SEN163A Fundamentals of Data Analytics

Assignment 2

Students: Reshma Joseph (5383595), Stijn Knoop(4608046), Boris van Overbeeke (4083164), Maxim Sachs (4236262)

1 Introduction

In this assignment, we have prepared a strategy for Groote National Bank to be able to enter into the mobile banking sector by mainly identifying the best four data centre locations that GNI Bank can use for hosting in the Europe. To be able to do this, we analysed the given Autonomous Systems (AS) and Probe datasets for limitations and elaborated on the assumptions that we made. We also included both IPV4 and IPV6 datasets instead of just either IPV4 or just IPV6, such that a more representative average latency can be obtained. This also helps increase the number of countries that are sampled from the probes within an AS. Taking into account this additional input information can help us identify the best strategy possible from the given data.

2 Process of finding the best strategy

This section addresses the five aspects that are asked by exploring and describing the datasets to be able to arrive at a viable strategy for hosting the datacentres in Europe.

2.1 Data Description

Of the two datasets that were provided, the AS dataset contains 60 122 AS. For each, it contains the autonomous system number (ASN), Country code, Name, number of IPs within the AS, and the type of AS. The type can be "business", "hosting", "isp", "education" or "unk" (unknown). The RIPE probe dataset contains the AS in which each probe ID is located. A total of 11 008 probes are contained within the dataset.

2.2 Assumptions and Limitations

Latency of RIPE representative for overall AS: We make the assumption that the average latency as seen by the RIPE probes within it are representative for the latency of the overall AS. It might be possible that the RIPE probes are attached to the network within the AS in such a way that their latency is unusually affected. By then assuming their performance representative for the whole AS error is being introduced.

Not all AS have probes: 0 or more RIPE probes can be located within an AS. As the analysis in 2.3 shows, not all AS have probe within a network. This introduces the second major limitation that roughly only 8.7 % of AS have a probe. 91.3 % of AS in the EU are therefore immediately excluded from the following analysis which might mean that the best hosting locations could be missed.

Not including number of people per country in optimisation: The optimisation process for AS selection does not take into account the number of people per country, but treats each country the same. A better approach might be to apply weights to each country according to population or potential clients expected to use the service, which would

result in an average EU latency which is more representative of what the people collectively experience.

Not all countries are samples by each RIPE probe: For each AS, there is at least one country that has not sampled by any probe within the AS. If we consider a country only serviced if it has at least one sample value from the RIPE ping data, then a potentially large sampling error is introduced. A more optimal solution might be found by allowing some flexibility of this requirement for the selection of a set of AS for hosting.

2.3 ASN analysis

The objective is to identify 4 possible hosting locations within the European union, such that the total latency to this AS can be minimised. Currently, the European Union is comprised of 27 nations. Their country codes are as follows:

```
["BE", "BG", "CZ", "DK", "DE", "EE", "IE", "GR", "ES", "FR", "HR", "IT", "CY", "LV",
 "LT", "LU", "HU", "MT", "NL", "AT", "PL", "PT", "RO", "SI", "SK", "FI", "SE"]
```

As a first step of the analysis, we identify which AS are located within a European country. A total of 18 495 ASNs were found to be located in the EU. Then we check which of these AS has at least 1 probe within it and are of type "hosting". 145 AS were found to have at least 1 probe located within their network. Table 1 shows the first and last 3 AS sorted by ascending ASN. For each the number of probes within that AS is shown in the "prb_count" column.

ASN	Name	Country	prb_count
AS5404	conova communications GmbH	AT	2
AS5430	freenet Datenkommunikations GmbH	DE	2
AS5521	PlusServer GmbH	DE	1
AS203944	NTT Luxembourg PSF S.A.	LU	1
AS203953	Hiper A/S	DK	6
AS205766	Jonas Pasche	DE	1

Table 1: First and Last 3 AS sorted by ascending ASN.

2.4 Latency Computation

Next, we iterate over each sample from the RIPE dataset and select all samples which have a probe_id within one of the AS selected by the previous step and also have a destination address within the EU. Both IPv4 and IPv6 samples are included. Then based on these samples we can compute the average latency from all probes within one AS to each EU country that was sampled by these probes. The result is stored in a matrix with 128 rows x 26 columns. There is one column for each European Country and one row per AS. Malta is missing due to lack of samples from RIPE probes with type hosting.

The lowest latency per country is shown in Table 2.

2.5 Optimising for four servers

GNI can only place four servers in Europe. The optimisation is done using a modified implementation of the k-centre problem algorithm. The requirements for the solution are to use at most 4 servers, which are placed in such a way to minimise the average total latency

	min latency	ASN	AS Name
Austria	6.504087	AS47692	Nessus GmbH
Belgium	1.993071	AS47692	Nessus GmbH
Bulgaria	0.730935	AS59729	ITL LLC
Croatia	10.978371	AS12637	SEEWEB s.r.l.
Cyprus	241.208538	AS16276	OVH SAS
Czechia	1.703906	AS39790	Web4U s.r.o.
Denmark	3.031476	AS203953	Hiper A/S
Estonia	4.098356	AS49604	Zone Media OU
Finland	2.115836	AS61189	Elkdata OU
France	3.330653	AS41653	Aqua Ray SAS
Germany	5.329887	AS29066	Host Europe GmbH
Greece	28.840419	AS47692	Nessus GmbH
Hungary	5.855697	AS29278	Deninet KFT
Ireland	3.506842	AS39122	Blacknight Internet Solutions Limited
Italy	1.155471	AS34971	Prometeus di Daniela Agro
Latvia	3.385624	AS12993	SIA Digitalas Ekonomikas Attistibas Centrs
Lithuania	1.069842	AS198651	HOSTLINE, UAB
Luxembourg	1.161574	AS203944	NTT Luxembourg PSF S.A.
Netherlands	2.61648	AS25542	Netspider Group B.V.
Poland	1.9643	AS48446	Hostersi Sp. z o.o.
Portugal	0.946163	AS62416	SAMPLING LINE-SERVICOS E INTERNET, LDA
Romania	8.777309	AS41653	Aqua Ray SAS
Slovakia	0.853356	AS42005	LightStorm Communications s.r.o.
Slovenia	10.045924	AS47692	Nessus GmbH
Spain	4.545234	AS15699	OGIC Informatica S.L.
Sweden	0.940146	AS42005	LightStorm Communications s.r.o.

Table 2: Autonomous System networks with lowest latency per EU country

within the European Union. The input data is the 128 x 26 matrix from the previous step.

The implemented algorithm begins by selecting the best AS within an arbitrary country and checks if this same AS can be reached by other countries below some selected maximum latency threshold. If this is the case, then the countries are considered to be served by this AS. For the remaining countries the process is repeated and for an arbitrary not-yet-served-country its lowest latency network is selected. This process continues until either all countries are served or 4 networks have been selected. Then the EU wide average latency for this set of selected AS is stored and the process is repeated a number of times with a different arbitrary country as starting point. This results in a dataframe of potential combinations of networks and their corresponding average latency.

As part of the algorithm, a parameter for the maximum number of countries not optimally served can be selected. If set to 0, the set of selected networks is required to have had a latency sample to each European country, but when set to 1, it may have an unknown latency to at most one European country. The advantage of this approach is that not all countries within the RIPE dataset have received the same number of samples. Especially Malta and Cypress have had few ping samples. Allowing the optimisation algorithm some flexibility in network selection might result in a more sensible choice of AS, as the average performance for the remainder of Europe can be improved. In terms of potential locations for a hosting server, Malta does not have any ASNs of type "hosting" so it cannot be

included in the analysis.

Table 3 present the results for different requirements considering the number of countries to be included. Indeed we see that the best result can be obtained when two or three countries can be left out of coverage. The best combination AS obtains a latency of about 20.1ms. When full coverage is required, the latency of obtained optimal combination increases to about 30.6 ms.

Selected Networks	Average latency (ms)	Countries missed
AS12993, AS203944, AS59729, AS39790	20.090194	Cyprus, Slovakia, Malta
AS12993, AS203944, AS12637, AS39790	20.441318	Slovakia, Cyprus, Malta
AS47692, AS12993, AS41653, AS39122	20.768172	Cyprus, Slovakia, Malta
AS59729, AS198651, AS47692, AS62416	20.862629	Slovakia, Cyprus, Malta
AS61189, AS12993, AS59729, AS39790	20.95742	Cyprus, Slovakia, Malta
AS12993, AS203944, AS39790, AS42005	20.394981	Cyprus, Malta
AS12993, AS59729, AS39790, AS42005	20.67563	Cyprus, Malta
AS12993, AS61189, AS39790, AS42005	20.876578	Cyprus, Malta
AS25542, AS12993, AS39790, AS42005	20.972702	Cyprus, Malta
AS12637, AS12993, AS39790, AS42005	21.012709	Cyprus, Malta
AS12993, AS39790, AS42005, AS16276	30.63495	Malta
AS203944, AS39790, AS16276, AS42005	31.433489	Malta
AS59729, AS39790, AS42005, AS16276	31.703343	Malta
AS61189, AS39790, AS16276, AS42005	31.896562	Malta
AS25542, AS39790, AS42005, AS16276	31.988989	Malta

Table 3: Potential optimal AS sets for different maximum number of countries excluded .

3 Conclusion and Recommendations

The assumption is made that the inconsistencies caused by Cyprus and Malta are a result of sampling error. So, then the best choice of set is represented in the top row of Table 3. Table 4 provides more information on the AS comprising that set.

Overall this analysis indicates that more RIPE data, especially for hosting AS in the countries of Malta, Cypress and Slovakia are required to find an optimal solution. Therefore more than just a single day of RIPE ping data should be used in the analysis.

ASN	Country	Name	type
AS12993	Latvia	SIA Digitalas Ekonomikas Attistibas Centrs	hosting
AS39790	Czechia	Web4U s.r.o.	hosting
AS42005	Slovakia	LightStorm Communications s.r.o.	hosting
AS16276	France	OVH SAS	hosting

Table 4: Information on potential optimal AS set