Cross Continental Access of Information
- And the Boarder Gateway Protocol

Korskontinental åtkomlighet av information och Boarder Gateway protokollet

Johan Frisk

Supervisor : Vengatanathan Krishnamoorthi
Examiner : Niklas Carlsson
Copyright

The publishers will keep this document online on the Internet - or its possible replacement - for a period of 25 years starting from the date of publication barring exceptional circumstances.

The online availability of the document implies permanent permission for anyone to read, to download, or to print out single copies for his/hers own use and to use it unchanged for non-commercial research and educational purpose. Subsequent transfers of copyright cannot revoke this permission.

All other uses of the document are conditional upon the consent of the copyright owner. The publisher has taken technical and administrative measures to assure authenticity, security and accessibility.

According to intellectual property law the author has the right to be mentioned when his/her work is accessed as described above and to be protected against infringement.

For additional information about the Linköping University Electronic Press and its procedures for publication and for assurance of document integrity, please refer to its www home page: http://www.ep.liu.se/.

© Johan Frisk
Abstract

The usage of technical equipment and the interconnection of all these devices has become essential to our every day life and to provide as good service as possible for the internet of today. However, many have the opportunity to inspect the data as it is being rushed across the globe to provide users with information at the press of a button.

This study examines the routing patterns from hundred thousands of trace routes going from user to server around the globe. Using this trace data, we demonstrate not only the number of companies involved in data transfers on other continents then where they are currently not residing, but also give some insight into the amount of routes pertaining to same continent data transfers which takes paths not contained within that same continent.

In doing so, we try to explain the inner workings of the protocols used within the process of data transfer, answer a number of security related questions regarding the protocols used, as well as discuss which other circumstances can affect the decision making process.
# Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>iii</td>
</tr>
<tr>
<td>Acknowledgments</td>
<td>iv</td>
</tr>
<tr>
<td>Contents</td>
<td>iv</td>
</tr>
<tr>
<td>List of Figures</td>
<td>vi</td>
</tr>
<tr>
<td>List of Tables</td>
<td>vii</td>
</tr>
<tr>
<td>1 Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Motivation</td>
<td>1</td>
</tr>
<tr>
<td>1.2 Aim</td>
<td>1</td>
</tr>
<tr>
<td>1.3 Research Questions</td>
<td>2</td>
</tr>
<tr>
<td>2 Theory</td>
<td>3</td>
</tr>
<tr>
<td>2.1 The Internet</td>
<td>3</td>
</tr>
<tr>
<td>2.2 The Border Gateway Protocol</td>
<td>6</td>
</tr>
<tr>
<td>3 Method</td>
<td>13</td>
</tr>
<tr>
<td>3.1 Refinement</td>
<td>13</td>
</tr>
<tr>
<td>3.2 Physical Placement Data</td>
<td>13</td>
</tr>
<tr>
<td>3.3 Extraction</td>
<td>14</td>
</tr>
<tr>
<td>3.4 Enumeration</td>
<td>15</td>
</tr>
<tr>
<td>4 Results</td>
<td>16</td>
</tr>
<tr>
<td>4.1 Multi-Continent Data Routes</td>
<td>16</td>
</tr>
<tr>
<td>4.2 Physical Multi-Continent Data Routes</td>
<td>17</td>
</tr>
<tr>
<td>5 Discussion</td>
<td>19</td>
</tr>
<tr>
<td>5.1 Results</td>
<td>19</td>
</tr>
<tr>
<td>5.2 Method</td>
<td>19</td>
</tr>
<tr>
<td>5.3 The Work in a Wider Context</td>
<td>19</td>
</tr>
<tr>
<td>5.4 Future Work</td>
<td>20</td>
</tr>
<tr>
<td>6 Conclusion</td>
<td>21</td>
</tr>
<tr>
<td>Bibliography</td>
<td>22</td>
</tr>
<tr>
<td>A Database Diagrams</td>
<td>24</td>
</tr>
<tr>
<td>A.1 EER Diagram</td>
<td>25</td>
</tr>
<tr>
<td>A.2 UML Diagram</td>
<td>26</td>
</tr>
<tr>
<td>B Code Examples</td>
<td>27</td>
</tr>
<tr>
<td>Section</td>
<td>Page</td>
</tr>
<tr>
<td>---------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>B.1 Route Parser Example</td>
<td>28</td>
</tr>
<tr>
<td>B.2 IP Lookup Example</td>
<td>29</td>
</tr>
<tr>
<td>B.3 Continent Adder Example</td>
<td>30</td>
</tr>
<tr>
<td>B.4 Route Finder</td>
<td>31</td>
</tr>
</tbody>
</table>
List of Figures

2.1 Car going from one house to another .................................................. 4
2.2 Packet from computer to computer ...................................................... 4
2.3 Car going from a house in one city to a house in another ........................ 5
2.4 Packet going from one computer to a computer on another subnet .......... 5
2.5 Car going by plane to another country .................................................. 6
2.6 Packet sent from one AS to another ..................................................... 6
2.7 BGP Finite State Machine .................................................................... 8
2.8 Peer connections without route reflectors ............................................. 9
2.9 Peer connections using route reflector on node C ................................. 10
2.10 BGP Confederations within an AS ....................................................... 10

3.1 MySQL query using ISP company continent data .................................. 15
3.2 MySQL query using IP geolocation data .............................................. 15

4.1 Percentage of multi-continent routes ...................................................... 16
4.2 Amount of multi-continent routes per continent ................................... 17
4.3 Percentage of physical multi-continent routes ...................................... 17
4.4 Amount of physical multi-continent routes per continent .................... 18
List of Tables

2.1 Breakdown of postal address .................................................. 4
1 Introduction

“In a world where information is power, the one controlling the medium controls all.”

1.1 Motivation

Everyone working on creating customer solutions today wants to make their products more user friendly, personal and accessible. In order to do that, many turn to the powerful tool, the internet. A vast and always expanding network of computers, routers and other systems all able to exchange information with each other.

For the individual user, the case that a product needs an internet connection is nothing new. The product is connected to the router they have at home. If asked, they will know that they are connected to their home router. In some cases they will know to where that router is connected, giving the name of their Internet Service Provider (ISP). However not many people know what happens to their data once it arrives at the ISP level of the internet, nor do they know how it is managed when travelling from their ISP to a computer at the other end of the world.

1.2 Aim

This study is going to examine a dataset collected for another paper published by Gustafsson et al., containing packet routes taken for packets going within and between the EU, Asia and the US.

Using the dataset to determine if some routes decided by the Boarder Gateway Protocol (BGP), for packets originating in the same continent as the end destination, results in the packets being routed through another continent and back.

Using the data this study will also try to give insight into if data has been routed through ISPs registered in a country not within the continent.
1.3 Research Questions

1. Does internet service providers registered outside of a continent have access to data routed within it?

2. Does data with a destination and an originating source within the same continent get routed outside of the continent?
2.1 The Internet

The Internet, used by almost everyone and rapidly becoming connected to everything, this chapter is going to explain how this world wide network works and how the companies responsible for maintaining it affect it.

As reported in the paper by Gustafsson et al. and by many others, the world wide network and packets sent within it is compared to the real world and the paper mailing system. Here, houses are used to represent computers, roads used to represent network connections and intersections switches. This paper is going to use the same analogy.

IP Address

Using the analogy of computers being houses and roads being connections, we can now use cars to represent packets going between different computers on the network. In order for cars to be able to reach the correct destination, they need to have an address for the house they want to end up at. This is true in a network as well, if a packet is going to be able to travel to another computer, it will need an address for that computer. The addresses used on the internet are called Internet Protocol addresses or simply an IP addresses.

So for example, if a car would want to go from Linköping University, 581 83 Linköping to Gärdsgatan 26, 587 24 Linköping it would have to follow the signed directions first to the right street and then to the right house number. This shown in Figure 2.1. The same would be done for a packet going over the network from one computer to another. It would first need the address for the computer, for example 192.168.0.2. It would then need to be routed the correct path, first to the right part of the network 192.168.0/24 and then to the right computer, which in this case would be number 2. This shown in Figure 2.2.
2.1. The Internet

Similarly to in the world, where houses are not arranged along one single highway, not all computers are hosted on the same giant network. Instead, just like communities, villages and cities make up a whole map. Several smaller networks come together to create the world wide network we usually call the internet. This would make addressing of houses in the different locations hard if they had to have addresses that where completely different. It would also make it hard for people to find their way to the different houses. Instead when addressing a house it is customary to make a whole address using several smaller parts. Using the example address above we can see this much more clearly.

Using Gårdsgatan 26, 587 24 Linköping as the whole address we can separate the different parts it is made up from. This is illustrated in Table 2.1 and Figure 2.3. Using these parts of the whole address people now have an easier time to navigate.

Table 2.1: Breakdown of postal address

<table>
<thead>
<tr>
<th>City</th>
<th>Postal Code</th>
<th>Street Name</th>
<th>House Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linköping</td>
<td>58724</td>
<td>Gårdsgatan</td>
<td>26</td>
</tr>
</tbody>
</table>
2.1. The Internet

In the same way a street address in the real world does not specify a house until the number in the very end, and uses the rest in order to guide you in to the right street in the right area of the right city. An IP address also only specifies the computer in question within the last digits of its long address. The rest of the address is instead used to guide the packet to the right subnetwork on the correct router with the correct part of the world wide network. An IP address is therefore broken down into two parts: prefix and ID. Here, the first part specifies the path of travel and the second specifies the computer to be reached on the This is illustrated in Figure 2.4.

![Figure 2.3: Car going from a house in one city to a house in another.](image1)

![Figure 2.4: Packet going from one computer to a computer on another subnet.](image2)

**Autonomous Systems**

Following the analogy along further we now add some distance to be driven by the car. Because of this we now also add some parts to the cars destination address. In the case when a car is to go from Belarus to Sweden it would have to be driven in to a plane going between the two countries. In order to find the correct route one would need to add the country name to the destination address. Using the above example, the full address would be Gårdsgatan 26, 587 24 Linköping, Sweden. This illustrated in Figure 2.5.

In the same way as adding Sweden to the address, in order to specify the country in which the certain address is located, one could see the use of Autonomous System(AS) numbers as adding the country of origin to an IP address.
2.2 The Boarder Gateway Protocol

In other words, an Autonomous System is an owner of a specified range of IP Addresses which, like countries, one by one makes up large parts of the world wide network. Though connections between each other they make it possible for every computer on the network to exchange information. This illustrated in Figure 2.6.

2.2 The Boarder Gateway Protocol

As visualized in the above section with the car taking a plane from Belarus to Sweden in order to go from one country to another. A packet also uses another routing protocol when going from an AS to another. Just as there are several ways to travel from one country to another, there are several protocols an AS could use to communicate to another AS. With that said there is only one protocol that is in wide use in the world today and that is the Boarder Gateway Protocol (BGP).

Leaving the analogy with the car, the rest of this chapter is going to go through the boarder gateway protocol (BGP).
2.2. The Border Gateway Protocol

What is BGP

The boarder gateway protocol (BGP) is a path vector protocol used to exchange path information between routers within an Autonomous System (AS). When used to exchange information between an AS own routers, it is often mentioned as IBGP where I stands for Internal.

BGP is also used in order to exchange information between different AS:s and while in use for this purpose often named EBGP where E stands for External.

These definitions are important as the BGP protocol often has various different configurations depending on if it is used to talk between AS:s or within a single AS. This because sending out certain path information provided from inside the AS network might badly cripple the surrounding AS:s, as well as every other AS if the wrong path information got out and got accepted as usable to pass data through.

Why it is Used

So given that the Boarder gateway protocol potentially could be, and has been in the center of bringing the internet to it’s knees in the past, why is it still in use? Because it is a very reliable protocol when properly configured and managed. Many of the faults that has involved BGP have had to do with mismanagement of sanity checks or misconfiguration due to typing errors by the engineers managing the systems.

BGP also have very good scalability in comparison to the other protocols that could be used in it’s place. This is why it also is used internally in large scale systems.

On top of it’s scalability BGP also supports multihoming, which is when several connections to different other AS:s can be set up in order to provide redundancy in routing between different AS:s.

How Does it Work

The initial setup and configuration for using BGP is a, in many ways, manual task. The routers that are going to be communicating over BGP are set up to talk to each other on standard port 179. Depending on if the routers that are going to talk to each other are placed within the same AS network, this communication does not have to be without proxy routers on the way between the two routers. In other words the routers talking to each other do not have to be neighbours (peers) if the communication is within the AS itself. However if the routers are edge routers, which means a router is to communicate with another AS, then the communication has to as if the router it is going to talk to is a peer.

The AS:s that want to exchange path information does not always have to be right next to each others networks. Because of this a number of proxy routers might need to be used in order for the edge routers to be able to communicate to each other. To solve the problem of having a proxy router when communicating EBGP a VPN tunnel is often used in order to secure communication by virtually creating a link between the two edge routers.

BGP Handshake

After the initial configuration and setup now starts the linking process or handshake of the BGP. This which is constructed in a way of a finite state machine.
As can be seen in Figure 2.7, there are six states which the router goes through when setting up the connection.

1. Idle
2. Connect
3. Active
4. OpenSent
5. OpenConfirm
6. Established

Going through these states as if setting up a new connection between two BGP talking routers. The router would start at the Idle state setting up the required systems.

After the required systems are ready the router will try to set up a Transmission control protocol (TCP) connection to a neighbouring router. This creates a transition to the Connect state.

If the router successfully creates a connection, it will then transition into the OpenSent state, if not it will start a connection retry timer and transition into the Active state. When the timer is done, it will then move back to the Connect state and retry setting up a connection.

When entered the OpenSent state, the router will send an open message which is then expected to be answered with a corresponding open message from the other. In the event of getting an open message back, the router then transitions over to the OpenConfirm state, where it then proceeds to start sending keep alive messages expecting the same from the corresponding router.

When the router receives keep alive messages back from the other router a connection is now considered established and the router transitions over to the Established state. It can now start sending and receiving information.
During this procedure other configuration options are also determined regarding recovery method to use for this session as well as multiprotocol extensions [15]. If multiprotocol extensions are negotiated, information regarding network layer reachability (NLRI) [12], is sent in order to enable different connection typology alternatives [2].

**BGP Connectivity Alternatives**

When storing routing information BGP uses a local routing table. This creates a scalability issue in the basic case when IBGP routing is created in a full mesh structure in which every node in the internal AS network is directly connected to each other. In order to amend the scalability problem other types of connection topologies are used.

**Route Reflectors**

Following the rules within the IBGP protocol, internal routers that receive new routing information, which is regarded as providing a better route, are to transmit this to its peers. The peers in turn are not to transmit this any further, because of this every node needs to have a direct connection to the other nodes within the AS. This illustrated in Figure 2.8

![Figure 2.8: Peer connections without route reflectors.](image)

By relaxing the rule of not retransmitting between IBGP peers, route reflectors (RR) are constructed to retransmit the new routing information from one peer to others. This removes the need for the first peer to have connections to the other ones [16]. Shown in Figure 2.9

![Figure 2.9: Peer connections using route reflector on node C.](image)

**BGP Confederations**

Just as route reflectors can be used in order to reduce scaling problems within AS:s, BGP
confederations can be used in order to create sub-AS:s inside a very large one, as shown in Figure 2.10. This making a group of geographically close peers look like one within the greater network.

Figure 2.10: BGP Confederations within an AS.

**BGP Security**

As mentioned before, the BGP protocol is not as secure as one would want it to be. Because of the system’s built-in functionality of automatic updates of routing tables, it becomes very susceptible to misinformation attacks. Certain configurations can be made to mitigate this, making the system less susceptible as it would sanity check updates of routes. Updates to the BGP protocol has also been done to further create a more secure updating procedure using public keys in order to validate routers before accepting new routing information. This update, going under the name secure BGP (S-BGP).

**Black Hole**

In the late 1990s, the United States government asked Peter Zatko, also known as Mudge within the hacker collective the L0pht, if it would be possible to take down the internet. Peter answered that it would take 30 minutes. This by targeting the BGP protocol bringing it down using attacks such as black holing.

This technique is executed by introducing new routing information into the network making it send information to a place which would then drop the incoming traffic. This makes the IP prefixes targeted by the attacks unreachable from the outside world, in effect cutting off that part of the internet until the routes are corrected. An example of this was when the Pakistani ISP by accident black holed the prefixes used by Youtube. The intention was to
2.2. The Boarder Gateway Protocol

make Youtube unreachable for the people of Pakistan but because of a misconfiguration it affected the outside world as well.

**Man In The Middle**

Another insecurity in the BGP protocol is the lack of transport layer security. The information transmitted is not encrypted in any sort of transport layer security. Because of this all the information transmitted that is not secured by the server client transport layer, which now has become a more standardized method when sending information on the internet, could be read along the way.

Because of the previously mentioned lack of security in the admitting of new routing information, an attacker could introduce new routes to the network, routing traffic to their own AS routers which would then be used to read all the information flowing through them. This was shown in a Defcon talk, where the person presenting the keynote rerouted all the traffic from the conference, while reading all the information going through their routers and executing obfuscation techniques in order to diffuse the fact when doing a traceroute.

**Prefix hijacking**

A starting point for any of the above mentioned attacks is a prefix hijack. This is constructed in a manner where an AS announces a prefix, which it does not own, in order to direct traffic intended for those prefixes its way. This is only possible because of the lack of interest in a large implementation of already existing security protocols for BGP. Many of these protocols has shown promising results in simulations done in other research.

**Misconfiguration**

There are also a certain level of lack in security when it comes to configuration. Other than active attacks using the BGP protocol, there are also problems with misconfiguration due to typing errors by system administrators. Even though this would not qualify as an attack, it would be a miss not to point it out as a security flaw in the whole of the system. It has been shown several times that what could be seen as a small configuration error, has big consequence on the world wide network.

**Controlling BGP**

As mentioned in the previous chapter, the BGP protocol is a strong protocol when correctly configured. Unfortunately, there are several cases when this is not done. Because of this, there have been cases when certain prefixes have been unreachable for the rest of the internet.

**Business Agreements**

There is also another side to the BGP control coin, including the money involved in keeping this geographically large and virtually complex system, up and running. There are large amounts on money in transmitting data over the world wide web. This is in part because the several companies owning the AS:s charging for transporting the data over their network. As AS owners does this, there is often the case that the different companies create a contract, in which they would both save money transmitting data between their networks for free or for a lower cost. Because of this, companies often configure their routers to use routes which would cost less money, even though this might not always be the best path in regards to performance or speed.
3 Method

3.1 Refinement

In order to answer the main research questions of this study a refinement and extraction of the dataset was needed, in order to ease the process of analysis moving forward.

Because of the way the original dataset was gathered, an extraction of the relevant information was going to be very hard. To mitigate this a decision was made to first refine the dataset into a database. This process was carried out according to the following steps.

Labeling the Dataset

As the given dataset was gathered in a text file with a hard analyzable format, steps needed to be taken to label the data. After this was done, the data could be reformatted in order to import it into a large database table.

Implementing a Database

After constructing a UML diagram (see appendix A.2), in accordance with earlier created EER diagram (see appendix A.1), the database could be implemented. This which was done within a virtual machine running Ubuntu server, running MYSQL as the database framework.

With the database implemented, the data could now be inserted for easier analysis and indexing.

3.2 Physical Placement Data

In order to be able to answer the research questions, additional data needed to be gathered to compliment the location data of the IP:s in the data.

IP to Location

Even though the original dataset had location data in regards to the registration countries and continents of the AS:s, this was not enough in order to fully answer the research questions, as it did not specify continent for each IP specifically. Neither did it base the continent data on real
world location data, but instead used the country in which the ISP company was registered to derive the continent.

**IP Lookup**

In order to make the data within the dataset IP specific, a program was created to extract all IP:s from the routes in the dataset (see appendix B.1). The IP:s were then placed in a list to use in a IP lookup program (see appendix B.2).

The created lookup program used a public application programming interface (API) from the domain ip-api.com in order to gather information listed below.

- AS number
- Geolocation latitude
- Geolocation longitude
- Region number
- IP
- Time zone
- Zip code
- ISP
- Organisation
- Region name
- Country code
- City

After updating the original dataset, all the required information except for country to continent correlation was gathered.

**Country to Continent**

To account for this another program was created which, by using another public API from the domain named restcountries.eu, correlated the countries to their corresponding continent (see appendix B.3). After the continent correlation programs alterations to the database it was now ready to be queried for results.

**The Usage of Public APIs**

The method of this paper relies heavily on the usage of public APIs. The reliability, both in accessibility and data of these, is completely up to the owner of the service. Because of this, also the API data must be viewed from a point of criticism. This is especially true for the IP lookup data, which through research has been shown not to be fully reliable in location accuracy.

**3.3 Extraction**

All the entries in the dataset were not relevant to answer the main research questions. In order to be deemed relevant the entry needed to originate and have a destination within the same region. This while having a chosen route going out from the originating region and back again.

Because of earlier work creating a database of the dataset and its updates, these criteria could be redefined into queries extracting the relevant information. This needed to be queried in two ways in order to answer the research questions. One extracting location the same way as the original dataset did for the whole routes (see Figure 3.1), and one extracting location through the geolocation data for the specific routers.
select A.ip, D.continent, A.asnum
from machine A inner join asys B on A.asnum = B.number
inner join company C on B.company = C.id
inner join country D on C.country = D.id
where A.ip = currentIp;

Figure 3.1: MySQL query using ISP company continent data.

select A.ip, E.continent, A.asnum
from machine A inner join asys B on A.asnum = B.number
inner join geolocation C on A.geoloc = C.id
inner join city D on C.city = D.id
inner join country E on D.country = E.id
where A.ip = currentIp;

Figure 3.2: MySQL query using IP geolocation data.

The results from these queries could then be analyzed for compliance with before mentioned criteria. This was done using another program taking in the IPs for every route comparing the IP:s continent data for alignment with the set requirements on relevant entries (see appendix B.4).

3.4 Enumeration

Using the program, enumeration of the routes aligning with the criteria was done. Using the relevant routes as starting data, the process of finding relevant routes of the same criteria was done using the query shown in Figure 3.2. The answers from this query were then used to answer the second research question.
4 Results

Within the method shown in Chapter 3 the dataset was modified into a labeled format. In order to add the information required to answer the outlined research questions, two APIs were used. Because of this, uncertainty was introduced into the dataset. This affected the result in regards to the second research question, giving a set of false positives.

4.1 Multi-Continent Data Routes

In regard to the first of the two research question posed in this study. There are several instances where routes took paths, using routers owned by non continent ISP:s. Non continent ISP:s meaning ISP companies registered in a country not pertaining to the continent where the data were sent. In the diagrams these routes goes under the name multi-continent routes.

![Figure 4.1: Percentage of multi-continent routes.](image)

As can be seen in Figure 4.1, 13\%(54,861) of all the routes available in the dataset(383,430), has a path which contains non continental ISP:s.
Out of all multi-continent routes taken, EU to EU was the most common by a landslide as shown in Figure 4.2.

Other continents did occur in the dataset but as the entries of these continents were too few they where deemed not sufficient to show in these diagrams. Worthy of note is that of all the routes in the dataset EU to EU ones were in majority. This would explain the multi-continent routes ending up being majority EU based ones as well. More generally since almost all routes started in Europe, due to the original selection of Planetlab nodes, we only have sufficient data to carefully analyze the EU-to-EU case.

### 4.2 Physical Multi-Continent Data Routes

In order to answer the question regarding routes containing the paths physically leaving the continent in which it is currently routed. Using the query show in Figure 3.2, Figure 4.3 is the results using the full dataset.

The percentage of routes physically leaving the continent in which they are routed only amount to 11% (48,371) of the total number of routes (383,430), shown in Figure 4.3. This being lower than the amount of routes aligning with the multi continent requirement above.
As shown in Figure 4.4, the study of physical multi-continent routes also has a greater representation of EU to EU entries. Again, most of these differences can be accounted for by the over representation of EU to EU routes overall, but does give the notion of how often inter-continent communications are routed outside of continent. Furthermore, at this point it shall be noted that our method relies heavily on the accuracy of the API databases used, and the number of false positives therefore is difficult to assess.
5 Discussion

5.1 Results

The argument could be made that the certainty of the result is somewhat compromised by the method used in regards to the API:s used. Care and caution must therefore be taken when interpreting the results. We therefore caution any presumption to be formed based on the results presented here. The focus of this study has instead been on developing good methods to calculate these structures using open API:s. As the accuracy of the open API:s, or other reliable API:s improve, so will the accuracy of the results presented here.

5.2 Method

As mentioned both in the above section and in the method itself the usage of public API:s is something to take in to account when viewing the results of the study, as these are only as accurate as the API used.

Aside from the usage of public API:s, the decision of using a database, even though it creates a very good structure to the future of working with the data, was a very time consuming part. The questions could have been answered without it, using tools already existing for working with text files.

In effect the large data framework reconstruction places a lot of pressure on the programmer, as to not corrupt the data during the process of reframing it. This could have been avoided using the alternative approach of using tools.

5.3 The Work in a Wider Context

As this work is a study regarding companies access to information sent in continents they do not reside in, the work also has a social and ethical aspect. It is very important to note that this study does not say that companies, in the examined dataset, view all the information or interact with the datastream improperly in any way.

The point of this study is to show that the opportunity exists for companies from other continents to interact with that data as there is nothing technically keeping them from altering their own hardware. This, even though not studied directly, infers an ethical as well as social aspect within the trust given to these companies.
5.4 Future Work

The results from this study show that there are routes taken that both use out of continent companies, and routes that physically go out of the continent on its path to the correct IP address. Future work regarding ISP to ISP relations would be interesting, as this would give a wider picture regarding the situation of why certain routes were taken.

Other than a purely technical study, it would also be of interest to know how the laws of countries effect the routes taken on a wide network level.

Hopefully the work creating a MySQL database will make the before mentioned future work subjects easier.

Future work regarding the, in the time on writing, very controversial subject of net neutrality and how it can come to effect the world wide network on a BGP level provides another interesting avenue to explore.
The results contained both routes supporting the notion that out of continent registered ISP:s have hardware used in within continent paths, and the notion that such routes physically goes out of the continent.

The conclusion can be made that to the question “Does internet service providers registered outside of a continent have access to data routed within it?” the answer is yes. ISP:s does, to a certain extent, have access to the data being sent within a continent. That is if their hardware is used within the path of transfer. This as one can see companies outside of the EU have a very good presence within it. Though the data does not show the same kind of involvement from EU registered companies outside of it.

In regards to the the second research question “Does data with a destination and an originating source within the same continent get routed outside of it?”. The answer could also be simplified in to a yes. It is true that the results support the notion that such routes exists. However because of the proven inaccuracy of public API:s done in other studies, this question can not be answered with the same certainty as the first question. Through discussion with the authors, this is the reason the first of these questions where the focus of the analysis of the original dataset.


Database Diagrams
A.1 EER Diagram
A.2 UML Diagram
Code Examples
import os
import sys

def parsefile(inputfilename):
    print('> Parsing input file...')
    inputfile = open(inputfilename,'r')
    outputarray = []
    for line in inputfile:
        wiparray = cleanline(line)
        for line in wiparray:
            outputarray.append(line)
    return outputarray

def output(array):
    print('> Printing results...')
    outputfile = open('outputfile.txt','a')
    for line in array:
        outputfile.write(line + '
')
    print('> Result printed to outputfile')

if __name__ == '__main__':
    inputfilename = sys.argv[1]
    array = parsefile(inputfilename)
    output(array)
    print('> Parse done, exiting script...')
sys.exit()
B.2 IP Lookup Example

```python
import sys
import os
import socket
import json
import requests

def send(payload, url, urlpath):
    fullurl = url+urlpath
    resp = requests.post(fullurl, data=payload)
    if resp.status_code == requests.codes.ok:
        print("> Request successfully sent")
    else:
        print("> ** ERROR sending request **")
        print(resp.status_code)
        print("> Received response")
    return resp

if __name__ == '__main__':
    ipfile = sys.argv[1]
    for i in range(0, DATASETSIZE):
        payload = parsefile(startnumber, endnumber, ipfile)
        response = send(payload, url, urlpath)

        parseresponse(response)
        output = open("metafile.txt",'a')
        output.write(response.text)
        output.write("\n")
        output.close()

        print(endnumber)
        print("going down to sleep")
        time.sleep(30)
        startnumber = endnumber+1
        endnumber += 100
    sys.exit()
```
B.3 Continent Adder Example

```python
# System libs
import os
import sys
import time

# Own libs
import sqldb
import country

if __name__ == '__main__':
    config = get_db_conf()
    country_conf = get_country_conf()

    myobj = sqldb.mysql(config)
    myobj.connect()

    idlist = myobj.query("select name,id from country;")
    conobj = country.Country(country_conf)
    conobj.populate_country_list(idlist)

    conamelist = conobj.get_namelist()
    for i in range(0, len(conamelist)):
        conobj.get_country_info(conamelist[i])
        time.sleep(10)
    for name in conamelist:
        if name in conobj.regiondic:
            region = conobj.regiondic[name]
        else:
            break
        array = []
        query = ("update country set continent = %s" "where name = %s")
        myobj.query(query,(region,name))
    myobj.close()
```
B.4 Route Finder

```python
#system libs
import os
import sys

# Own imports
import routeparser
import printer
import sqldb
import arraystripper

configfilename = "configfile.txt"
inputfilename = "../datasets/dataset-ippath.txt"

if __name__ == "__main__":
    parser = routeparser.parse()
    array = []
    inputfile = open(inputfilename,'r')
    for line in inputfile:
        array.append(parser.toArray(line))
    contarray = []

    dbcon = sqldb.mysql(configfilename)
    dbcon.connect()

    query = ("select A.ip, D.continent, A.asnum "
             "from machine A inner join asys B on "
             "A.asnum = B.number "
             "inner join company C on B.company = C.id "
             "inner join country D on C.country = D.id "
             "where A.ip = %s")
```
for line in array:
    wiparray = []
    for word in line:
        cursor = dbcon.dbcon.cursor()
        cursor.execute(query,(word,))
        for (ip, continent, asnumber) in cursor:
            wiparray.append((ip, continent, asnumber))
        contarray.append(wiparray)
arstrip = arraystripper.strip()
arstriparray = arstrip.one_to_one(contarray)

routeprinter = printer.routeprinter()
routeprinter.print_file(5, arstriparray)
dbcon.close()