5-2018

Quantifying Irregular Geographic Exposure on the Internet

Jordan Alexander Holland
University of Tennessee, jholla19@vols.utk.edu

To the Graduate Council:

I am submitting herewith a thesis written by Jordan Alexander Holland entitled "Quantifying Irregular Geographic Exposure on the Internet." I have examined the final electronic copy of this thesis for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Master of Science, with a major in Computer Science.

Maxfield J. Schuchard, Major Professor

We have read this thesis and recommend its acceptance:

Mark E. Dean, Audris Mockus

Accepted for the Council:
Dixie L. Thompson

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

Quantifying Irregular Geographic Exposure on the Internet

A Thesis Presented for the

Master of Science

Degree

The University of Tennessee, Knoxville

Jordan Alexander Holland

May 2018

Acknowledgments

First, I would like to thank my parents for their unwavering support. I would also like to thank my advisor, Max Schuchard, as this would not have happened without his guidance. Finally, I would like to thank the Volsec Security lab at the University of Tennessee for putting up with me.

Abstract

In this work, we examine to what extent the Internet's routing infrastructure needlessly exposes network traffic to nations *geographically* irrelevant to packet transmission. We quantify what countries are *geographically logical* to see on a network path traveling between two nations through the use of convex hulls circumscribing major population centers, and then compare that to the nation states observed in over 14.5 billion measured paths. Our results show that 49% of paths unnecessarily expose traffic to at least one nation. We further explore what nations, regions, and ASes expose and benefit from this geographically illogical traffic. As an example, we see that 23% of source/destination pairs located outside of the United States send their traffic through the US, but only 8% of those paths are geographically logical. Finally, we examine what happens when countries exercise both legal and physical control over ASes transiting traffic, gaining access to traffic outside of their geographic borders, but carried by organizations that fall under a particular country's legal jurisdiction. When considering both the physical and legal countries that a path traverses, our results show that over 57% of paths expose traffic to a geographically irrelevant country.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

The Internet is comprised of independent networks called Autonomous Systems (ASes), which depend on each other for inter-network connectivity. Network traffic must often traverse multiple ASes in order to reach its final destination. Any adversarial transit AS situated between sender and receiver, termed a *path based adversary*, can degrade network availability, violate data integrity, and undermine confidentiality. While a single malicious transit AS is a powerful adversary, nation states represent an even stronger path based adversary. A nation state has the ability to exercise control over both ASes that physically operate networking infrastructure within its borders and ASes whose corporate governance falls within the nation state's legal jurisdiction. A motivated nation state could coerce multiple ASes into acting as colluding path based adversaries on the nation state's behalf.

Revelations in recent years have demonstrated the extent to which nation states are willing to exert pressure on private entities in an effort to execute national cyber policy on network traffic. For example, the United States, Great Britain, and other members of the so called Five Eyes intelligence alliance have integrated dragnet surveillance into core Internet transit links that reside within their borders [7, 24, 8]. Additionally, there exist recorded instances of censoring nation states such as China applying domestic censorship policies to network traffic that neither originates nor is bound for domestic sources [25, 11].

Due to the potential security concerns exposure to additional nation states might present, it is reasonable to expect that the Internet's routing infrastructure minimizes such exposures. In scenarios where sender and receiver reside inside the same or adjacent nation states, one

would expect that these are the only jurisdictions traffic is exposed to. Even in scenarios where the sender and receiver are located in non-adjacent nations, and exposure to third party nation states is a physical necessity, one might assume that exposure is limited to a minimal set of nations required to build a physical link between sender and receiver. However, data routing decisions focus on the *logical* network topology rather than the *geographic* and *political* topology. This can result in paths exposing traffic to nations which do not lie between the geographic locations of the sender and receiver. This excess exposure needlessly increases the power of certain nation states if they elected to coerce ASes into serving as path based adversaries on their behalf.

In this work, we examine to what extent the Internet's routing infrastructure increases the capacity of certain nation states to undermine security properties of network traffic beyond what would be predicted. Specifically, we quantify how often network traffic is exposed to additional nation states beyond those residing along a *geographically logical* path between sender and receiver. In order to do this, we examine more than 14.5 billion paths observed from traceroutes conducted by CAIDA's Ark measurement framework [ark] from January 1, 2017 to October 31, 2017. To explore **physical exposure** of data to nation states we compare the set of nations where the network infrastructure traversed by a given path physically resides to the set of nations we might *geographically expect* to see on a physical path between the sender and receiver. In order to build the set of geographically logical countries between sender and receiver we use a novel technique based on computing a population biased convex hull between the sender and receiver's home nation states and constructing the set of all countries that at least partially reside inside the convex hull. We also explore what happens when nation states, in addition to taking advantage of physical exposure, additionally exercise control over ASes whose corporate governance operates within their legal jurisdiction, and leverage that pressure to observe all data handled by such ASes, something we term **legal exposure**.

Overall, we find that roughly half, 49%, of examined paths physically expose traffic to *at least* one unexpected nation state. We also find that the degree to which a user can expect their network traffic to not be physically exposed to additional nation states varies widely. For more than half of our observed source countries more than 60% of their paths contain

additional nations, and a full third of nations see 75% or more of their paths exposed to unexpected countries. We see similar asymmetry when it comes to which geographically illogical nations appear on paths. In our measurements the top ten offending countries account for 83% of all of the instances of geographically irregular countries seen in paths. We also see how certain countries disproportionately benefit from this additional exposures, allowing them to observe and impact more traffic than expected. As an example, we find that the United States resides on 55% of paths that both start and end in foreign countries, but is geographically expected to only appear on 8% of those paths.

When we expand to consider paths which either physically or legally expose traffic to unexpected nation states the situation gets worse. More than 57% of expose traffic to at least one nation state.

The remainder of this paper is presented as follows. In Section 2 we will cover both relevant background on how logical routing decisions are made, along with providing motivation for our study. In Section 3 we will present our methodology for collecting paths, labeling countries that the path either physically or legally exposes data to, and lastly how we build a quantifiable measure of what countries are geographically logical to observe between source and destination. In Section 3.0.3 we will cover the basic properties of our resultant datasets. Section 4.1 examines the extent to which nations physically exposure their traffic to other countries, along with how often that exposure is geographically illogical. Section 4.2 expands this analysis to additionally consider exposure to entities that could be legally coerced by a particular country. Lastly we draw our final conclusions in Section 5.

Chapter 2

Background

2.0.1 Path Selection On the Internet

The Internet is comprised of a collection of independently administered networks called Autonomous Systems, or ASes. ASes provide network connectivity to hosts inside of their network, enabling those hosts to connect to other host located either inside the same AS or in a remote AS. To this end, ASes deploy special purpose networking equipment called routers whose job it is to compute the best path between data sender and receiver and forward data to the next hop (router) in that path. In order to compute best paths, routers execute routing protocols which facilitate the exchange of topology information between routers and then use that information to compute best paths. Routing protocols do not compute the best path to every individual IP address. Instead, routing protocols compute best paths to blocks of IP addresses, and forward traffic addressed to any host inside that block along the same path. Routing protocols are typically divided into two categories, inter-AS and intra-AS routing protocols. Inter-AS routing protocols compute the sequence of ASes data will travel when moving between remote ASes. Intra-AS routing protocols compute the best path traffic takes inside of a given AS. Intra-AS routing protocols are responsible for both getting traffic to a end destination inside the current AS and delivering traffic to a boarder router that will transfer traffic to the next AS along a multi-AS path.

The Border Gateway Protocol [20], or BGP, is the de-facto standard inter-AS routing protocol. This single standard routing protocol is a result of the demand for interoperability

between independently managed organizations that often compete with each other. BGP is a path vector routing protocol with policies. The policy portion of BGP allows network administrators to select paths based on arbitrary criteria, rather than simple the shortest paths. Commonly, ASes utilize their business relationship with neighboring ASes to make a first pass routing decision. ASes that have direct connectivity with each other typically form customer-provider relationships, where the customer pays the provider for all traffic flowing between the two ASes in exchange for connectivity to remote ASes via the provider. ASes will generally follow a routing policy termed "Valley Free Routing", where they prefer routes that are more economically advantageous, however recent measurement studies have shown that this is not always the case [9].

Because intra-AS decisions only need to be computed over infrastructure held by one organization, removing the demand for interoperability, there are a myriad of intra-AS routing protocols deployed. Examples include link state protocols such as IS-IS [19] and OSPF [18], along with path vector variants such as EIGRP [22]. Most intra-AS protocols include the ability to include network policy rather than simple network distance as part of the routing decision making process. This policy is often expressed as an "administrative distance" giving the algorithm hints as to the adminsitrator's preferences.

2.0.2 Measuring Utilized Paths

Predicting the exact path data will travel between source and destination is challenging. In the case of inter-domain routing, AS relationships are closely held secrets. AS relationships can be inferred with some degree of accuracy based on publicly available BGP mirrors, sources of information on the current state of the BGP global routing table. However, as shown most recently by Anwar et. al. [9], using inferred relationships to predict inter-AS routing paths can be inaccurate. Predicting intra-AS routes is even more challenging. First, *which* intra-AS routing protocol an AS is using is a corporate secret, and difficult to detect. Second, the particular configuration and administrative preferences that factor into the intra-AS routing protocol's decision making process are difficult to both infer and to use as a predictive model accurately.

An alternative, but more accurate, method for determining the utilized path between hosts is to execute a `traceroute` [17] between the two hosts. Traceroute takes advantage of how networking equipment commonly behaves when it encounters a packet with an expired Time-To-Live (TTL) field. Often, but not always, routers will respond with an ICMP message to the sending host, informing the host that the packet expired, and the IP address at which the packet expired. By sending packets with small, but incrementally increasing, TTLs, and recording the IP addresses that respond, a host can map the sequence of routers a packet traverses on its path. While traceroute provides exceptional accuracy, there are several shortcomings. Most obviously, the source must be under the control of the measuring party. To overcome this shortcoming, there are several distributed measurement test-beds that either conduct traceroutes at regular intervals to large portions of the Internet or conduct traceroutes to specified hosts. Examples of such test-beds include CAIDA's Ark Infrastructure [ark] and RIPE's Atlas Infrastructure [rip]. Another major issue is that network infrastructure is not obligated to respond when a packet's TTL expires. In this case gaps in the full path to the host will result.

2.0.3 Path Based Adversaries

The security properties of many distributed systems can be impacted by the adversarial AS that transmits data, what we refer to as a *path based adversary*. As an example, a path based adversary can trivially violate the confidentiality of any unencrypted traffic. Despite this obvious threat vector, a recent study by the EFF found that only about half of web traffic is actually encrypted [13]. More complex attacks undermining confidentially are also possible. An example of such an attack is an AS that wishes to attempt to de-anonymize users of Tor, an anonymous communication system. Feamster and Dingledine [12] first pointed out that an AS that appears on the path between a user and their entry into the Tor network, as well as appearing between their traffic's exit point from the Tor network and its final destination, could undermine the anonymity properties of the Tor network. The integrity of distributed systems can also be disrupted by adversaries who lie along a utilized path. Apostolaki et. al. [10] demonstrated that adversaries capable of observing traffic between 900 IP blocks could control and edit interactions between a majority of the computational power

dedicated to mining Bitcoin, opening up the possibility of double spending via forced forking of the block chain. Additionally, in 2014 attackers utilized compromised BGP speakers in an effort to hijack communications between Bitcoin miners and their pool servers [14], resulting in the theft of Bitcoins current worth approximately $1.3 million USD.

There exist both academic studies and real world examples of adversaries that can control multiple ASes, becoming exceptionally wide reaching path based adversaries. Johnson et. al. [15] first expanded the AS level Tor adversarial model to include such powerful adversaries when they explored the capacity of Internet Exchange Points to de-anonymize Tor users. Revelations from whistle blowers including both the Snowden leaks [7] and earlier revelations by the AT&T contractor Mark Klein [8, 24] demonstrated the willingness of the NSA and other spy agencies such as GCHQ to integrate surveillance devices, and even systems which actively inject data into network streams inside core network infrastructure located within their respective nations. In addition to the attacks outlined above, documents have revealed a complex infrastructure for violating user confidentiality by building relationship graphs based solely on linking data senders and receivers. Furthermore, there exists evidence of censoring nation states: for example China either accidentally or intentionally applying censorship policies to traffic that neither originates in nor is bound to domestic hosts [25, 11].

Chapter 3

Experimental Methodology

Data potentially falls under the legal jurisdiction of any nation it either physically crosses or that can exert legal influence over a transit AS utilized en-route to its destination. From a security perspective, exposing traffic to nations is potentially problematic because it increases the *possibility* for malicious activity by every country crossed. Examples of potential malicious activity includes dropping the data, eavesdropping on the data, and even changing the contents of the data. While some exposure is unavoidable, any extraneous nations traffic is exposed to, i.e. those not physically necessary for the propagation of traffic, needlessly increase this security risk. Our goal is to accurately measure the fraction of paths which expose network traffic to nations not required for the actual transmission of data, and to quantify how much of the traffic a nation state can observe and control results from geographically illogical paths.

The ideal situation for data traveling from nation A to nation B is a path that consistently moves *towards* nation B. Phrased differently, we would logically expect network level paths between two nations to approximately, but not exactly, traverse the shortest path between those nations. While this definition is simplistic, it does highlight certain path selection choices that are illogical and should not be observed. For example, a path that goes the *opposite* direction than its destination should be considered illogical. Additionally, data with a destination that lies within the same nation it originated from should *almost never* leave that nation.

In order to quantify how many paths follow geographically illogical routes, and which countries can exert additional control over traffic as a result, we need to build two datasets. First, we must establish a set of expected countries traffic could be exposed to during transit between a particular source/destination pair based on geographic realities. We term countries inside of this set *geographically normal* for a given source and destination. Second, we must establish the actual paths data takes from sources to destinations and establish which countries the traffic is exposed to. We compute both which countries physically host the network infrastructure traveled, termed *physical exposure*, and which countries can exert legal pressure on the ASes appearing along the path, termed *legal exposure*. By comparing the set of countries a path exposes traffic to with the set of geographically normal countries, we can label the path as either *normal*, no excess exposure, or a *bad path*, containing geographically illogical countries. We can also quantify the number of times a country is a *benefactor* of a bad path, specifically the number of additional source destination pairs it can observe as the result of bad paths.

3.0.1 Computing Geographically Logical Paths

Defining the geographically normal countries for a source/destination pair was done using a novel technique based on *convex hulls*. The *convex hull* of a set of points S in n dimensions is the intersection of all convex sets containing S. For N points *p1, ..., pN*, the convex hull C is then given by the expression:

$$C = \sum_{j=1}^{N} \lambda_j \rho_j : \lambda_j \geq 0 \ \forall j \ and \ \sum_{j=1}^{N} \lambda_j = 1 \tag{3.1}$$

A more intuitive way to think about the definition of a convex hull is: given a set of points, what is the shape a stretched rubber band takes when encompassing all of them.

Using Equation 3.0.1, we can build convex hulls containing *both* the set of points that define the country containing the source and a set of points that define the country containing the destination. These convex hulls are computed taking the spherical nature of the Earth into account. Note that when a country is the source and destination of traffic, we *only* consider that country as normal and do not build a convex hull for the single country. This

convex hull construction efficiently defines all points that lie between source and destination countries. Source and destination were considered at the granularity of nation state due to potential limitations in the accuracy of GeoIP location use later; it should be noted that this coarser granularity only *increases* the number of countries considered geographically normal. We then compute the set of geographically normal countries by finding all countries that either fully or partially fall within this convex hull. In order to detect countries that *fully* reside inside the convex hull, we test if any of the 15 largest cities in the given country resides inside the convex hull. To detect countries that lie only partially inside of the convex hull, we test if any point along the edge of the convex hull lies inside the borders of a particular country.

One option for defining the set of points that make up a country is to utilize the nation's political borders, provided in the Matic dataset [sha]. In order to test this approach, we utilized shapefiles which contain points that define polygons of the actual borders of each country. This approach tends to result in convex hulls between two nations that contain countries which do not lie in the path between those nations. One factor contributing to this is countries with non-contiguous territories or remote territorial holdings. An examples of this is the United State's convex hull when including Alaska, Hawaii, Guam, and other remote territories, as the resulting convex hull covered roughly one quarter of the earth's *total* surface area. Additionally, the political borders of a country do not necessarily reflect where bulk the Internet infrastructure of the country is located; as this generally lies in the more populated areas. Relevant examples of this include China and Russia. To address, this we built a separate definition of points describing a country using the latitude and longitude of the top 15 most populous cities in that country [pop].

Figure 3.1 shows an example of the two construction techniques for the path between China and Mongolia. The population based convex hull results in a stricter version of a normal path between two countries and accurately reflects the fact that 83% of China's population, including all of its major cities, reside in the eastern portion of the country. The border based convex hull includes countries in the wrong cardinal direction, such as India and Vietnam, a result of China's concave shape. For these reasons we chose to use the city based construction of a convex hull for the measurements contained inside this work.

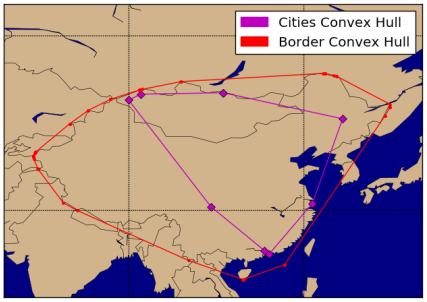

Figure 3.1: A comparison of the border based convex hull and population biased convex hull between China and Mongolia. Note that over 83% of China's population lives on its eastern coast.
s

3.0.2 Computing Actual Physical and Legal Exposure

To establish the actual path from one IP address to another we utilize a distributed traceroute measurement framework. As mentioned in Section 2.0.2, there are two widely accepted options for such a framework. The first, the Ripe Atlas measurement ecosystem, allows for users to spend credits to do measurements from a set of over 1000 geographically distributed probes to user chosen endpoints. These measurements are then made publicly available. This system has been used to appropriately measure subsets of the geographic topology of the internet in related work by Karlin et. al. [16], which examined paths bound to specific destinations rather than the Internet as a whole. However, RIPE Atlas's "on demand" nature means that it does not uniformly sample the IP space, but rather over samples certain destinations. Since our work considers the entire topology of the internet another, more appropriate option exists: CAIDAs Archipelago Measurement Infrastructure (ARK) [ark]. ARK is comprised of 180 monitors that work in teams to send traceroutes to a random IP address in each block of globally addressable IPs [1] every 48 hours. Note that this does not imply that each monitor probes each prefix every 48 hours, but that at a single destination prefix will be probed by *one* monitor each cycle. This unique measurement system gives us a uniform sample across the set of possible destinations and taken from a diverse collection of geographic locations, making it better suitable for measuring the geographic topology of the internet *as a whole*.

[1] Specifically each /24 prefix.

We collected traceroutes over a 10 month period from January 1, 2017 to October 31, 2017. To process each traceroute in our dataset, we needed to be able to map each IP address that appears in the path back to the country it is physically located in as well as the AS it is owned by. Mapping each IP address back to its correct country was done using the Geolite IP geolocation database [geo]. While the accuracy of geolocation is at times limited in its precision, it has been shown to be highly accurate at a country level [23]. To best build the mapping between IP address and owning AS, we consulted snapshots of BGP routing tables from RouteViews [RouteViews] taken the same day that the traceroute measurement was done, and assigned ownership to the AS which originated the path for the block of IP addresses that day. In order to assign which legal jurisdiction has control over that AS, we simply map the AS back to the country it is registered in based on the IANA registry [ian].

Processing a traceroute, visualized in Figure 3.2, involved converting an IP level path into an aggregate path of mapped country, AS tuples. When building this new path, we *do not* add any hops from the original traceroute measurement where we do not know the AS or country that the IP address belongs to, nor do we include hops which do not respond with an ICMP message to the traceroute. This implies that our measurements is a *lower bound on the countries and ASes that the path exposes traffic to.* After building this new path, we compress repeated instances of the same tuple down to a single instance. Finally, to expand on the number of source/destination paths, we inferred the path from each hop contained in the built path to the destination, rather than simply that of the originating node, an inference made possible because of routing routing protocols making forwarding decisions based only on the destination of the packet. We then label paths as either geographically normal or "bad" based on sets of expected countries to appear on the path constructed in Section 3.0.1.

3.0.3 Dataset Overview

The result was a dataset of over 14.5 billion paths from January 1, 2017 to October 31, 2017 to test against our definition of normal. The set of paths involved 13221 distinct ASes and 218 different countries. Figure 3.3 shows the distribution of the lengths of the paths measured, where length is defined as the number of AS/country tuples rather than the number of IP

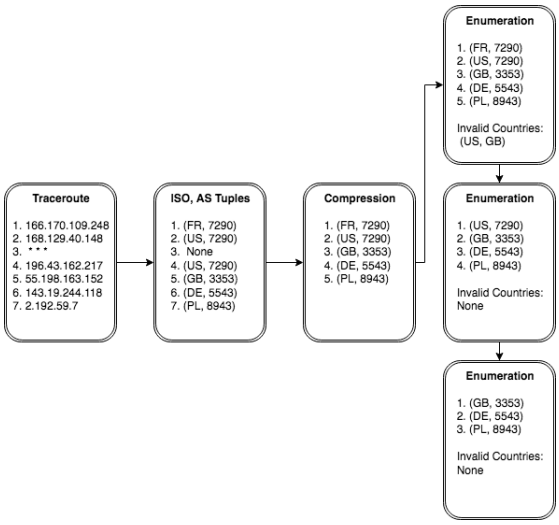

Figure 3.2: Visualizing the process of processing a single traceroute measurement

hops in the unprocessed path. Note that we expect to see more paths of shorter length due to the fact that when we see a path of length n, we always also see paths of length $n-1$, $n-2, \ldots, 1$ as a result of our enumeration process. Following an approach utilized in the rest of the paper, we will briefly examine the high level properties of the dataset at the national, AS, and regional level.

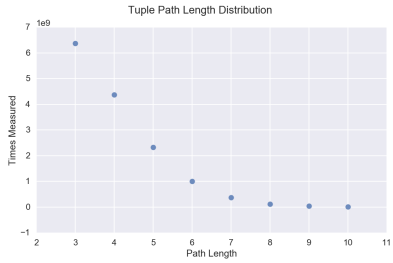

Figure 3.3: Distribution of tuple path lengths in our dataset.

When we look into the distribution of data at a country level, we see that on average, a country is involved in approximately 179.8 million paths. This distribution is not uniform however, as an example, as shown in Table 3.1 we see that each of the top 9 countries is involved in an *order of magnitude* more paths than the average case. Furthermore, the United States is actually physically present in *over 80%* of the paths in our dataset.

If we consider the dataset at an AS level, we find that the average number of paths each unique AS is seen in is approximately 3.6 million. Table 3.2 shows the top 10 ASes in terms of unique paths seen in. Something interesting to note is that only 6 of the top 10 most seen

Table 3.1: Countries Involved in the Most Paths

Country	Number of Paths Seen in
United States	11,791,492,339
Great Britain	3,800,257,024
China	1,661,946,147
France	1,348,298,693
Germany	1,338,891,831
Japan	1,255,653,909
Ireland	1,100,471,082
Australia	1,019,086,757

ASes actually contain ARK monitors. This highlights the importance of the other 4 ASes in them in terms of quantity of internet traffic transited. However, it is also important to clarify that an AS that is adjacent to a monitor AS which is overwhelmingly utilized by the monitor's AS, will be seen in *more* traffic than the monitor AS, due to the way we process paths, as shown in Figure 3.2.

Table 3.2: ASes Involved in the Most Paths. A majority of which are legally registered in the United States

ASN	Number of Paths Seen in	Legally Registered In
3356	2,962,654,510	US
2914	1,695,105,059	US
174	1,684,565,281	US
6939	1,417,839,004	US
3257	1,197,307,452	DE
6453	1,118,512,068	US
209	951,915,100	US
4134	777,856,768	CN
7018	728,254,273	US
6762	680,694,661	IT

Another point of interest in Table 3.2 is that 8 of the top 10 ASes in terms of most unique paths seen are legally registered in the United States. While this coincides with the United States physically being involved in over 80% of the paths we have measured, it is important to note that many ASes actually have physical infrastructure outside of the country they are legally registered in. Of the 13,221 ASes we have seen in our measurements, 1,880 ASes have been observed to have infrastructure in more than 1 country. Figure 3.4

shows these 1,880 ASes and the number of unique countries that are observed hosting their physical infrastructure. Here we see that over 800 ASes have infrastructure in more than two countries, and one AS was actually measured to have infrastructure in 127 different countries. We further investigate this trend of the legal exposure vs physical exposure in Section 4.2.

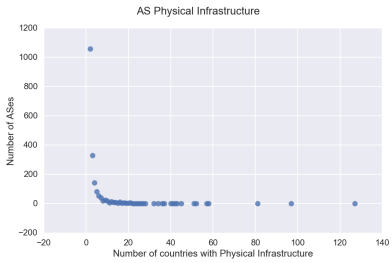

Figure 3.4: Of the 1,880 ASes we see in our measurements, over 800 ASes have infrastructure in more than 2 countries.

Finally, we look at the dataset through the lens of regional, at a continent level, paths. In Table 3.3 we see that both the Americas and Europe are physically present in over half the paths measured. Furthermore, from what we learned in Table 3.1 we see that the United States is actually present in more paths *than any entire region*, with the exception of its own.

Table 3.3: Number of unique paths we see each region in

Region	Number of Paths Seen In
Africa	1,865,890,628
Americas	12,180,787,552
Asia	5,595,364,972
Europe	8,696,139,141
Oceania	1,075,526,183

Chapter 4

Evaluation

4.1 Physical Exposure

We want to quantify the amount of needless geographic exposure from one entity to another. As a metric of normalcy, we have defined the **degree of normality (DoN)** between a particular source and destination as:

$$DoN = \frac{total\ "normal\ paths"\ seen}{total\ paths\ seen} \tag{4.1}$$

Due to the trivial nature of paths length less than two being correct, we only only considered paths containing three or more AS, country tuples in our measurements. In this Section we consider only physical exposure of traffic, in Section 4.2 we explore what occurs when we additionally consider legal pressure nation states can exert.

Over the entirety of the paths we examined, the total DoN was 0.510. We start our investigation of DoN by looking at the entirety of the measurements from a few different angles. First, we examine DoN given the length of the compressed AS, country tuple path. Figure 4.1 shows that as the number of hops in the path increases, the the DoN continually degrades. Only paths with 3 hops have a DoN above the global average and any paths containing 6 or more hops have a DoN 0.2 and below. We do see a negligible number of paths beyond length 10, which tend to be inflated in length due to alternating hops between

one AS in different countries or two ASes in the same country, which could indicate a routing loop when the measurement was conducted.

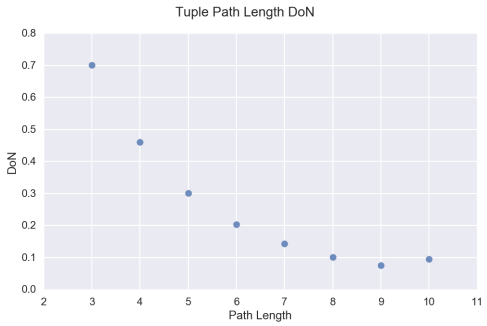

Figure 4.1: Path Length DoN

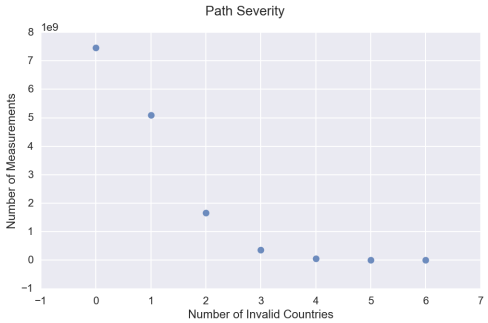

Figure 4.2: Severity of Non-Normal Paths

Figure 4.2 gives us insight into the number of geographically illogical countries appearing on each wrong path that was observed. We term any country that is not geographically normal between sender and receiver, but appears in a transit capacity along the path as a *benefactor*. We see that the majority of paths expose data to one or two benefactors. However, a non-negligible amount of paths still get exposed to traffic to more than 2 benefactors. Keep in mind that this is only one possible metric of path error severity. One might also wish to consider "intangibles" related to the relationship between, source, destination, and benefactor nations. For instance, one country may want to avoid exposing internet traffic to specific countries for political reasons. Another potential metric for path error severity might consider distance from the last country to the extraneous country. As an example, consider the path shown in Figure 4.3. This path could be considered more severely wrong since it crosses the Atlantic Ocean and as opposed to a path which slightly deviated inside the European continent.

Figure 4.3: Visualizing of a geographically illogical path from France to Germany. This path is contained inside a *single* AS. Both the United States and the United Kingdom benefit from this path.
s

We split the rest of the examining of DoN into the same three levels as in Section 3.0.3 section: country, regional, and AS. Additionally, we split scenarios for each three entities based on the role of the entity in the path: the data source, the destination, or neither the source or destination (a transit entity). At the country level we also investigate specific interesting case studies including the United States, Great Britain, and the Five Eyes, a known global intelligence sharing agency comprised of the United States, Canada, Australia, New Zealand, and Great Britain.

4.1.1 Country DoN

When examining country DoN specifically, we filter out countries who are not in a total of 1,000,000 paths in all types of DoN (source, transit, and destination). This leaves us with a total of 70 countries to examine at a nation state level. Figure 4.4 shows CDFs of the DoN of countries given their role in a path. Immediately, we see that the DoN for countries follow the same curve given any role in the path. Furthermore, the average standard deviation of the three types of DoN for countries is only 0.094, suggesting that the curves are not matching up by chance but instead because countries lie in approximately the same area on all three data series. We do see exceptions to this trend. Of particular note is Great Britain, who has a near global average source and destination DoN (0.490 and 0.545) while having a transit DoN of just 0.215. Another country of interest is China, who actually has a higher than global average transit DoN at 0.593, but a much lower source and destination DoN, at 0.384 and 0.417 respectively.

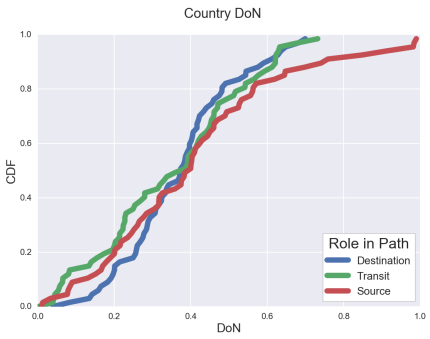

Figure 4.4: DoN at a country level

While the global average DoN is 0.51, we can see from the CDF that this average is biased by a minority of countries with more paths than others which are more successful in avoiding needless exposures. Only roughly a quarter of countries have on average at or above the global DoN. The North American countries represent an example of countries that have exceptionally high DoN. We see in Table 4.1 all three countries have extremely high source and destination DoN, which could be contributed to the vast amount of infrastructure in the United States and the proximity of the two countries. We do see, however, that the United States has a much lower transit DoN than its neighbors. We further investigate this trend in Section 4.1.3, where we look at the United States specifically splitting the cases where it is transiting traffic that addressed to a destination inside the United States vs traffic bound for foreign destinations. On the other hand, several countries preform quite poorly. Half the countries have a DoN of 0.40 or less, with a third of countries having a DoN of less than 0.30 in some roles.

Table 4.1: Every DoN in North America is over the global average, with the exceptions of the United States exhibiting a significantly lower Transit DoN than its neighbors

Country	Source DoN	Destination DoN	Transit DoN
Canada	0.645	0.628	0.623
Mexico	0.698	0.649	0.732
United States	0.741	0.700	0.513

A visualization of which countries are the biggest benefactors of non-normal paths can be seen in Figure 4.5. Through this visualization we immediately see that the United States and Great Britain benefit the most from erroneous paths. In fact, both countries benefit from at least 3 times the amount of erroneous paths than the next closest country.

Countries which most often appear as a transit provider are powerful potential path based adversaries, but how many of the observed paths in the largest path based adversaries are geographically logical? Table 4.2 shows the countries who transit the most traffic, along with what fraction of the observed paths is normal. What we see is 8 of the 10 countries who transit the most traffic have a transit DoN of below the global average DoN, with only China and the United States being above the average. *None* of the European countries in 4.2 have a transit DoN over 0.300.

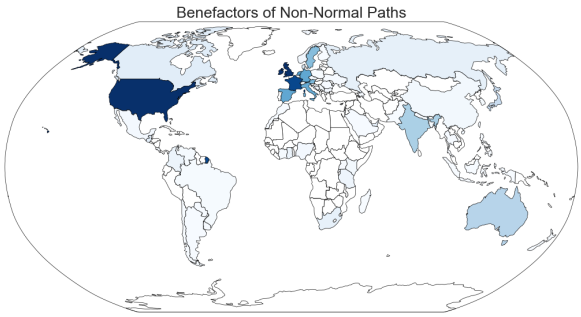

Figure 4.5: The United States and Great Britain benefit from *over 3 times* the amount of non-normal paths than the next closest country

Table 4.2: The United States transits over 4 times as much traffic as the next closest country.

Country	Ratio of All Paths Transited	Transit DoN
US	0.639	0.512
GB	0.186	0.215
FR	0.058	0.214
JP	0.055	0.471
IE	0.050	0.134
DE	0.044	0.329
SG	0.041	0.268
ES	0.040	0.155
CN	0.039	0.593
NL	0.039	0.228

While it is important to see what the raw magnitude nations benefit from the paths they are in, it is also valuable to consider the ratio of the number of paths a nation benefits from to the total amount of traffic the nation transits. Figure 4.6 visualizes this ratio. We see that many countries that have disproportionately small DoNs relative to their network footprint

are coastal nations, which could suggest that many of the paths they benefit result from oceanic cable landing points.

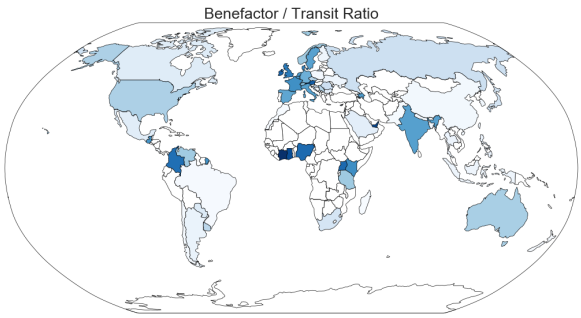

Figure 4.6: In coastal countries, we see a much higher ratio of irregular paths to total paths transited.

Finally, we consider the case where an entity is only transiting traffic on a path: that is, *they are not the source or the destination*. This is a particularly important case due to the fact that many of the paths inside larger countries can end up never leaving the country, leaving them to be the source, transit, and destination entity in the path. What we find when considering this foreign-transit-only DoN is that many of the countries who previously had relatively high transit DoN drop *drastically*. Table 4.3 shows us just how bad these countries drop in DoN when being a foreign transit entity on a path. Particularly, we see the United States go from having a transit DoN of 0.512, to a transit only DoN of 0.082. *The United States can see over 23% of the traffic on the internet that is not starting or ending within its borders, and only 8% of it is considered geographically normal.*

4.1.2 Regional DoN

Next, we examine DoN at a regional level, where we see in Table 4.4 that certain regions have a better degree of normality when the path is to them than from them. For instance, the Americas (North, South, and Central) have a higher than average DoN when the path ends or starts there, but a much lower DoN when they are found transiting data. Part of this could be explained by the smaller number of countries in the Americas, particularly North America. When having more adjacent countries, such as in Europe, there are more choices of countries to route through, and could naturally bring down the DoN for the region. Furthermore, we see that Asia has a better transit DoN than paths that start or end in the

Table 4.3: When considering foreign transit, *every* country performs incredibly poorly. The United States can see over 23% of foreign traffic and only 8% of *that* traffic is geographically normal

Country	Transit Only / Total Paths	Transit Only DoN
US	0.234	0.082
GB	0.152	0.129
FR	0.049	0.151
IE	0.043	0.064
DE	0.031	0.206
IT	0.027	0.157
NL	0.027	0.101
ES	0.025	0.068
MU	0.025	0.001
SE	0.025	0.245

region. Finally, we see that as a whole, the only region that has *any* above average DoN is the Americas, with Africa actually performing more poorly than the Oceanic region.

Table 4.4: Most regions have close to the same source, destination, and transit DoN.

Region	Source DoN	Destination Don	Transit DoN
Africa	0.174	0.192	0.182
Americas	0.692	0.644	0.522
Asia	0.421	0.412	0.445
Europe	0.400	0.459	0.267
Oceania	0.384	0.296	0.322

Table 4.5 examines DoN on a region to region basis. When staying inside a region, only the Americas and Oceania have above the overall average DoN. Interestingly, a traceroute traveling from Europe to the Americas has a better chance of following a normal path than a path staying inside Europe. Table 4.5 also shows that the DoN from one region to another is highly symmetrical; the DoN traversing *from* region 1 to region 2 is typically close to the DoN when traversing from region 2 *to* region 1.

4.1.3 Case Studies

We now present case studies of some of the most interesting countries in our measurements: starting with the United States. Of the over 14 billion examined paths, the United States

Table 4.5: We see that DoN from region to region is highly symmetrical

To \ From	Africa	Americas	Asia	Europe	Oceania
Africa	0.423	0.151	0.043	0.218	0.009
Americas	0.217	0.814	0.546	0.590	0.640
Asia	0.067	0.649	0.439	0.177	0.174
Europe	0.195	0.642	0.202	0.386	0.133
Oceania	0.016	0.323	0.382	0.102	0.878

showed up in roughly 80% of all of them. Even though they have a transit DoN of above the global average, the raw amount of traffic that they get exposed to makes them the biggest benefactor of non-normal paths in terms of raw volume. The United States ends up showing up in 3,004,163,418 paths that are geographically illogical. Further examining this trend, Of the 4,246 different source/destination country pairs we have looked at, the United States benefits from at least one path in 3,848 of them.

The next case study we examine is the Great Britain. Great Britain appears in roughly 25% of all the paths we have examined. Great Britain is a particularly interesting nation in that the DoN *to* and *from* the United Kingdom is over *double* the DoN of paths that transit DoN. Great Britain actually benefits from being in 1,885,585,255 paths that we do not expect them to be in. To further investigate this abnormally low transit DoN versus source and destination DoN, we visualize the sources countries that great Britain benefits from in from in Table 4.6. The table shows us that while many of the main benefactors lie in Europe, their reach is not contained to the region. The top 10 countries they benefit from includes the United States and portions of Africa and Asia. Great Britain's influence in paths in Africa partially explains the exceptionally low regional DoN we see from them as a region.

We move away from single nations for our next case study and look at the Five Eyes, an intelligence alliance including Australia, Canada, New Zealand, the United Kingdom, and the United States. This is a particularly interesting case study given allegations in the past several years over massive surveillance, including Five Eyes nations spying on citizens of other Five Eyes nations in order to circumvent legal protections. Table 4.7 shows the disparity between the transit DoN of the Five Eyes versus the source and Destination. a path is *33%* less likely to be normal when transiting these countries than when it originates

Table 4.6: Great Britain benefits from countries in virtually all regions of the globe

Source Country	Number of Paths Benefited From
United States	856,030,652
South Africa	246,732,302
Mauritius	206,726,729
Spain	205,730,419
Italy	161,752,654
China	154,269,999
Denmark	144,920,379
Rwanda	113,164,771
Japan	104,166,904
France	94,228,973

or ends there. Recall from Figure 4.4 that generally the source, destination, and transit DoN for a country are approximately equal. Also of note, the DoN for paths from a five eyes country to a five eyes country is 0.820, which is *60%* higher than the global DoN and much higher than any type of DoN of the member nations.

Table 4.7: Comparing DoN of the Five Eyes as one entity

Source	Transit	Destination
0.664	0.446	0.659

To further investigate the Five Eyes, we look at what countries the alliance benefits from the most. Figure 4.7 visualizes what countries needlessly expose their traffic to the Five Eyes. Immediately we see that their influence encompasses the entire globe, which can be explained both by the massive amount of traffic that the entity sees as a whole, and their distributed geographical presence. Considering the Five Eyes is an intelligence sharing agency, it is worth noting that 3 of the 5 countries comprising the five eyes lie in the top 10 countries that the entity befits from the most, and *all 5* are in the top 16 countries in the same list.

4.1.4 AS DoN

Finally, we look at DoN at the AS level. Since ASes comprise the entities that actually are making routing decisions, this helps illustrate what is happening at a network level.

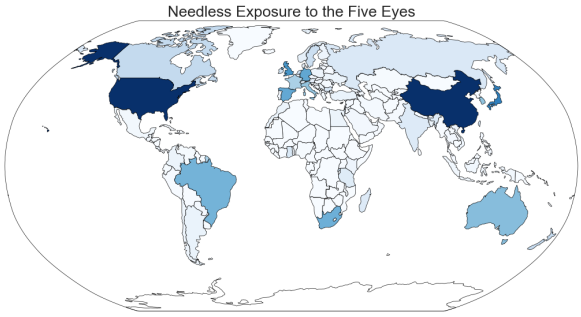

Figure 4.7: Countries the Five Eyes Benefits From

Here, we only consider ASes that are in over 50,000 paths in all types of DoN (source, destination, and transit). This leaves us with 560 core transit ASes outside the Internet's default free zone to consider. Figure 4.8 shows that in general, the DoN for paths transiting most ASes and ending in most ASes follow the same curve. However, the curve for paths *from* ASes demonstrates a trend of higher DoN, suggesting that a minority of the sources, by AS, contribute to poor DoN. We also, as one would expect, see a higher average standard deviation in AS DoN than countries, at 0.148.

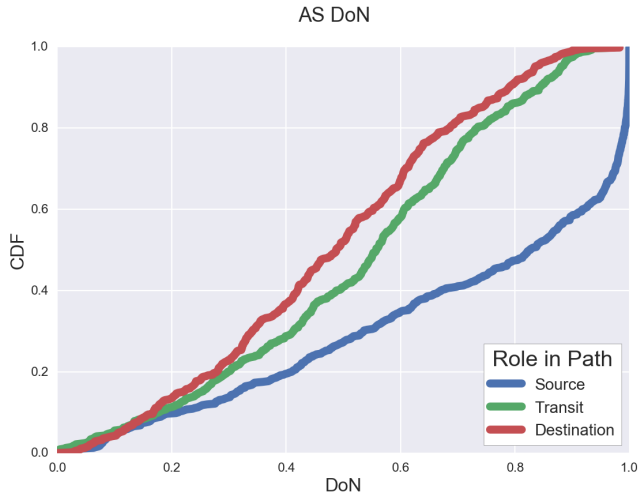

Figure 4.8: General AS DoN

Next we want to examine if, when an AS introduces transit infrastructure that is not located in a geographically normal country into a path, is that illogical country consistent for the AS, or do ASes expose traffic to a diverse collection of countries? It turns that of the 560 ASes we consider, 383 (68%) of them expose traffic to a single, consistent, illogical country. Of course outliers exist, with one AS actually seeing non-normal data in 32 countries. We

are also interested in, when an AS exposes traffic to a single country, is that country where they are legally registered at? In all 383 countries that benefit from a single country those ASes only benefit the country they are registered in.

To further investigate DoN at an AS level, we examined when erroneous countries show up in the path, attempting to answer the question: do "bad" countries get visited on the boundaries of ASes, or when traveling inside an AS? It turns out that the majority of the errors that occur in paths happen on the edge of ASes. Of the over 10 billion errors we have examined, 6,445,982,120 (64%) happen on the edge of ASes. Certainly in some of these instances geographically illogical decisions are being made for financial gain due to AS relationships, however, in 44% of the appearances of illogical countries, the error happens *inside* an AS, where there is not an immediately clear business reason to make such a decision.

Next, we look at these internal AS errors in one specific case, single AS paths. In Figure 4.3 we saw an example of path with a single AS that was considered geographically non-normal. We attempt to further investigate this specific trend here. Of the 560 ASes we consider in this Section, 424 of them were actually not involved in a single non-normal path that only included themselves. It is important to note that this could also mean that they were never involved in a path that only included themselves. This means that the 94,123,432 non-normal paths that involved a single AS were influenced by only 136 total ASes. Figure 4.9 looks at the distribution of these ASes in terms of how many single AS paths they were involved in that were non-normal. There we see that even of the remaining 136 ASes involved in these paths, a small minority of them comprise the majority of the paths.

Now let us examine the countries that benefit the most from these single AS paths. We see in Table 4.8 that the top 10 countries in terms of the number of single AS paths they benefit from. Immediately, we once again see the United States at the top of the list. Furthermore, we see many European countries, which could be explained due to ASes having infrastructure in multiple European countries as well as the general low DoN of Europe as a Region.

Finally, we look at what source/destination country pairs are involved in these non-normal single AS paths. Table 4.9 shows 4 of the top 5 source/destination pairs in terms of

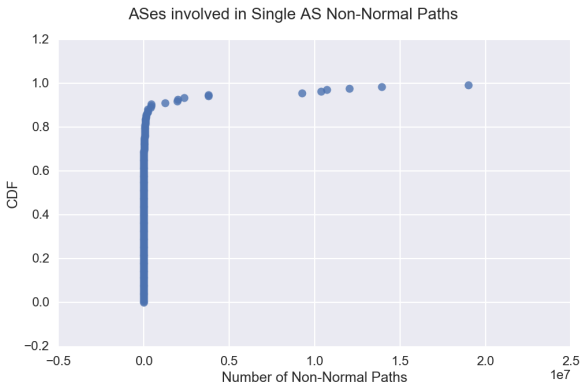

Figure 4.9: Number of non-normal single AS paths ASes were involved in

Table 4.8: Countries which benefit from geographically illogical single-AS paths.

Country	Number of Single AS Paths Benefited From
United States	27,166,822
Mauritius	12,115,955
Ireland	10,882,008
Austria	8,567,360
France	7,284,654
Singapore	6,222,979
Netherlands	5,468,196
Denmark	4,488,637
Canada	3,397,747
India	2,610,547

number of paths end up in the same country they started in. This is particularly worrying, due to the fact that a single AS has full autonomy over the path, the path *should* never leave the nation it started in, yet we see that these paths do end up outside of their source and destination.

4.2 Legal Exposure

While looking at paths from a purely physical standpoint is important, we also want to consider how geographically normal a path when additionally considering countries that can exert legal control over an AS transiting traffic. To do this, we take every physical AS, country tuple path we have and map the as back to the country it was registered in. If we

Table 4.9: 4 of the top 5 source/destination pairs for single AS non-normal paths are the same country. This is particularly worrying, as the AS has full control over the path and still leaves the country.

Source Country	Destination Country	Non-Normal Single AS Paths
United States	United States	15,706,670
Spain	Spain	7,405,173
Great Britain	Great Britain	3,769,939
Tanzania	France	3,243,382
Netherlands	Netherlands	2,998,219

do not know the country of registration for *any* of the ASes on the path, we simply skip the path and do not consider it.

In doing this, we still ended up with a set of over 14.5 billion paths to consider the normality of from a legal standpoint. If we ignore the countries where infrastructure is physically located, and purely legal control over the AS, the global DoN is 0.506. This is only slightly lower than the physical DoN. Considering many ASes only have infrastructure in one country, this DoN almost mirroring the physical DoN makes sense.

Perhaps more interesting, is when we consider the *union* of physical and legal exposure. To do this, we consider every path as if they traversed both the legal and physical countries on the path and compare that set of countries to our set of normal countries for the physical location of the source/destination pair. Figure 4.10 shows how many countries were typically added to the path when considering the union rather than simply the physical exposure. We see that about 80% of the paths actually traverse the same set of countries, however around 20% of path add one or more countries when legal pressure is also considered.

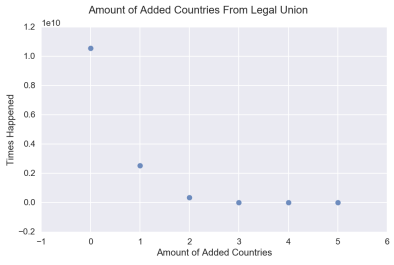

Figure 4.10: Number of Added Countries by Unioning the Legal and Physical Countries traversed

Finally, we consider the DoN of the above unioned set. In other words, we consider what fraction of the total paths are normal when we consider both the legal and physical countries they traverse. As one would expect after looking at Figure 4.10, the overall DoN of these unioned paths is 0.425, almost 10% lower than the physical and legal DoN. Figure 4.11 looks at this union DoN as the length of the path grows. There, we see that as the path gets longer, the DoN drops quickly. This makes sense due to the fact that every added hop on the path could potentially add 2 different countries to the countries traversed. Finally, we consider the above union DoN at a country level, breaking down further into when the country is the physical source and the physical destination of the path. Figure 4.12 looks exceptionally similar to Figure 4.4.

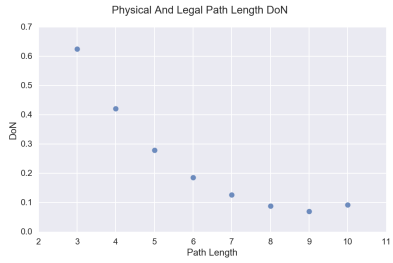

Figure 4.11: DoN as a function of path length.

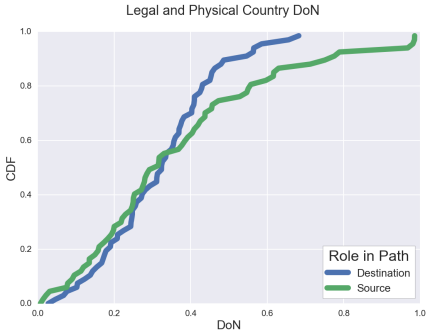

Figure 4.12: CDF of DoN at a country level when the union of physical and legal exposure is considered.

While one might automatically assume that the added legally bad paths can be attributed to the United States due to the number of ASes registered to the US. When we examine who the added benefactors of legal paths are this turns out to not be true. Table 4.10 shows us the top 10 countries in terms of paths benefited from legally. We immediately see that this

does *not* mirror the physical benefactors of paths, instead many European countries jump to the top of the list, with the US appearing in position number 6.

Table 4.10: The nations who benefit the most legally from paths do not necessarily match the nations who benefit physically.

Country	Legal Benefactor Count
Germany	684,705,398
Bulgaria	222,291,061
Netherlands	213,295,028
Norway	170,369,826
United States	152,841,248
Hong Kong	151,969,558
Mauritius	145,004,667
Israel	140,153,468
Nepal	76,406,543

Chapter 5

Conclusion

In this work, we have examined the extent the Internet's routing infrastructure needlessly exposes network traffic to different nations. We have developed a unique infrastructure for doing so by creating defining what a "normal" geographic path is between two countries through the use of convex hulls. Next, we quantified the amount of normal and irregular traffic between two entities through the use of our defined degree of normality. In the over 14.5 billion paths we have examined, 49% of them unnecessarily expose network traffic to at least one nation. Furthermore, we have examined what nations, regions, and ASes benefit and expose this network traffic to geographically irrelevant countries. Finally, we examined the legal countries that each measured path traversed to determine what countries have legal jurisdiction over the traversed ASes in each measured path. When considering both the physical and legal countries that each path crosses, over 57% of paths expose traffic to at least one nation.

Future Work In the future, we plan to expand our measurements to look at countries that see temporary, but marked, changes in their DoN, and attempt to establish the root cause of such changes. We are interested in examining if adversarial actions could result in a temporarily reduced DoN for nations, or if particular nations could inordinately benefit from adversarial reductions in DoN. Lastly, we wish to examine if nations can adjust their routing policies in an effort to increase their DoN, effectively reducing their exposure to nation state level path adversaries.

Bibliography

[ark] The CAIDA ucsd ipv4 routed /24 topology dataset : 1 jan 2017 - 31 oct 2017. http://www.caida.org/data/active/ipv4_routed_24_topology_dataset.xml. 2, 6, 11

[geo] Geolite database. https://dev.maxmind.com/geoip/legacy/geolite/. 12

[pop] The geonames database. http://download.geonames.org/export/dump/. 10

[ian] The internet assigned numbers authority. https://www.iana.org/numbers. 12

[sha] The matic mapping. http://thematicmapping.org/. 10

[rip] The ripe atlas dataset. https://atlas.ripe.net/. 6

[7] (2013). The NSA uses powerful toolbox in effort to spy on global networks. http://www.spiegel.de/international/world/the-nsa-uses-powerful-toolbox-in-effort-to-spy-on-global-networks-a-940969.html. 1, 7

[8] Anderson, N. (2006). Att engineer: Nsa built secret rooms in our facilities. https://arstechnica.com/uncategorized/2006/04/6585-2/. 1, 7

[9] Anwar, R., Niaz, H., Choffnes, D., Cunha, Í., Gill, P., and Katz-Bassett, E. (2015). Investigating interdomain routing policies in the wild. In *Proceedings of the 2015 ACM Conference on Internet Measurement Conference*, pages 71–77. ACM. 5

[10] Apostolaki, M., Zohar, A., and Vanbever, L. (2016). Hijacking bitcoin: Routing attacks on cryptocurrencies. *arXiv preprint arXiv:1605.07524*. 6

[11] Ereche, M. (2010). Odd behaviour on one node in i root-server. https://lists.dns-oarc.net/pipermail/dns-operations/2010-March/005260.html. 1, 7

[12] Feamster, N. and Dingledine, R. (2004). Location diversity in anonymity networks. In *Proceedings of the 2004 ACM workshop on Privacy in the electronic society*, pages 66–76. ACM. 6

[13] Gebhart, G. (2017). We're halfway to encrypting the entire web. https://www.eff.org/deeplinks/2017/02/were-halfway-encrypting-entire-web. 6

[14] Greenberg, A. (2014). Hacker redirects traffic from 19 internet providers to steal bitcoins. https://www.wired.com/2014/08/isp-bitcoin-theft/. 7

[15] Johnson, A., Wacek, C., Jansen, R., Sherr, M., and Syverson, P. (2013). Users get routed: Traffic correlation on tor by realistic adversaries. In *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*, pages 337–348. ACM. 7

[16] Karlin, J., Forrest, S., and Rexford, J. (2009). Nation-state routing: Censorship, wiretapping, and bgp. *arXiv preprint arXiv:0903.3218*. 11

[17] Malkin, G. (1993). Traceroute using an ip option. RFC 1393, RFC Editor. 6

[18] Moy, J. (1998). Ospf version 2. STD 54, RFC Editor. http://www.rfc-editor.org/rfc/rfc2328.txt. 5

[19] Oran, D. (1990). Osi is-is intra-domain routing protocol. RFC 1142, RFC Editor. 5

[20] Rekhter, Y., Li, T., and Hares, S. (2006). A border gateway protocol 4 (bgp-4). RFC 4271, RFC Editor. http://www.rfc-editor.org/rfc/rfc4271.txt. 4

[RouteViews] RouteViews. RouteViews Dataset. http://www.routeviews.org/. 12

[22] Savage, D., Ng, J., Moore, S., Slice, D., Paluch, P., and White, R. (2016). Cisco's enhanced interior gateway routing protocol (eigrp). RFC 7868, RFC Editor. 5

[23] Shavitt, Y. and Zilberman, N. (2011). A geolocation databases study. *IEEE Journal on Selected Areas in Communications*, 29(10):2044–2056. 12

[24] Zetter, K. (2015). What we know about the nsa and att's spying pact. https://www.wired.com/2015/08/know-nsa-atts-spying-pact/. 1, 7

[25] Zmijewski, E. (2010). Accidentally importing censorship. https://dyn.com/blog/fouling-the-global-nest/. 1, 7

Vita

Jordan is a 5 year MS-BS Candidate in Computer Science. His research interests currently focus on privacy and network security, specifically on the Internet. As a graduate student Jordan has worked on projects ranging from investigating the activity of global botnets to attacking real-world cryptocurrency mining pools.

Jordan also received his BS in Computer Science from the University of Tennessee. While he was an undergrad at UT, Jordan worked in both industry and research environments. Specifically, he interned at OSIsoft as a Sophomore, creating an automated platform for data verification on remote servers. He also spent time at Oak Ridge National Laboratory, working in the Vehicle Security Laboratory looking for security vulnerabilities in multiple vehicles.

In his spare time, Jordan spends the warmer months of the year backpacking, with his longest trip being the 487 mile Colorado Trail running from Denver to Durango, Colorado. He has also spent time backpacking the Appalachian Trail through Great Smoky Mountain National Park.

In the fall of 2018 Jordan is starting a PhD program at Princeton University, where he plans to continue researching the intersection of network security and privacy