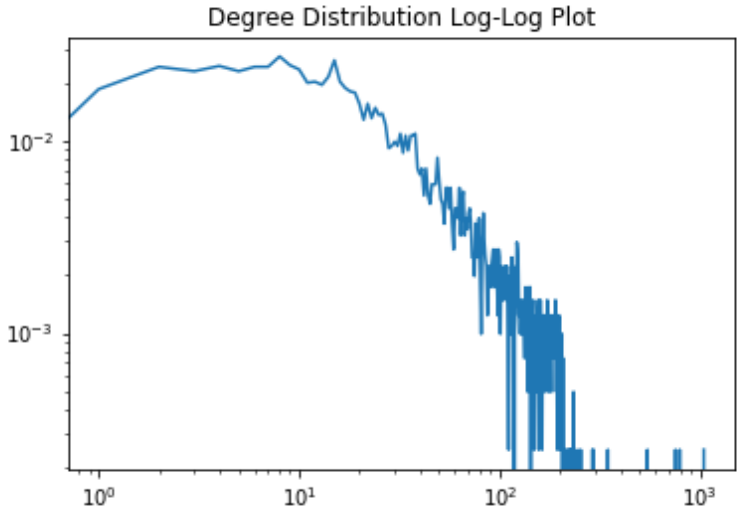
[15 points]
Plot the degree distribution on a log-log plot.
Does the degree distribution obey the power law well?
If so, what is the power-law coefficient? (Recall that a power law says that the prob(k) of degree k is proportional to k^{-gamma}. Gamma is known as the power-law coefficient (it also has other names).



Hint: On a log-log plot, the power-law is linear. You can use linear regression (such as in sklearn) to deduce gamma.

Its Not obey the power law well.

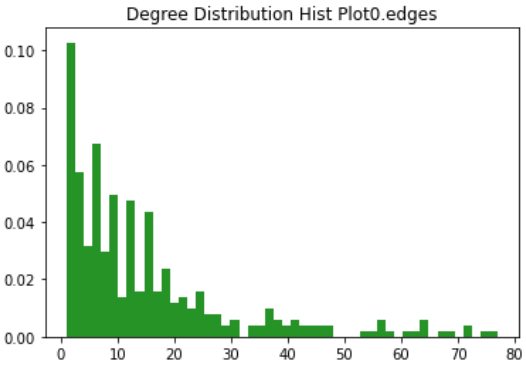
[5 points]
How many friends does the node with the highest degree have?
What is the clustering coefficient of that node?
Interpret the result: do you believe this individual is a member of a cohesive and tightly knit group or more like a central figure that people are connecting to for some reason

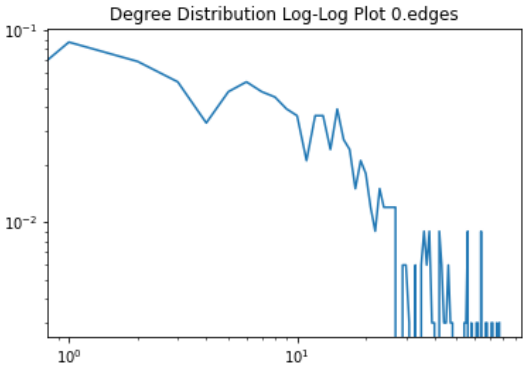
ANS:

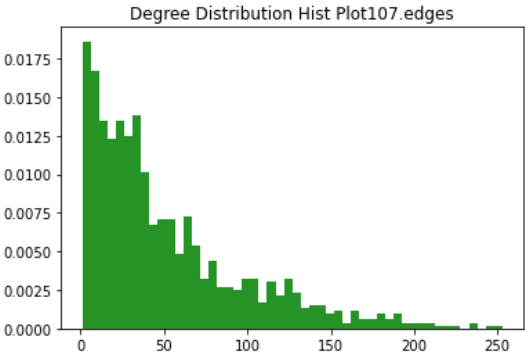
The highest degree node is: 107
It has 1045 friends
Its clustering coefficient is: 0.049038479165520905

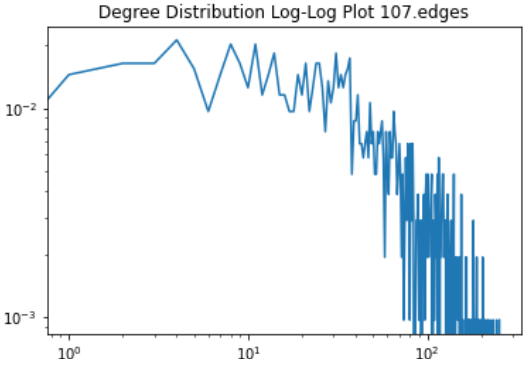
Clustering coefficient tells how well connected the neighborhood of the node is. The coefficient of this node is pretty close to 0 which means that there are hardly any connections in the neighborhood. So this individual could be a celebrity's account, some brands' official account, etc.

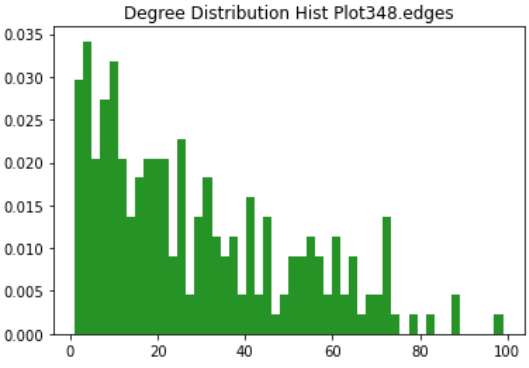
[10 points extra credit]
Separately report the degree distributions of the ten individual networks on the dataset webpage under file facebook.tar.gz.

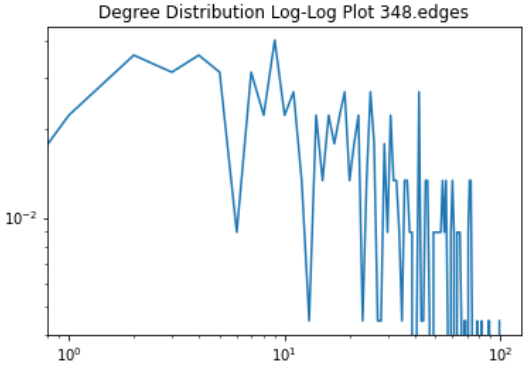


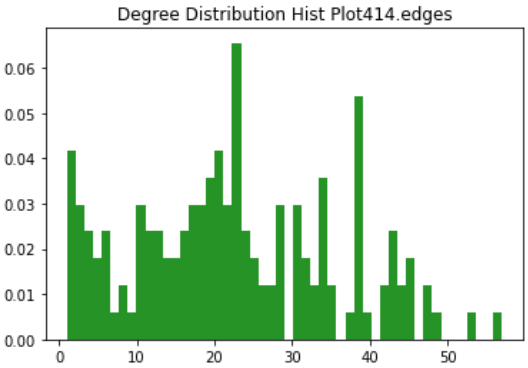


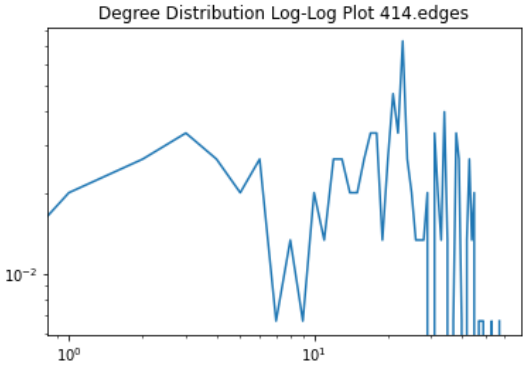


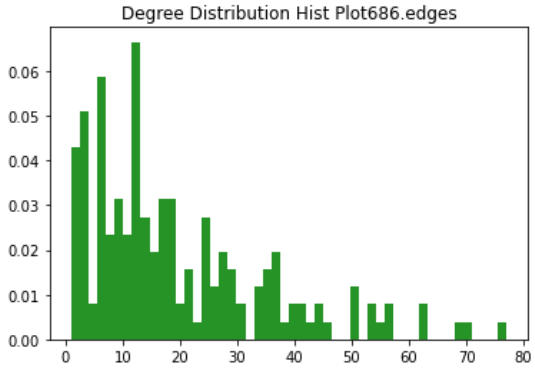


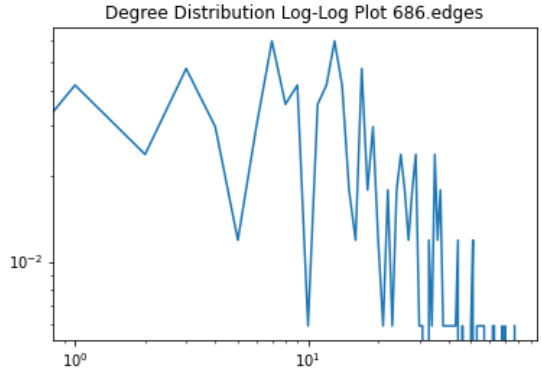


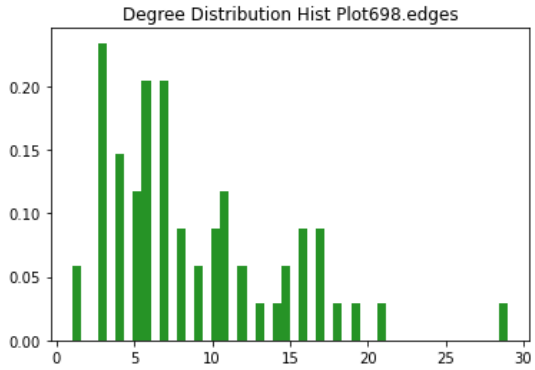


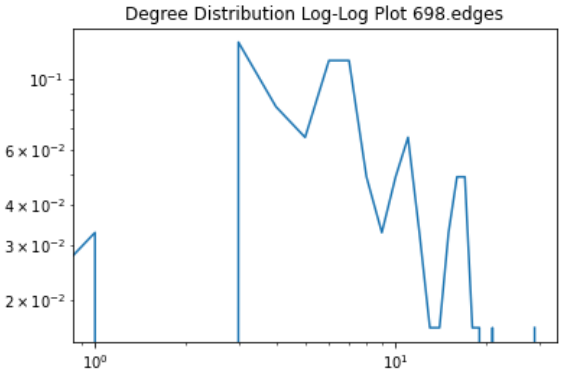


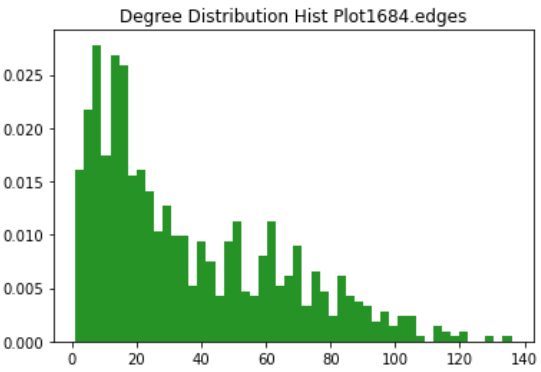


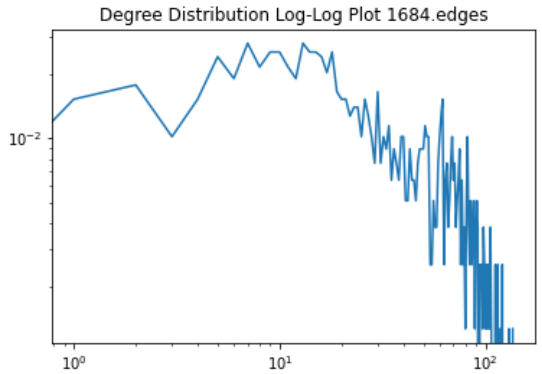

Degree Distribution Hist Plot686.edges
0.06
0.05
0.04
0.03
0.02
0.01
0.00
0
10
20
30
40
50
60
70
80

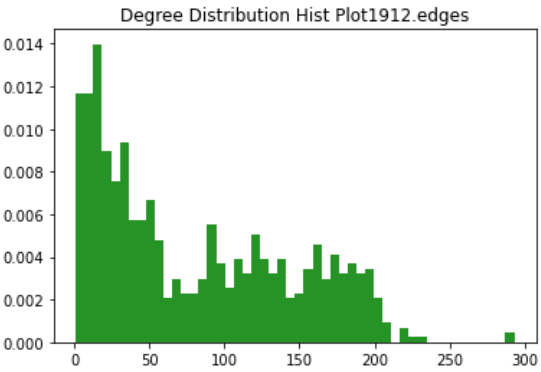

Degree Distribution Log-Log Plot 686.edges
10^{-2}
10^{0}
10^{1}

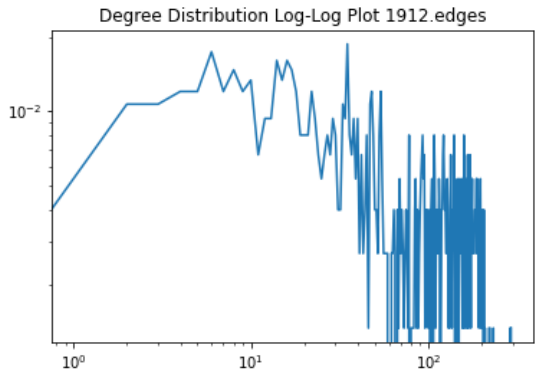

Degree Distribution Hist Plot698.edges
0.20
0.15
0.10
0.05
0.00
0
5
10
15
20
25
30

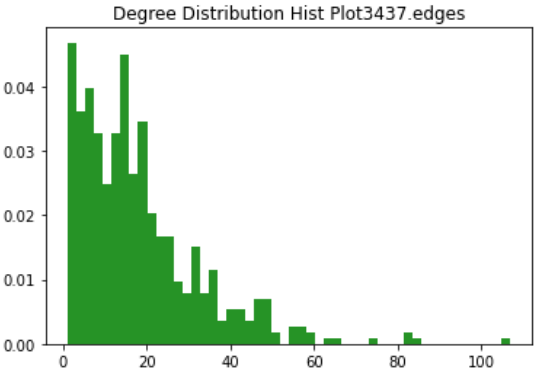

Degree Distribution Log-Log Plot 698.edges
10^{-1}
6×10^{-2}
4×10^{-2}
3×10^{-2}
2×10^{-2}
10^{0}
10^{1}

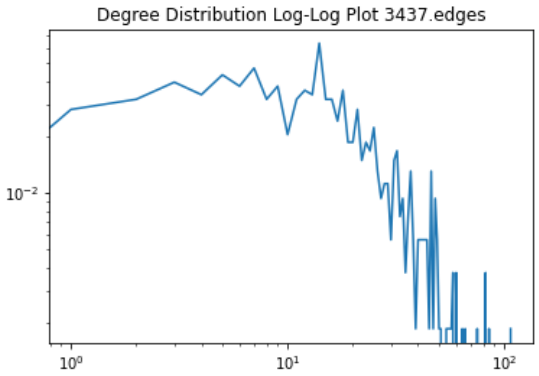

Degree Distribution Hist Plot1684.edges
0.025
0.020
0.015
0.010
0.005
0.000
0
20
40
60
80
100
120
140

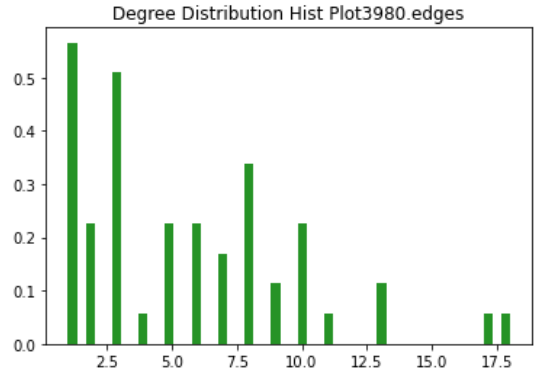

Degree Distribution Log-Log Plot 1684.edges
10^{-2}
10^{0}
10^{1}
10^{2}

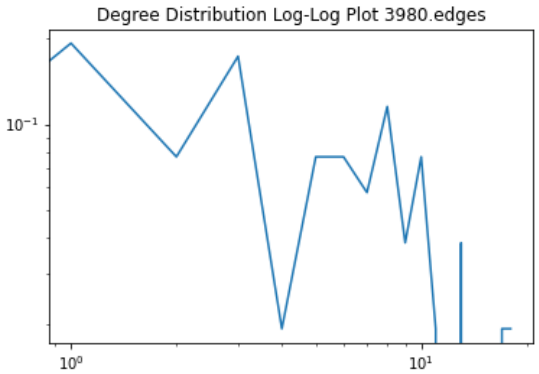

Degree Distribution Hist Plot1912.edges
0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
0
50
100
150
200
250
300


Degree Distribution Log-Log Plot 1912.edges
10^{-2}
10^{0}
10^{1}
10^{2}


Degree Distribution Hist Plot3437.edges
0.04
0.03
0.02
0.01
0.00
0
20
40
60
80
100


Degree Distribution Log-Log Plot 3437.edges
10^{-2}
10^{0}
10^{1}
10^{2}


Degree Distribution Hist Plot3980.edges
0.5
0.4
0.3
0.2
0.1
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5


Degree Distribution Log-Log Plot 3980.edges
10^{-1}
10^{0}
10^{1}

[40 points]
Use the function above (including any others you need) to randomly sample edges from G to produce four graphs with 15000, 30000, 45000 and 60000 edges respectively. Re-compute the average clustering coefficient, fraction of closed triangles and diameter on these graphs.

In the report, tabulate these results and provide a succinct conclusion.
Is there a sample size at which the network starts looking like the original network? Which (if any) of the quantities you have computed above are reliable and at what sampling threshold? For example, you may conclude, 'fraction of closed triangles starts approaching the level in the original network even in the 15000 edge sample, but the diameter continues to be untrustworthy till we see at least 45000 edges...'.

Edges	15000	30000	45000	60000
Average Clustering Coefficient (STD)	0.0763942416315 (0.0021638)	0.15514566575629 (0.003241)	0.22293201295407 (0.003393)	0.282039319878259 (0.003651)
Closed Triangles. (STD)	0.01472 (0.000988)	0.07997 (0.00410)	0.2025 (0.0116)	0.3718 (0.01487)
Diameter. (STD)	999999999. (0)	999999999 (0)	999999999 (0)	999999999 (0)

Conclusion:
Average clustering coefficient continues to be untrustworthy until 60000 edges;
Fraction of closed triangles starts approaching the level in the original network in 45000 edge sample;
Diameter remain untrustworthy even at 60000 edges.

[10 points extra credit].
You are trying to sample x edges, and the metric of success is the degree distribution. At what point does your sample bring you 'close enough' to the original?

ANS:
At 5002 edges the sample is close enough to the original