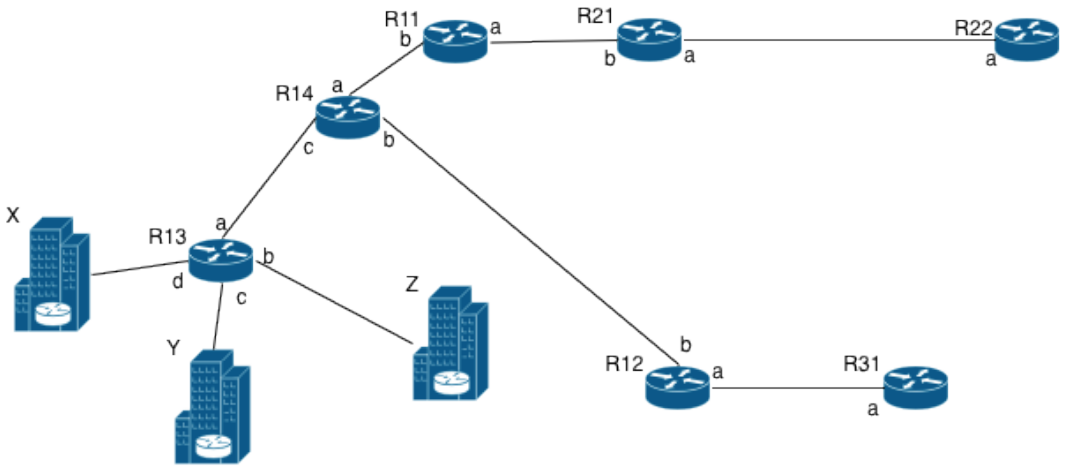
2.2.1 Internet Topology Levels

The study of Internet topology can occur at any of the seven layers of the OSI model as depicted in Figure 2.2, with each layer inferring a different graphical representation of the Internet. In this section, we provide examples of Internet topology at the IP layer. At the IP layer, we depict four granularity levels commonly used in network science: subnet-level, interface-level, router-level, subnet-level, and AS level. In our research, we focus on the interface level as the method used in data collection; the interface level can be reduced to router, subnet or AS level. This occurs at the IP layer of the OSI model, providing representations of each device through its network interface (or possible multiple interfaces).

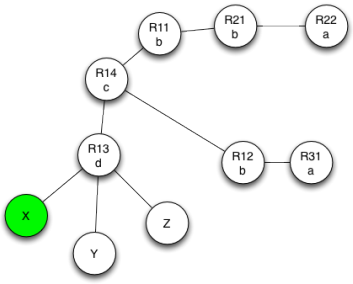
Interface-Level Topology. Internet topology at the interface level depicts connections between interfaces. Each interface is represented by a vertex or node, with the direct physical or wireless link represented by an edge. While a single router can contain several interfaces, each connection represents a separate edge. Figure 2.4 [4] illustrates the detail provided by interface-level topology as seen from various vantage points. Note that a router with multiple interfaces would be represented by multiple vertices.

Router-level topology. A router-level mapping involves the use of IP Alias Resolution [1] to represent a router and all of its interfaces as one node in a graph. Edges between vertices represent established connections between routers; however, the accuracy of IP Alias Resolution limits

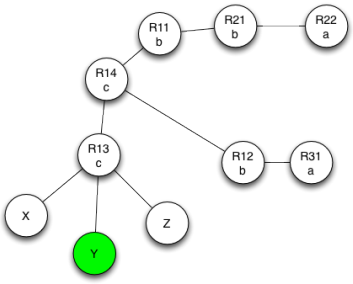
[1] This is a process which resolves IP addresses to host routers.



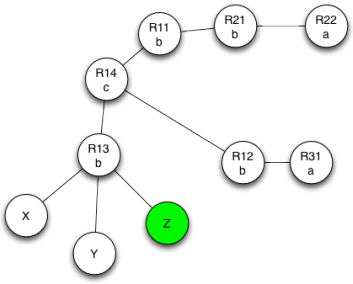
(a) Network map.



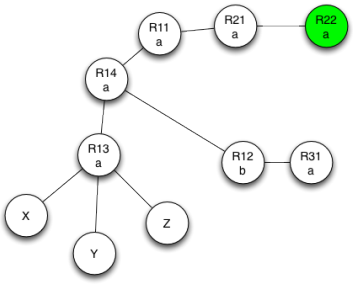
(b) Graph of interfaces as seen from X.



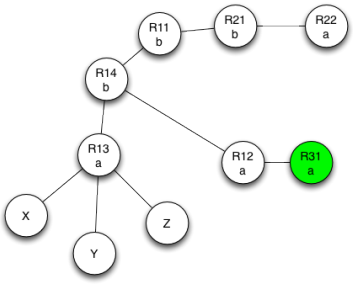
(c) Graph of interfaces as seen from Y.



(d) Graph of interfaces as seen from Z.



(e) Graph of interfaces as seen from R22.

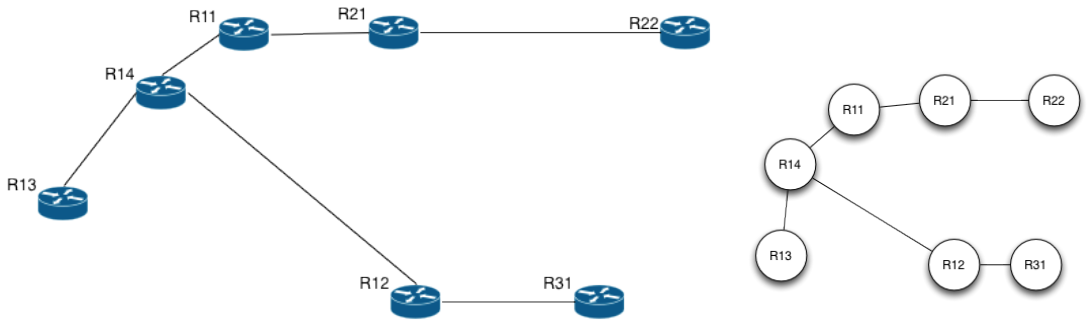


(f) Graph of interfaces as seen from R31.

Figure 2.4: Interface-level representations, from [4].

the granularity available when considering logical connections or links. Figure 2.5a illustrates a router-level representation of a network [4].

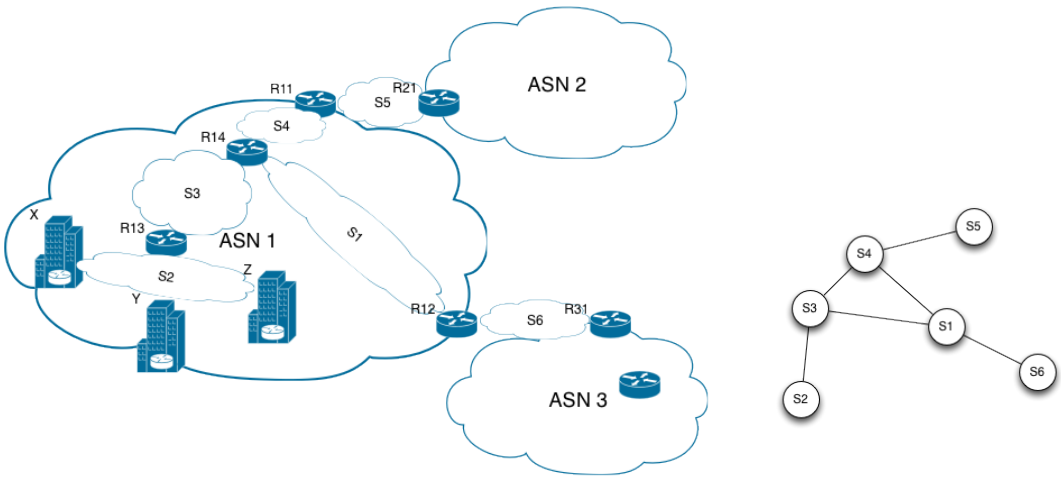
Subnet-level topology. Internet topology viewed at the subnet-level includes IP addresses



(a) Network map. (b) Graph.

Figure 2.5: Router-level representation, from [4].

hosted within the same subnet [8]. Network operators create subnets through connections established between interfaces. In this case, a graphical representation of a subnet-level mapping depicts subnets as vertices, and the links between subnets as edges. These links typically represent the logical connection between the subnets within a router configuration. An example of a subnet-level topology is Figure 2.6a, and its graph representation in Figure 2.6b [4].

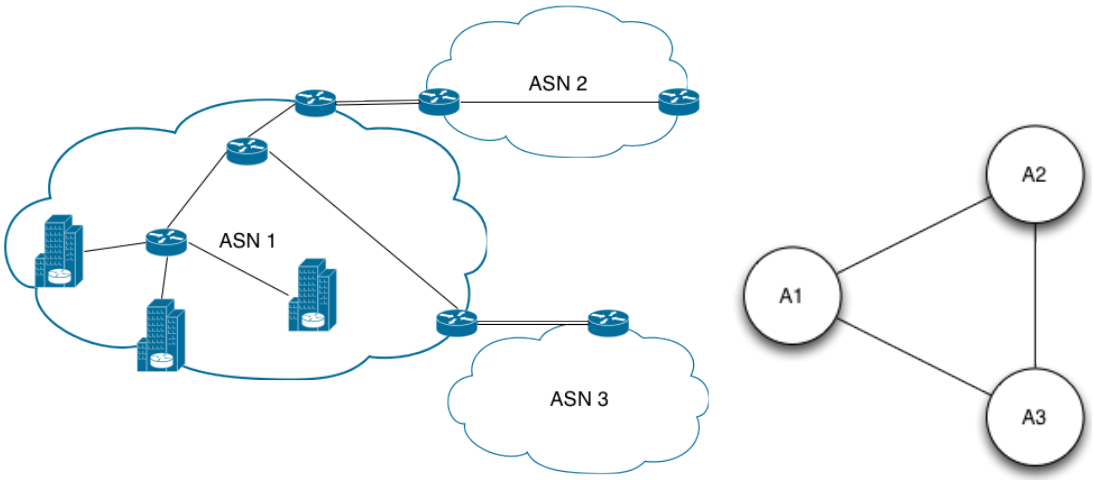


(a) Network map. (b) Graph.

Figure 2.6: Subnet-level representations of network map, from [4].

AS-level topology. An AS level representation of Internet topology portrays the physical and logical components of the network, routers and their corresponding subnets, as one node. This high level representation characterizes relationships between customers and their providers or

peering relations between ASes [9]. The relations depicted in an AS level topology include commercial and contractual agreements between ISPs and their customers, both influenced by economic factors. The economic impact reflects in the routing policies across domains such as bandwidth utilization and prioritizing data in queues. Figure 2.7a is an example of an AS level representation of a network, modeled by the graph in Figure 2.7b [4].



(a) Network map. (b) Graph.

Figure 2.7: AS-level representation of a network, from [4].

2.2.2 Obtaining Network Topology

The current state of network security vulnerabilities and the impact of exploits encourage the exploration of the use of network science as a network defense tool. While well-known exploits expose user accounts and financial information from large customer ASes, the ability to control and understand the evolution of complex networks like the Internet could minimize the impact of malicious software and organizational insiders seeking retaliation. One challenge to controlling a complex network as large as the Internet is the development of algorithms that quickly and efficiently capture a representative sample of the ground truth, or the actual network topology, via Network Topology Capture (NTC) algorithms. The implementation of defense-in-depth using firewalls, access control lists, and other hardening techniques, limit the granularity of the

results of the algorithm when compared to ground truth. Ground truth is very difficult to obtain since the availability of the topological maps of an AS to outsiders would be a vulnerability; hackers could use the information contained in such a map in exploits or use them to infer the physical layout of an organizational campus. Thus, it is indeed challenging to compare to true topology, unless a virtual network would be created for this purpose. Reference [10] contains examples of network data developed from information made public by network operators, which is the closest we have to ground truth.

Many of the current NTC algorithms are time consuming, a limiting factor to capturing a network's topology attributed to the size and scale of the Internet. In [11], active measuring techniques employing previous data and knowledge of subnets, improve the runtime for the Interface Set Cover (ISC) algorithm. The efficient use of discovery probes also minimizes the possibility that algorithm's traces will appear as a denial of service (DoS) attack, where the number of traces overwhelm the network, appearing to degrade or deny authorized access to networked resources. Bourgeau's paper also uses accelerated probing, which employs information from previous traceroutes, to capture network dynamics, maintaining network coverage in the process [12]. The data from CAIDA, the data used in our research, employs (1) active measurements, which introduces traffic on the probed network, and (2) passive measurements, which passively observe existing traffic without modification, in the collection of datasets. Our research uses only the active traceroute data.

2.3 Traceroute

Traceroute [13] is a diagnostic tool for computer networks that shows the time delays and forward router interface path of an IP packet. The tool uses the packet's Time To Live (TTL) to create a route history that details the list of nodes traversed during delivery and return. TTL limits the life of data in transit; when the TTL reaches zero, the router sends a response back to

the sender, allowing the reconstruction of the route history. For our research, we only use the list of IPs from the traceroute, discarding the TTL.

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 3:

Measures

In this chapter, we review mathematical concepts that facilitate a view of complex networks through a mathematical lens. Specifically, we will measure the Internet by applying different measurements to the Internet, an example of a complex network. These measures in conjunction with descriptive statistics will allow us to draw conclusion about the topology of the Internet. The purpose of employing these concepts is their relation to the underlying structure of the Internet and the ability to generate statistics that serve as indicators to the behaviors we study. In this chapter, we describe existing network measures that can be used on the graph representing the Internet.

We will measure the Internet by translating the interconnections between nodes into graphs. Some translations require the incorporation of set theory on the vertex V(G) or edge E(G) sets of the graph. The main measures used in this thesis indicate a percentage of change between two graphs, G_1 and G_2, each representing a snapshot of the Internet at a given time. Because of the large scale of the Internet, it can be difficult to infer changes in the vertex or edge set based solely on these two measures. Therefore, we incorporate statistical measures to discover any discernible changes to the graph, thereby indicating a change over time.

3.1 Set Theory

The definitions and concepts described in this section are from [14].

The **difference** between two sets A and B, $A \setminus B$, is the set of elements of A not in B, denoted by the following:

$$A \setminus B = \{x \in A | x \notin B\}$$

The **union** of sets A and B, $A \cup B$, is the set that contains elements either in A or B, or both. The following denotes $A \cup B$:

$$A \cup B = \{x | x \in A \vee x \in B\}$$

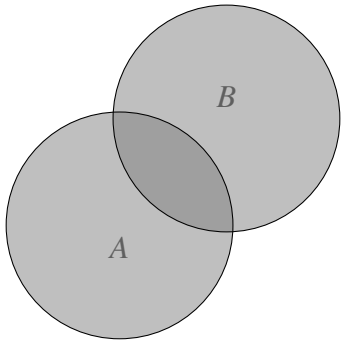


Figure 3.1: Union of A and B, or $A \cup B$, from [15].

The **intersection** of sets A and B, $A \cap B$, is the set containing the elements in both A and B. The following denotes $A \cap B$:

$$A \cap B = \{x | x \in A \wedge x \in B\}$$

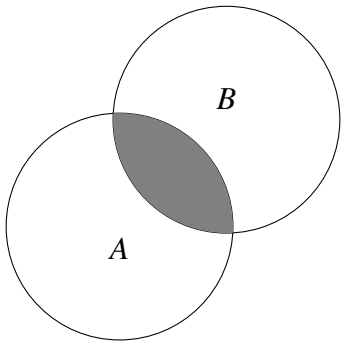


Figure 3.2: Intersection of A and B, or $A \cap B$, from [15].

The **symmetric difference** of A and B, $A \oplus B$, is the set containing elements in either A or B, but not in both A and B. Similarly, it is the set which contains the elements in exactly one of A

or B, or the union of A and B without the intersection.The following denotes $A \oplus B$:

$$A \oplus B = \{x | (x \in A \wedge x \notin B) \vee (x \notin A \wedge x \in B)\}$$

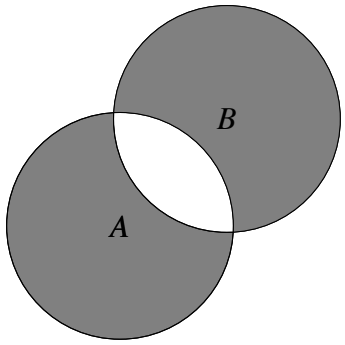


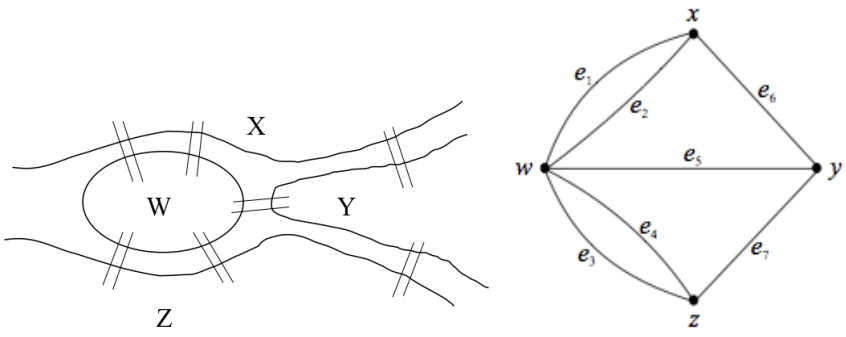
Figure 3.3: Symmetric Difference of A and B, or $A \oplus B$, from [15].

The **cardinality** of a set A is the number of elements in the set, denoted by $|A|$. All of these definitions generalize to more than two sets.

3.2 Graph Theory

One can trace the origins of graph theory to a problem posed by Leonhard Euler in 1735, The Seven Bridges of Königsberg [16]. In this problem, citizens sought a route that crosses each bridge in Königsberg exactly once and returns to the starting point. Figure 3.4 illustrates Euler's problem [4]. Earlier studies in graph theory focused on small, simple graphs that were static, allowing researchers to have complete information in the form of exact values for the characteristics of the graphs under study. We highlight some of these characteristics below from [17].

A **graph** G **= (V,E)** consists of a set of vertices V(G) and a set of edges E(G). The vertex set



(a) Königsberg Bridges. (b) Graphical Representation.

Figure 3.4: Seven Bridges of Königsberg, from [18].

V(G) is the set of vertices of graph G. The edge set E(G) is a set of 2-element subsets of V, such as the edge $\{v_1, v_2\} = e \in E(G)$. The 2-element subsets indicate the **endpoints** of the edge. **Multiple edges** are edges that share the same two endpoints. A **loop** is an edge with matching endpoints, or an edge between a node and itself. A **simple graph** is a graph that does not include loops or multiple edges.

In our research, we only consider simple graphs, as logically, a loop would indicate the connection of an interface to itself, which may be useful for troubleshooting, but not for the purposes of this research. We also decided not to include multiple edges in this research.

One can characterize a simple graph by its vertex set and edge set, with the edge set consisting of a set of unordered pairs of vertices such as the edge $e = uv = vu$, where u and v are endpoints. The uv notation means "u is adjacent to v." Our analysis involves data collected from bidirectional probes; as a result, our graphs contain undirected edges. Below we list the **main classes** of graphs in Graph Theory.

3.2.1 Graph Classes

Complete Graph A graph G is complete if every two distinct vertices of G are adjacent. We denote an unlabeled complete graph with n vertices as K_n.

Bipartite Graph A graph G is a bipartite graph if $V(G)$ can be partitioned into two subsets A and B such that every edge of G joins a vertex of A and a vertex of B.

Complete Bipartite Graph A complete bipartite graph is a bipartite graph with partite sets A and B such that every vertex of A is adjacent to every vertex of B. This is denoted by $K_{a,b}$, where a and b are the sizes of sets A and B respectively.

Path A path P_n is a simple graph whose vertices can be ordered so that two vertices are adjacent if and only if they are consecutive on the list. For example, the edge set of P_5 is $E(P_5) = \{v_1v_2, v_2v_3, v_3v_4, v_4v_5\}$.

Cycle A cycle is a graph that consists of a sequence of different vertices (except the starting and ending vertex which must be the same), with each two consecutive vertices in the same sequence adjacent to each other in the graph.

Erdős Rényi (ER) Random Graph Paul Erdős and Alfréd Rényi introduced a model for generating random graphs. In their model, a graph $G(n,p)$ is a simple graph with n possible vertices. Each edge between those vertices occurs with equal probability p. Their model and its probabilistic properties are used as the default type of graph to determine if a property holds for arbitrary graphs. In our research, there seems to be some randomness to the appearance of nodes or edges at any given hour of the day.

There are two models of the ER random graph [19].

- $G(n,M)$ model. From a class of all graphs with n nodes and M edges, one graph is chosen uniformly at random.
- $G(n,p)$ model. Construct a graph by randomly connecting nodes via edges with an independent probability p. All graphs with n nodes and M edges have equal probability of $p^M(1-p)^{\binom{n}{2}-M}$.

Symmetric Difference Graph If G and H are graphs with vertex set V, then the symmetric difference $G \triangle H$ is the graph with vertex set V that contains the set of vertices in either G or H and not in the intersection $G \cap H$. This is not the symmetric difference used in this paper.

3.2.2 Graph Measures

The parameters below illustrate measurements typically employed in graph theory to study a graph and its properties. There are two types of measures on graphs - Type I and Type II. Type I measures are about the graph, including diameter, vertex count, and average degree. Type II measures are about the vertices of the graph, including degree, neighborhood degree, or clustering coefficient per vertex. Since our graphs are so large, and our goal is to study the big picture, we use almost exclusively, Type I measure. We are particularly interested in comparing graphs, so we use graph properties (rather than vertex properties) to do so. While these parameters describe the graph, they are not sufficient to compare two graphs. Our research involves data comparisons through statistical measures and graphical representations of the Internet to determine similarities and differences at various periods of time. The terminology below is from [20], and the examples consider the graphical representation of the Seven Bridges of Königsberg from Figure 3.4b, reproduced below.

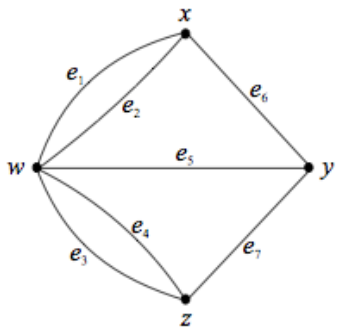


Figure 3.5: Graphical Representation of the Seven Bridges of Königsberg, from [18].

Degree. The degree of a vertex v, denoted $deg(v)$, is the number of edges incident to v. Similarly, it is the number of vertices adjacent to v. Given the Konigsberg example in Figure 3.5,

the degree of vertex w is 5, or $deg(w) = 5$.

Average Degree. The average degree of a graph is the number of edges in the graph per vertex, or:

$$avedegree = \frac{\sum_{i=1}^{n} deg}{n} = \frac{2m}{n},$$

where m is the number of edges and n is the number of vertices. In Figure 3.5, the average degree is $\frac{14}{4} = \frac{7}{2}$.

Distance. The distance from u to v, $d(u,v)$, is the least number of edges in a uv path in G. lf G has no such path, then $d(u,v) = \infty$. In Figure 3.5, $d(x,z) = 2$.

Diameter. The diameter of a graph G, *diam G*, is the maximum distance between any two vertices in graph G. Equivalently, it is the longest, shortest path between two vertices, i.e., $diam\ G = max_{u,v \in V(G)} d(u,v)$. In Figure 3.5, $diamG = 2$.

Eccentricity. The eccentricity of a vertex u, $\varepsilon(u)$, is the maximum distance from u to all vertices in the graph, denoted as $\varepsilon(u) = max_{v \in V(G)} d(u,v)$. It is also the largest distance between u and any other vertex in the graph. The eccentricity of the vertex x in the Konigsberg graph is $\varepsilon(x) = 2$, while $\varepsilon(y) = 1$.

Radius. The radius of a graph G, *rad(G)*, is the smallest eccentricity among all eccentricities in the graph, i.e., $rad(G) = min_{u \in V(G)} \varepsilon(u)$. In the Konigsberg example in Figure 3.4, $rad(G) = 1$.

3.3 Complex Network Measures

Interests in representing various networks as graphs expanded tremendously over the past two decades. Some of the current networks explored through research include transportation, utilities, biological, and neural networks. The increased proliferation of the Internet spawned additional areas of interest including online social networks (Facebook, LinkedIn), email networks,

and the World Wide Web graph. One characteristic that is common to all of these types is these networks evolve over time, unlike the simple networks studied in the earlier days of graph theory. In contrast to simpler graphs, complex networks have a larger number of components which may not have well-defined roles, present self-emergent properties, and exhibit organizational behaviors not necessarily influenced by well established principles. An example of a trait common to many complex networks is the *small world* phenomenon, pioneered by Watts and Strogatz [21]. They found that some self-organizing networks like the Internet tend to be highly clustered with small path lengths. Researchers hypothesize that for the Internet, the lack of central control suggests it may follow some random structure. This would also suggest a Gaussian degree distribution, which we believe is not the case based on experiments [19]. Data about the Internet shows that the degree distribution follows a power law asymptotically x^k, where the bounds on k are typically, but not limited to $2 < k < 3$. Barabasi suggested that the preferential attachment in complex networks can be achieved through *preferential attachment growth* [22]. Websites on the WWW link to each other in a method consistent with preferential attachment. We present some complex network measures that we employ in this research.

3.3.1 Vertex Symmetric Difference (VSD)

In this section we describe a way to compare the vertex and edge sets of two graphs. Lee, in [4], introduced the concept and the paper contains additional information beyond the scope covered herein.

Considering only the vertices, we compare two graphs G and H having two vertex sets, V(G) and V(H), respectively.

Definition 3.3.1. *[4] Given graphs G and H, the* ***vertex symmetric difference, vsd(G, H),*** *is*

$$vsd(G,H) = \frac{|V(G) - V(H)| + |V(H) - V(G)|}{|V(G)| + |V(H)|}.$$

Generally, the vertex symmetric difference counts the vertices that are in one graph and not the other and then it normalizes the count for interpretation as a percent. Given graphs G and H, if $V(G) = V(H)$, then $vsd(G,H) = 0$. If $G \cap H = \emptyset$, then $vsd(G,H) = 1$. The upper and lower bounds allow us to determine if the difference between the vertex sets of two graphs is significant.

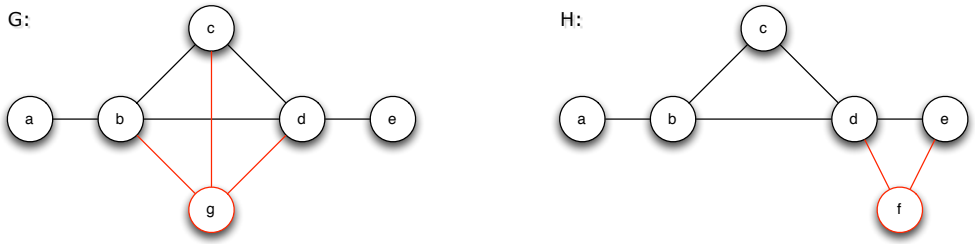


Figure 3.6: Example to illustrate *vsd* and *esd* between two graphs, from [4].

Comparing the two graphs in Figure 3.6 [4], $g \in V(G)$, but $g \notin V(H)$; likewise, $f \notin V(H)$, but $f \in V(G)$. Therefore, for graphs G and H,

$$vsd(G,H) = \frac{|V(G) - V(H)| + |V(H) - V(G)|}{|V(G)| + |V(H)|} = \frac{1+1}{6+6} = \frac{2}{12} = 16.7\%$$

3.3.2 Edge Symmetric Difference

The same concept can apply to the edge set of a graph, referred to as the **edge symmetric difference** [4]. In our study of Internet topology, it applies to links established between vertices during a snapshot of a traceroute in a given hour. As each graph contains the IP address as a vertex label, those labels help to compose the distinct edges in the edge set, where an edge is an established connection between two interfaces.

Definition 3.3.2. *[4] Given two graphs, G and H, the* edge symmetric difference $esd(G,H)$ *is*

defined as

$$esd(G,H) = \frac{|E(G) - E(H)| + |E(H) - E(G)|}{|E(G)| + |E(H)|}.$$

Generally, the edge symmetric difference counts the vertices that are in one graph and not the other and then it normalizes the count for interpretation as a percent. Given graphs G and H, if $E(G) = E(H)$, then $esd(G,H) = 0$. If $G \cap H = \emptyset$, then $esd(G,H) = 1$. The upper and lower bounds allow us to determine if the difference between the edge sets of two graphs is significant.

Recalling the two graphs in Figure 3.6,

$$esd(G,H) = \frac{|E(G) - E(H)| + |E(H) - E(G)|}{|E(G)| + |E(H)|} = \frac{3+2}{8+7} = \frac{5}{15} = 33.3\%$$

In the context of our analysis, we consider graphs G and H with different vertex and edge sets. We use the "Vertex Symmetric Difference" and "Edge Symmetric Difference" on hourly snapshots of the Internet, comparing one hour to all other hours in a probing cycle.

3.4 Statistical Measurements

As mentioned before, the inferred network topology will be modeled by a network (or graph) in order to facilitate measuring the Internet. That is, the interface-level maps from traceroutes will be represented by graphs, with vertices denoting the interfaces and undirected edges denoting the pair-wise connection between the interfaces. Given the large number of vertices and edges collected in a CAIDA probing cycle as described in Section 4.1.1, the traditional graph theory measurements, which lend themselves to a graphical view of the network, can be complemented. Additionally, some traditional graph metrics are infeasible to compute given the size of the graphs considered in this research. Thus, we use the statistical measures below

to determine similarity in graphs. These definitions from [23] consider the data collected by CAIDA as samples taken from the Internet.

Mean. The mean, $\bar{x}$, refers to the average or center of the distribution of the data and is given by the following equation:

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n},$$

where x_i is the i^{th} element of the data set and n is the cardinality of the data set. Given a data set $V_0 = \{154769, 158565, 196052, 199607, 219048, 234695, 247291\}$, the mean of V_0 would equal: $\bar{x}_{v_0} = \frac{199607+219048+247291+196052+234695+158565+154769}{7} = 201432.40$.

Median. The median, $\tilde{x}$, is the middle value when data is ordered from smallest to largest and is found by:

$$\tilde{x} = (\tfrac{n+1}{2})^{th} \text{ ordered value if } n \text{ is odd, or}$$
$$\tilde{x} = \text{average of } (\tfrac{n}{2})^{th} \text{ and } (\tfrac{n}{2}+1)^{th} \text{ values if } n \text{ is even.}$$

After ordering the data set v_0 from smallest to largest, $\tilde{x}_{v_0} = 199607$.

Standard Deviation [24]. The standard deviation, s, is the standard measure of spread or the average distance of the data from the mean, found by:

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x - \bar{x})^2}.$$

For V_0, $s_{v_0} = 35513.12$.

Boxplot. A boxplot is a picture summary used to describe a data set's prominent features which include its median, spread, symmetry, and outliers. The **outliers** are observations that

are unusually far from the main body of data. An example of a boxplot for the v_0 dataset is Figure 3.7.

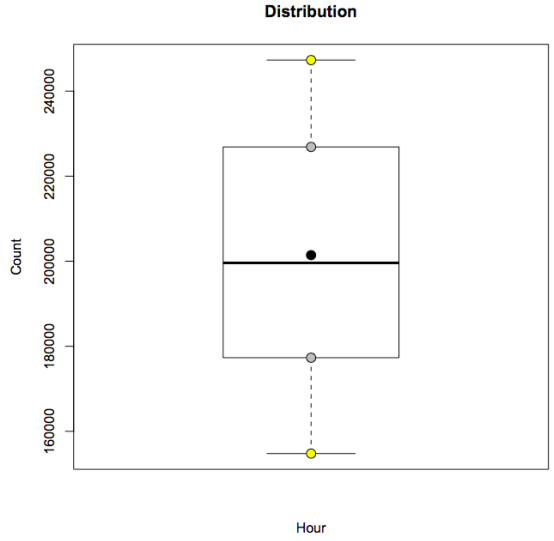


Figure 3.7: Boxplot of V_0 data.

In Figure 3.7, the bold horizontal line is the median, $\tilde{x}_{v_0}$, and the black circle is the mean, $\bar{x}_{v_0}$. The horizontal limits of the box, denoted by gray circles, indicate the Interquartile Range (IQR), everything between them representing the middle half of the data. The lower quartile of the IQR is the median of the lower half of data, while the upper quartile of the IQR is the median of the upper half of the data. The dashed lines represent the whiskers of the boxplot. The range of the whiskers is 1.5 times the IQR, measured from the median. The solid horizontal lines at the ends of the whiskers indicate the end of their range. Any values outside of these limits are referred to as outliers. In this example, there are no outliers; if so, they would appear outside the whiskers of the boxplot.

Confidence Interval. The confidence interval (CI) is an estimate for the interval containing a parameter. The confidence level, α, indicates how frequently the interval contains the parameter. Given a small sample size n, where $n < 20$, and a sample standard deviation s, the small sample confidence interval for the mean $\bar{x}_{v_0}$ is denoted by

$$\bar{x}_{v_0} \pm t_{\frac{\alpha}{2},n-1} \frac{s}{\sqrt{n}},$$

where $t_{\frac{\alpha}{2},n-1}$ is the interval width. Because our research examines the data from 7 cycles, we use a t-distribution with $n-1$ degrees of freedom. In our research, the confidence level is $\alpha = 0.05$, or $(1-\alpha) = 95\%$. Given $\bar{x}_{v_0}$, the Confidence Interval (CI) for the mean of the data set V_0 would be:

$$201432.40 \pm 2.447 \frac{35513.12}{\sqrt{7}} = 201432.40 \pm 32845.34.$$

We can also express the CI as $[168587.06, 234277.74]$.

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 4:

Data and Methodology

In this chapter, we discuss how the aforementioned mathematical concepts apply to our study of the Internet. In doing so, we consider the collection and subsequent analysis of the data.

4.1 Source of Data

The source of data for our research and analysis is CAIDA, which employs active and passive measurements to capture the topology of the Internet at a given time through the use of probes from over 90 vantage points across the world, spanning all continents with the exception of Antarctica [25]. These Archipelago (Ark) monitors, which are small form-factor computers, are the source of active measurements contained in CAIDA data. The locations of the monitors are in Figure 4.1.

4.1.1 CAIDA data

In order to collect data from the Internet, CAIDA employs scamper, an active measurement tool that probes the Internet for topology analysis and performance [26]. Scamper uses network diagnostic tools, such as traceroute and ping, to probe networks supporting Internet Protocol version 4 (IPv4) and Internet Protocol version 6 (IPv6). The scamper tool is part of the Ark infrastructure of active measurement monitors located at each vantage point across the world, serving as a collection station for probes sent using scamper. The intent of the Ark infrastructure is to increase the efficiency of large-scale measurements and facilitate collaboration with others who perform measurement tasks [25]. The data used in this research is the IPv4 Routed /24 Topology Dataset [2] [27]. In an effort to improve efficiency and speed of the collection process,

[2] An IPv4 address is a 32-bit integer value arranged in four octets or bytes. /24, which refers to the first three octets, is the prefix of the IPv4 network starting at a given address. The remaining 8 bits are for device addressing.

Figure 4.1: Locations of CAIDA Monitors, from [25].

Ark groups the monitors into three teams (each team gets a complete probing cycle independent of the other teams), facilitating traceroute measurements of the probed /24 networks. The probing and measurement period typically lasts 2-3 days for each team, which we refer to as a probing cycle. One probing cycle represents probing an address in each /24 network in the entire Internet.

The data is a result of probes sent from randomly selected vantage points to destination addresses in an IPv4 /24 prefix. Given the number of vantage points and possible destinations in the /24 prefix, the data collected, which includes start and stop markers as well as metadata

for one CAIDA cycle, can exceed three GB and is in a *warts* file format. File parsing tools such as sc_analysis_dump convert the data into a textual format readable by additional scripts that output the results of each trace within the probing cycle [4]. This output contains traceroute information including the interfaces traversed and the delay of its response, as well as the metadata mentioned earlier.

4.2 Data Selection and Preparation

In this section, we detail the data content collected by CAIDA. We parse the initial CAIDA data for each probing cycle into 24 1-hour partitions. That is, we partition the contents of a probing cycle into hours of the day in Greenwich Mean Time (GMT). The amount of data contained in one cycle for one team averages over 900,000 vertices and over 2 million edges, where the vertices represent interfaces and edges represent established connections between those interfaces. The data used in our research is from probing cycles that occurred in February and December 2013.

4.2.1 Preparation

The output from the sc_analysis_dump tool described by Lee in [4] contains a list of partial data from a traceroute, listing transit delay measurements and a record of the packet's route history and a Round-trip time (RTT) from each router encountered along the path. As our focus is only on the interfaces traversed, not the time elapsed during the traversal, we remove the RTT measurements, resulting in a sequence of interfaces in a fixed order. This sequence contains the path from a source IP address, with interfaces encountered along the path to a destination IP address. We remove the last hop or destination as our interests lie in routers and links; removing the last hop also minimizes variance resulting from probes to devices that may not sustain a continuous connection to the network.

When converting this data into a graph, we represent each interface as a vertex. The order of

the IP sequence represents connections established between interfaces; thus, we represent the sequence order as edges. Identifying the common IPs as a single IP within the same probing cycle results in a graph, which is a representation of the interfaces probed during that cycle as well as the connections between them. The graph also gives a network map of the Internet as depicted by the probes in that team's cycle. The *vsd* and *esd* are metrics for the comparison of two of these graphs.

4.2.2 Challenges

While the traceroute tool mentioned in 2.3 serves as a useful utility for network operators, it also serves as an attack vector for hackers who seek to employ DoS attacks on an AS or an interface within an AS. As a network hardening technique, some network operators configure interfaces on their routers to respond to traceroutes in various ways. In the event a trace reports back all intermediate interfaces between the source and destination that respond, we refer to these traces as complete. Some interfaces, which we refer to as non-responding, may not respond to requests but forward the packets; while others may drop the packet completely without sending a reply, which we refer to as probe-dropping [4]. The value $'q'$ denotes "anonymous" interfaces, or no response at that particular hop in the trace. While we obtain incomplete traces in either case, each of the two options yields a different output. An example of these different interface behaviors is in Figure 4.2.

In Figure 4.2, there are three traceroute results for the same source and destination at three different times. The varying outputs illustrate the unreliability of traceroute probes, though one can infer the route based on the combined results of the three traces. For example, there is a non-response on the second intermediate interface in one trace, that actually responded (at a different time) in the other two traces.

[2]These outputs are extracted from fields 14 and onwards of recorded traceroutes. Refer to [4] for details.

203.181.248.60	203.181.248.60	203.181.248.60
203.181.249.21	203.181.249.21	q
203.181.102.129	203.181.102.129	203.181.102.129
118.155.197.1	118.155.197.1	118.155.197.1
203.181.100.126	203.181.100.126	203.181.100.126
59.128.2.210	59.128.2.210	59.128.2.210
65.19.143.9	65.19.143.9	65.19.143.9
72.52.92.37	72.52.92.37	72.52.92.37
72.52.92.233	72.52.92.233	72.52.92.233
216.66.77.102	216.66.77.102	216.66.77.102
209.152.158.18	209.152.158.18	209.152.158.18
q	209.152.158.18	
q		
q		
209.152.158.18		

Figure 4.2: Comparison of traceroutes with same source (203.181.248.60) and destination (209.152.158.18) addresses, from [4].

4.3 Analysis

Recall from Section 4.1.1 that CAIDA selects probing destinations randomly from each /24 prefix. To mitigate the effects of random destinations, we exclude the IP address of the destination interface from our analysis.

4.3.1 Temporal

In our analysis of network snapshots over time, we parse the data contained in one CAIDA cycle into 24 time periods, ranging from 0000 hours to 2359 hours GMT, with each period

containing one hour. Each graph contains a union of the data corresponding to that graph from each vantage point used during the cycle. The resulting periods contain all of the IPs within that given hour from any of the vantage points that probed during that hour.

For our analysis, we parse the data in the same fashion for seven cycles. We then consider the network and statistical measures presented in Chapter 3. We compare the vertex and edge counts for each hourly graph as well as the *esd* and *vsd* to contrast the data collected during each time period.

4.3.2 Spatial

To account for abnormalities in network behavior, we can consider the geographic location of the AS by obtaining the country code for the AS from the Regional Internet Registry (RIR), which manages the allocation and registration of IP addresses and ASes within a country or region. In our research, we consider the interfaces encountered during a traceroute probe that pertain to a specific AS as internal interfaces, and refer to the interfaces encountered outside of the AS during the probe as external interfaces [4]. In Chapter 5, we consider the number of IP addresses traversed by country in various sets of data. For example, in instances where the *vsd* or *esd* are inconsistent, we explore the causes for inconsistency by constructing vertex and edge set containing the IP addresses that occur in the hourly partitions with higher *vsd* or *esd*. We also use the MaxMind GeoIP database [28] to identify the geographic location of IP addresses unique to an hourly partition. The MaxMind database improves upon the data contained in a typical *whois* or reference information lookup of an IP address's organization, AS, or ASN.

CHAPTER 5:

Results

In this chapter, we analyze the results obtained by applying the measures described in Chapter 3 to the inferred graphs that resulted from traceroute probes as described in Chapter 4. We began by observing the results of various graph measures for seven probing cycles from February and December 2013. To gain a deeper understanding of the behaviors reflected in the graph measures, we considered complex network measures as well as statistical measures in an effort to compare data parsed by the hours, to analyze the correlation between the time of day and the probing data obtained.

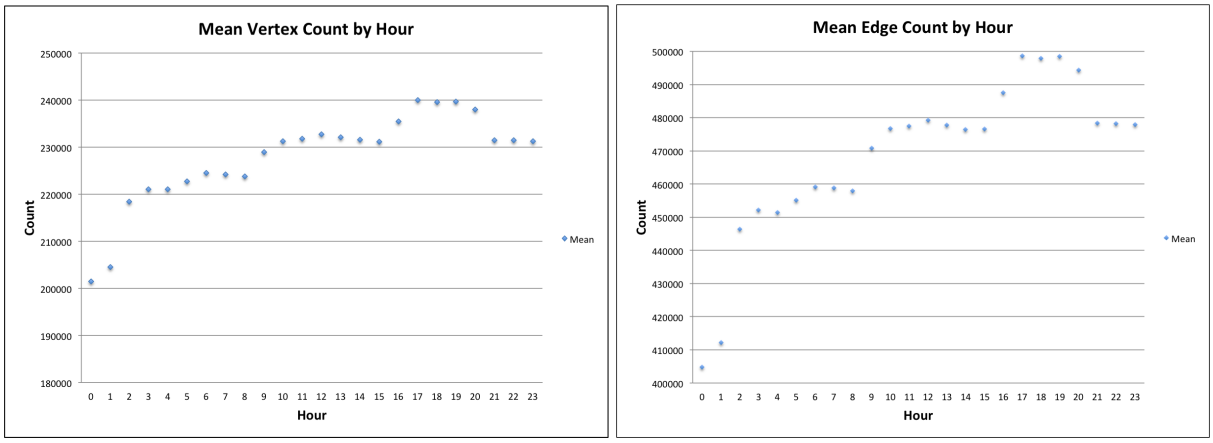
5.1 Graph Measures

To determine if the time of day is an influence factor in probing data, we considered the number of vertices and edges encountered during each probe over a 24-hour period, with each day's worth of data parsed into 24 segments representing each hour. In Table 5.1, we observed the mean for the vertex and edge counts by probing cycle varied for most of the hourly values for each probing cycle. In Table 5.1, the vertex count for hours 00 and 01 in probing cycle-20130215 were noticeably lower than the other 22 hours in the cycle. The results were similar in Table 5.2, which represents edge counts.

The same pair of hours in the other six probing cycles revealed a similar disparity. As a result, we considered the mean of the vertex counts of each hour by probing cycle. In Figure 5.1a, the low counts in hours 00 and 01 relative to subsequent hours give cause for further exploration. As expected, when we considered the edge counts, the trends continued, as indicated in Figure 5.1b.

Hour	cycle-20130215	cycle-20130217	cycle-20131201	cycle-20131203	cycle-20131205	cycle-20131207	cycle-20131209	Mean
0	199607	219048	247291	196052	234695	158565	154769	201432
1	207639	236611	247034	194990	235292	155963	154103	204519
2	237394	237065	247780	195697	236171	155456	219705	218467
3	238532	238663	247496	195749	235975	156096	235001	221073
4	238949	238360	248322	196124	233770	156163	235391	221011
5	238539	238133	248387	196406	220714	181902	235179	222751
6	238595	239112	248451	196633	176592	236223	236189	224542
7	239122	236613	247677	196861	177338	235427	236272	224187
8	238758	236223	247807	196243	177093	235310	235195	223804
9	239529	235636	248842	176481	230392	235757	236321	228994
10	238426	236073	247591	174914	248573	236529	237107	231316
11	238579	237688	247684	177375	249650	235769	235468	231745
12	239241	237447	241617	188347	249401	236973	236232	232751
13	239692	235925	177062	249792	249533	236414	236330	232107
14	238847	235962	176694	248505	249829	235629	235968	231633
15	237569	235948	177244	248742	248474	234936	235293	231172
16	237495	233782	217745	246693	244263	234488	234079	235506
17	236963	233238	246399	247092	245469	234535	236248	239992
18	234905	233215	246139	246645	246566	233069	236586	239589
19	235772	233923	246241	246465	245384	233670	236459	239702
20	235574	233815	246991	246417	246040	232257	225253	238050
21	235870	233037	246763	246761	245997	232754	179380	231509
22	235527	232807	246650	246084	245900	234428	178641	231434
23	234878	233279	247021	245540	247583	232943	177947	231313
Mean	234833	235067	237122	216692	234196	216302	220797	

Table 5.1: Vertex Counts by Probing Cycle.



(a) Mean of Vertex Counts by Hour.

(b) Mean of Edge Counts by Hour.

Figure 5.1: Mean Vertex and Edge Counts by Hour.

5.2 Statistical Measures

We examined the distribution of the vertex and edge counts by hour through the use of boxplots to determine if the traceroute results in each cycle were similar. In Figure 5.2, we see the data in hours 00 and 01 spreads over a much larger range than the subsequent hours, indicating higher

Hour	cycle-20130215	cycle-20130217	cycle-20131201	cycle-20131203	cycle-20131205	cycle-20131207	cycle-20131209	Mean
0	412194	457556	511713	387405	481539	296406	286700	404788
1	431961	501915	510625	384798	481400	289343	284998	412149
2	504682	502153	512973	387904	483691	288387	444710	446357
3	506705	504798	513469	386369	482849	289965	481603	452251
4	505450	504721	514495	386436	477627	289951	481432	451445
5	505861	502573	514454	387002	444417	349801	481690	455114
6	505088	503912	514524	386488	340942	480161	482902	459145
7	505465	498524	514248	388671	342496	478869	483140	458773
8	505577	496990	513729	387058	342490	478953	480731	457933
9	506645	496880	516163	339865	471133	481586	483650	470846
10	505080	497322	515864	336736	515665	481999	484721	476770
11	505735	499907	515532	342627	516307	480084	481793	477426
12	506867	499104	497015	368153	517238	482690	483563	479233
13	507205	497217	342131	517152	517133	479754	483713	477758
14	505182	498853	341320	514270	515231	478897	481841	476513
15	504510	498564	341998	516159	513719	478520	483288	476680
16	503843	494575	440217	512552	503700	477414	480483	487541
17	502744	493703	511568	512909	506984	478019	484414	498620
18	499383	493277	510642	510736	509458	474791	487600	497984
19	500616	494993	511244	511480	506967	477354	487132	498541
20	501525	495347	512363	510116	509094	474709	457019	494310
21	502139	493785	511016	511290	509330	474666	346845	478439
22	502222	492314	510643	510719	509432	477320	345224	478268
23	500777	493022	512145	509203	512228	473801	344044	477889
Mean	497394	496334	487920	437754	479628	433893	446802	

Table 5.2: Edge Counts by Probing Cycle.

variance. The location of the median is close to the mean in hours 00 and 01, whereas the difference in the two summary statistics tends to increase in hours 02 through 14, indicating the outliers have a significant impact on the mean vertex and edge counts mentioned in Figure 5.1. As a result, we determined the mean did not indicate similarity among the hourly partitions. Considering subsets of hours, we see hours 02 through 08 are similar, as are hours 09 through 15 and hours 16 through 23; however, these subsets are not similar to hours 00 and 01. The larger variance in hours 00 and 01 as indicated in Figure 5.2 also encourages further study to determine if there are events of note occurring in those hours that are different from the subsequent hours. We will do so in the next section.

5.3 Complex Network Measures

The visible differences in hours 00 and 01 are consistent throughout each method used to study the data. By using complex network measures, we seek to determine if the hour makes a difference in the data captured, through the study of the vertex set and edge set for an arbitrary

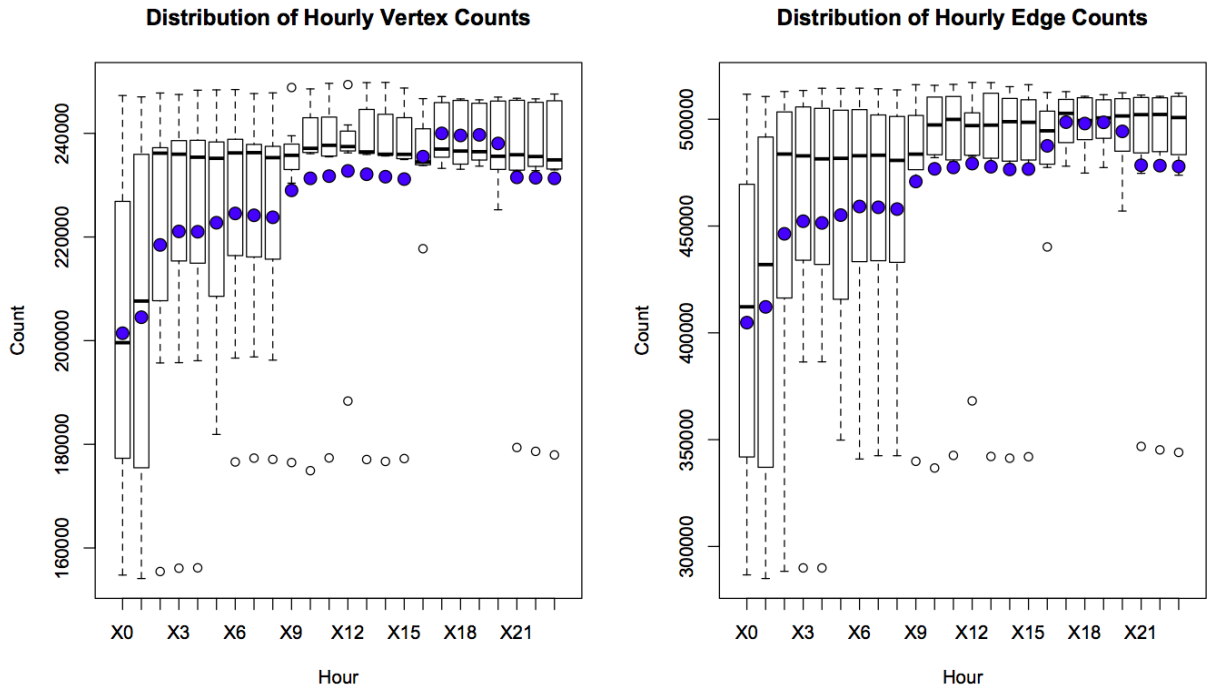


Figure 5.2: Distribution of Vertex and Edge Counts for seven probing cycles in 24-hour partitions.

probing cycle. For the following analysis, we chose probing cycle-2013_02_15.

First, we considered the pairwise inter-hour *vsd* for cycle-2013_02_15. In this representation, we see zeros along the diagonal, which we expect, as a comparison of a graph to itself. In Table 5.3, we see the *vsd* between hour 00 and the other hours (and similarly for hour 01) suggests greater changes in the captured data during hours 00 and 01 as compared to subsequent hours. Next, we considered *esd* measurements to determine if there were similar behaviors with the edge sets. In Table 5.4, the *esd* between hour 00 and the other hours (and similarly for hour 01) again exhibit a difference from subsequent sets as encountered earlier with the *vsd* values contained in Table 5.3.

2013_02_15	hour00	hour01	hour02	hour03	hour04	hour05	hour06	hour07	hour08	hour09	hour10	hour11	hour12	hour13	hour14	hour15	hour16	hour17	hour18	hour19	hour20	hour21	hour22	hour23
hour00	0	0.387	0.388	0.388	0.39	0.389	0.388	0.389	0.389	0.389	0.388	0.389	0.388	0.389	0.388	0.388	0.389	0.388	0.388	0.387	0.387	0.387	0.387	0.387
hour01	0.387	0	0.384	0.384	0.384	0.384	0.385	0.386	0.385	0.386	0.385	0.385	0.385	0.386	0.385	0.385	0.385	0.385	0.385	0.385	0.385	0.385	0.385	0.384
hour02	0.388	0.384	0	0.376	0.376	0.377	0.377	0.377	0.378	0.377	0.377	0.377	0.377	0.378	0.377	0.378	0.378	0.377	0.379	0.378	0.378	0.379	0.378	0.378
hour03	0.388	0.384	0.376	0	0.376	0.376	0.376	0.376	0.376	0.378	0.375	0.377	0.376	0.376	0.377	0.376	0.379	0.378	0.378	0.378	0.377	0.378	0.379	0.378
hour04	0.39	0.384	0.376	0.376	0	0.376	0.376	0.377	0.377	0.376	0.376	0.378	0.377	0.377	0.377	0.378	0.379	0.377	0.379	0.379	0.378	0.378	0.379	0.379
hour05	0.389	0.384	0.377	0.376	0.376	0	0.377	0.377	0.377	0.377	0.378	0.378	0.376	0.377	0.377	0.377	0.378	0.378	0.38	0.378	0.378	0.378	0.378	0.379
hour06	0.388	0.385	0.377	0.376	0.376	0.377	0	0.377	0.377	0.377	0.376	0.377	0.377	0.379	0.377	0.378	0.378	0.378	0.379	0.378	0.379	0.378	0.378	0.379
hour07	0.389	0.386	0.377	0.376	0.377	0.377	0.377	0	0.377	0.377	0.377	0.376	0.376	0.378	0.379	0.378	0.378	0.378	0.379	0.378	0.378	0.379	0.379	0.378
hour08	0.389	0.385	0.378	0.376	0.377	0.377	0.377	0.377	0	0.376	0.376	0.376	0.376	0.376	0.378	0.378	0.378	0.378	0.379	0.379	0.378	0.379	0.378	0.38
hour09	0.389	0.386	0.377	0.378	0.376	0.377	0.377	0.377	0.376	0	0.376	0.376	0.376	0.377	0.377	0.378	0.377	0.378	0.379	0.378	0.378	0.379	0.379	0.379
hour10	0.388	0.385	0.377	0.375	0.376	0.378	0.376	0.377	0.376	0.376	0	0.376	0.375	0.377	0.375	0.376	0.378	0.376	0.377	0.377	0.377	0.378	0.378	0.377
hour11	0.389	0.385	0.377	0.377	0.378	0.378	0.377	0.376	0.376	0.376	0.376	0	0.376	0.376	0.377	0.378	0.378	0.377	0.378	0.378	0.378	0.379	0.378	0.379
hour12	0.388	0.385	0.377	0.376	0.377	0.376	0.377	0.376	0.376	0.376	0.375	0.376	0	0.375	0.375	0.376	0.376	0.377	0.378	0.378	0.377	0.378	0.378	0.378
hour13	0.389	0.386	0.378	0.376	0.377	0.377	0.379	0.378	0.376	0.377	0.377	0.376	0.375	0	0.377	0.377	0.376	0.377	0.378	0.378	0.376	0.379	0.378	0.378
hour14	0.388	0.385	0.377	0.377	0.377	0.377	0.377	0.379	0.378	0.377	0.375	0.377	0.375	0.377	0	0.376	0.377	0.378	0.378	0.378	0.379	0.378	0.378	0.378
hour15	0.388	0.385	0.378	0.376	0.378	0.377	0.378	0.378	0.378	0.378	0.376	0.378	0.376	0.377	0.376	0	0.377	0.376	0.379	0.377	0.377	0.379	0.378	0.378
hour16	0.389	0.385	0.378	0.379	0.379	0.378	0.378	0.378	0.378	0.377	0.378	0.378	0.376	0.376	0.377	0.377	0	0.376	0.377	0.377	0.376	0.378	0.378	0.377
hour17	0.388	0.385	0.377	0.378	0.377	0.378	0.378	0.378	0.378	0.378	0.376	0.377	0.377	0.377	0.378	0.376	0.376	0	0.376	0.377	0.377	0.376	0.378	0.377
hour18	0.388	0.385	0.379	0.378	0.379	0.38	0.379	0.379	0.379	0.379	0.377	0.378	0.378	0.378	0.378	0.379	0.377	0.376	0	0.377	0.376	0.378	0.377	0.378
hour19	0.387	0.385	0.378	0.378	0.379	0.378	0.378	0.378	0.379	0.378	0.377	0.378	0.378	0.378	0.378	0.377	0.377	0.377	0.377	0	0.376	0.377	0.376	0.377
hour20	0.387	0.385	0.378	0.377	0.378	0.378	0.379	0.378	0.378	0.378	0.377	0.378	0.377	0.376	0.379	0.377	0.376	0.377	0.376	0.376	0	0.376	0.375	0.376
hour21	0.387	0.385	0.379	0.378	0.378	0.378	0.378	0.379	0.379	0.379	0.378	0.379	0.378	0.379	0.378	0.379	0.378	0.378	0.378	0.377	0.376	0	0.376	0.377
hour22	0.387	0.385	0.378	0.379	0.379	0.378	0.378	0.379	0.378	0.379	0.378	0.378	0.378	0.378	0.378	0.378	0.378	0.378	0.377	0.376	0.375	0.376	0	0.377
hour23	0.387	0.384	0.378	0.378	0.379	0.379	0.379	0.378	0.38	0.379	0.377	0.379	0.378	0.378	0.378	0.378	0.377	0.377	0.378	0.377	0.376	0.377	0.377	0

Table 5.3: Data of probing cycle 2013_02_15: *vsd* comparison by hour.

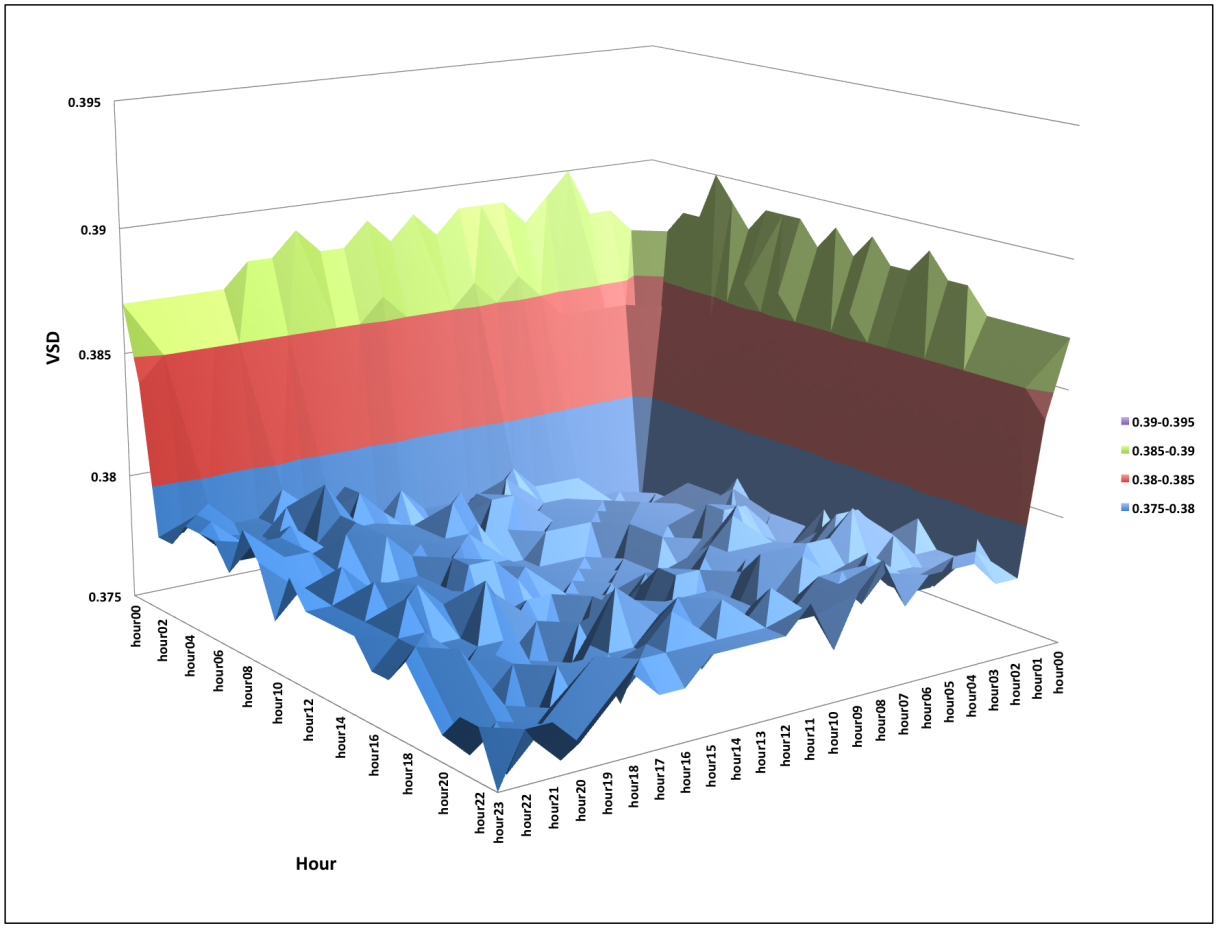


Figure 5.3: A visualization of the *vsd* comparison (24 hrs x 24 hrs) for probing cycle 2013_02_15.

2013_02_15	hour00	hour01	hour02	hour03	hour04	hour05	hour06	hour07	hour08	hour09	hour10	hour11	hour12	hour13	hour14	hour15	hour16	hour17	hour18	hour19	hour20	hour21	hour22	hour23
hour00	0.000	0.463	0.462	0.462	0.464	0.463	0.463	0.463	0.463	0.463	0.463	0.463	0.462	0.463	0.462	0.462	0.462	0.461	0.461	0.459	0.459	0.459	0.458	0.459
hour01	0.463	0.000	0.455	0.457	0.457	0.458	0.458	0.459	0.459	0.459	0.459	0.459	0.459	0.459	0.458	0.459	0.458	0.458	0.457	0.456	0.457	0.457	0.458	0.457
hour02	0.462	0.455	0.000	0.443	0.445	0.446	0.446	0.447	0.448	0.447	0.446	0.447	0.447	0.448	0.447	0.447	0.447	0.447	0.447	0.447	0.446	0.447	0.446	0.447
hour03	0.462	0.457	0.443	0.000	0.443	0.444	0.445	0.446	0.446	0.446	0.446	0.445	0.445	0.446	0.446	0.447	0.448	0.446	0.446	0.446	0.446	0.446	0.447	0.447
hour04	0.464	0.457	0.445	0.443	0.000	0.444	0.445	0.446	0.447	0.447	0.447	0.447	0.447	0.448	0.447	0.448	0.448	0.447	0.448	0.448	0.448	0.447	0.448	0.448
hour05	0.463	0.458	0.446	0.444	0.444	0.000	0.445	0.446	0.446	0.446	0.447	0.446	0.446	0.447	0.447	0.448	0.447	0.447	0.448	0.447	0.447	0.447	0.447	0.448
hour06	0.463	0.458	0.446	0.445	0.445	0.445	0.000	0.445	0.446	0.446	0.446	0.447	0.447	0.448	0.447	0.447	0.447	0.447	0.448	0.446	0.447	0.446	0.448	0.448
hour07	0.463	0.459	0.447	0.446	0.446	0.446	0.445	0.000	0.445	0.445	0.447	0.445	0.446	0.448	0.448	0.448	0.448	0.447	0.447	0.447	0.447	0.447	0.448	0.448
hour08	0.463	0.459	0.448	0.446	0.447	0.446	0.446	0.445	0.000	0.445	0.447	0.446	0.447	0.447	0.448	0.448	0.448	0.447	0.447	0.448	0.448	0.448	0.448	0.449
hour09	0.463	0.459	0.447	0.446	0.447	0.446	0.446	0.445	0.445	0.000	0.444	0.444	0.445	0.446	0.446	0.447	0.446	0.447	0.447	0.447	0.447	0.447	0.447	0.448
hour10	0.463	0.459	0.446	0.446	0.447	0.447	0.446	0.447	0.447	0.444	0.000	0.443	0.444	0.446	0.446	0.447	0.447	0.446	0.446	0.446	0.447	0.447	0.447	0.447
hour11	0.463	0.459	0.447	0.445	0.447	0.446	0.447	0.445	0.446	0.444	0.443	0.000	0.444	0.444	0.445	0.446	0.446	0.446	0.446	0.446	0.447	0.447	0.446	0.447
hour12	0.462	0.459	0.447	0.445	0.447	0.446	0.447	0.446	0.447	0.445	0.444	0.444	0.000	0.444	0.444	0.446	0.446	0.446	0.446	0.445	0.446	0.446	0.447	0.448
hour13	0.463	0.459	0.448	0.446	0.448	0.447	0.448	0.448	0.447	0.446	0.446	0.444	0.444	0.000	0.445	0.445	0.446	0.446	0.446	0.446	0.445	0.447	0.447	0.448
hour14	0.462	0.458	0.447	0.446	0.447	0.447	0.447	0.448	0.448	0.446	0.446	0.445	0.444	0.445	0.000	0.444	0.445	0.445	0.445	0.445	0.446	0.446	0.446	0.447
hour15	0.462	0.459	0.447	0.447	0.448	0.448	0.447	0.448	0.448	0.447	0.447	0.446	0.446	0.445	0.444	0.000	0.444	0.444	0.445	0.444	0.445	0.445	0.445	0.447
hour16	0.462	0.458	0.447	0.448	0.448	0.447	0.447	0.448	0.448	0.446	0.447	0.446	0.446	0.446	0.445	0.444	0.000	0.442	0.443	0.444	0.444	0.444	0.445	0.445
hour17	0.461	0.458	0.447	0.446	0.447	0.447	0.447	0.447	0.447	0.447	0.446	0.446	0.446	0.446	0.445	0.444	0.442	0.000	0.442	0.443	0.443	0.443	0.445	0.445
hour18	0.461	0.457	0.447	0.446	0.448	0.448	0.448	0.447	0.447	0.447	0.446	0.446	0.446	0.446	0.445	0.445	0.443	0.442	0.000	0.442	0.443	0.443	0.444	0.444
hour19	0.459	0.456	0.447	0.446	0.448	0.447	0.446	0.447	0.448	0.447	0.446	0.446	0.445	0.446	0.445	0.444	0.444	0.443	0.442	0.000	0.442	0.442	0.443	0.444
hour20	0.459	0.457	0.446	0.446	0.448	0.447	0.447	0.447	0.448	0.447	0.447	0.447	0.446	0.445	0.446	0.445	0.444	0.443	0.443	0.442	0.000	0.442	0.442	0.443
hour21	0.459	0.457	0.447	0.446	0.447	0.447	0.446	0.447	0.448	0.447	0.447	0.447	0.446	0.447	0.446	0.445	0.444	0.443	0.443	0.442	0.442	0.000	0.442	0.443
hour22	0.458	0.458	0.446	0.447	0.448	0.447	0.448	0.448	0.448	0.447	0.447	0.446	0.447	0.447	0.446	0.445	0.445	0.445	0.444	0.443	0.442	0.442	0.000	0.441
hour23	0.459	0.457	0.447	0.447	0.448	0.448	0.448	0.448	0.449	0.448	0.447	0.447	0.448	0.448	0.447	0.447	0.445	0.445	0.444	0.444	0.443	0.443	0.441	0.000

Table 5.4: Data of probing cycle 2013_02_15: *esd* comparison by hour.

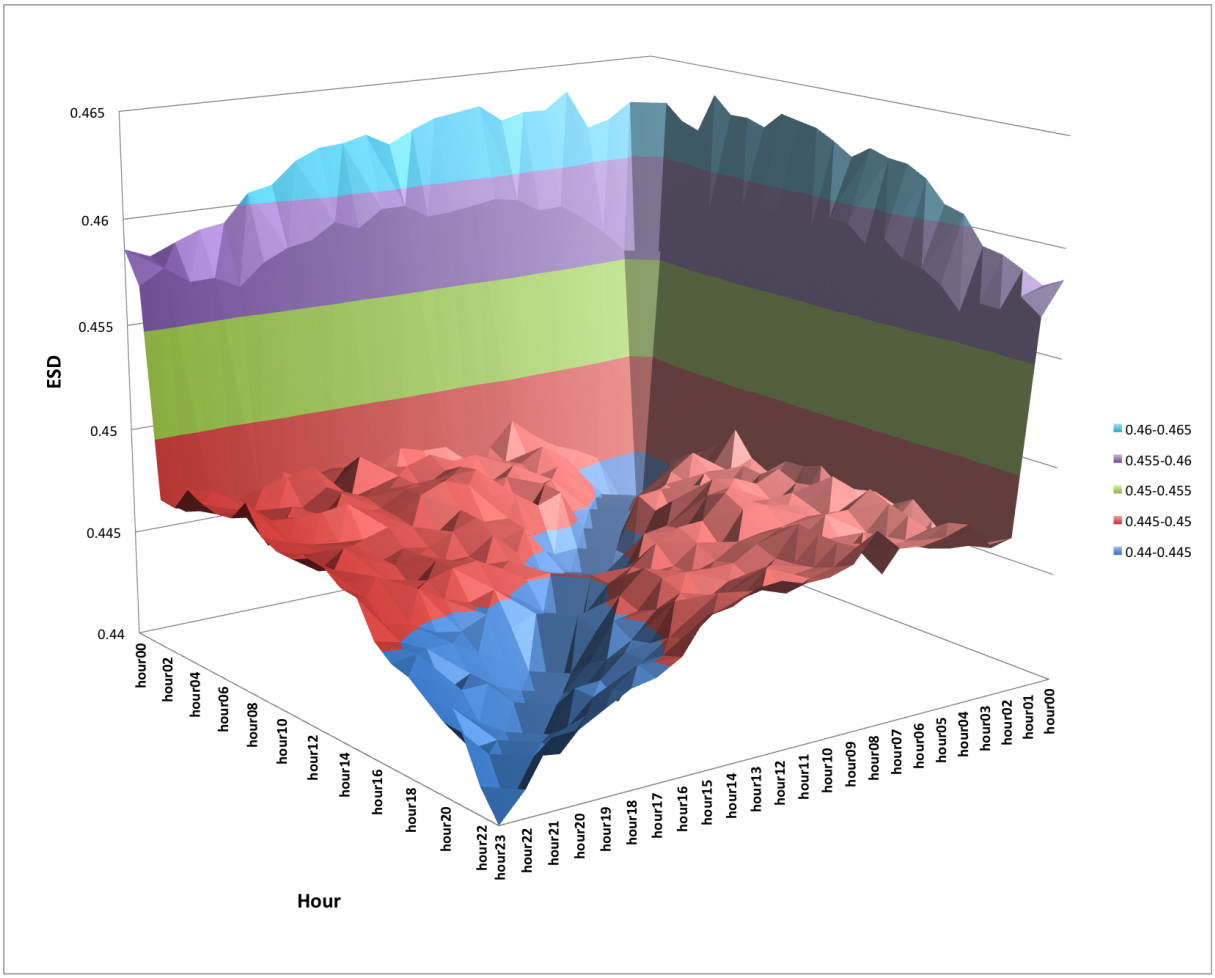


Figure 5.4: A visualization of the *esd* comparison (24 hrs x 24 hrs) for probing cycle 2013_02_15.

In Figure 5.4, we see the *esd* between hour 00 and the other hours (and similarly for hour 01) behaves similarly to the *vsd* values for the same probing cycle-2013_02_15, revealing the percentage of elements in the edge set for hours 00 and 01 that are not present in subsequent hours. We included a more detailed illustration of the *esd* values in Figure 5.4 to emphasize the difference, while small, between the first two hours and the subsequent hours.

As all the measurements to this point indicated a difference in the vertex and edge sets for hours 00 and 01 when compared to subsequent hours, we investigated the elements of the vertex and edge sets of these two hours and compare them to the vertex and edge sets for the subsequent hours as a whole. We did so by determining the union of the vertex and edge sets for three different sets: all 24 hours including hours 00 and 01, denoted as $G_{\cup_{all}}$, all 23 hours: 01,02,...,23 (excluding hour 00), denoted as $G_{\cup_{00}}$, and all 23 hours: 00,02,03,04,...,23 (excluding hour 01), denoted as $G_{\cup_{01}}$. We then performed a difference among these three sets; the results are indicated in Table 5.5. Let G_{00} and G_{01} be the graphs representing hours 00 and 01 respectively. In Table 5.7, $G_{\cap_{all}}$ is the graph representing the intersection of all 24 graphs. $G_{\cap_0}$ is the graph representing the intersection of 23 hours, excluding hour 00. In Table 5.8, $G_{\cap_0}$ is the graph representing the intersection of 23 hours, excluding hour 01. The results of the union and intersection of graphs is indicated in the four tables to follow: Table 5.5 through Table 5.8. In Table 5.5, the 12781 vertices in G^*_{00} represent the vertices that are unique to G_{00}. G^*_{00} represents the difference between the vertex and edge sets. In an effort to determine if there is a geographical relationship between these unique vertices, we performed a geolocation on each IP address in G^*_{00}; the results are in Figure 5.5.

	G_{00}	$G_{\cup_{all}}$	$G_{\cup_{00}}$	$G^*_{00} = G_{00} - G_{\cup_{00}}$	$G_{\cup_{00}} - G_{00}$	$G^*_{00} \cap G^*_{01}$	$G_{00} - G_{\cup_{all}}$	$G_{\cup_{all}} - G_{00}$
Vertex Count	199607	962897	950116	12781	763290	124858	0	763290
Edge Count	412194	2220021	2192721	182969	1714877	226755	0	1972048

Table 5.5: Vertex and Edge Set Differences for Hour 00, probing cycle 2013_02_15.

	G_{01}	$G_{\cup_{all}}$	$G_{\cup_{01}}$	$G^*_{01} = G_{01} - G_{\cup_{01}}$	$G_{\cup_{01}} - G_{01}$	$G^*_{00} \cap G^*_{01}$	$G_{01} - G_{\cup_{all}}$	$G_{\cup_{all}} - G_{01}$
Vertex Count	207639	962897	949073	13824	755258	124858	0	755258
Edge Count	431961	2220021	2190236	192420	1614782	226755	0	1961495

Table 5.6: Vertex and Edge Set Differences for Hour 01, probing cycle 2013_02_15.

	G_0	$G_0 \cap G_1$	$G_{\cap_{all}}$	$G_{\cap_0}$	$G_{\cap_0} - G_0$	$G_0 - G_{\cap_0}$	$G_{\cap_{all}} - G_0$	$G_0 - G_{\cap_{all}}$
Vertex Count	199607	124858	70304	71829	1525	129303	0	129303
Edge Count	412194	226755	101747	104395	45048	352847	41985	352432

Table 5.7: Vertex and Edge Set Intersections for Hour 00, probing cycle 2013_02_15.

	G_1	$G_0 \cap G_1$	$G_{\cap_{all}}$	$G_{\cap_1}$	$G_{\cap_1} - G_1$	$G_1 - G_{\cap_1}$	$G_{\cap_{all}} - G_1$	$G_1 - G_{\cap_{all}}$
Vertex Count	207639	124858	70304	71554	1250	137335	0	129303
Edge Count	431961	226755	101747	104218	74058	372487	71357	401571

Table 5.8: Vertex and Edge Set Intersections for Hour 01, probing cycle 2013_02_15.

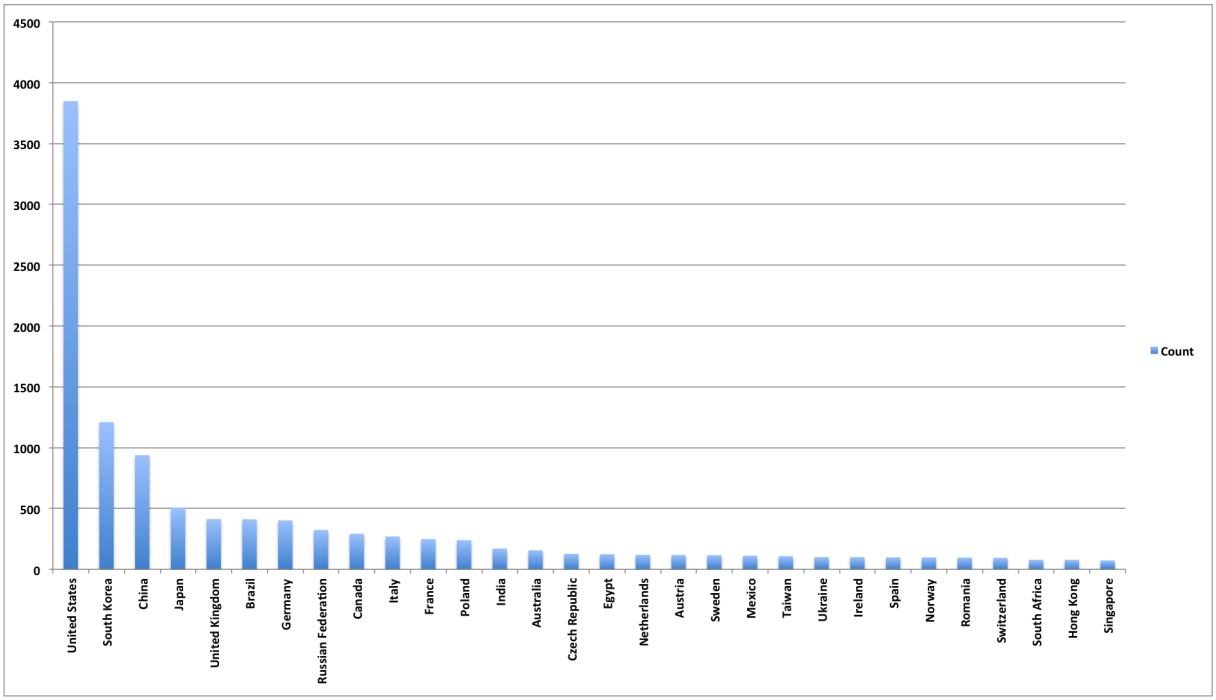


Figure 5.5: Data of G^*_{00}: vertex count by geographic location.

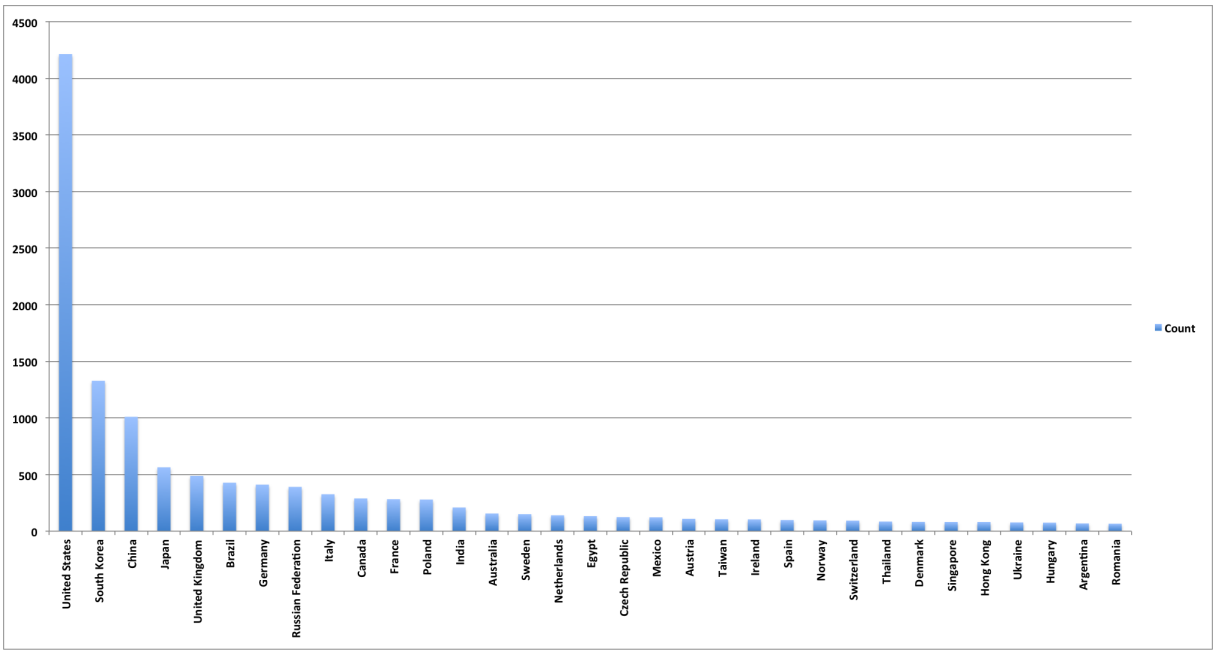


Figure 5.6: Data of G^*_{01}: vertex count by geographic location.

In Figure 5.5, we depict 30 of the countries with the highest number of allocated IPv4 addresses in G^*_{00}. The distribution of vertices across multiple countries does not suggest a unique country or AS as the attributing factor for the increased *vsd* between G_{00} and the other hours. In Table 5.9, a comparison to the percentage of total allocated IPv4 address space from [29] shows the proportion of unique vertices in G^*_{00} are generally consistent with the distribution of IPv4 addresses across the Internet. In Figure 5.6, the results are similar for G^*_{01}.

Let $\bar{G}_n$ be the graph representing the difference of $G_n \backslash G_{\cap_{all}}$, where n is the hour. We computed the *vsd* of $\bar{G}_n$, where $n \in \{0,1,2,...,23\}$, comparing the $\bar{G}_{00}$ graph to all other $\bar{G}_n$ graphs for all values of n. We performed the *vsd* comparison for every pair of graphs of $\bar{G}_n$, resulting in Table 5.10. The large *vsd* values, between 50%-60%, in Table 5.10 suggest all graphs contain a similar number of vertices that are unique to that hour's graphical representation.

Country	%IP Space	G00*	G01*
United States	35.9%	30.1%	30.5%
China	7.7%	7.3%	7.3%
Japan	4.7%	4.0%	4.1%
United Kingdom	2.9%	3.2%	3.5%
Germany	2.8%	3.1%	3.0%
South Korea	2.6%	9.5%	9.6%
France	2.2%	1.9%	2.0%
Canada	1.9%	2.3%	2.1%
Italy	1.2%	2.1%	2.4%
Brazil	1.1%	3.2%	3.1%
Australia	1.1%	1.2%	1.1%
Netherlands	1.1%	0.9%	1.0%
Russian Federation	1.0%	2.5%	2.8%
India	0.8%	1.3%	1.5%
Taiwan	0.8%	0.8%	0.8%

Country	%IP Space	G00*	G01*
Sweden	0.70%	0.9%	1.1%
Spain	0.70%	0.8%	0.7%
Mexico	0.60%	0.9%	0.9%
Poland	0.50%	1.9%	2.0%
Switzerland	0.50%	0.7%	0.7%
South Africa	0.50%	0.6%	0.5%
Norway	0.40%	0.8%	0.7%
Indonesia	0.40%	0.6%	0.4%
Turkey	0.40%	0.5%	0.4%
Austria	0.30%	0.9%	0.8%
Denmark	0.30%	0.5%	0.6%
Hong Kong	0.30%	0.6%	0.6%
Ukraine	0.30%	0.8%	0.6%
Argentina	0.30%	0.5%	0.5%
Romania	0.30%	0.8%	0.5%

Table 5.9: Percentage of IPv4 allocation space for G^*_{00} and G^*_{01} vertices by country, from [29].

$\bar{G}_n$	$\bar{G}_{00}$	$\bar{G}_{01}$	$\bar{G}_{02}$	$\bar{G}_{03}$	$\bar{G}_{04}$	$\bar{G}_{05}$	$\bar{G}_{06}$	$\bar{G}_{07}$	$\bar{G}_{08}$	$\bar{G}_{09}$	$\bar{G}_{10}$	$\bar{G}_{11}$	$\bar{G}_{12}$	$\bar{G}_{13}$	$\bar{G}_{14}$	$\bar{G}_{15}$	$\bar{G}_{16}$	$\bar{G}_{17}$	$\bar{G}_{18}$	$\bar{G}_{19}$	$\bar{G}_{20}$	$\bar{G}_{21}$	$\bar{G}_{22}$	$\bar{G}_{23}$
$\bar{G}_{00}$	0.000	0.591	0.572	0.571	0.574	0.573	0.572	0.572	0.572	0.573	0.571	0.573	0.571	0.572	0.572	0.573	0.573	0.572	0.574	0.572	0.571	0.572	0.571	0.573
$\bar{G}_{01}$	0.591	0.000	0.561	0.561	0.561	0.561	0.562	0.563	0.561	0.563	0.562	0.563	0.562	0.563	0.562	0.563	0.562	0.563	0.564	0.564	0.563	0.564	0.564	0.563
$\bar{G}_{02}$	0.572	0.561	0.000	0.534	0.533	0.535	0.535	0.535	0.536	0.534	0.535	0.535	0.534	0.536	0.535	0.536	0.537	0.537	0.540	0.538	0.537	0.539	0.538	0.538
$\bar{G}_{03}$	0.571	0.561	0.534	0.000	0.533	0.533	0.534	0.533	0.534	0.535	0.532	0.534	0.533	0.533	0.534	0.534	0.537	0.536	0.538	0.537	0.536	0.538	0.539	0.538
$\bar{G}_{04}$	0.574	0.561	0.533	0.533	0.000	0.533	0.533	0.535	0.534	0.533	0.533	0.536	0.533	0.534	0.534	0.536	0.537	0.536	0.539	0.538	0.538	0.537	0.539	0.539
$\bar{G}_{05}$	0.573	0.561	0.535	0.533	0.533	0.000	0.534	0.535	0.534	0.534	0.536	0.536	0.532	0.533	0.534	0.535	0.536	0.536	0.540	0.537	0.538	0.538	0.537	0.539
$\bar{G}_{06}$	0.572	0.562	0.535	0.534	0.533	0.534	0.000	0.534	0.535	0.535	0.534	0.535	0.534	0.536	0.534	0.536	0.536	0.536	0.538	0.538	0.539	0.537	0.538	0.539
$\bar{G}_{07}$	0.572	0.563	0.535	0.533	0.535	0.535	0.534	0.000	0.534	0.533	0.534	0.533	0.532	0.535	0.536	0.536	0.537	0.536	0.538	0.537	0.537	0.538	0.538	0.538
$\bar{G}_{08}$	0.572	0.561	0.536	0.534	0.534	0.534	0.535	0.534	0.000	0.533	0.533	0.534	0.533	0.532	0.536	0.536	0.536	0.537	0.539	0.539	0.537	0.539	0.538	0.540
$\bar{G}_{09}$	0.573	0.563	0.534	0.535	0.533	0.534	0.535	0.533	0.533	0.000	0.533	0.532	0.532	0.533	0.534	0.535	0.535	0.536	0.539	0.537	0.537	0.539	0.539	0.538
$\bar{G}_{10}$	0.571	0.562	0.535	0.532	0.533	0.536	0.534	0.534	0.533	0.533	0.000	0.532	0.532	0.534	0.532	0.534	0.536	0.534	0.537	0.536	0.536	0.537	0.538	0.537
$\bar{G}_{11}$	0.573	0.563	0.535	0.534	0.536	0.536	0.535	0.533	0.534	0.532	0.532	0.000	0.532	0.533	0.534	0.537	0.536	0.535	0.538	0.537	0.538	0.539	0.538	0.539
$\bar{G}_{12}$	0.571	0.562	0.534	0.533	0.533	0.532	0.534	0.532	0.533	0.532	0.532	0.532	0.000	0.531	0.531	0.534	0.534	0.535	0.537	0.536	0.536	0.536	0.536	0.538
$\bar{G}_{13}$	0.572	0.563	0.536	0.533	0.534	0.533	0.536	0.535	0.532	0.533	0.534	0.533	0.531	0.000	0.534	0.534	0.533	0.535	0.538	0.537	0.535	0.538	0.537	0.537
$\bar{G}_{14}$	0.572	0.562	0.535	0.534	0.534	0.534	0.534	0.536	0.536	0.534	0.532	0.534	0.531	0.534	0.000	0.534	0.536	0.536	0.538	0.537	0.538	0.537	0.537	0.537
$\bar{G}_{15}$	0.573	0.563	0.536	0.534	0.536	0.535	0.536	0.536	0.536	0.535	0.534	0.537	0.534	0.534	0.534	0.000	0.536	0.534	0.539	0.536	0.536	0.539	0.538	0.538
$\bar{G}_{16}$	0.573	0.562	0.537	0.537	0.537	0.536	0.536	0.537	0.536	0.535	0.536	0.536	0.534	0.533	0.536	0.536	0.000	0.534	0.537	0.537	0.535	0.538	0.538	0.536
$\bar{G}_{17}$	0.572	0.563	0.537	0.536	0.536	0.536	0.536	0.536	0.537	0.536	0.534	0.535	0.535	0.535	0.536	0.534	0.534	0.000	0.536	0.536	0.536	0.535	0.538	0.537
$\bar{G}_{18}$	0.574	0.564	0.540	0.538	0.539	0.540	0.538	0.538	0.539	0.539	0.537	0.538	0.537	0.538	0.538	0.539	0.537	0.536	0.000	0.538	0.537	0.539	0.538	0.539
$\bar{G}_{19}$	0.572	0.564	0.538	0.537	0.538	0.537	0.538	0.537	0.539	0.537	0.536	0.537	0.536	0.537	0.537	0.536	0.537	0.536	0.538	0.000	0.535	0.537	0.537	0.537
$\bar{G}_{20}$	0.571	0.563	0.537	0.536	0.538	0.538	0.539	0.537	0.537	0.537	0.536	0.538	0.536	0.535	0.538	0.536	0.535	0.536	0.537	0.535	0.000	0.536	0.535	0.537
$\bar{G}_{21}$	0.572	0.564	0.539	0.538	0.537	0.538	0.537	0.538	0.539	0.539	0.537	0.539	0.536	0.538	0.537	0.539	0.538	0.535	0.539	0.537	0.536	0.000	0.537	0.538
$\bar{G}_{22}$	0.571	0.564	0.538	0.539	0.539	0.537	0.538	0.538	0.538	0.539	0.538	0.538	0.536	0.537	0.537	0.538	0.538	0.538	0.538	0.537	0.535	0.537	0.000	0.537
$\bar{G}_{23}$	0.573	0.563	0.538	0.538	0.539	0.539	0.539	0.538	0.540	0.538	0.537	0.539	0.538	0.537	0.537	0.538	0.536	0.537	0.539	0.537	0.537	0.538	0.537	0.000

Table 5.10: Data of $\bar{G}_n$: *vsd* comparison by hour.

Given the results of Figure 5.2, we considered the possibility that each of the remaining six probing cycles would exhibit similar behavior as revealed in probing cycle-2013_02_15. We began with a *vsd* comparison for the remaining six probing cycles as indicated in Figure 5.7. All of the graphs display a maximum *vsd* between hours with the largest difference in vertex counts as indicated in Table 5.1. For example, in Figure 5.8, hour 00 has over 17000 less vertices than hour 01. This difference is consistent with comparisons between hour 00 and subsequent hours as well, as depicted in Figure 5.8. The remaining figures reveal a similar relationship between the hour with the lowest vertex count and the maximum *vsd*; the maximum *vsd* indicated in the figure corresponds to the hour within the probing cycle with the lowest vertex count. From Figure 5.7 to Figure 5.13, we see the results are similar for every probing cycle, which could be the explanation of the slightly bigger *vsd* values.

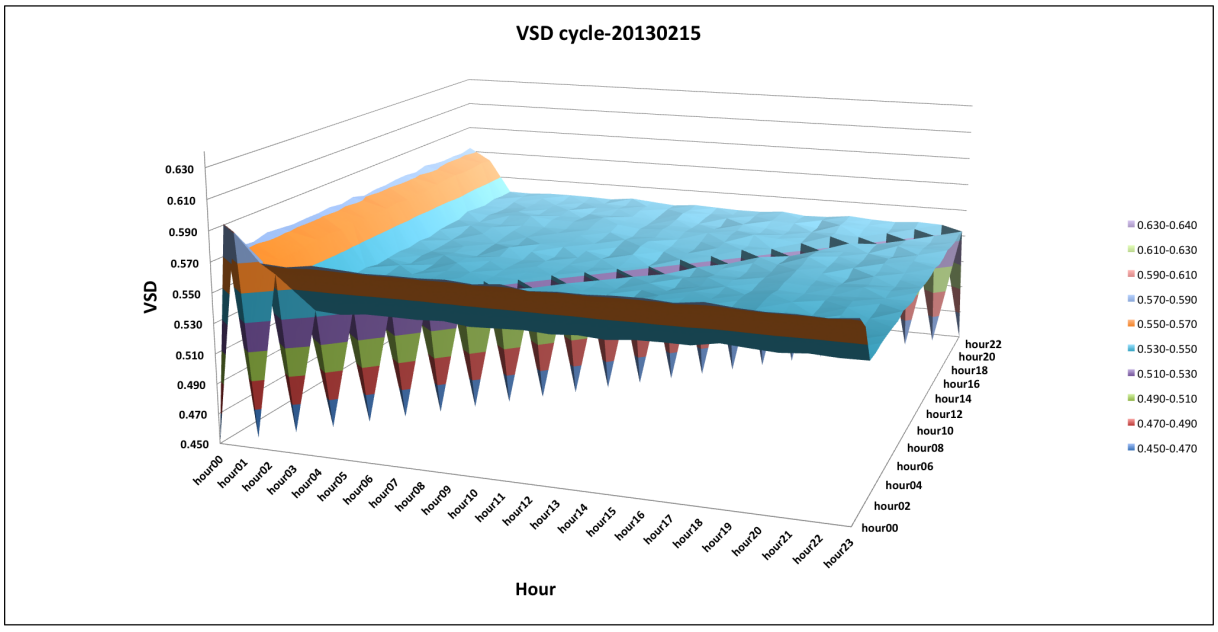


Figure 5.7: A visualization of the *vsd* comparison (24 hrs x 24 hrs) for probing cycle 2013_02_15.

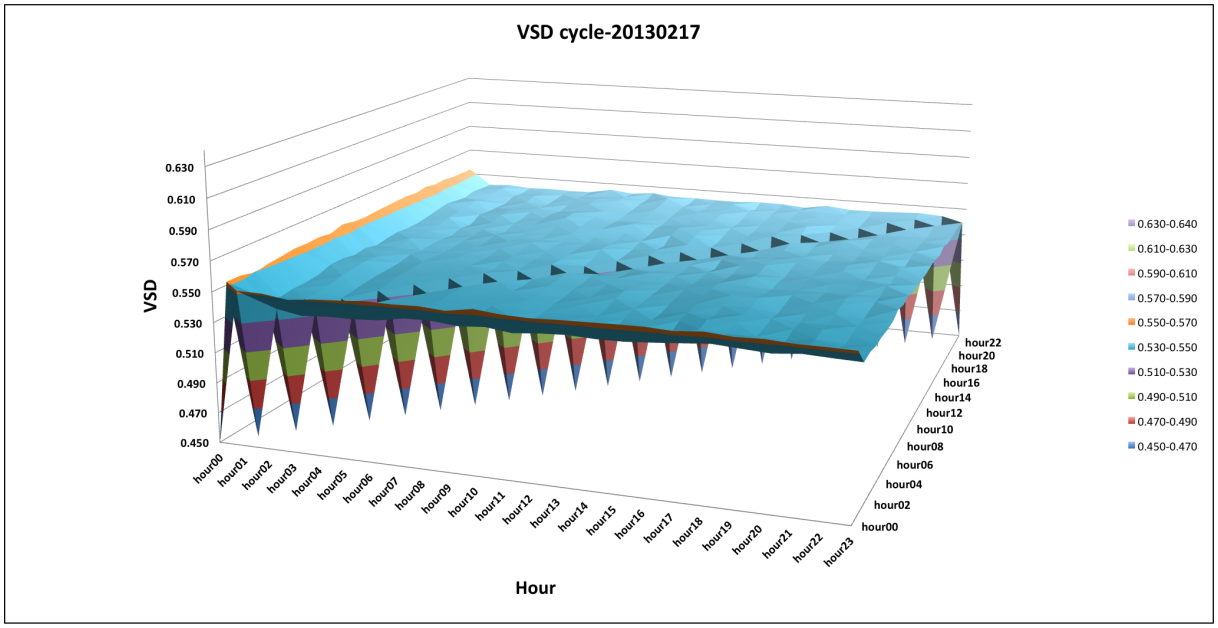


Figure 5.8: A visualization of the *vsd* comparison (24 hrs x 24 hrs) for probing cycle 2013_02_17.

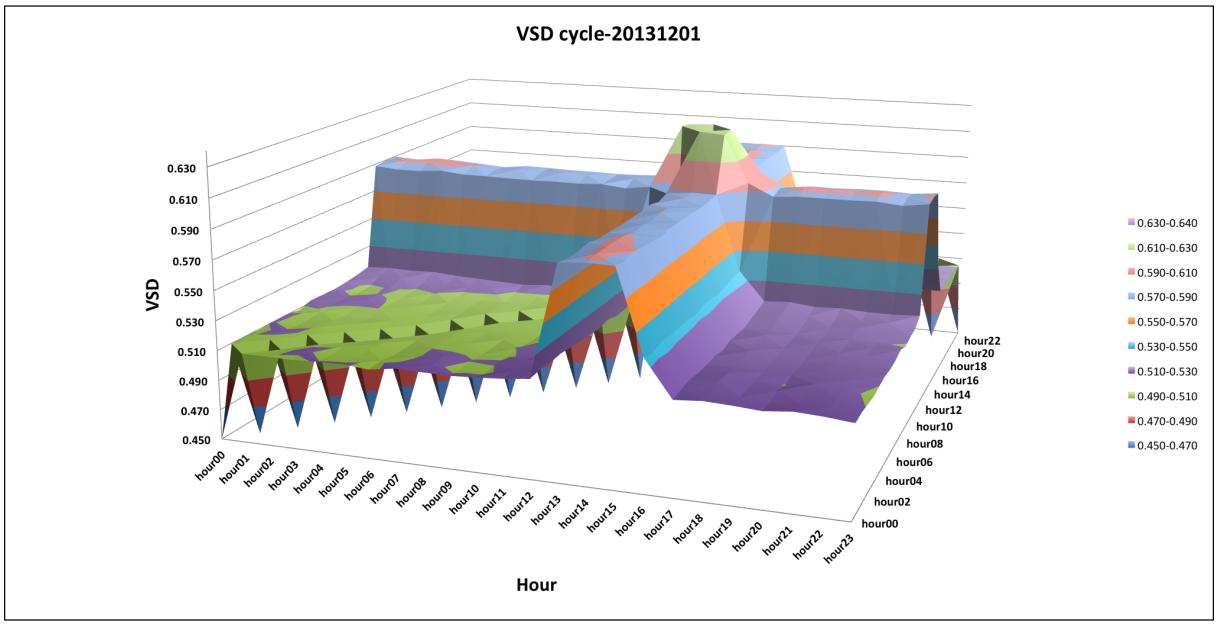


Figure 5.9: A visualization of the *vsd* comparison (24 hrs x 24 hrs) for probing cycle 2013_12_01.

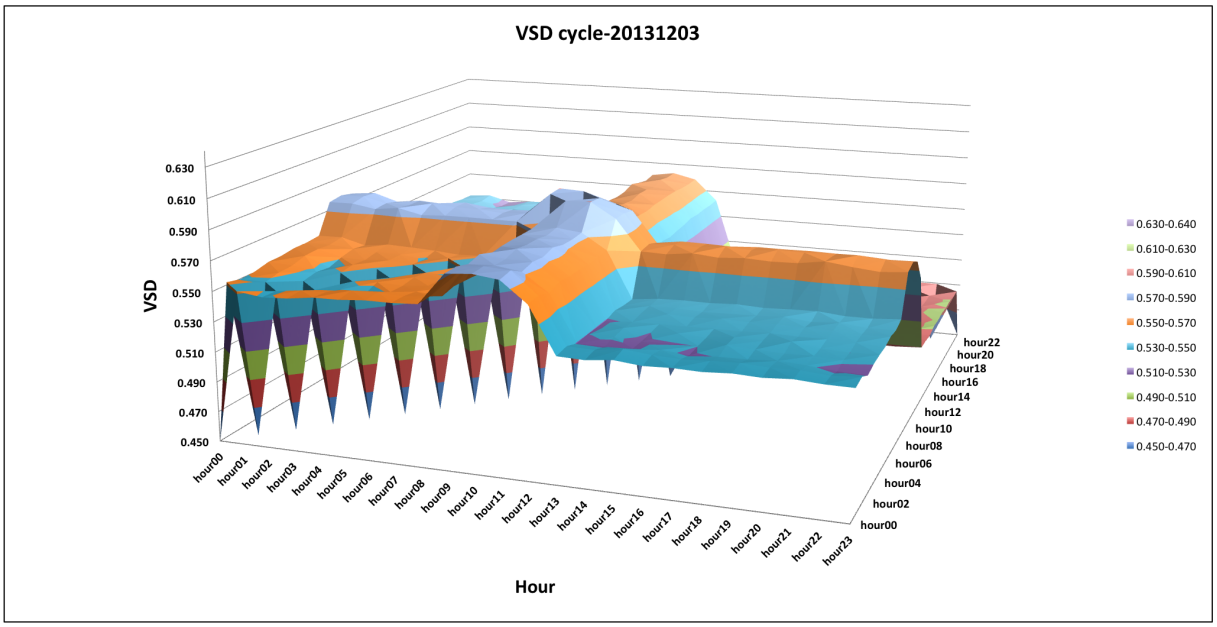


Figure 5.10: A visualization of the *vsd* comparison (24 hrs x 24 hrs) for probing cycle 2013_12_03.

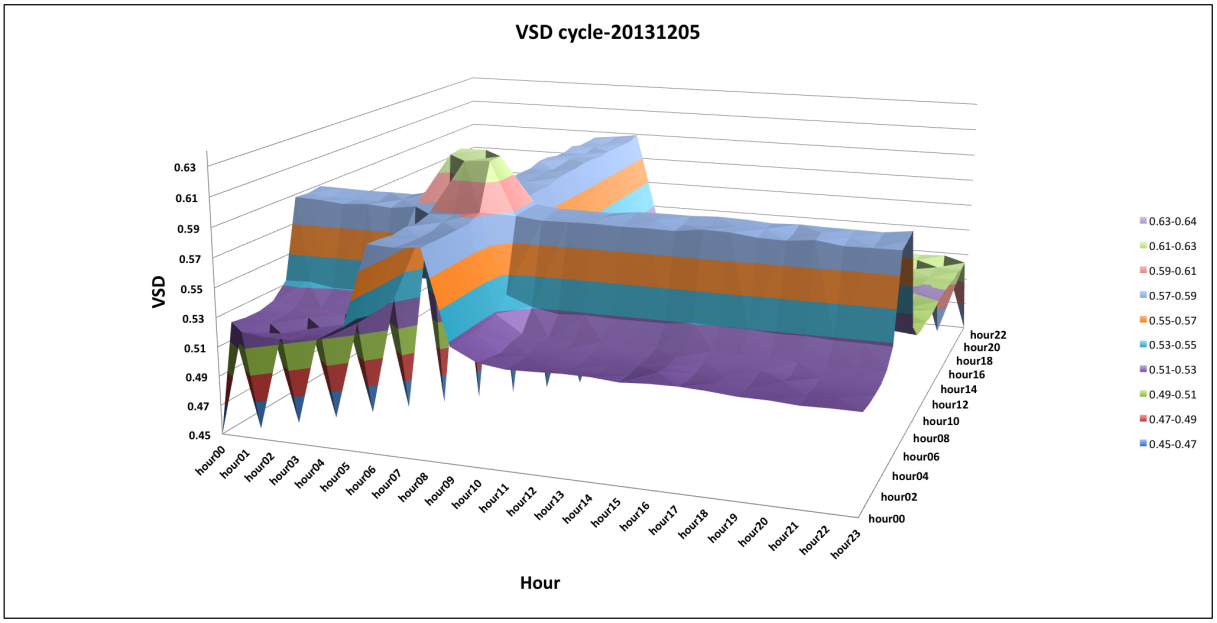


Figure 5.11: A visualization of the *vsd* comparison (24 hrs x 24 hrs) for probing cycle 2013_12_05.

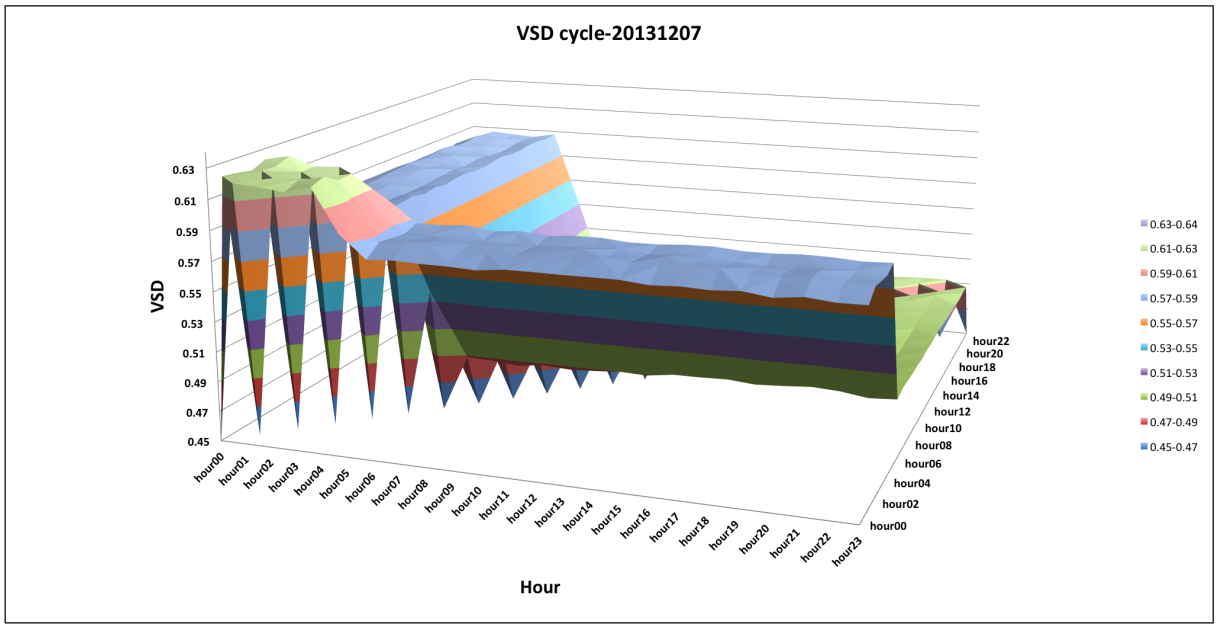


Figure 5.12: A visualization of the *vsd* comparison (24 hrs x 24 hrs) for probing cycle 2013_12_07.

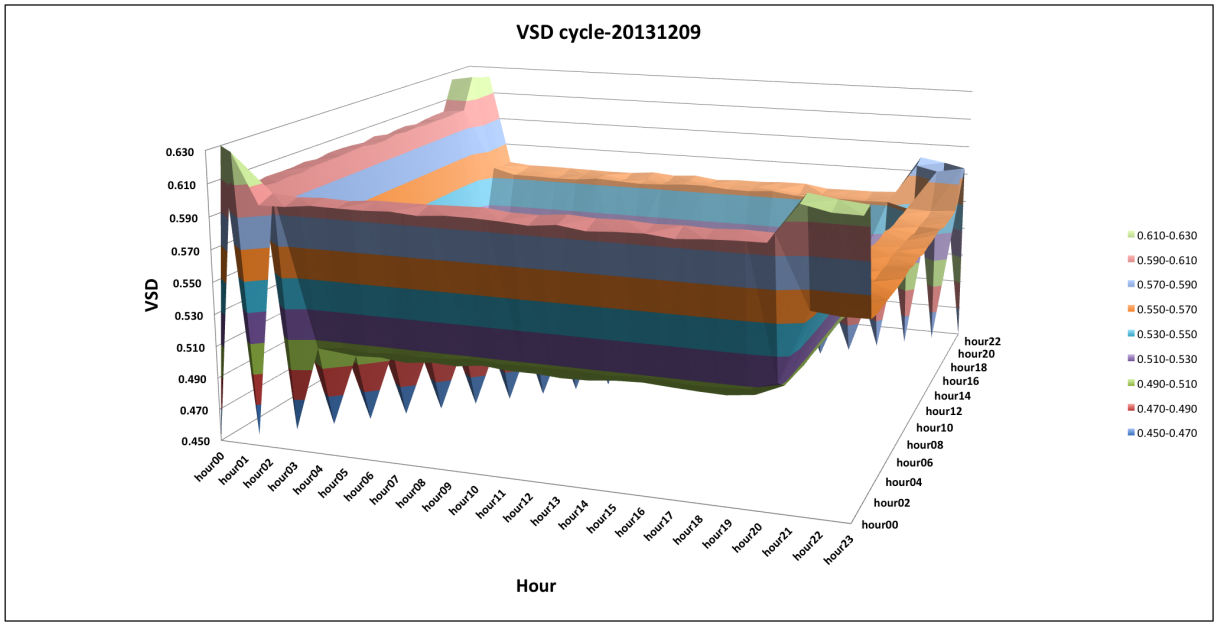


Figure 5.13: A visualization of the *vsd* comparison (24 hrs x 24 hrs) for probing cycle 2013_12_09.

CHAPTER 6:

Future Work and Conclusion

In this chapter, we present our findings and provide insights into areas that might require further research.

6.1 Summary

The intent of our research was to determine if the time of day is a factor when probing the Internet for measurement. Given the results from our analysis of seven probing cycles, there is no indication that time is a factor. The graph measures of vertex and edge count played a significant role in determining our outcome; however, the use of graph measures alone is not sufficient. While the statistical measures allowed for quantitative comparisons of each hourly partition, the small sample size of seven probing cycles was not enough to employ more robust statistical analysis. The boxplot in Figure 5.2 identified a difference in hours 00 and 01 when compared to the subsequent sets. The processing time required to convert the large files containing CAIDA probing cycles limited our ability to infer a difference with certainty. We believe there is a relationship between low vertex counts and higher *vsd* values. The use of all three measures reinforced the reasoning and analysis that resulted in our outcome.

6.2 Future Work

A limitation of our research is the large size of the warts files representing a probing cycle from CAIDA. Given file sizes over 3GB per cycle, the time required to format the data for analysis limited the focus of our research to seven probing cycles. Using a larger sample size, future work in the areas below would expand our research.

- The vertex and edge counts serve as data for the statistical and complex network measures. Using additional graph measures such as clustering coefficient, radius, and diameter may reveal insights not addressed in this research.
- The statistical measures offer the ability to employ hypothesis testing on the mean of hourly partitions to determine if the means are similar. One could apply a multiple comparisons test to determine if the means in Figure 5.1 are similar. Combining the hourly partitions of multiple cycles may reveal additional insights as well.
- The visualizations of the December 2013 probing cycles illustrate a relationship between the vertex count and *vsd* along the diagonal of the figure. As the day count increases, the maximum *vsd* shifts towards hour 00 compared to the other probing cycles. Further study to include additional probing cycles would determine if the phenomena is unique to those probing cycles or indicative of other properties, such as the beginning of a probing cycle.

6.3 Conclusion

The use of graphical, statistical, and complex network measures gave no indication of time as a factor in probing the Internet for measurement. The three measures were not sufficient individually; however, relationships between the measures contributed to our outcome, particularly the relationship between large variations in vertex counts and larger *vsd*. The variations in vertex counts result in significant changes in the vertex set for each graph, increasing the *vsd*.

REFERENCES

[1] M. Waldrop, "DARPA and the Internet Revolution: 50 years of bridging the gap," *DARPA*, 78–85, Apr. 2008.

[2] E. Nygren, R. Sitaraman, and J. Sun, "The akamai network: a platform for high-performance Internet applications," *ACM SIGOPS Operating Systems Review*, vol. 44, no. 3, pp. 2–19, July 2010.

[3] S. Floyd and V. Paxson, "Difficulties in simulating the Internet," *IEEE/ACM Transactions on Networking (TON)*, vol. 9, no. 4, pp. 392–403, Aug. 2001.

[4] D. Lee, "Toward large-graph comparison measures to understand Internet topology dynamics," M.S. Thesis, GSEAS, Naval Postgraduate School, Monterey, CA, 2013.

[5] B. Cheswick and H. Burch, Internet mapping project, Lumeta Corporation, 2000.

[6] *Network and Security: OSI and Network Layers* [Online]. Tri-Tel, Elgin, IL, 2013. Available: http://tri-tel.com/2013/10/25/structured-cabling-defined-part-3-network-and-security/

[7] Guidelines for creation, selection, and registration of an autonomous system (AS), RFC 1930, 1996.

[8] M. H. Gunes and K. Sarac, "Inferring subnets in router-level topology collection studies," in *Proceedings of the 7th ACM SIGCOMM Conf. on Internet measurement*, San Diego, CA, 2007, pp. 203–208.

[9] L. Gao, "On inferring autonomous system relationships in the Internet," in *IEEE Global Telecommunications Conference (GLOBECOM)*, San Francisco, CA, 2000, pp. 387–396.

[10] S. Knight, H. X. Nguyen, N. Falkner, R. Bowden, and M. Roughan, "The Internet topology zoo," *Selected Areas in Communications, IEEE Journal*, vol. 29, no. 9, pp. 1765–1775, Oct. 2011.

[11] R. Beverly, A. Berger, and G. Xie. "Primitives for active internet topology mapping: Toward high-frequency characterization," presented at the Tenth ACM SIGCOMM/USENIX Internet Measurement Conference (IMC), Melbourne, Australia, Nov. 1–3, 2010.

[12] T. Bourgeau and T. Friedman, "Toward fast and efficient IP-level network topology capture," in *Proc. of the 2012 ACM conference on CoNEXT student workshop*, 2012, pp. 5–6.

[13] V. Jacobson, Traceroute software [Online]. Lawrence Berkeley Laboratories, 1989. Available: ftp://ftp.ee.lbl.gov/traceroute.tar.gz.

[14] K. Rosen, *Discrete Mathematics and Its Applications*, New York, NY: McGraw-Hill, 2011, pp. 589-610.

[15] T. Tantau, TikZ Examples: Venn diagram [Online]. Available: http://www.texample.net/tikz/examples/venn-diagram/.

[16] L. Euler, "Leonhard Euler and the Königsberg bridges," *Scientific American*, vol. 189, pp. 66–70, July 1953.

[17] G. Chartrand and P. Zhang, *A First Course in Graph Theory*, Mineola, NY: Dover, 2012.

[18] A. Dickson, "Introduction to Graph Theory," unpublished.

[19] M. E. J. Newman, "The structure and function of complex networks," *SIAM Review*, vol. 45, no. 2, pp. 167–256, Jan. 2003.

[20] D. B. West, *Introduction to graph theory*, Englewood Cliffs, NJ: Prentice Hall, 2001.

[21] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, pp. 440–442, Jun. 1998.

[22] A. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, pp. 509–512, Oct. 1999.

[23] D. Wackerly, W. Mendenhall, and R. Scheaffer, *Mathematical Statistics with Applications*, Belmont, CA: Thomson, 2008.

[24] L. Gonick and W. Smith, *The cartoon guide to statistics*, New York, NY: Harper Collins, 1993.

[25] Y. Hyun, CAIDA Monitors: The Archipelago Measurement Infrastructure [Online], Available: http://www.caida.org/data/monitors/monitor-map-ark.xml.

[26] *Scamper* [Online], Cooperative Association for Internet Data Analysis. Available: http://www.caida.org/tools/measurement/scamper/.

[27] The IPv4 Routed /24 Topology Dataset [Online], CAIDA. Available: http://www.caida.org/data/active/ipv4_routed_24_topology_dataset.xml.

[28] GeoIP [Online], LLC MaxMind. Available: https://www.maxmind.com/en/geolocation_landing

[29] T. Tantau, Allocation of IP Addresses by Country [Online]. Available: https://www.countryipblocks.net/allocation-of-ip-addresses-by-country.php.

Initial Distribution List

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California