An Empirical Analysis and Evaluation of Internet Robustness

Michele Stampanoni

STMMIC001

Submitted to the University of Cape Town in partial fulfillment of the requirements for the degree: **Master of Philosophy specializing in Financial Technology**.

Faculty of Commerce: School of Economics

University of Cape Town

Supervisor: Associate Professor Co-Pierre Georg

December 2022

Abstract

The study of network robustness is a critical tool in the understanding of complex interconnected systems such as the Internet, which due to digitalization, gives rise to an increasing prevalence of cyberattacks. Robustness is when a network maintains its basic functionality even under failure of some of its components, in this instance being nodes or edges. Despite the importance of the Internet in the global economic system, it is rare to find empirical analyses of the global pattern of Internet traffic data established via backbone connections, which can be defined as an interconnected network of nodes and edges between which bandwidth flows. Hence in this thesis, I use metrics based on graph properties of network models to evaluate the robustness of the backbone network, which is further supported by international cybersecurity ratings. These cybersecurity ratings are adapted from the Global Cybersecurity Index which measures countries' commitments to cybersecurity and ranks countries based on their cybersecurity strategies. Ultimately this empirical analysis follows a three-step process of firstly mapping the Internet as a network of networks, followed by analysing the various networks and country profiles, and finally assessing each regional network's robustness. By using TeleGeography and ITU data, the results show that the regions with countries which have higher cybersecurity ratings in turn have more robust networks, when compared to regions with countries which have lower cybersecurity ratings.

1 Introduction

The objective of this thesis is to provide an analysis of the robustness of the Internet in relation to the importance of different nodes that govern the functioning of the Internet, which is measured by high graph-theoretical robustness metrics as well as high levels of protection the nodes have against cyberattacks. From this, the primary contribution of the thesis is a detailed disaggregated analysis of country level impacts to the robustness of the Internet as a whole, through analysing the criticality of the Internet backbone from a graph-theoretical perspective. The objective and contribution differ from the previous literature in that the core of this thesis makes use of real-world Internet geography and global cybersecurity data upon applying network and graph theory analysis. Similar to Yan et al. (2010), the main theme of this thesis is to perform a criticality analysis of the Internet infrastructure at a national level, with the difference being the use of different robustness methodologies and cybersecurity ratings as opposed to applying various traffic modelling techniques.

Network robustness is defined as a measure of a network's ability to continue functioning when part of the network is naturally damaged or targeted for attack. In the case of the Internet, the ability of the network to maintain its function means transferring data between pairs of nodes. The focus of network robustness is always placed on topological robustness properties and the corresponding topological robustness graph metrics, which mainly abstract from technical or organizational details of the Internet by the use of mathematical graph theory and particular graph models (Oehlers & Fabian, 2021).

Given the Internet represents the global infrastructure for transporting products of the digital economy – information, business processes and communications – more individuals, firms and countries have come to rely on this global network infrastructure. While the Internet is a system that efficiently creates wealth in the economy ($8 trillion is exchanged each year)(McKinsey Global Institute, 2011), it can also be used to promote the spread of cyberattacks, which as of 2016 cost the global economy between $445 billion and $600 billion (IBM, 2020). Hence by using a series of empirical studies such as formal network analysis and graph theory, it is possible to quantify the robustness and vulnerability of such a scale-free network in the face of an attack. The interconnectedness of globalization is further enhanced by the fact that two billion people are connected to the Internet (McKinsey Global Institute, 2011), and the reason opportunities for cyberattacks are so prevalent stems from the TCP/IP protocols inherent in the functioning of the Internet.

The Internet is a packet switched network, implying that a node using the Internet does not have a dedicated part of the network exclusively to themselves. As a result, all the data flowing along an edge between two nodes is split up into packets and aggregated with all the data packets from other nodes on the network (Pastor-Satorras & Vespignani, 2007). Thus, during this transmission stage, the original data that is sent is vulnerable as it is only fully compiled when it reaches the specified, destination IP address. In this instance, individual countries (nodes) act as the basis of the backbone network since they facilitate global control in the information exchange system by transporting the data packets along regional and international routes (edges). Therefore, the backbone of the internet constitutes of two entity types, nodes and edges, with the bandwidth data being sent along the edges between the nodes.

Unlike most empirical studies which focus on the autonomous systems (AS) layer of the internet, this thesis uses graph models to study Internet robustness at the backbone level. The analysis takes place on the Internet backbone layer as this forms the core of the Internet, since it provides the shortest paths through high-capacity links and routers that are owned and operated by major Internet service providers. Moreover, given that cyberattacks are increasingly borderless, cybersecurity remains a transnational issue due to the increasing in3terconnection and correlated infrastructures between nations, and thus the degree of intactness of the international Internet backbone infrastructure determines how robust the virtual Internet network is.

With regards to robustness metrics, a combination of local and global metrics are used, where local metrics measure the individual node's impact on network robustness (its vulnerability to connectivity failures to the rest of the network), and global metrics capture features about the entire network graph. In this thesis both types of metrics are accounted for as local metrics often depend on a smaller information set, provide more detailed insight and need less calculation time, while global metrics are more meaningful for assessing the state of the entire network and also allow for the comparison of different graphs (Oehlers & Fabian, 2021). Robustness metrics are then further classified as static and dynamic. Static metrics take a snapshot of the key characteristics that influence the robustness of a network, while dynamic metrics assess network behaviour under arbitrary removal strategies (Oehlers & Fabian, 2021). In this regard, the thesis focuses on only static robustness measures.

Supporting the thesis objectives, the results from the robustness analysis show that some metrics differ when identifying the most robust countries in networks, and so using only one metric is not sufficient to measure the network robustness, which is also supported by findings from Oehlers & Fabian (2021). Therefore, this is the reason that a set of varying metrics is considered to calculate the robustness scores and used to compare the results. Essentially the objectives support the results in that certain countries with higher robustness scores and higher cybersecurity ratings have a greater contribution to the overall robustness of the Internet backbone when compared to regions which have poorer robustness metrics and a weaker cybersecurity status. This is attributable to more developed regions having high degree countries which in turn have better GCI ratings due to their influence as key nodes in their specific networks. And this is further supported by the fact that high degree countries also tend to have high Clustering Coefficient, Betweenness Centrality, and Eigenvector Centrality scores, all alluding to a more robust Internet backbone in that particular region.

The remainder of this thesis is structured as follows: Section 2 outlines the Related Literature, Section 3 explores the Methodology and Data, Section 4 presents the Results and Discussion, and Section 5 closes with the Conclusion.

2 Literature Review

An important design feature of the Internet is its robustness, which can be described as the ability of the network to provide and maintain an acceptable level of service in the face of various faults and challenges to its normal operation. Several graph models have been used to imitate the Internet structure, the most popular being the classical network modelling approaches such as the Erdos and Renyi (ER) graph model (Erdös & Renyi, 1959) and the scale-free BA model developed by Barabsi and Albert (Newman, 2003).

In contrast, this thesis uses Oehlers & Fabian (2021) as a starting point for studying network robustness, given that they provide a comparative overview of an extensive set of robustness metrics in six major categories, and discuss the respective advantages and drawbacks of each metric, as well as the main results obtained from analysing the robustness of theoretical Internet graphs. Furthermore, they compare and assess the suitability of each metric in detecting crucial backbone structures and measure important robustness aspects of Internet topology. However, unlike Oehlers & Fabian (2021), this research applies the robustness metrics to real world Internet data. A similar approach to this is seen by Rueda, Calle & Marzo (2017), who conduct a robustness analysis of fifteen real telecommunication networks under multiple failure scenarios, and by expanding on the taxonomy of structural and centrality robustness metrics, they analyse the common topological proprieties that group networks with similar robustness behaviour. The core of this thesis parallels Rueda, Calle & Marzo (2017) by using network analysis and graph theory which are a set of research procedures for identifying specific structures within interconnected systems, based on relationships among various components of the systems.

Further supporting the empirical literature on this topic are Baumann & Fabian (2015a), who use a large integrated dataset describing the Internet as a complex graph and develop a multi-dimensional Connectivity Risk Score that acts as a topological indicator of single Autonomous Systems (AS). They essentially develop an analytical scoring method that helps risk managers assess the potential vulnerability of an organization beyond its network perimeters. Comparably, Baumann & Fabian (2015b) investigate Internet robustness at the AS-level, and they too use empirical data and graph analysis to develop a graph model of the Internet in order to conduct a global assessment of Internet robustness. Yan et al. (2010) implement an AS path inference algorithm to derive inter-domain paths that are used on the real Internet, and similar to this thesis, they also study the geographical topology from an Internet backbone perspective on a national level, except their paper focuses only on Internet traffic in the US by using the skitter dataset. Following form this, the mapping of the TeleGeography network data in this thesis follows a similar narrative to the *Internet Atlas* described in Durairajan et al. (2013), where the *Atlas* is a visual geographical representation of the physical Internet including nodes and edges connecting the nodes. While visualising the physical backbone layer of the Internet is not the focus in this thesis, the common ground of analysing the interconnection structure of the Internet has been the subject of a large number of studies over the past decade, which concentrate on router-level (CAIDA, 2017; Madhyastha et al., 2006), PoP-level (Shavitt & Shir, 2005; Spring et al., 2004) and AS-level (Mao et al., 2003; Zhang et al., 2005) Internet graphs. In addition and similar to Zhang et al. (2005), this thesis accumulates information from data over time, as Zhang et al. (2005) emphasize the importance of both collecting from data sources such as the TeleGeography and ITU registries, as well as accumulating the findings over time in order to obtain more complete connectivity information.

Graph-theoretical analysis studies the structural properties of a geographical network derived from the Internet backbone topology, including its Degree distribution, Clustering structure, and also its Betweenness and Eigenvector Centrality measures (Yan et al., 2010). These four centrality measures that are widely used in network analysis (Newman, 2008) are also used as the robustness metrics that form the basis of this thesis, which can then be categorised into four buckets – Adjacency, Clustering, Throughput and Spectral measures.

The Adjacency and Clustering categories build on each other and describe the general structure of graphs, hence why the Degree and Clustering Coefficient measures, Equations 1 and 2 respectively in the Methodology section, are considered to be structural metrics (Oehlers & Fabian, 2021; Rueda, Calle & Marzo, 2017). Degree refers to how connected a node is, namely how many links a node has with other nodes: the degree of node i in a given network G is defined by the number of edges connected to it. The higher the degree, the more central the node is, and this can be an effective measure since nodes with high degrees may have high centrality by other measures (Hansen, Shneiderman & Smith, 2011). A node degree of 15 indicates that 15 edges are connected to the node in the Internet network, however one pitfall of the degree metric is that it only indicates how many connections a node has, and not the quality of the connection. The Degree measure is used in most of the classical results on Internet robustness as it is the simplest measure of nodal centrality and is determined by the number of neighbours connected to a node. For example, the removal of a node with the highest degree could cause large damage to the network: the larger the degree, the more important the node is, the more difficult it is to disconnect the graph, and hence the higher the robustness of the network (Oehlers & Fabian, 2021). However, if a node with a high degree fails, potentially higher numbers of connections are also prone to being affected. Therefore, the degree of a node has influence on network robustness beyond only assessing the connectivity of a node to its direct neighbours, as it is rather important for capturing the topological connectivity of the Internet since the Internet cannot be seen as an isolated unit but is interconnected with a massive, interrelated network structure (Pastor-Satorras & Vespignani, 2007). Furthermore, relying on Degree alone for estimating the importance of a node could be misleading, hence additional measures are needed.

Continuing with the definitions of the key variables, the Clustering Coefficient measures the probability that two nodes having a common neighbour are neighbours themselves (Hansen, Shneiderman & Smith, 2011). For instance, a Clustering Coefficient of 1 indicates that all of the possible connections between neighbouring nodes are actually realized in the network. The Clustering Coefficient essentially refers to the extent of overlap between connections, where the proportion of a node's neighbours are connected with each other, and the higher the Clustering Coefficient, the more triangles there are (since triangle count in a network is an important property to characterize and analyse network graphs) and hence the higher the robustness of the network (Oehlers & Fabian, 2021). The clustering of an undirected graph can be quantitatively measured by means of the Clustering Coefficient (Watts & Strogatz, 1998). From Equation 2, a high c_i (Clustering Coefficient) denotes a node whose removal will impact the network's robustness, as the many potential routes between its neighbours will cease to exist. As mentioned above, this is because the Clustering Coefficient captures the presence of triangles formed by a set of three nodes, and compares the number of possible triangles that can be formed to the number of actual triangles formed in the network (Rueda, Calle & Marzo, 2017). Hence the higher the Clustering Coefficient, the more triangles there are, and the higher the robustness of the network.

Next, the Throughput category focuses on concepts that are crucial for communication networks, such as approximating the concrete Internet routing processes via shortest-paths and accounting for Internet-specific link capacity restrictions (Oehlers & Fabian, 2021). A centrality measure like the Betweenness metric, given by Equation 3 in the Methodology, is calculated based on the whole network structure, meaning that changing arbitrary nodes in the network will have an influence on the robustness of the entire network (Singer, 2006). Betweenness Centrality refers to how important a node is in terms of connecting other nodes, and essentially measures a node's ability to pass information from one node to another node on a different part of the network, by calculating the number of shortest paths passing through a certain node and roughly approximating the potential traffic flow through that node (Du, 2019). Nodes with a high degree tend to have large betweenness scores since they are important for routing traffic, hence these nodes are important as they form the centre of the network and provide short routes through the entire network for other nodes (Singer, 2006). Furthermore, a node may have a high Betweenness Centrality while being connected to only a small number of other nodes, which are not necessarily central themselves. This is due to the fact that nodes acting as bridges between other nodes typically have a high value, so these nodes play a key role in the network and are important in information diffusion (Du, 2019). Given this measure detects the amount of influence a node has over the flow of information in a network, a node with a high betweenness score has a high probability to occur on a randomly chosen shortest path between two randomly chosen nodes (Rueda, Calle & Marzo, 2017; Singer, 2006). As a result, Betweenness Centrality has a natural connection to graph robustness since it accounts for the network effects of how a node's actions may impact the actions of its neighbours by measuring the traffic flow among other nodes in the network: the greater the information flow, the more central the node, the higher the betweenness score, and the more robust the network (Oehlers & Fabian, 2021).

Consequently, some aspects of these three classes mentioned above are combined together with the Spectral methods which include more sophisticated matrix-calculation schemes such as random walks (Du, 2019; Oehlers & Fabian, 2021). As laid out in Equations 4 and 5, Eigenvector Centrality is proportional to the sum of the centrality scores of its neighbours, where the centrality corresponds to the largest eigenvector of the adjacency matrix, thus, it can take a large value either due to the node being connected to many other nodes or due to it being connected to a small number of important nodes (Rueda, Calle & Marzo, 2017). Interestingly, Choi, Barnett & Chon (2006) find there is a high positive correlation between Eigenvector Centrality and Degree, as well as with the shortest path traffic loads, essentially known as Betweenness Centrality. This is because there is an inherent circularity in the calculation of Eigenvector Centrality scores, and so a node that is connected to more central nodes has its own centrality boosted, which indicates a more robust network. In this thesis, these strong, positive correlations are evident in the results.

Overall, this thesis differs from previous work in its objectives in that it investigates Internet robustness at the backbone level, and over and above the network robustness analysis conducted in this thesis, it incorporates a cybersecurity element of ranking countries and regions based on their cybersecurity status. Shafqat & Masood (2016) focus on a related concept, except the authors only analyse and compare national cybersecurity strategies (based on legal, operational, technical, and policy-related measures) for 20 countries. Their research specifies and recommends best practices for improving the state of national cybersecurity and resilience, by evaluating the different cybersecurity trends, measures and approaches outlined in the respective publicly available strategy documents, and while the same factors as Shafqat & Masood (2016) are used to rank the cybersecurity profiles, 38 countries are analysed in this thesis. In a similar manner to Shafqat & Masood (2016), Yarovenk et al. (2020) discuss the formation of effective cybersecurity strategies on a country level where these strategies cover areas such as cybersecurity policy making, developing appropriate legal frameworks and powerful technical systems, and investing in cybersecurity research and educational programs. These areas are also reflected in the Global Cybersecurity Index (GCI) scores which is the basis on which this thesis ranks the cybersecurity status of countries. This final variable in the analysis, the GCI, is a composite index of indicators that monitor the level of cybersecurity commitment in the five pillars (Legal, Technical, Organizational, Capacity Building and Cooperative) of the Global Cybersecurity Agenda (International Telecommunication Union (ITU), 2015, 2017, 2018, 2020). The GCI score provides a country specific ranking of the following: The type, level, and evolution of cybersecurity commitments within countries and relative to other countries; the progress in cybersecurity commitment of countries from a global perspective; the progress in cybersecurity commitment from a regional perspective; and the cybersecurity commitment divide (the difference between countries in terms of their level of engagement in cybersecurity initiatives). In developing a national cybersecurity strategy, Yarovenk et al. (2020) argue it is important to understand which aspects of a country's cybersecurity need to be strengthened, and which already have a strong basis and require sustained support. They state this can be assessed by determining the ranking of countries which is dependent on the level of cybersecurity, and this is exactly what is done in this thesis. There are several sources used to rank countries, including the National Cyber Security Index (NCSI) quoted in Yarovenk et al. (2020) and the Global Cybersecurity Index (ITU, 2015, 2017, 2018, 2020) used in this thesis, which assess the level of readiness of countries to counter cyber threats. Both of these sources make use of the same indicators – Legal, Organizational, Technical, and Educational – that relate to the various aspects of cybersecurity, and after combining the points across these indicators, an average score is calculated which is used to rank the country’s overall rating. As is stipulated in Yarovenk et al. (2020), Kshetri and Murugesan (2013) and additionally in this thesis, a country’s cybersecurity ranking serves as a means to assess its national and global impact and to highlight key elements of its national cybersecurity strategy.

Ultimately, to develop a complete and empirical analysis of Internet robustness, this thesis combines the quantitative elements of graph theory and network analysis as well as the qualitative factors relating to the cybersecurity status of countries.

The current limitations and drawbacks of the metrics used in this thesis are highlighted by Yan et al. (2010) in that although theoretically appealing and appropriate for the application to Internet backbone analysis, the graph-theoretical analysis ignores the hierarchical routing scheme of the real Internet and due to its simplicity only uses the shortest path routing scheme. Instead, Yan et al. (2010) apply route-based analysis which models realistic inter-domain and intra-domain routing schemes used on the Internet and then identifies those facilities that appear most frequently on paths in the Internet backbone topology, in addition to using traffic-based analysis which weighs each path by its respective traffic demand. Future work in this regard could make use of consequence-based analysis to evaluate the importance of each node by measuring the amount of traffic lost after it is removed from the topology.

With regards to the limitations of the cybersecurity rankings, while determining the ranking of countries by their level of cybersecurity is quite relevant and helps to obtain an adequate assessment of the country in terms of its ability to withstand cyberattacks, Yarovenk et al. (2020) confirm this approach used in this thesis does not take into account the importance of other indicators in forming an overall rating. This generalized approach does not respond to cases where there are different amplitudes of values and does not consider the use of additional characteristics that would better identify deviations from actual scores. However, the use of multi-attribute decision-making methods allows to solve several problems relating to the dimensionality of data and the determination of indicator weights, essentially accounting for the diversity of values of indicators and their fundamental differences. Therefore, the use of different approaches such as multicriteria analysis of decisions would cater for a more detailed assessment of cybersecurity ratings since they eliminate the above-mentioned shortcomings.

3 Methodology

3.1 Data

3.1.1 Data Overview

In this section, the two types of data used in this thesis will be described in depth, with the first being Internet geography network data, and the second being global cybersecurity ratings. This section will also cover the limitations and introduce the descriptive statistics for both data types.

3.1.2 Data Description

A) Network Data:

The thesis uses Global Internet Geography data from TeleGeography (TeleGeography, 2020), which offers a complete source of data on international Internet capacity. The original data are broken into two sections: country-by-country international Internet capacity and international Internet traffic. However, this analysis uses only Internet traffic data as opposed to Internet capacity data, because International Internet bandwidth (capacity) represents point-to-point rather than end-to-end relationships and shows only the routes available between those points, whereas international Internet traffic (capacity usage) travels between any two points anywhere on the network. Consequently, the Internet traffic data represents the network topologies of several hundred Internet providers operating international Internet links such as routers or switches that directly connect across an international border.

These links comprise the public Internet, which carries general Internet traffic travelling in both directions between the countries of each route. Ultimately, the bidirectional averages of both average and peak traffic over a link are used as the measures of traffic between two nodes. Average traffic is defined as the sum of all traffic across a link divided by the number of seconds in the month, and peak traffic is measured at the 95th percentile, which is calculated by dividing one month’s traffic into 5-minute increments, ranking the traffic levels of each increment, and removing the top 5 percent. More notably, the link enabling traffic to flow between nodes is the basis for a common metric – link utilization – used to measure network performance, as it quantifies how much of a given link’s capacity is used over a given period of time. Link utilization is an important measure, where over a period of years it can be indicative of actual Internet traffic growth between two nodes. In essence, the link utilization rate, or weight, of a link is the traffic level divided by the capacity of the link, and the greater the weight of a link, the more influential that route is in the network. Overall, this Internet traffic data is applied to a network topology in order to map the Internet as a network of interconnected countries and is subsequently analysed to determine its overall robustness.

In summary, the TeleGeography data is structured such that 10 routes for each region (Africa, Asia, Europe, Latin America, Middle East, North America and Oceania) are reported, where these 10 routes are kept constant year on year with only the level of international Internet traffic between the routes changing. This network data is reported in its raw and tabular format for each of the seven regions used in this thesis. The data is composed of countries (nodes) and the route between two countries (edges), with the Internet traffic flowing between edges measured in Gbps. Each node has been assigned an attribute – the node's respective cybersecurity status given by the International Telecommunication Union GCI data, and each edge has been assigned an attribute – the edge's respective weight of Internet traffic given by the proportion of bandwidth sent over that route. This weight is calculated as a simple fraction of an edge's traffic relative to the entire network's traffic. The network data of each region is then aggregated to make up an international dataset spanning the five years 2016 – 2020, to produce a total of 38 nodes and 61 edges.

B) Cybersecurity Data:

In addition to the network data, there is data representing each node's cybersecurity defence resources. The International Telecommunication Union (ITU) publishes a Global Cybersecurity Index (GCI) that ranks countries based on their cybersecurity strategies. The GCI measure presents factual representations of each country's level of cybersecurity, as the GCI is a composite index combining 25 indicators into one benchmark. This ranking applies to 193 ITU Member States in all regions: Africa, Americas, Arab States, Asia-Pacific, and Europe. The index aims to quantify the type, level, and evolution of cybersecurity policies in countries and relative to other countries, as well as the progress in cybersecurity methodologies of all countries from a global perspective. Furthermore, the index accounts for progress in cybersecurity strategies from a regional perspective, and accounts for the difference between countries in terms of their level of engagement in cybersecurity initiatives.

The following five designated areas of the Global Cybersecurity Agenda form the basis of the 25 indicators that the GCI statistic is comprised of.

First off, legal measures are based on the existence of legal institutions and frameworks dealing with cybersecurity and cybercrime. The legal framework sets the minimum foundation of compliance on which further cybersecurity capabilities can be developed. The key objective is to have sufficient legislation in place to synchronize compliance on a national and global level, in order to reduce cybercrime.

Second, the technical approaches are evaluated on the existence of technical elements and practical mechanisms dealing with cybersecurity. Technology is the primary tool against combatting cyberattacks, as without suitable technical skills countries are vulnerable. Therefore, for effective ICT development and utilization to hold, countries need to ensure that minimum security criteria are built for software applications and systems.

Thirdly, organizational factors are based on the existence of policy coordination institutions and strategies for cybersecurity development at the national level. These factors involve agencies both setting strategic cybersecurity targets, as well as evaluating the outcomes. The organization of guidelines is important to ensure that the various industries are aligned with national cybersecurity objectives.

Fourth is capacity building, where these aspects are evaluated on the existence of research and development, education and training programs, and certified professionals and public sector agencies promoting the legal, technical, and organizational pillars. While cybersecurity is primarily resolved via technical measures, there are socio-economic and political implications which require human and institutional capacity building to generate systematic solutions.

Lastly, cooperation includes measures dependent on the existence of partnerships, collaborative frameworks and information sharing networks. Given cybercrime is a global problem, a multi-stakeholder approach is necessary for increased cooperation which gives rise to greater resources for building a better defence network, ultimately deterring a higher amount of online attacks.

These five pillars essentially shape the inherent building blocks of a national cybersecurity culture, and as a result, the GCI rankings can be used to describe each node's cybersecurity status, with the added advantage that the GCI measure acts as a proxy for a node's vulnerability for all the countries that the TeleGeography data represents.

The cybersecurity GCI statistic is bound between 0 and 1, where a score closer to 0 indicates a country has a poor cybersecurity status, while a score closer to 1 indicates a country has a sound cybersecurity status. Each of the four versions (2015, 2017, 2018, 2020) of the cybersecurity data has GCI scores for all the 38 TeleGeography countries, and hence the rankings are averaged over the four years to form a mean GCI Score. However, data is unavailable for years 2016 and 2019, hence GCI scores from the previous iterations (2015 and 2018, respectively) are imputed for these two missing years. Each iteration of the reports used in this thesis pinpoint the exact score and rank for all 38 countries.

3.1.3 Data Limitations:

This section highlights the several data limitations in both the Internet traffic data and in the cybersecurity data.

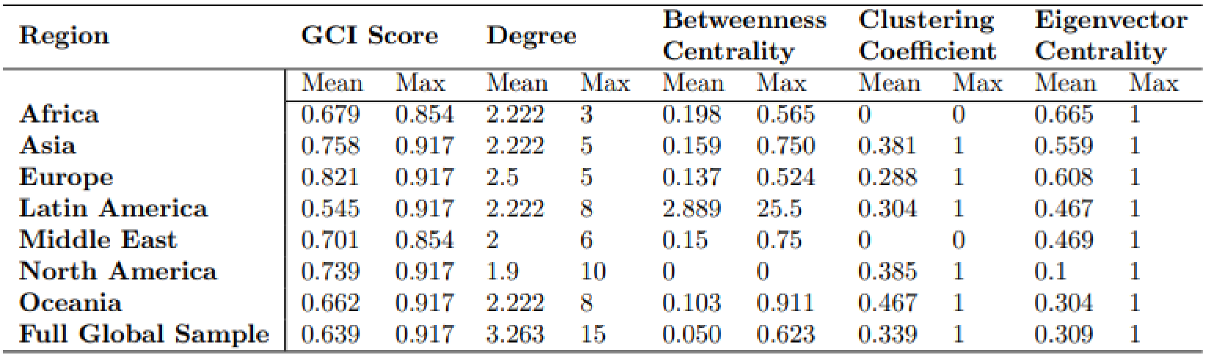
With regards to the TeleGeography dataset, the first data limitation is that the Global Internet Geography data includes international Internet bandwidth statistics for 92 countries, however the international Internet traffic statistics are available for only 38 of the 92 countries. Therefore, the full Internet traffic dataset used in the thesis is for 38 countries, aggregated over the time period from 2016 to 2020. Secondly, on routes that have imbalanced traffic flows, the data underreport the true average and peak demands placed on the link, since it examines the average traffic flow from both directions. The Internet traffic data reflects statistics obtained directly from service providers, and although the participating providers operate in a variety of regions and represent a large portion of all international Internet capacity, the data do not include statistics from all of the hundreds of international Internet operators in the world. To compensate for this incomplete dataset, TeleGeography developed estimates using proxies on backbone deployment and capacity utilization trends, including the carrier type and route type.

Regarding the data limitations in the cybersecurity data, the ITU has published only four iterations of the GCI report (2015, 2017, 2018 and 2020), and as a result two years of data, namely 2016 and 2019, are missing when compared to the TeleGeography data. Hence for the years that do not have corresponding GCI data, the previous year's data will be used.

3.1.4 Descriptive Statistics:

The Descriptive Statistics section provides an overview of the topological properties of each of the seven regional networks, as well as for the fully complete international network. This section serves as an introduction to and summary of the network and cybersecurity data by using specific methods such as the mean, which measures the centre or central tendency of a set of data, and the max, to calculate and summarize the data. The figures in Table 1 are based on the TeleGeography and International Telecommunication Union datasets, which were manually cleaned, and the resulting data was then transformed into the summary statistics all using Excel. Since the Descriptive Statistics are reported numerically, descriptions are given below to provide meaningful insight on the data. This section differs from the Results and Discussion section in that the Descriptive Statistics merely introduces a high-level overview of the robustness metrics and the cybersecurity GCI scores. Furthermore, the statistics reported in this section do not pertain to individual countries but to overall regions, and hence a brief evaluation of the overall networks is given, where a more in-depth analysis of the individual country nodes takes place in the Results and Discussion.

Summary Statistics of Static Robustness Measures:



Region	GCI Score		Degree		Betweenness Centrality		Clustering Coefficient		Eigenvector Centrality	
	Mean	Max	Mean	Max	Mean	Max	Mean	Max	Mean	Max
Africa	0.679	0.854	2.222	3	0.198	0.565	0	0	0.665	1
Asia	0.758	0.917	2.222	5	0.159	0.750	0.381	1	0.559	1
Europe	0.821	0.917	2.5	5	0.137	0.524	0.288	1	0.608	1
Latin America	0.545	0.917	2.222	8	2.889	25.5	0.304	1	0.467	1
Middle East	0.701	0.854	2	6	0.15	0.75	0	0	0.469	1
North America	0.739	0.917	1.9	10	0	0	0.385	1	0.1	1
Oceania	0.662	0.917	2.222	8	0.103	0.911	0.467	1	0.304	1
Full Global Sample	0.639	0.917	3.263	15	0.050	0.623	0.339	1	0.309	1

Table 1: The above Descriptive Statistics table provides the average and highest values for the topological properties of each of the seven regional networks, and the fully complete international network.

At the upper level of Internet robustness, the more developed regions dominate the Degree metric, however the less developed regions dominate the Betweenness Centrality measure. Based on Table 1 above, the European sample has the highest average number of links with other countries (2.5) but has the second lowest value of absolute links (5). The North American sample has the highest maximum number of links with other countries (10) but the lowest average links (1.9). This indicates that regional connectivity among countries in Europe is more evenly distributed compared to North America, where in that region the concentration of links occurs mainly in the United States. Interestingly, the Latin American region has the highest Betweenness Centrality score (2.889), indicating it has the most robust regional Internet network, with Africa (0.198), Asia (0.159) and the Middle East (0.15) also at the upper tail of the robustness scale. At the lower end of the Degree measure the developing regions prove to be less robust than the developed regions, although with regards to the Betweenness metric, the developed regions are less robust than the developing regions in this robustness measure. From the Descriptive Statistics in Table 1, Africa, Asia, Latin America, and Oceania all have the same number of average links among their respective countries (2.222), where these figures were calculated using the built-in "Degree" algorithms in Cytoscape and Gephi. The reason why all these regional networks have the same average number of connections is because the African, Asian, Latin American, and Oceanic networks all have 9 nodes and 10 edges, hence averaging to 2.222 across all these four regions. However, Africa and Asia are the least well connected, with a maximum of 3 and 5 links respectively. Table 1 also shows that the North American region has the lowest Betweenness Centrality score (0), indicating it has the least robust regional Internet network in this regard. Oceania and Europe are slightly more robust than North America, but are still the second and third least robust Internet networks in terms of Betweenness Centrality given their scores are 0.103 and 0.137, respectively.

Therefore, with regards to the Degree metric, there is a pattern among the more developed regions having a higher Internet robustness compared to the lesser developed regions. While the Betweenness Centrality measure shows an opposite and unexpected pattern which is also reported in (Singer, 2006), as seen by the less developed Latin American and African regions having a more robust Internet backbone network.

Given that a higher Clustering Coefficient implies higher robustness, the Descriptive Statistics table (Table 1) shows that on average, countries in the Oceanic region overlap the most compared to countries from any other region, with countries in North America and Asia following suit. The African and Middle Eastern regions have the lowest Clustering Coefficients and are the only regions that do not have a maximum Clustering Coefficient of 1. Thus, no countries in Africa or the Middle East have a complete connectivity overlap with their neighbouring countries. The Clustering metric follows the same pattern as the Degree metric, where the more developed regions showcase better network robustness compared to lesser developed regions. However, from a global perspective the Clustering Coefficient is 0.339, indicating the global Internet backbone cannot be considered well connected, as only 33.9% of all possible connections among countries are realized.

Following from this, the Eigenvector Centrality in Table 1 displays interesting insights. Africa has the highest average Eigenvector Centrality measure (0.665), suggesting that countries in this region are well-connected to other countries from other regions, which are well-connected themselves. This is a surprising finding as one would expect the more developed regions such as Europe, North America and Oceania to have the most influential connections with countries in other regions. However, while Europe ranks second (0.608) in terms of Eigenvector Centrality, the opposite is true for North America and Oceania, which are both at the tail end of the ranking, scoring the lowest (0.1) and second lowest (0.304), respectively. Just as the more developed regions tend to be more robust in terms of the Degree and Clustering Coefficient metrics, the less developed regions tend to be more robust in terms of the Betweenness Centrality and Eigenvector Centrality measures.

The overall pattern of cybersecurity levels for each region can be inferred from the Descriptive Statistics table in Table 1. The European region has the highest cybersecurity rating (0.821), with the Asian and North American regions trailing in second (0.758) and third (0.739) rank respectively. Latin America has the lowest GCI ranking (0.545) with Africa at the third lowest rank (0.679). Therefore, the more developed regions score higher in terms of cybersecurity ranking relative to the less developed regions. Generally, each of the regions, excluding Latin America, score a higher rank when compared to the GCI cybersecurity ranking of the combined global network.

To summarize the underlying trend in the descriptive statistics, North America, Europe and Oceania tend to have more robust Internet networks when the Degree and Clustering Coefficient metrics are considered, as well as higher cybersecurity rankings; while Latin America, Africa and the Middle East have higher Internet robustness when the Betweenness and Eigenvector Centrality metrics are considered, however these regions have lower cybersecurity rankings. The Asian region displays high Internet robustness in both the Clustering Coefficient and Betweenness Centrality measures, and also scores the highest out of all the regions in terms of cybersecurity status. Ultimately, the more developed regions fare well in terms of cybersecurity status and the structural robustness measures (Degree and Clustering Coefficient) compared to the lesser developed regions, while the developing regions fare well in the centrality robustness measures (Betweenness Centrality and Eigenvector Centrality) but under-perform in cybersecurity status.

3.2 Robustness Metrics

The Methodology section expands on the Robustness Metrics that are used on the TeleGeography network data, where all metrics were calculated with the help of graph analysis software Cytoscape and Gephi.

The resulting networks focus on four key node measures, of which half are structurally based - Degree and Clustering Coefficient - and the other half are centrality based - Betweenness Centrality and Eigenvector Centrality. These four structural and central measures are then also applied to evaluate the static robustness of the regional networks, where the metrics analyse the global Internet topology, and so measure the topological robustness of networks which are modelled by undirected, weighted graphs. In fact, the analysis combines these several metrics to take advantage of their multiple benefits which outweigh their combined disadvantages. Furthermore, each of the four Robustness Metrics (Degree, Betweenness Centrality, Clustering Coefficient and Eigenvector Centrality) pertain to broader robustness classes, namely Adjacency, Throughput, Clustering and Spectral, respectively.

By taking these several connectivity-based aspects into account, the analysis is considered multidimensional, which helps to cope with the high complexity of the Internet graphs. The Robustness Metrics used in this thesis are then all applied to the TeleGeography network dataset, where the original data was already reported per regional network. Since the data was initially split into the seven regions, it only required to be converted into seven separate CSV files to fit the formatting for importing into Cytoscape, where Cytoscape has a built-in application called NetworkAnalyzer that for every node in a network computes its Degree, its Clustering Coefficient, and a variety of other parameters such as Betweenness Centrality. The seven network files were loaded into Cytoscape and the built-in formulae for calculating Degree, Clustering Coefficient, Betweenness Centrality and Eigenvector Centrality returned the values for every individual country in each of the seven regional networks. Hence the below metrics were applied to the existing datasets via the built-in Cytoscape NetworkAnalyzer formulae (Max Planck Institute for Informatics, 2018), which are represented by the equations that follow. The results were then collected and reported in the Results and Discussion section that follows the detailed metric equations below.

A) Adjacency:

The assessment of node adjacency is one of the first and easiest approaches for investigating network robustness. The intuition is that a node with many edges, a high-degree node, could be more important to the overall graph structure than a low-degree node. The average node Degree is the first indicator of overall network robustness, and of resilience against high-degree attacks.

The Degree k_i of node i is defined as the number of edges in the graph incident on the node i:

$$k_i = \sum_{j \in V} a_{ij} = \sum_{j \in V} a_{ji} \tag{1}$$

For an undirected graph, with a symmetric adjacency matrix, $k_{in,i} = k_{out,i} \equiv k_i$: The in-degree of node *i* is the number of incoming edges and the out-degree is the number of outgoing edges.

This Degree metric was applied to the existing data via Cytoscape, where the built-in NetworkAnalyzer calculates the Degree of node *i* as being the number of edges linked to node *i*, where this degree distribution gives the number of edges, or degree '*k*' for *k* = 0,1,…,*n* (Diestel, 2017; Max Planck Institute for Informatics, 2018).

B) Clustering:

Clustering metrics aim to provide a detailed overview of the structure of the Internet in order to gain an understanding of its dynamic organization. The clustering of a graph refers to the tendency observed in many networks that form cliques based on specific nodes in the neighbourhood of a given network. Clustering implies the property that if node i is connected to node j, and at the same time j is connected to l, then there is some probability i is also connected to l.

The Clustering Coefficient is commonly applied to measure the interconnectedness of nodes with the same neighbour. Given node i, with degree k_i, and e_i as the number of edges existing between the k_i neighbors of i, the Clustering Coefficient, c_i, of node i, is defined as the ratio between the actual number of edges among its neighbors, e_i, and its maximum possible value, $k_i(k_i - 1)/2$:

$$c_i = \frac{2c_i}{k_i(k_i - 1)} \tag{2}$$

Thus, the Clustering Coefficient c_i measures the average probability that any two neighbors of node i are also connected to each other. Cytoscape's NetworkAnalyzer applies the Clustering Coefficient metric to the TeleGeography data by computing the ratio $\frac{N}{M}$ for each network, where *N* is the actual number of edges between the neighbours of node *i*, and *M* is the maximum number of edges that could possibly exist between the neighbours of *i* (Max Planck Institute for Informatics, 2018). Ultimately, the clustering coefficient of a node always ranges between 0 and 1, with a value closer to 1 indicating high clustering (Watts & Strogatz, 1998; Barabási & Oltvai, 2004).

C) Throughput:

Throughput measures consider the fact that the capacity for information forwarding is limited, and these measures also calculate the maximum workload of network entities, or their likelihood to become overloaded when other entities fail. Evaluating path redundancy and distance increases is not enough for assessing network robustness, as traffic loads induced by flows along shortest paths and their capacity constraints must also be considered.

To go from one node to another in the graph, following the shortest path, a certain sequence of nodes is visited. By counting all the nodes visited by the shortest paths between all the possible pairs of nodes in the graph, some key nodes will be visited more often than others. This fact can be quantitatively measured by the Betweenness, b_i, of node i, defined as the total number of shortest paths between any two nodes in the graph that pass through node i. More precisely, if $L_{h,j}$, is the total number of shortest paths from h to j and $L_{h,i,j}$ is the number of these shortest paths that pass through node i, the Betweenness is defined as:

$$b_i = \sum L_{h,i,j} \;/\; L_{h,j} \tag{3}$$

Where the sum is taken over all h, j pairs with $j \neq h$.

Cytoscape applies the Betweenness Centrality metric to the existing Internet geography data by ensuring that the betweenness value for each node *i* is normalized by dividing by the number of node pairs excluding *i*: $\frac{(N-1)(N-2)}{2}$, where *N* is the total number of nodes in the connected component that *i* belongs to (Max Planck Institute for Informatics, 2018). Thus, the Betweenness Centrality of each node is a number between 0 and 1. Furthermore, the NetworkAnalyzer uses the fast algorithm by Brandes (2001) for the computation of node Betweenness Centrality, where this algorithm has a complexity function $O(NM)$, with *N* being the number of nodes and *M* the number of edges in the network. Essentially the Betweenness metric counts the number of times a node acts as a bridge on the shortest path between two other nodes and it measures the total bandwidth flows which have to be redirected if node i fails, thus quantifying how much control i has over network traffic.

D) Spectral:

Spectral methods make use of random walks in their calculations, which provide a useful way to estimate node importance as their role in an alternative, non-shortest path metric.

The Eigenvector Centrality of node i is given by the i^{th} entry of the eigenvector of the adjacency matrix **A** corresponding to the largest eigenvalue:

$$\mathbf{A} \cdot u_v = \lambda_v u_v \tag{4}$$

$$u_v(i) \ = max\ t\frac{\sum_{j\in V_i} u_{v,t-1}\ (j)}{\sum_{k\in V} u_{v,t}(k)}, with\ u_{v,0}(i) = 1 \tag{5}$$

With adjacency matrix **A**, its derivatives and corresponding eigenpairs (λ_i, u_i), and paths between nodes i and j having length k.

The manner in how the NetworkAnalyzer in Cytoscape applies the Eigenvector Centrality metric to the existing dataset, is that initially all nodes have equal centrality values, and at each step t the value of each node is set to the sum of the values of its neighbours V_i. Then all values are normalized so their sum is 1 and the process is repeated until it converges. The vector entries reflect how many and how important the neighbors of a given node are, implicitly accounting for the whole network topology (Max Planck Institute for Informatics, 2018). Accordingly, Eigenvalue Centrality is based on the notion that a node should be classified as important if it is linked to other important nodes.

4 Results and Discussion

The results of the regional structural and centrality robustness metrics in a static scenario are presented alongside a regional cybersecurity comparison in this section. This group of metrics attempts to identify which countries in a network are the most central as these countries are better equipped to transport data in the network, and hence they are crucial in preventing the network from collapsing. These metrics also define the network centralization as a measure of how central the most central country is in relation to how central all the other countries are. Tables 2 to 8 show the structural and centrality robustness metrics and the cybersecurity status for the defined set of the regional Internet networks.

A) Degree:

Taking Degree, given by Equation 1 above, as the first structural robustness metric, each regional network has different countries with a differing level of connectivity, and a resulting pattern between certain regions is seen: the higher income regions (North America and Oceania), with the exception of the Middle East, are better connected than the lower income regions, Africa and Latin America.

Table 2 shows the African network has Algeria and Morocco each with a maximum Degree of 3 in the network, followed by the Latin American network in Table 5 with Argentina having the maximum of 3 connections in this respective network as well. Interestingly, the Asian network (Table 3) and European network (Table 4) are not at the upper end of the connectivity scale, as evidenced by Singapore and Germany each with the highest of only 5 connections in their respective regions.

Slightly better connected is Turkey representing the Middle Eastern network (Table 6) with a maximum Degree of 6. All of these regions are comparably less well connected than the Oceanic (Table 8) and North American (Table 7) networks, which feature Australia and the United States having the highest Degrees of 8 and 10 respectively, essentially being connected to the majority of countries in their regions.

From the GCI report, the above-mentioned countries that have a high connectivity (Turkey, Australia and the United States) all excel in four out of the five Global Cybersecurity Agenda pillars, whereas the countries with a lower connectivity (Algeria, Morocco and Argentina) excel in only one out of the five cybersecurity pillars (International Telecommunication Union, 2020).

When compared to the literature, Rueda, Calle & Marzo (2017) find that in many networks only a small number of nodes have high degrees, however in this analysis it is clear there are many regional Internet networks with high degree countries.

B) *Clustering Coefficient:*

In Tables 8, 7 and 3, the second structural robustness metric – the Clustering Coefficient (CC) – shows that Oceania, North America and Asia are the most robust countries in the network for this specific measure. The following countries, Australia, United States, Japan and Vietnam, all have a Clustering Coefficient of 1, meaning they realize 100% of all possible connections in their respective regional networks. In other words, these countries are the most interconnected with their neighbouring countries compared to countries in Africa and the Middle East, where the Clustering Coefficient is 0 for all individual nations in these regions. This is due to the slower pace of Internet diffusion in developing countries and Least Developed Countries (LDCs), which is related to their overall level of development (Creese, Dutton & Esteve-González, 2021).

Again, Europe falls between the extreme sides of regional connectivity, alongside Latin America, with only the Netherlands, Brazil and Chile featuring a Clustering Coefficient of 1 and the majority of the remaining countries in these regions with a CC of 0.

Ultimately, the reason behind a high Clustering Coefficient is that there are many alternative paths (triangles) when forming a link between countries and their connections, and with developed countries having more absolute connections (as seen by the Degree results above) as well as more wealth, it is expected that these nations would rank higher in this particular robustness metric. This result is supported by the findings of Creese, Dutton & Esteve-González (2021), who emphasize the impact of national development being anchored to regional Internet connectivity.

C) *Betweenness Centrality:*

With regards to the centrality-based metrics, Betweenness Centrality measures the network centralization and the importance of a vertex in a graph (Yan et al., 2010). As explained in the methodology, network centralization is used to analyse network robustness, which equates to the difference between the centrality of the most central node and that of all other nodes, which is reflected in Equation 3 above.

From a regional perspective, the results indicate that Latin America (Table 5), Africa (Table 2) and Asia (Table 3) tend to have the highest Betweenness scores, whereas North America, Oceania and Europe have the lowest Betweenness scores. This is surprising given that for the other structural robustness measures (Degree and Clustering Coefficient) Latin America and Africa are two of the worst performing regions, whereas North America and Oceania are the two best performing regions. However, upon closer inspection of the individual country values, Latin America is an outlying result due to the United States' pronounced influence in the region as seen by its extreme Betweenness Centrality score of 25.5. This is because bandwidth traffic between the United States and the other countries in the Latin American region flows from the United States to the other countries, and not from the other countries to the United States. Following from this, 78% and 90% of the respective countries in Oceania and North America have Betweenness scores of 0, which provides emphasis on how outlying results such as these may affect a network's robustness.

Accordingly, the results from the robustness analysis show that some metrics differ when identifying the most robust countries in networks, and so using only one metric is not sufficient to measure the network robustness (Oehlers & Fabian, 2021). Therefore, this is the reason that a set of varying metrics is considered to calculate the robustness and compare the results.

D) *Eigenvector Centrality:*

The final robustness metric, Eigenvector Centrality, depends on both the number of a country's neighbours as well as the quality, which in this case is measured by the cybersecurity level, of a country's connections.

The Asian, European, Middle Eastern and North American regions all feature countries (Singapore, Germany, Turkey and the United States) which have maximum Eigenvector scores of 1. While the countries in the remaining regions (Africa, Oceania and Latin America) do not score an Eigenvector value of 1, they score in the upper end of the range: Morocco with 0.939, New Zealand with 0.667 and Argentina with 0.626. Again, the same pattern of regions and countries are being repeated when it comes to scoring highly in the robustness metrics. Countries in the more developed North American, Asian and Oceanic regions tend to have higher scores in the robustness metrics compared to countries in the African and Latin American regions, while the European and Middle Eastern countries tend to be more consistent with having above-average robustness scores, falling in between the extreme values.

E) CGI Scores:

The concluding part of the results section involves analysing the Global Cybersecurity Index for all the countries in the seven regional networks. The International Telecommunication Union's GCI Reports (International Telecommunication Union (ITU, 2015, 2017, 2018, 2020) classify each individual country as developed or developing and outline the key cybersecurity areas in which the specific country needs to improve upon. Countries that rank lower in the Global Cybersecurity Index are more likely to be LDCs as they are more likely to face resource challenges in bridging their cyber-capacity gap, which includes a lack of institutional knowledge, policy limitations and skills shortages to protect their cybersecurity systems, both physically and virtually.

One of the various facets underlying the GCI ranking - Computer Incident Response Teams (CIRTs) - is considered a primary factor within the technical measures pillar, as it enables countries to respond to incidents at a national level, using a centralized contact point that promotes quick and systematic action. For a country to effectively deter targeted cyber threats and incidents, it is essential to have technical teams that efficiently disseminate threat information to the concerned authorities and provide cyber protection and resilience capabilities (Shafqat & Masood, 2016). Such teams include CIRTs, and while many countries have made progress in implementing CIRTs, LDCs especially face significant barriers in establishing CIRTs due to a lack of resources, technological knowledge, and prioritization of an integrated cybersecurity ecosystem (International Telecommunication Union, 2018 & 2020).

This is evident in the network results from Tables 2 to 8, where the five countries with the lowest GCI scores are Iraq (0.136), Papua New Guinea (0.145), Fiji (0.191), Algeria (0.302) and Ecuador (0.362). Unsurprisingly, Iraq, Algeria and Ecuador are classified as developing nations, Fiji as a LDC and Papua New Guinea as a Small Island Developing State (SIDS). Over half of LDCs do not have a national CIRT in place, and 60% of LDCs have yet to start the process of developing a National Cybersecurity Strategy (International Telecommunication Union, 2015 & 2017).

Contrarily, the five countries with the highest GCI scores are the United States (0.971), Singapore (0.871), Canada (0.870), Australia (0.863) and the United Kingdom (0.854), all of which are classified as developed nations by the GCI report. This result relating to regional cybersecurity status is further supported by the fact that Africa as a continent only has 19 countries with a national CIRT, whereas the North American region has 21 CIRTs and the European region has only six countries that lack national CIRTs (International Telecommunication Union, 2017 & 2020).

F) Correlations:

From the discussion above the key countries responsible for maintaining the structure of the Internet backbone network when considering robustness metrics and cybersecurity levels are: United States, Singapore, Australia, Germany, Japan, Canada, United Kingdom, Algeria, Morocco, Turkey, Argentina, Brazil, Chile, Vietnam, Netherlands and New Zealand. These countries feature the highest robustness graph theory metrics as well as the highest GCI ratings. Consequently, correlations among these specific countries were computed to determine their dependency and how sufficiently robust they are in the face of potential cyberattacks. Tables 9 to 14 reflect the findings of the correlations of the four robustness metrics (Degree, Betweenness Centrality, Clustering Coefficient, Eigenvector Centrality) and the GCI cybersecurity scores between these key countries.

First off, Table 9 features the correlations between the key countries in the Africa region with Algeria and Morocco having a strong positive correlation of 0.990 since both countries have a Degree of 3 in the region. Further contributing to this strong correlation is the fact that Morocco and Algeria have almost identical Eigenvector Centrality scores of 0.939 and 0.937 respectively, and that Africa is one of the highest scoring regions for the Betweenness Centrality metric, hence both Algeria and Morocco have relatively high Betweenness Centrality scores – 0.19 and 0.565. Even though Algeria scores as one of the lowest ranked countries in terms of GCI ratings (0.302), the results imply that the Degree, Betweenness Centrality and Eigenvector Centrality metrics are fundamental in maintaining the African segment of the Internet backbone.

Next is Table 10 which reflects the correlation scores among the nodes in the Asia region. Japan and Vietnam boast a very strong and positive correlation of 0.975 since both countries have a Clustering Coefficient of 1. Additionally, Singapore has an Eigenvector Centrality of 1 and Vietnam with 0.701, and Singapore and Japan both score highly for the GCI rating, which leads to high correlations of 0.814 and 0.799 respectively. It is evident these three countries have high values for the Clustering Coefficient and Eigenvector Centrality robustness metrics as well as high GCI values, which are fundamental to the robustness of the Internet, and so it is not surprising these key nodes are well connected and correlate highly with one another.

The European countries in Table 11 show very positive correlations among one another, with the United Kingdom and Germany almost being perfectly correlated at 0.995 with regards to the robustness and cybersecurity metrics. This can be explained by the fact that Germany has the highest number of links in Europe, 5, with the United Kingdom in second place with 4 links, as well as that Germany has an Eigenvector Centrality score of 1 with the United Kingdom scoring 0.924. The Netherlands also has a high Eigenvector score of 0.859 which leads to a correlation of 0.934 with Germany. Furthermore, the strong correlation between the Netherlands and the United Kingdom is given by a Clustering Coefficient of 1 for the Netherlands and 0.5 for the United Kingdom, and is supported by the United Kingdom ranking as the fifth highest GCI score of 0.854 with the Netherlands slightly below at 0.823. These three developed countries in Europe prove to score and rank highly across most of the graph theoretical and cybersecurity metrics which implies these nodes are important in contributing to the robustness of the global Internet backbone.

Table 12 highlights the most prominent countries in the Latin American regional network, namely Argentina, Brazil and Chile. Given both Brazil and Chile have a Clustering Coefficient of 1, there is a correlation of 0.991 between them. Argentina has 3 links and an Eigenvector Centrality score of 0.626, and Brazil consists of 2 links and an Eigenvector Centrality score of 0.506, resulting in a lower correlation than that of itself and Chile however still a very strong and positive 0.887 between itself and Argentina. The last pair in the Latin American segment of the Internet backbone is Argentina and Chile, where both countries have almost identical GCI scores of 0.451 and 0.477 respectively, and hence result in a strong correlation of 0.908.

With regards to the Middle Eastern region, Turkey is the only country that features in achieving any sort of robust scores as seen by a Degree of 6, and an Eigenvector Centrality score of 1, making it one of the most highly connected out of all the countries. The other nodes in the Middle Eastern network do not display any sort of robustness qualities; for instance Iraq has the lowest GCI score (0.136) of all 38 countries in the analysis and each country in the Middle Eastern network, including Turkey, has a

Clustering Coefficient of 0. Hence Turkey has no other robust countries in its regional network that it correlates with and is the sole key node in the Middle Eastern portion of the Internet backbone which is why there is no correlation table for the Middle East.

Table 13 displays the North American nodes that are key to maintaining a robust structure of the Internet backbone topology. Although the United States scores highly across every single robustness measure: a Degree of 10, a maximum score of 1 for the Clustering Coefficient, Betweenness Centrality and Eigenvector Centrality, and is at the top of the GCI rankings with a rating 0f 0.917, it interestingly has a low correlation of 0.657 with its key counterpart regional node, Canada. This is due to the fact that the United States is an outlier in achieving the highest scores for the robustness metrics, as Canada only fares well in being third highest in the cybersecurity rankings with a rating of 0.870, however only has 1 link, has a CC score of 0.316 and yields 0 for both the Betweenness and Eigenvector Centrality metrics. Nonetheless, these two countries still remain central to the robustness of the Internet in the North American region.

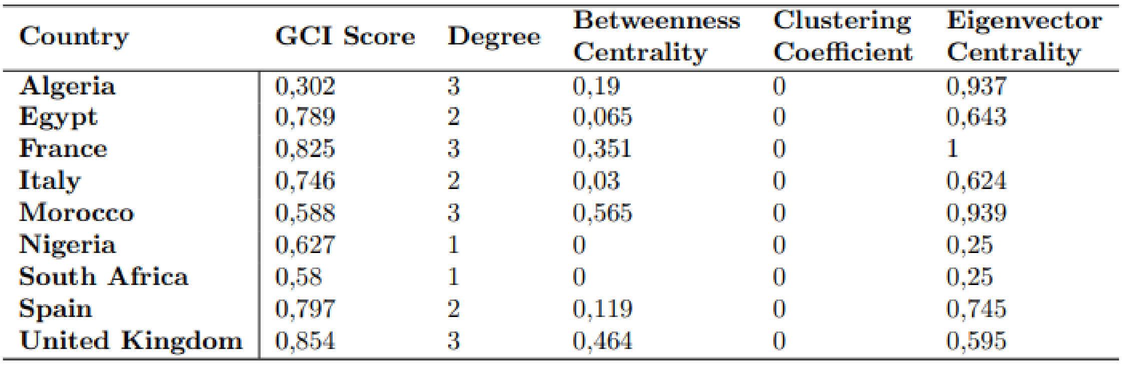
Lastly, the Oceanic country correlations are displayed in Table 14. The two robust nodes in the Oceanic region are Australia and New Zealand, with a positively strong correlation of 0.952. This is supported by the fact that Australia has a Degree of 8 and a Clustering Coefficient of 1, with New Zealand’s CC being 0.626. Furthermore, this high correlation is given by Australia ranking fourth highest in the GCI rating with New Zealand relatively close behind at 0.771, as well as the fact that these are the only two nodes in the Oceanic network that yield non-zero Betweenness Centrality scores. As a result, Australia and New Zealand are the two countries supporting the Oceanic portion of the global Internet backbone.

Overall, countries that do not have metrics for assessing their cybersecurity risk at the national level make it more difficult for themselves to assess current risks, prioritize cybersecurity interventions, and track progress, and ultimately, a weaker cybersecurity status corresponds to a weaker and more vulnerable Internet network. Given the objectives of this thesis are to investigate the importance of different nodes that govern the functioning of the Internet, the results support the proposed research objectives in that certain countries with higher robustness scores and higher cybersecurity ratings have a greater contribution to the overall robustness of the Internet backbone when compared to regions which have poorer robustness metrics and a weaker cybersecurity status. Ultimately, these key countries are part of more developed regions with high robustness metrics which in turn have better GCI ratings due to their influence as key nodes in their specific networks.

The results also tie in with the overall thesis contribution of a being a detailed analysis of country level impacts to the robustness of the Internet as a whole. Through analysing the criticality of the Internet backbone from a graph-theoretical perspective, the results make it evident that some metrics differ when identifying the most robust countries in networks, and that using only one metric is not sufficient to measure the network robustness. Therefore, this is the reason that a set of varying metrics is considered to calculate different robustness scores and used to compare the results. Furthermore, by using real-world Internet geography and global cybersecurity data this thesis contributes more than just a theoretical approach of using and implementing different robustness methodologies to the Internet backbone topology, but also fills an analysis gap in the study of networks by applying graph theory metrics to global Internet traffic and cybersecurity data.

Table of Results for All Regional Networks:

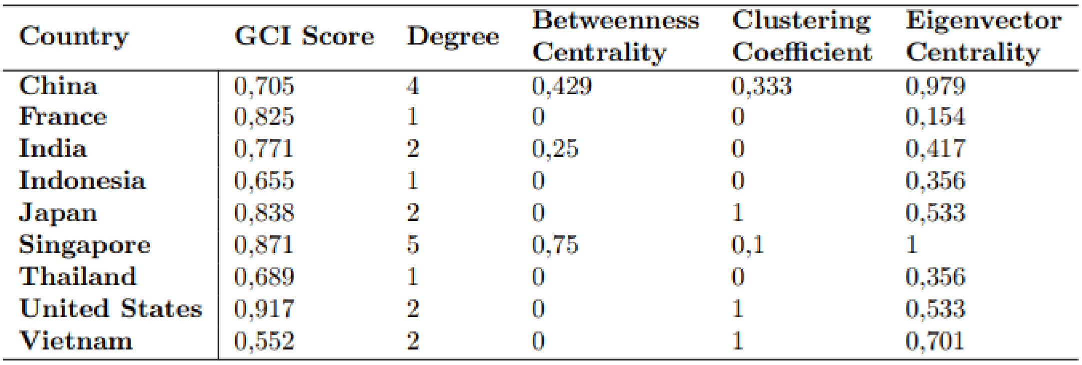
Table of Results showing the African Network Robustness Scores



Country	GCI Score	Degree	Betweenness Centrality	Clustering Coefficient	Eigenvector Centrality
Algeria	0,302	3	0,19	0	0,937
Egypt	0,789	2	0,065	0	0,643
France	0,825	3	0,351	0	1
Italy	0,746	2	0,03	0	0,624
Morocco	0,588	3	0,565	0	0,939
Nigeria	0,627	1	0	0	0,25
South Africa	0,58	1	0	0	0,25
Spain	0,797	2	0,119	0	0,745
United Kingdom	0,854	3	0,464	0	0,595

Table 2: The above results table highlights the robustness and cybersecurity scores for each country within the African network. Countries in the African network ranked at the top in the Betweenness Centrality and Eigenvector Centrality robustness metrics.

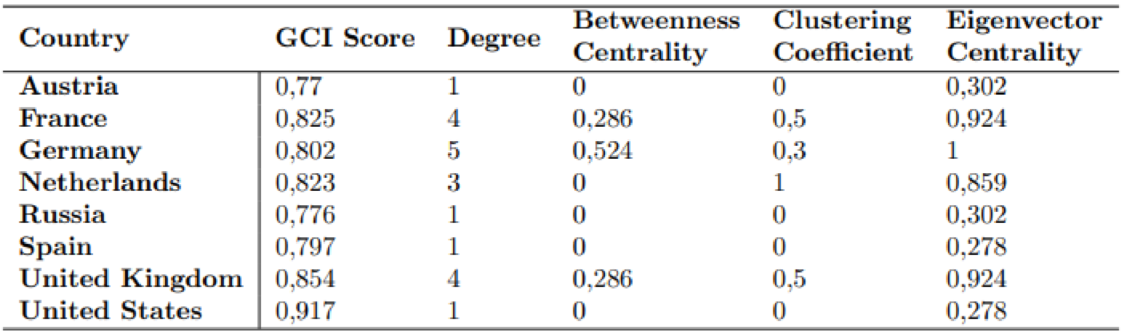
Table of Results showing the Asian Network Robustness Scores



Country	GCI Score	Degree	Betweenness Centrality	Clustering Coefficient	Eigenvector Centrality
China	0,705	4	0,429	0,333	0,979
France	0,825	1	0	0	0,154
India	0,771	2	0,25	0	0,417
Indonesia	0,655	1	0	0	0,356
Japan	0,838	2	0	1	0,533
Singapore	0,871	5	0,75	0,1	1
Thailand	0,689	1	0	0	0,356
United States	0,917	2	0	1	0,533
Vietnam	0,552	2	0	1	0,701

Table 3: The above results table highlights the robustness and cybersecurity scores for each country within the Asian network. Countries in the Asian network scored highly with regards to cybersecurity, as well as with the Betweenness Centrality and Clustering Coefficient metrics.

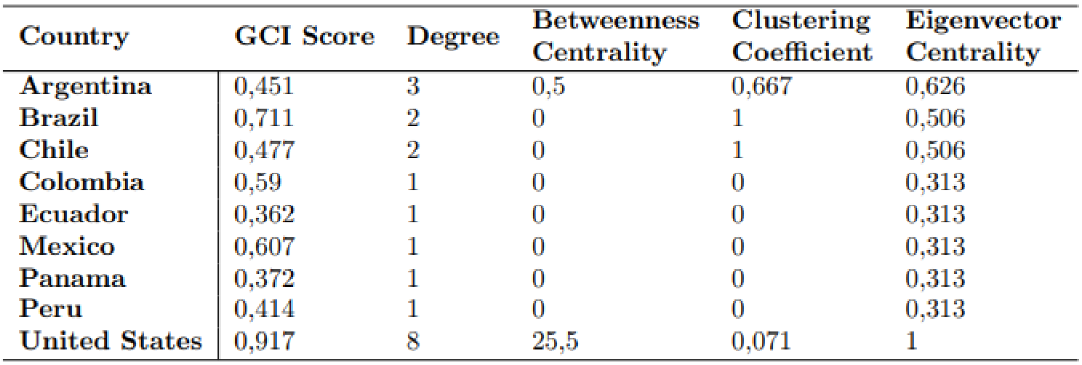
Table of Results showing the European Network Robustness Scores



Country	GCI Score	Degree	Betweenness Centrality	Clustering Coefficient	Eigenvector Centrality
Austria	0,77	1	0	0	0,302
France	0,825	4	0,286	0,5	0,924
Germany	0,802	5	0,524	0,3	1
Netherlands	0,823	3	0	1	0,859
Russia	0,776	1	0	0	0,302
Spain	0,797	1	0	0	0,278
United Kingdom	0,854	4	0,286	0,5	0,924
United States	0,917	1	0	0	0,278

Table 4: The above results table highlights the robustness and cybersecurity scores for each country within the European network. Countries in the European network fared well in the Degree and cybersecurity measures.

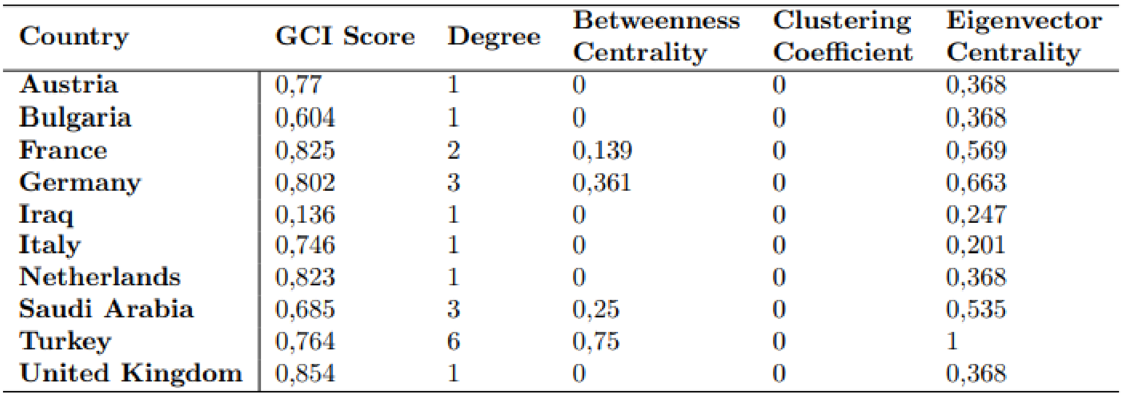
Table of Results showing the Latin American Network Robustness Scores



Country	GCI Score	Degree	Betweenness Centrality	Clustering Coefficient	Eigenvector Centrality
Argentina	0,451	3	0,5	0,667	0,626
Brazil	0,711	2	0	1	0,506
Chile	0,477	2	0	1	0,506
Colombia	0,59	1	0	0	0,313
Ecuador	0,362	1	0	0	0,313
Mexico	0,607	1	0	0	0,313
Panama	0,372	1	0	0	0,313
Peru	0,414	1	0	0	0,313
United States	0,917	8	25,5	0,071	1

Table 5: The above results table highlights the robustness and cybersecurity scores for each country within the Latin American network. Countries in the Latin American network performed well in the Betweenness and Eigenvector Centrality measures.

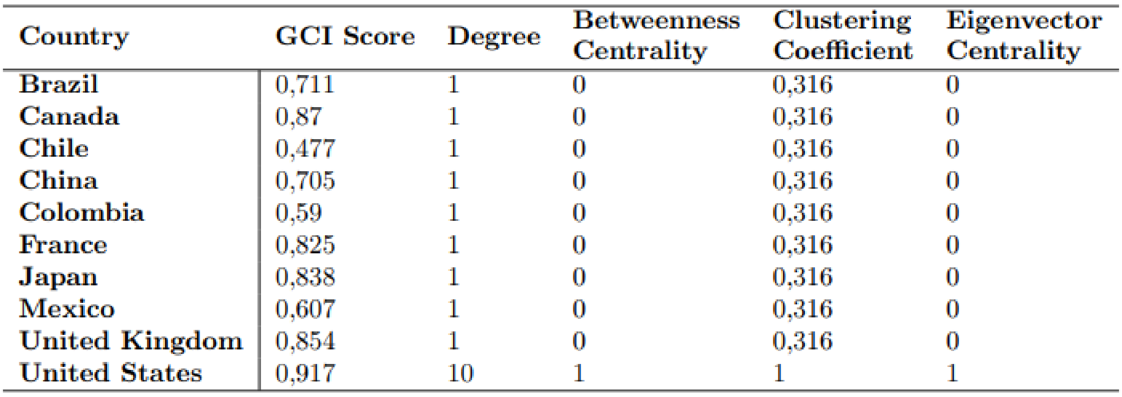
Table of Results showing the Middle Eastern Network Robustness Scores



Country	GCI Score	Degree	Betweenness Centrality	Clustering Coefficient	Eigenvector Centrality
Austria	0,77	1	0	0	0,368
Bulgaria	0,604	1	0	0	0,368
France	0,825	2	0,139	0	0,569
Germany	0,802	3	0,361	0	0,663
Iraq	0,136	1	0	0	0,247
Italy	0,746	1	0	0	0,201
Netherlands	0,823	1	0	0	0,368
Saudi Arabia	0,685	3	0,25	0	0,535
Turkey	0,764	6	0,75	0	1
United Kingdom	0,854	1	0	0	0,368

Table 6: The above results table highlights the robustness and cybersecurity scores for each country within the Middle Eastern network. Countries in the Middle Eastern network only score highly in the Eigenvector Centrality metric.

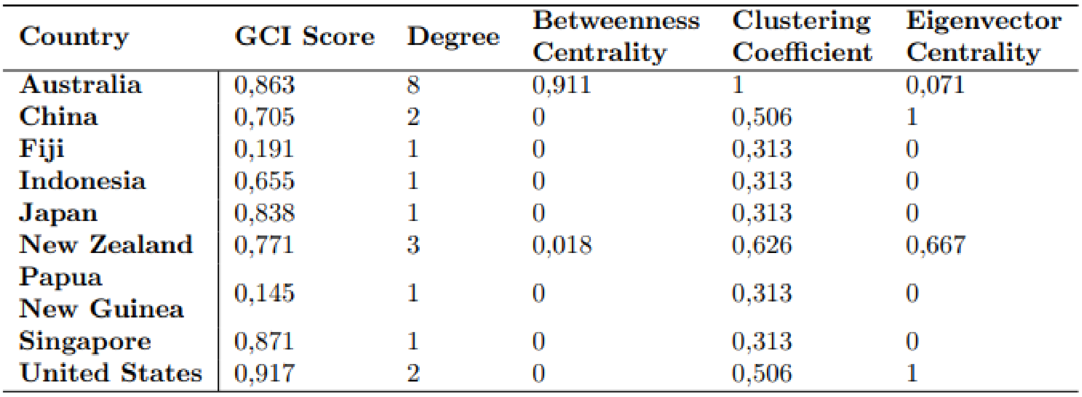
Table of Results showing the North American Network Robustness Scores



Country	GCI Score	Degree	Betweenness Centrality	Clustering Coefficient	Eigenvector Centrality
Brazil	0,711	1	0	0,316	0
Canada	0,87	1	0	0,316	0
Chile	0,477	1	0	0,316	0
China	0,705	1	0	0,316	0
Colombia	0,59	1	0	0,316	0
France	0,825	1	0	0,316	0
Japan	0,838	1	0	0,316	0
Mexico	0,607	1	0	0,316	0
United Kingdom	0,854	1	0	0,316	0
United States	0,917	10	1	1	1

Table 7: The above results table highlights the robustness and cybersecurity scores for each country within the North American network. Countries in the North American network ranked at the top in terms of Degree, Clustering Coefficient and cybersecurity.

Table of Results showing the Oceanic Network Robustness Scores

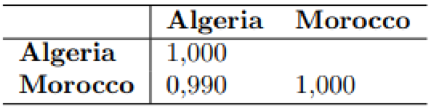


Country	GCI Score	Degree	Betweenness Centrality	Clustering Coefficient	Eigenvector Centrality
Australia	0,863	8	0,911	1	0,071
China	0,705	2	0	0,506	1
Fiji	0,191	1	0	0,313	0
Indonesia	0,655	1	0	0,313	0
Japan	0,838	1	0	0,313	0
New Zealand	0,771	3	0,018	0,626	0,667
Papua New Guinea	0,145	1	0	0,313	0
Singapore	0,871	1	0	0,313	0
United States	0,917	2	0	0,506	1

Table 8: The above results table highlights the robustness and cybersecurity scores for each country within the Oceanic network. Countries in the Oceanic network scored highly when looking at the Clustering Coefficient metric and the cybersecurity status.

Table of Results reflecting the Correlations of Graph Metrics between Robust Countries:

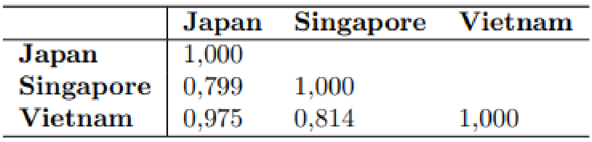
Table of Results displaying the Robust African Node Correlations:



	Algeria	Morocco
Algeria	1,000	
Morocco	0,990	1,000

Table 9: The above results table highlights the correlations between the robustness and cybersecurity scores for the key nodes within the African network. Algeria and Morocco scored identically for Degree and almost identically Eigenvector Centrality hence the very high positive correlation.

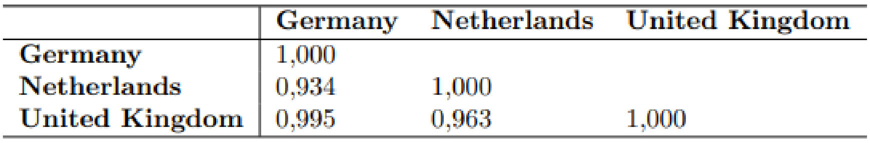
Table of Results displaying the Robust Asian Node Correlations:



	Japan	Singapore	Vietnam
Japan	1,000		
Singapore	0,799	1,000	
Vietnam	0,975	0,814	1,000

Table 10: The above results table highlights the correlations between the robustness and cybersecurity scores for the key nodes within the Asian network. Japan, Singapore and Vietnam score highly in the CC, Eigenvector and GCI metrics which yield these high correlations.

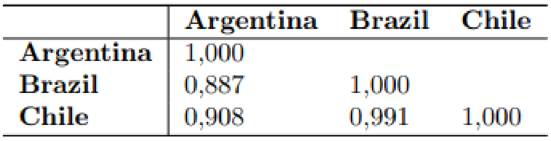
Table of Results displaying the Robust European Node Correlations:



	Germany	Netherlands	United Kingdom
Germany	1,000		
Netherlands	0,934	1,000	
United Kingdom	0,995	0,963	1,000

Table 11: The above results table highlights the correlations between the robustness and cybersecurity scores for the key nodes within the European network. Germany, Netherlands and the United Kingdom each scored highly and evenly across all robustness metrics: Degree, Clustering Coefficient, Eigenvector Centrality and GCI score, and in the mid-range for Betweenness Centrality, ultimately producing high correlations across all three key countries.

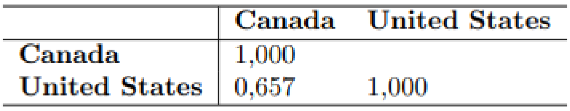
Table of Results displaying the Robust Latin American Node Correlations:



	Argentina	Brazil	Chile
Argentina	1,000		
Brazil	0,887	1,000	
Chile	0,908	0,991	1,000

Table 12: The above results table highlights the correlations between the robustness and cybersecurity scores for the key nodes within the Latin American network. The high correlations among Argentina, Brazil and Chile are due to each scoring similarly but low for Degree, and similarly high for Clustering Coefficient and Eigenvector Centrality.

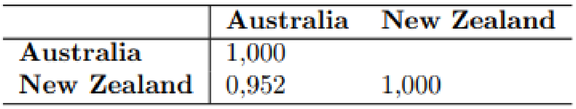
Table of Results displaying the Robust North American Node Correlations:



	Canada	United States
Canada	1,000	
United States	0,657	1,000

Table 13: The above results table highlights the correlations between the robustness and cybersecurity scores for the key nodes within the North American network. Canada and the United States did not correlate highly as the United States scores excessively well across all robustness metrics however Canada only fares well in the GCI metric, hence the low correlation.

Table of Results displaying the Robust Oceanic Node Correlations:



	Australia	New Zealand
Australia	1,000	
New Zealand	0,952	1,000

Table 14: The above results table highlights the correlations between the robustness and cybersecurity scores for the key nodes within the Oceanic network. The high correlation between Australia and New Zealand is because both scored highly for the GCI rating and the CC, and both have non-zero values for the Betweenness Centrality.

5 Conclusion

The study of network robustness is a critical tool in the understanding of complex interconnected systems such as the Internet, which due to digitalization, gives rise to an increasing prevalence of cyberattacks. Despite the importance of the Internet in the global economic system, it is rare to find empirical analyses of the global pattern of Internet traffic data established via backbone connections. This thesis uses metrics based on graph properties of network models to evaluate the robustness of the Internet backbone network. I chose one metric (Degree, Clustering Coefficient, Betweenness Centrality and Eigenvector Centrality) from each of the four respective robustness categories (Adjacency, Clustering, Throughput and Spectral Methods), to analyse seven regional Internet networks, where each metric is applicable to at least one graph type and is a direct measure of network robustness. Added to these measures is the International Telecommunication Union's Global Cybersecurity Index, which ranks countries based on their cybersecurity status.

The analysis consists of a three-step process of firstly mapping the Internet backbone as a network of networks using TeleGeography Global Internet traffic data, followed by analysing the various network and country cybersecurity profiles by using the GCI data. The final step assesses each regional network's robustness, where a combination of robustness measures are considered since the application of a single metric or metrics of the same category would not be sufficient for effectively evaluating Internet robustness.

Finally, the results show that the regions with countries which have higher cybersecurity ratings in turn have more robust networks, when compared to regions with countries which have lower cybersecurity ratings: the more economically developed regions have higher network robustness, while the less economically developed regions have lower network robustness. This is attributable to more developed regions having high degree countries which in turn have better GCI ratings due to their influence as key nodes in their specific networks. This is further supported by the fact that high degree countries also tend to have high Clustering Coefficient, Betweenness Centrality, and Eigenvector Centrality scores.

The limitations of the current work done in this thesis are that only static robustness measures are used to analyse Internet robustness, and that only a brief overview of attack and defence models is given. Therefore, for future research, there should be consideration for using dynamic robustness measures such as arbitrary node removal strategies to analyse how robust the different components of the network would be depending on different parameters of the attacks such as origin and strength. Additionally, there could be scope for investigating the impact of simulated and calibrated cyberattacks on different nodes of the Internet by using various attack and defence models, and determining what the residual equilibrium network would be. In terms of the cybersecurity limitations, the current scoring system does not account for the dimensionality of data nor the differences in the diversity of indicator values, and hence for future research, multi-attribute decision making methods could be used to better rank the level of a country’s cybersecurity status.

6 References

Barabási, A.L. & Oltvai, Z.N. 2004. Network biology: Understanding the cell's functional organization. *Nature Reviews Genetics*. 5(2):101–113. DOI: 10.1038/nrg1272.

Baumann, A. & Fabian, B. 2015a. Vulnerability Against Internet Disruptions – A Graph-based Perspective.

Baumann, A. & Fabian, B. 2015b. How robust is the internet? – insights from graph analysis. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. 8924:247–254. DOI: 10.1007/978-3-319-17127-2_18.

Brandes, U. 2001. A Faster Algorithm for Betweenness Centrality. Available: http://kops.ub.uni-konstanz.de/volltexte/2009/7188/.

CAIDA. 2007. The Skitter Project. http://www.caida.org/tools/measurement/skitter/.

Choi, J.H., Barnett, G.A. & Chon, B.S. 2006. Comparing world city networks: A network analysis of Internet backbone and air transport intercity linkages. *Global Networks*. 6(1):81–99. DOI: 10.1111/j.1471-0374.2006.00134.x.

Creese, S., Dutton, W.H. & Esteve-González, P. 2021. The social and cultural shaping of cybersecurity capacity building: a comparative study of nations and regions. *Personal and Ubiquitous Computing*. DOI: 10.1007/s00779-021-01569-6.

Diestel, R. 2017. *Graph Theory*. V. 173. (Graph Theory). Springer Berlin Heidelberg.

Doyle, J.C., Low, S.H., Paganini, F., Vinnicombe, G., Willinger, W. & Parrilo, P. 2005. Robustness and the internet: theoretical foundations. *Robust Design: A Repertoire of Biological, Ecological, and Engineering Case Studies*. 273–285.

Du, D. 2019. Social Network Analysis: Centrality Measures. Faculty of Business Administration, University of New Brunswick.

Durairajan, R., Ghosh, S., Tang, X., Barford, P. & Eriksson, B. 2013. Internet Atlas: A Geographic Database of the Internet. *HotPlanet'13*.

Hansen, D.L., Shneiderman, B. & Smith, M.A. 2011. *Analyzing Social Media Networks with NodeXL: INSIGHTS FROM A CONNECTED WORLD*. DOI: 10.1016/c2018-0-01348-1.

IBM. 2020. *Cost of a Data Breach Report-2020*. Available: https://www.ibm.com/downloads/cas/ZBZLY7KL.

International Telecommunication Union (ITU). 2015. *Global Cybersecurity Index & Cyberwellness Profiles*.

International Telecommunication Union. 2017. *Global Cybersecurity Index (GCI) 2017*. Available: https://www.itu.int/dms_pub/itu-d/opb/str/D-STR-GCI.01-2017-PDF-E.pdf.

International Telecommunication Union (ITU). 2019. *Global Cybersecurity Index 2018*.

International Telecommunication Union. 2020. *Global Cybersecurity Index (GCI)*. Available: https://www.itu.int/dms_pub/itu-d/opb/str/D-STR-GCI.01-2017-PDF-E.pdf.

Klau, G.W. & Weiskircher, R. 2005. Robustness and resilience. *In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. V. 3418 LNCS. Springer Verlag. 417–466. DOI: 10.1007/978-3-540-31955-9_15.

Madhyastha, H. v, Isdal, T., Piatek, M., Dixon, C., Anderson, T., Krishnamurthy, A. & Venkataramani, A. 2006. iPlane: An Information Plane for Distributed Services.

Mahajan, R., Spring, N. & Wetherall, D. n.d. Measuring ISP topologies with Rocketfuel.

Max Planck Institute for Informatics. 2018. *NetworkAnalyzer Online Help*. Available: https://med.bioinf.mpi-inf.mpg.de/netanalyzer/help/2.7/index.html#settings

McKinsey Global Institute. 2011. *Internet matters: The Net's sweeping impact on growth, jobs and prosperity*.

Morley, Z., Rexford, M.J., Wang, J. & Katz, R.H. 2003. Towards an Accurate AS-Level Traceroute Tool.

Newman, M.E.J. 2003. The structure and function of complex networks. *SIAM Review*. 45(2):167–256. DOI: 10.1137/S003614450342480.

Newman, M.E.J. 2008. Mathematics of networks. *The New Palgrave Encyclopedia of Economics*.

Oehlers, M. & Fabian, B. 2021. Graph metrics for network robustness—a survey. *Mathematics*. 9(8):1–48. DOI: 10.3390/math9080895.

Pastor-Satorras, R. & Vespignani, A. 2007. *EVOLUTION AND STRUCTURE OF THE INTERNET: A Statistical Physics Approach*. DOI: 10.1017/CBO9781107415324.004.

Rueda, D.F., Calle, E. & Marzo, J.L. 2017. Robustness Comparison of 15 Real Telecommunication Networks: Structural and Centrality Measurements. *Journal of Network and Systems Management*. 25(2):269–289. DOI: 10.1007/s10922-016-9391-y.

Shafqat, N. & Masood, A. 2016. Comparative Analysis of Various National Cyber Security Strategies. *International Journal of Computer Science and Information Security (IJSIS)*. 14(1):129–136.

Shavitt, Y. & Shir, E. 2005. DIMES: Let the Internet Measure Itself.

Singer, Y. 2006. Dynamic measure of network robustness. *IEEE Convention of Electrical and Electronics Engineers in Israel, Proceedings*. (1):366–370. DOI: 10.1109/EEEI.2006.321105.

TeleGeography, Inc. (2020) Global Internet Geography 2020, Washington, DC.

Watts, D.J. & Strogatz, S.H. 1998. Collective dynamics of 'small-world' networks. *Nature*. 393(6684):440–442. DOI: 10.1111/cobi.13031.

Yan, G., Eidenbenz, S., Thulasidasan, S., Datta, P. & Ramaswamy, V. 2010. Criticality analysis of Internet infrastructure. *Computer Networks*. 54(7):1169–1182. DOI: 10.1016/j.comnet.2009.11.002.

Yarovenko, H., Kuzmenko, O. & Stumpo, M. 2020. Strategy for Determining Country Ranking by Level of Cybersecurity. *Financial Markets, Institutions and Risks*. 4(3):124–137. DOI: 10.21272/fmir.4(3).124-137.2020.

Zhang, B., Liu, R., Massey, D. & Zhang, L. 2005. Collecting the Internet AS-level Topology. Available: http://irl.cs.ucla.edu/topology/.

7 Appendices

7.1 Appendix A: Student Outline DMP

AN EMPIRICAL ANALYSIS AND EVALUATION OF INTERNET ROBUSTNESS

A Data Management Plan created using UCT DMP

Creator: Michele Stampanoni

Affiliation: University of Cape Town

Template: UCT Student Generic DMP

Project abstract:
The study of network robustness is a critical tool in the understanding of complex interconnected systems such as the Internet, which due to digitalization, gives rise to an increasing prevalence of cyberattacks. Robustness is when a network maintains its basic functionality even under failure of some of its components, in this instance being nodes or edges. Despite the importance of the Internet in the global economic system, it is rare to find empirical analyses of the global pattern of Internet traffic data established via backbone connections, which can be defined as an interconnected network of nodes and edges between which bandwidth flows. Hence in this thesis, I use metrics based on graph properties of network models to evaluate the robustness of the backbone network, which is further supported by international cybersecurity ratings. These cybersecurity ratings are adapted from the Global Cybersecurity Index which measures countries' commitments to cybersecurity and ranks countries based on their cybersecurity strategies. Ultimately this empirical analysis follows a three-step process of firstly mapping the Internet as a network of networks, followed by analyzing the various networks and country profiles, and finally assessing each regional network's robustness. By using TeleGeography and ITU data, the results show that the regions with countries which have higher cybersecurity ratings in turn have more robust networks, when compared to regions with countries which have lower cybersecurity ratings.

ID: 3918

Start date: 03-08-2020

End date: 13-01-2023

Last modified: 08-12-2022

AN EMPIRICAL ANALYSIS AND EVALUATION OF INTERNET ROBUSTNESS - STUDENT OUTLINE DMP

1. GENERAL GUIDELINES

PURPOSE OF THIS TEMPLATE - The purpose of the Outline DMP is to indicate your initial plans for how your data will be collected, shared and stored, and to give you a chance to think about these data-focused aspects of the research process. As your begin doing your research, your data process may change, and it is perfectly acceptable to change your data management plan to accommodate the changes in your research process. Indicate below that you understand the purpose of completing this Outline DMP template.

I understand the Outline DMP template is a projection of my anticipated data management planning requirements and should be updated as my project develops.

2. AUTHORS AND SUPERVISORS

PROJECT NAME - Replicate the title of your project, dissertation or thesis exactly as it appears in your proposal document.

An Empirical Analysis and Evaluation of Internet Robustness.

PERSONAL DETAILS - Indicate the name(s) and student number(s) of the student(s) who will be involved in this project, dissertation or thesis.

Michele Romolo Stampanoni - STMMIC001

SUPERVISOR(S) DETAILS - Indicate who will supervise this project, dissertation or thesis. If you do not yet have a supervisor, leave this section blank.

Co-Pierre Georg

3. DATA COLLECTION/GENERATION

COLLECTION OF ORIGINAL DATA - Indicate whether or not you intend to gather/produce original data for your study, and provide a brief description of the kind of data you think you will collect. If you are unsure at this time, indicate what you think you are most likely to collect. If you are not intending to gather or collect your own data, declare that here.

- I d not intend to collect original data.

USE OF EXISTING DATA - Indicate if you intend to re-use existing data, either from online searches or from datasets provided by your supervisor, lab, or funder. If you are not intending to re-use existing data, declare that here.

- I intend to reuse existing data in my study (described below).

I intend to use existing datasets provided by TeleGeography Inc, for which I have obtained permission by TeleGeography Inc to use their data. Furthermore I also intend to use existing and publicly made available data from the International Telecommunication Union. In terms of storage, the TeleGeography data is stored on the TeleGeography Inc repository and both the TeleGeography and the International Telecommunication union data are stored on Google Drive and an external hard drive.

DATA SHARING - Indicate whether or not you are intending to publish your research data. If you are, indicate where you are intending to publish your data and under what licensing conditions, such as Creative Commons. If you are not intending to publish your data, provide reasons and reference the appropriate ethical considerations, commercial applications/patenting ambition, or data re-use agreements that prevent you from publishing your data.

- I intend to share my data (details below).

My research data is made open by default as there is no specific reason for my data not to be shared. The data from the International Telecommunication Union is already publicly available, however, TeleGeography has requested that all data be referenced upon publishing.

4. DATA STORAGE

ANTICIPATED DATASET SIZE - Indicate the estimated size of your completed dataset, and indicate whether or not you will need to access additional data storage facilities. If such storage is not provided by your unit or department, you may need to factor in the cost of purchasing additional storage space.

- 20GB or less

No additional data storage facilities are required as I have sufficient storage space for both the TeleGeography data as well as the International Telecommunication Data.

DATA BACKUPS - Indicate how you plan to ensure your data is secure and retrievable in case of errors or hardware failure. Describe what procedures you will put in place to back-up copies of your data and where they will be stored.

- I intend to backup my data using a commercial service provider.

During my data collection and analysis phase, I will backup my data each month to my personal Google Drive account as well as to an external hard drive. I will do a final backup when I submit my final draft for examination.

5. DATA CENTRE(S)/REPOSITORIES

DATA CENTRES/REPOSITORIES - Once your project, dissertation or thesis is complete, it is advisable to curate and archive your completed dataset with an established data centre or repository. Note that you should archive your data even if you are not intending to publish it. Check with your supervisor or funder if you are required to deposit your data in a specific repository, or declare that you will deposit the data in ZivaHub (see the Guidance section).

- At the end of my study, I will deposit my data on ZivaHub.

I am not required to store my data on a specific subject repository, and hence I will be using the UCT secure data repository called ZivaHub.

METADATA - Metadata is descriptive information that others will need to make sense of your dataset. Metadata includes things like study descriptions or abstracts, study instruments (sample collection schedules, codebooks for variables, survey instruments, etc.), subject codes, and keywords. Indicate what metadata will accompany your curated dataset.

The completed dataset will be accompanied by keywords, a short description taken from my dissertation abstract and relevant paragraphs on the data process taken from my data and methodology sections.

6. BUDGET

BUDGET - Indicate any costs specifically relating to the management and curation of your data, such as purchasing additional storage space, digitisation of physical media, data storage or curation charges, and data audits. Most student research will be able to make use of free options provided by UCT and will not have to budget for data costs.

- I do not anticipate any data costs as my data is less than 10GB, and I will be using a storage system provided by UCT (UCT GoogleDrive, UCT OneDrive, Netstorage, ZivaHub, etc.) to curate my data.

UCT’s ZivaHub platform and Google Drive do not charge for usage in the case of small datasets such as the one I will be using, and hence no budget will be required for data storage, data audits or data curation.

7.2 Appendix B: Student Full DMP

AN EMPIRICAL ANALYSIS AND EVALUATION OF INTERNET ROBUSTNESS

A Data Management Plan created using UCT DMP

Creator: Michele Stampanoni

Affiliation: University of Cape Town

Template: UCT Student Generic DMP

Project abstract:
The study of network robustness is a critical tool in the understanding of complex interconnected systems such as the Internet, which due to digitalization, gives rise to an increasing prevalence of cyberattacks. Robustness is when a network maintains its basic functionality even under failure of some of its components, in this instance being nodes or edges. Despite the importance of the Internet in the global economic system, it is rare to find empirical analyses of the global pattern of Internet traffic data established via backbone connections, which can be defined as an interconnected network of nodes and edges between which bandwidth flows. Hence in this thesis, I use metrics based on graph properties of network models to evaluate the robustness of the backbone network, which is further supported by international cybersecurity ratings. These cybersecurity ratings are adapted from the Global Cybersecurity Index which measures countries' commitments to cybersecurity and ranks countries based on their cybersecurity strategies. Ultimately this empirical analysis follows a three-step process of firstly mapping the Internet as a network of networks, followed by analyzing the various networks and country profiles, and finally assessing each regional network's robustness. By using TeleGeography and ITU data, the results show that the regions with countries which have higher cybersecurity ratings in turn have more robust networks, when compared to regions with countries which have lower cybersecurity ratings.

ID: 3918

Start date: 03-08-2020

End date: 13-01-2023

Last modified: 08-12-2022

AN EMPIRICAL ANALYSIS AND EVALUATION OF INTERNET ROBUSTNESS - STUDENT FULL DMP

1. PROJECT DETAILS

PROJECT NAME - Replicate the title of your project, dissertation or thesis exactly as it appears in your proposal document.

An Empirical Analysis and Evaluation of Internet Robustness

PERSONAL DETAILS - Indicate the name(s) and student number(s) of the student(s) who will be involved in this project, dissertation or thesis.

Michele Romolo Stampanoni - STMMIC001

SUPERVISOR(S) DETAILS - Indicate who will supervise this project, dissertation or thesis. If you do not yet have a supervisor, leave this section blank.

Co-Pierre Georg

2. PROJECT SUMMARY

RESEARCH SUMMARY - Briefly summarise your study. Include the study's objectives, design, and methods.

The study of network robustness is a critical tool in the understanding of complex interconnected systems such as the Internet, which due to digitalization, gives rise to an increasing prevalence of cyberattacks. Robustness is when a network maintains its basic functionality even under failure of some of its components, in this instance being nodes or edges. Despite the importance of the Internet in the global economic system, it is rare to find empirical analyses of the global pattern of Internet traffic data established via backbone connections, which can be defined as an interconnected network of nodes and edges between which bandwidth flows. Hence in this thesis, I use metrics based on graph properties of network models to evaluate the robustness of the backbone network, which is further supported by international cybersecurity ratings. These cybersecurity ratings are adapted from the Global Cybersecurity Index which measures countries' commitments to cybersecurity and ranks countries based on their cybersecurity strategies. Ultimately this empirical analysis follows a three-step process of firstly mapping the Internet as a network of networks, followed by analyzing the various networks and country profiles, and finally assessing each regional network's robustness. By using TeleGeography and ITU data, the results show that the regions with countries which have higher cybersecurity ratings in turn have more robust networks, when compared to regions with countries which have lower cybersecurity ratings.

3. DESCRIPTION OF THE DATA

DATA REUSE DESCRIPTION - If you re-used data from third-party sources in your study, record pertinent details here such as the source of the data, the extent of the data, usage rights or restrictions pertaining to the data, and how it was incorporated into your study.

I have used existing data in my study.

I used data from TeleGeography Inc, consisting of 7 Excel spreadsheets of regional Internet geography data .This data was used as a baseline for my thesis's in depth network analysis. I signed a contract with TeleGeography obtaining their permission to use the data, provided I reference their data. In terms of the cybersecurity data, I used the International Telecommunication Union dataset which consists of four PDF documents of cybersecurity ratings for all 38 countries. This cybersecurity data is readily and publicly available on the internet for all to use.

DATA DESCRIPTION - Describe the data you have gathered for your study. Briefly describe the nature, scope and scale of the data you have produced.

The TeleGeography data is structured such that 10 routes for each region (Africa, Asia, Europe, Latin America, Middle East, North America and Oceania) are reported, which is aggregated to make up an international dataset spanning the five years 2016 – 2020, to produce a total of 38 nodes and 61 edges. This internet Geography data is quantitative in nature with the seven Excel spreadsheets totalling to a size of 89kb. In addition to this, the Cybersecurity data from the International Telecommunication Union reports cybersecurity ratings for each of the 38 countries analysed in the thesis. This is also quantitative data and the four PDF documents total to 21.9 MB in size.

4. FORMATS AND QUALITY CONTROL

QUALITY CONTROL - Describe what measures you took to ensure the data you collected were of high-quality.

The TeleGeography data was used directly from their repository, where TeleGeography developed estimates using proxies on backbone deployment and capacity utilization trends, including the carrier type and route type. Furthermore, I did general data cleaning activities such as checking for null values and outliers.

Regarding the ITU data, the cybersecurity GCI statistic is bound between 0 and 1, and each of the four versions (2015, 2017, 2018, 2020) of the cybersecurity data has GCI scores for all the 38 TeleGeography countries, and thus I averaged the rankings over the four years to form a mean GCI Score, Upon my general EDA, I noticed that data was unavailable for years 2016 and 2019, and hence I imputed GCI scores from the previous iterations (2015 and 2018, respectively) for these two missing years. Each iteration of the reports used in this thesis are provided, pinpointing the exact score and rank for all 38 countries.

FILE FORMATS - Indicate the formats in which your data will be collected and processed. Clarify whether these formats require specialised proprietary software to access or if they will be produced in or converted to more open, accessible formats for long-term accessibility and preservation. In the case of physical data objects (such as artworks or models) indicate whether these will be digitised or otherwise preserved for accessibility.

The TeleGeography Internet Data is all in Excel format and the International Telecommunication Union Cybersecurity data is all in PDF format. I then combined these datasets into tabular Latex tables. Essentially all the data is accessible in open formats.

5. DATA MANAGEMENT, DOCUMENTATION AND CURATION

CURATION (MANAGING AND STORING) DATA - Describe how you organise and manage your data. Specify any file-naming conventions or community data standards you have adopted.

I organised and managed my data by storing it in its own folder on my personal computer as well as in its own folder on Google Drive and on the external hard drive. I named the Internet Geography data by region and I named the Cybersecurity data by year.

BACKUP AND STORAGE - Describe how your data is being stored and backed-up. If you are using a data service provider, provide details on for how long they will retain the data.

I stored all my data on my personal computer, as well as on Google Drive and additionally I made back ups of all my data on an external hard drive. This includes the raw data, as well as the analysed data.

METADATA STANDARDS AND DATA DOCUMENTATION - Articulate what metadata and documentation you have produced about the data you have generated, collected or re-used.

The TeleGeography data is structured such that 10 routes for each region (Africa, Asia, Europe, Latin America, Middle East, North America and Oceania) are reported, where these 10 routes are kept constant year on year with only the level of international Internet traffic between the routes changing. The data is composed of countries (nodes) and the route between two countries (edges), with the Internet traffic flowing between edges measured in Gbps. Each node has been assigned an attribute – the node's respective cybersecurity status given by the International Telecommunication Union GCI data, and each edge has been assigned an attribute – the edge's respective weight of Internet traffic given by the proportion of bandwidth sent over that route. This weight is calculated as a simple fraction of an edge's traffic relative to the entire network's traffic. The network data of each region is then aggregated to make up an international dataset spanning the five years 2016 – 2020, to produce a total of 38 nodes and 61 edges. The International Telecommunication Union (ITU) publishes a Global Cybersecurity Index (GCI) that ranks countries based on their cybersecurity strategies. The GCI measure presents factual representations of each country's level of cybersecurity, as the GCI is a composite index combining 25 indicators into one benchmark. This ranking applies to 193 ITU Member States in all regions: Africa, Americas, Arab States, Asia-Pacific, and Europe. The index aims to quantify the type, level, and evolution of cybersecurity policies in countries and relative to other countries, as well as the progress in cybersecurity methodologies of all countries from a global perspective. Furthermore, the index accounts for progress in cybersecurity strategies from a regional perspective, and accounts for the difference between countries in terms of their level of engagement in cybersecurity initiatives. The ITU has published four iterations of the GCI report (2015, 2017, 2018 and 2020) all of which have been averaged and used in my thesis.

6. DATA SECURITY AND CONFIDENTIALITY OF POTENTIALLY DISCLOSIVE INFORMATION

SECURITY - Indicate to what extent your data can be considered sensitive or at-risk. Describe how you will control access to your data. Indicate whether you anticipate a need for encryption or password-controlled access, and if so, how you will enforce that access.

My data is not considered to be sensitive, as it only deals with bandwidth usage and cybersecurity ratings. The TeleGeography data is stored on the TeleGeography website, and all both the TeleGeography and ITU data is being stored on my personal computer, Google Drive and an external hard drive. I downloaded copies to my computer for active cleaning and analysis, where my computer is password-secured.

ETHICS AND PRIVACY - Describe, as per your Ethics Clearance form or other similar documentation, any ethical or privacy issues that your data are subject to (if any). Summarise the main risks to the confidentiality and security of information related to human participants, the level of risk, and how this risk will be managed. If your project did not require ethical clearance, you may ignore this section.

My data on Global Internet Geography and Cybersecurity ratings contain no ethical issues as there are no human or animal participants.

7. DATA SHARING AND OPEN ACCESS

DATA OWNERSHIP - If you have used existing datasets, note down any restrictions the data providers have indicated regarding data sharing. Otherwise, leave blank.

- I have used existing data in my study and I have noted down the relevant restrictions as pertains to data sharing(details below).

I am using data from TeleGeography, and I have signed a contractual agreement with them granting me a non-exclusive license to use the data. I am also using data from the International Telecommunications Union however this data has been made publicly available.

DATA LICENCE - Indicate under which licence you intend to share your research data. If you are not sharing your data, provide the appropriate justification as per the UCT Research Data Management guidelines.

- CC BY

I will share my data from my study under a CC BY licence. This license allows me to distribute, remix, adapt, and build upon the material in any medium or format, so long as attribution is given to the creator. The license allows for commercial use however credit must be given to the creator.

DATA PUBLICATION - Indicate where you intend to publish your research data at the end of your project.

I will share my data on ZivaHub at the end of my project.

8. RELEVANT INSTITUTIONAL OR STUDY POLICIES

Indicate the relevant departmental, unit, or institutional policies that influence your data management activities.

Not applicable.