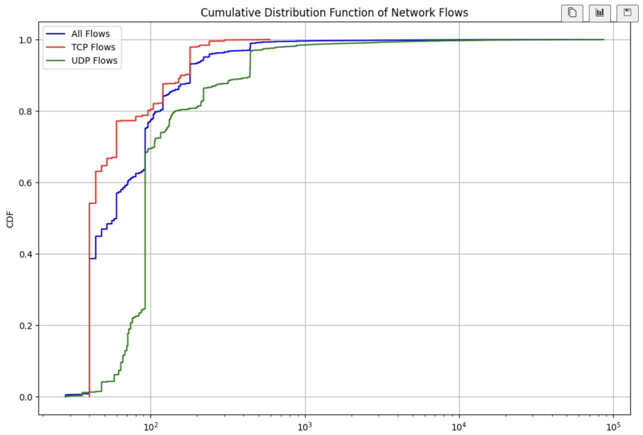
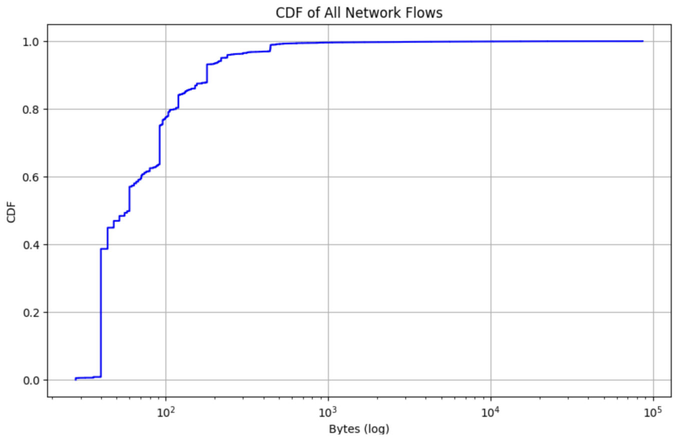
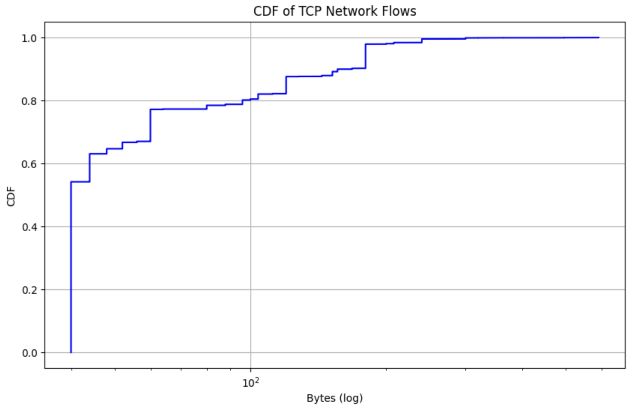
Analysis

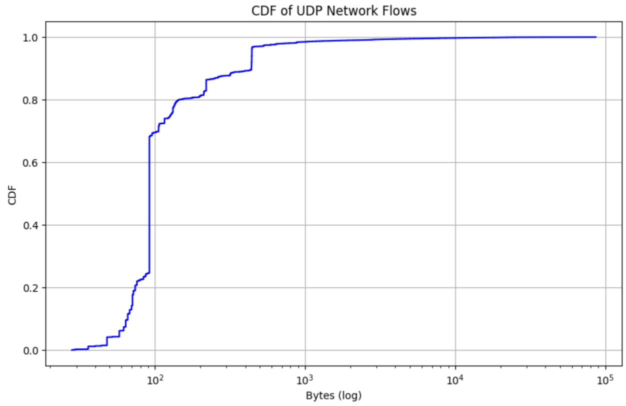
The code is written in Python and contained within a Jupyter notebook file. To run the code, please ensure that the files [50813A_code.ipynb], [bgp_route.csv], [bgp.update.csv], and [netflow.csv] are placed in the same folder. It also utilizes some Python libraries; if your current environment does not have these libraries installed, use pip to install them before running the code.

Question 1.1


Cumulative Distribution Function of Network Flows


CDF of All Network Flows


CDF of TCP Network Flows

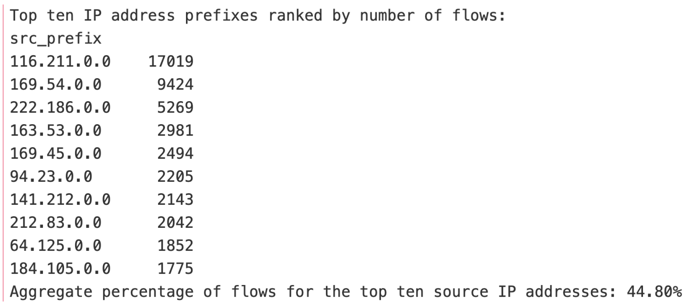

CDF of UDP Network Flows

Observation:

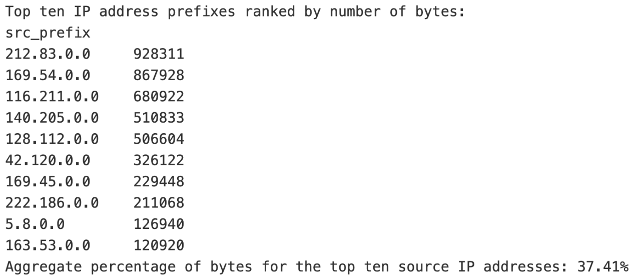
1. TCP flows which is in red line, have a steep increase in the graph and spread in a large range of size which perhaps because TCP is used for a wide range of applications, ranging from small to substantial payloads, such as file transfers, web browsing, and email communication.

2. UDP flows which is in green line is at the higher byte size range, which indicates the UDP flows has larger size， which aligns with the protocol's common use in applications such as online game and video streaming that prioritize smooth, uninterrupted data flow over precise delivery guarantees, necessitating the transmission of larger data packets.

Question 1.2



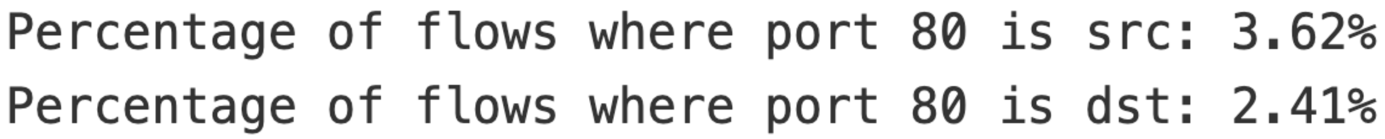
```
Top ten IP address prefixes ranked by number of flows:
src_prefix
116.211.0.0    17019
169.54.0.0      9424
222.186.0.0     5269
163.53.0.0      2981
169.45.0.0      2494
94.23.0.0       2205
141.212.0.0     2143
212.83.0.0      2042
64.125.0.0      1852
184.105.0.0     1775
Aggregate percentage of flows for the top ten source IP addresses: 44.80%
```



```
Top ten IP address prefixes ranked by number of bytes:
src_prefix
212.83.0.0     928311
169.54.0.0     867928
116.211.0.0    680922
140.205.0.0    510833
128.112.0.0    506604
42.120.0.0     326122
169.45.0.0     229448
222.186.0.0    211068
5.8.0.0        126940
163.53.0.0     120920
Aggregate percentage of bytes for the top ten source IP addresses: 37.41%
```

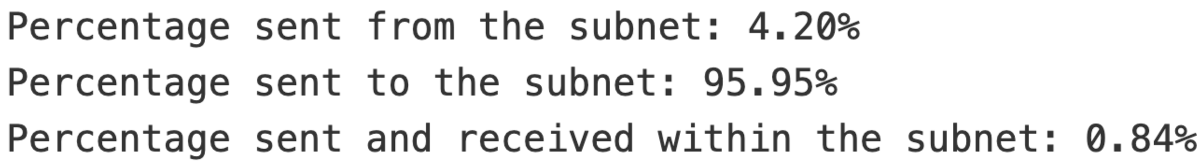
Question 1.3

PORT80: transmit Hypertext Transfer Protocol (HTTP) data.



```
Percentage of flows where port 80 is src: 3.62%
Percentage of flows where port 80 is dst: 2.41%
```

Question 1.4



```
Percentage sent from the subnet: 4.20%
Percentage sent to the subnet: 95.95%
Percentage sent and received within the subnet: 0.84%
```

Observation:

1. The percentage of bytes sent from subnet is really low, which indicates that the main work of device in this subset maybe is receive the information or it just contact in the subnet.

2. The percentage of bytes sent to the subnet about 95.95% which indicates that the device in this subnet is the main target of the receving data.

3. Only 0.84% bytes send and received at same time, this percentage

indicates that there are only relative few device contact with each other.

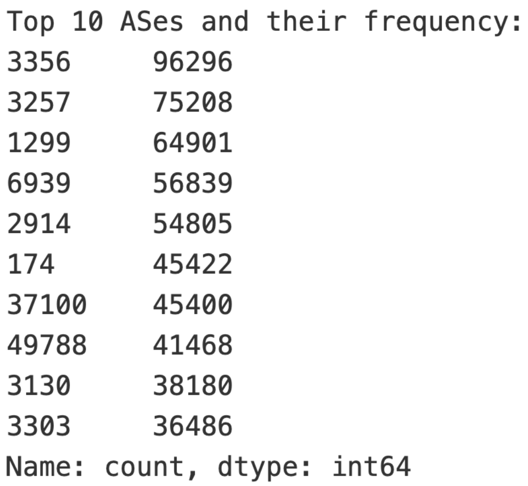
Question 1.5

The total traffic sum will increase, encompassing both TCP and UDP. Users in cafes may engage more frequently in social media activities such as watching videos or streaming music, potentially leading to an increase in the average size of UDP flows.

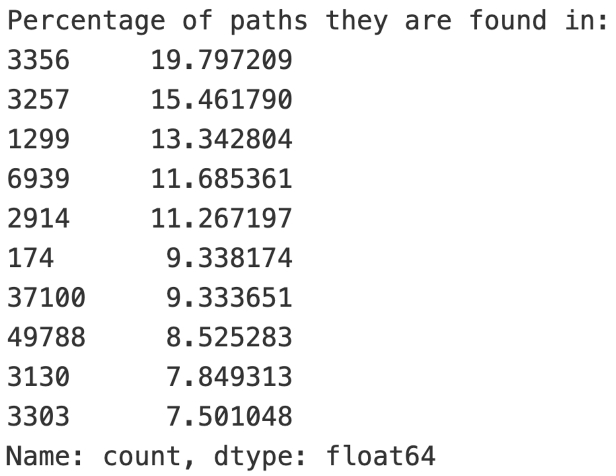
The usage of port 80 as both source and destination will likely surge, with the percentage rising significantly. This is because users in busy public cafes may browse websites and access social media platforms more frequently, all of which typically utilize port 80.

The percentage of bytes sent from the subnet is expected to rise, as cafe users may engage in a wider range of activities beyond the subnet. Conversely, the bytes sent into the subnet may decrease, as users may explore a more diverse array of websites rather than concentrating on a single subnet. Additionally, the percentage of bytes both sent and received within the subnet may increase due to the diverse range of network activities taking place in public cafes.

Question 2.1



```
Top 10 ASes and their frequency:
3356     96296
3257     75208
1299     64901
6939     56839
2914     54805
174      45422
37100    45400
49788    41468
3130     38180
3303     36486
Name: count, dtype: int64
```



```
Percentage of paths they are found in:
3356     19.797209
3257     15.461790
1299     13.342804
6939     11.685361
2914     11.267197
174       9.338174
37100     9.333651
49788     8.525283
3130      7.849313
3303      7.501048
Name: count, dtype: float64
```

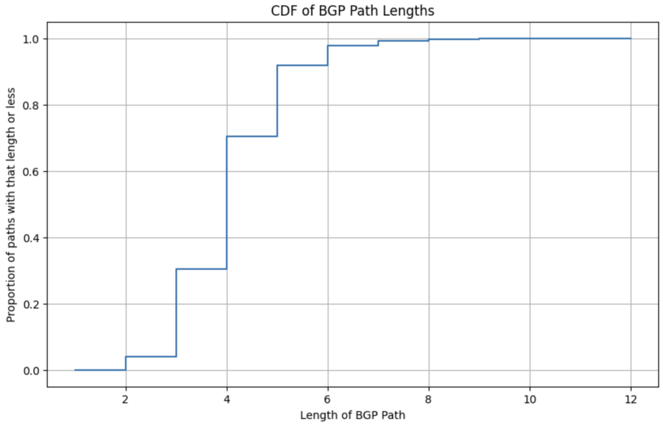
Searching result

3356	LEVEL3, United States
3257	GTT-BACKBONE GTT, United States
1299	TWELVE99 Arelion, fka Telia Carrier, Sweden
6939	HURRICANE, United States
2914	NTT-LTD-2914, United States
174	COGENT-174, United States
37100	SEACOM-AS, Mauritius
49788	NEXTHOP, Norway
3130	RGNET-SEA RGnet Seattle Westin, Estonia
3303	SWISSCOM Swisscom Switzerland Ltd, Switzerland

potential problem

When data flow is concentrated through a few AS, any issues within these AS can lead to widespread internet disruptions and performance degradation. Moreover, due to the high dependency on specific AS, the route choices may not be optimal, resulting in reduced efficiency and increased costs. Additionally, frequent appearances of an AS in routing paths can lead to collisions due to inconsistencies in their own policies.

Question 2.2



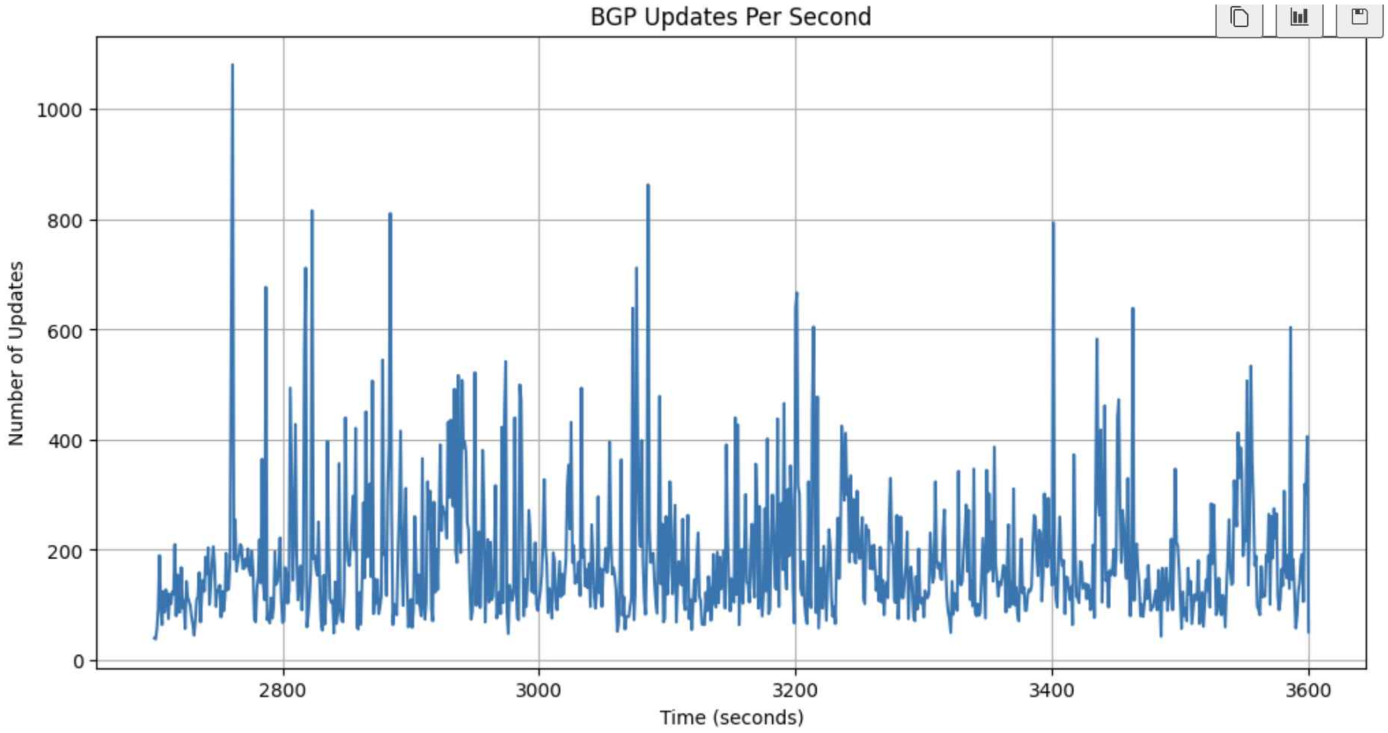
About BGP route lengths: From this graph, we can observe that approximately 80% of the paths involve four or fewer AS, with a noticeable sharp increase at four. This suggests that a significant number of paths are concentrated around four AS, indicating a certain degree of network centralization.

About a packet’ s travel across the internet: The shorter path lengths imply that packets traverse fewer AS before reaching their destination, resulting in fewer hops and faster transmission speeds. Additionally, this might suggest a high dependency on certain AS within the network connections, where any disruptions could significantly impact many data flows.

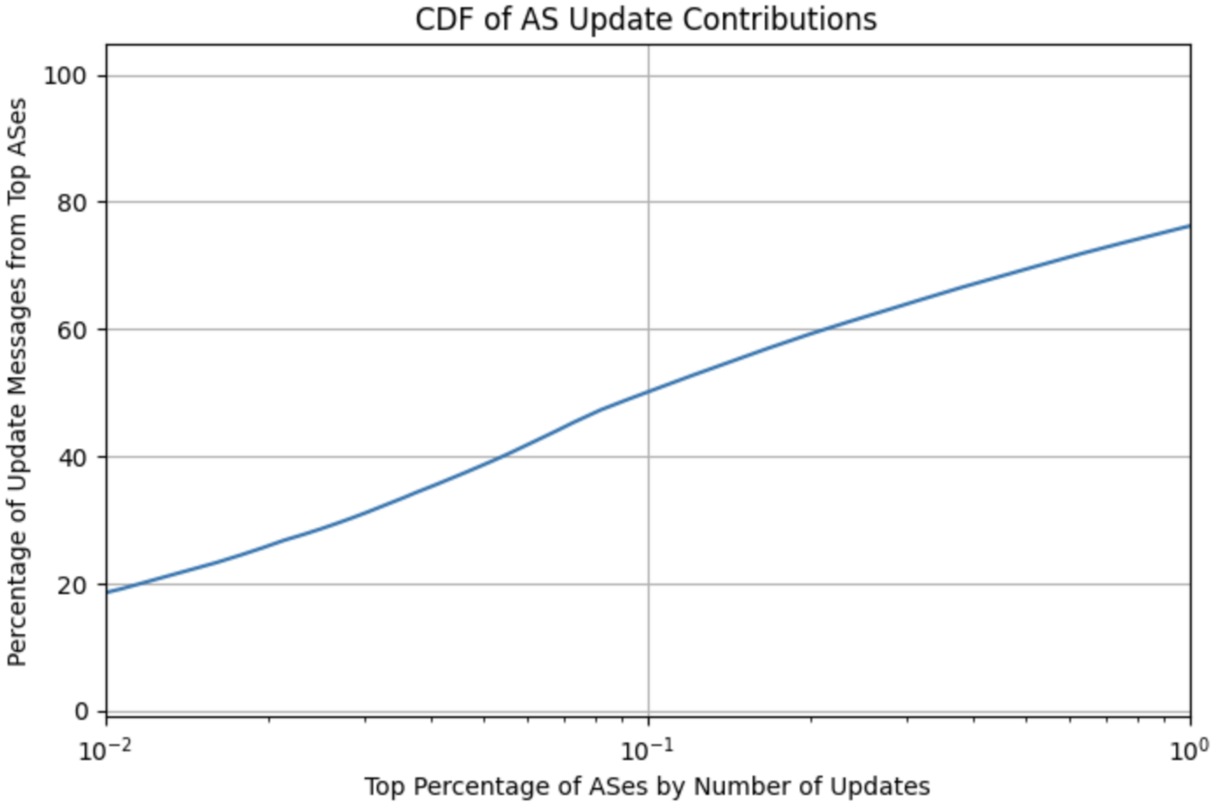
Question 2.3

Average BGP updates per minute: 11083.933333333332

(I made the time start from 45:00:00 and 00:00:00 means the new hour)



Question 2.4



The graph indicates a small fraction of ASes is responsible for a large portion of update messages. Specifically, the top 10% of ASes are responsible for more than 60% of update messages, and approximately 1% of ASes contribute up to 20%. This suggests a high level of centralization in BGP updates.

In terms of network stability, this implies that a few ASes might be changing their routes quite frequently, which can reflect an unstable network condition. A small change in these ASes' routing policies or network configurations can potentially trigger a disproportionately large number of BGP updates.

The graph also shows that the rate of increase in update contributions slows down as we look at a larger percentage of ASes, indicating that the remaining ASes contribute relatively fewer updates. This can also be interpreted as the BGP update process being dominated by a minority of the ASes, potentially leading to greater vulnerability to network disruptions if these key ASes experience issues.