DATA.ML.430-2020-2021-1 Complex Networks

Tuomo Jussila Finland 252980 tuomo.jussila@tuni.fi

April 2021

1 Dataset

The data is ready-made email network from some European research institution. Stanford university provides this data set here: (https://snap.stanford.edu/data/email-Eu-core.html). This is directed network, a link from persons A to B was created if person A sent an email to person B. The data also contains departments where people work. Every node/person has one department. Nodes and departments are marked as numbers to make this data set anonymous.

There are 1004 nodes, 25571 directed links and 42 departments. For further analysis, I need undirected network and one giant component. So, I took reciprocal links(A links to B and B links to A) as undirected links. This approach should remove mailing lists. Also I took biggest component of this network, so all nodes can reach any other node in this component. After these modifications this network contains 776 nodes and 9447 links.

There isn't other attributes than departments. The distribution of departments isn't even. Some departments have over hundred employees and some departments have few employees.

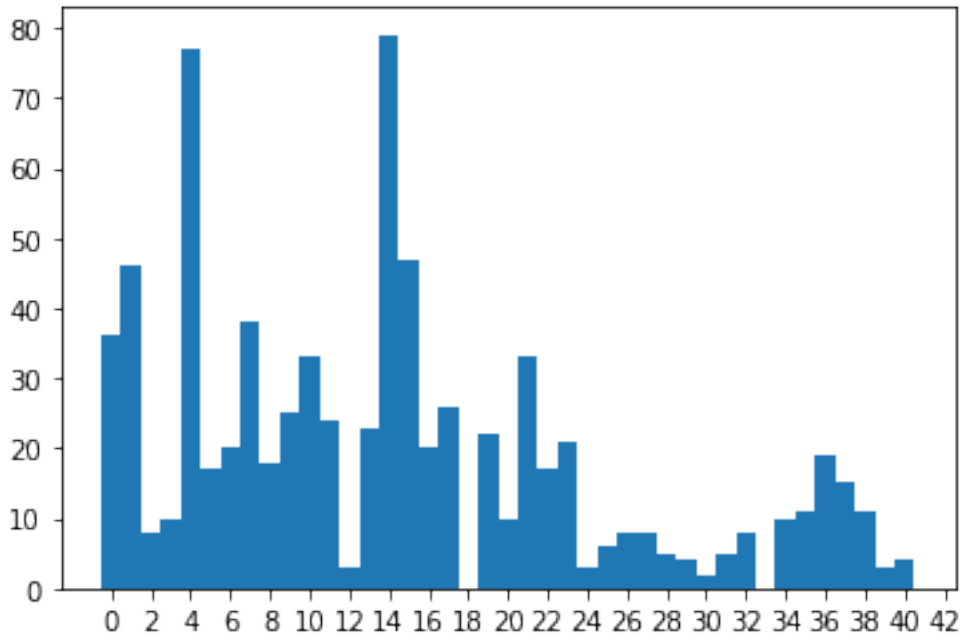

Figure 1: Histogram of connected component's departments

Figure 1 shows the distribution of departments. This histogram shows only nodes of connected component. There isn't departments 18,33 and 41 in the connected component. There is some large departments, maybe they will appear in community analysis.

2 Basic properties of Email network, Erdos-Renyi Network and Barabasi-Albert Network

This chapter introduces Email network,Erdos-Renyi network and Barabasi-Albert network.

2.1 Introduction

Email network is project's crawled network. For further analysis I use undirected and fully connected network. It contains 776 nodes, 9447 links and 39 departments(3 departments dropped out from filtering).

Erdos-Renyi network is random network where nodes are linked for given probability. In this case the probability is same as Email network's density. The amount of links differs slightly from Email network since creating links is random process.

Barabasi-Albert network is also random network but network is scale-free. It's density will be same as density of Email network.

Table 1: Densities and average degrees

Network	Nodes	Links	Density	Average degree
Email	776	9947	0.0314	24.35
Erdos-Renyi	776	9454	0.0314	24.37
Barabasi-Albert	776	9234	0.0307	23.80

Table 1 shows some basic properties of all three networks. They differ slightly but it should be fine.

2.2 Distribution of degrees

Four figures below presents degree distributions of networks. Distributions was plotted into logarithm scale for better visibility. Also distributions were computed using one minus cumulative distribution because it is usually smoother than other ways.

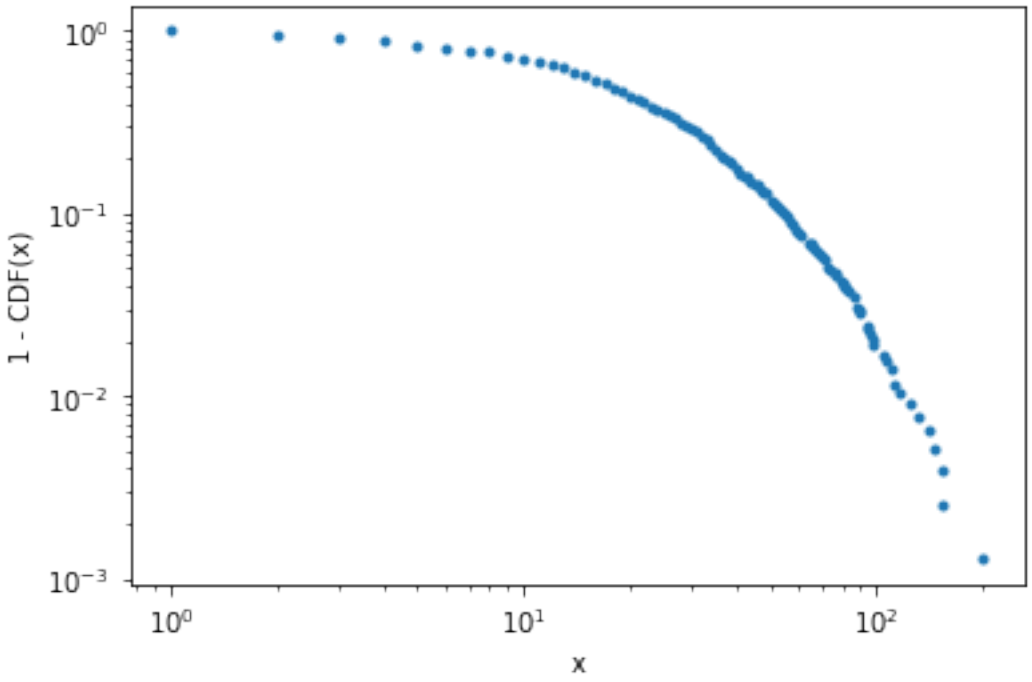

Figure 2: Degree distribution of Email network

Figure 2 shows distribution of degrees for Email network. Seems that distribution isn't fully scale free since there is curvature in the middle of curve. There is a lot of nodes having small degree. However after 10 degrees distribution is linear and I would say that distribution here is scale free.

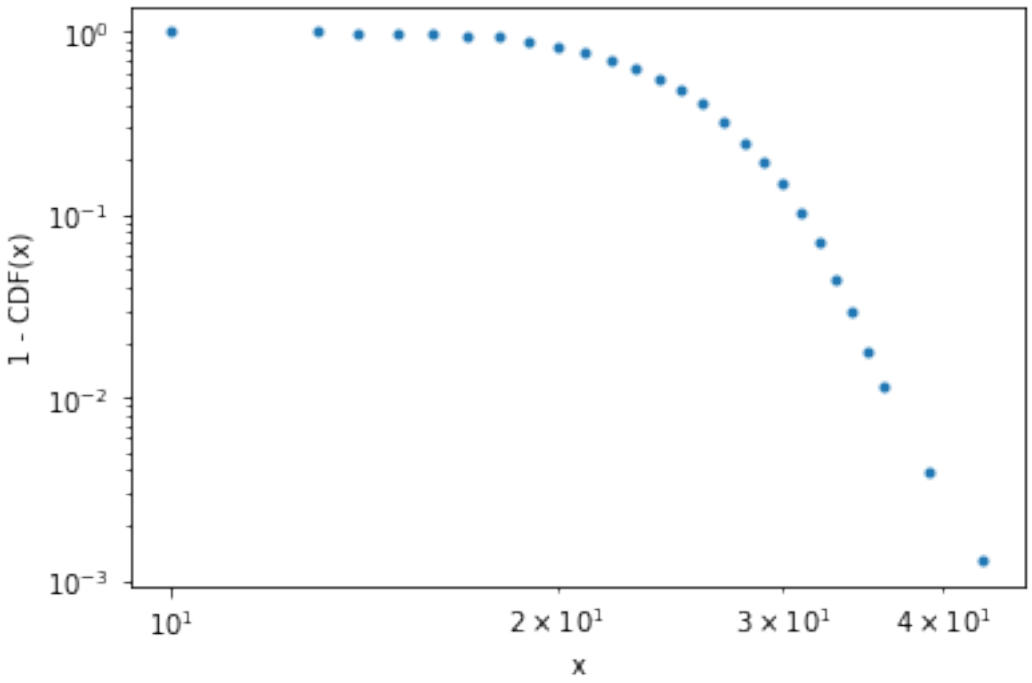

Figure 3: Degree distribution of Erdos-Renyi network

Figure 3 shows distribution of degrees for Erdos-Renyi network. Shape of distribution seems to similiar to Email distribution but there is significant difference in magnitude. Most nodes of this network has small degree and only a few nodes has more than degree of 20.

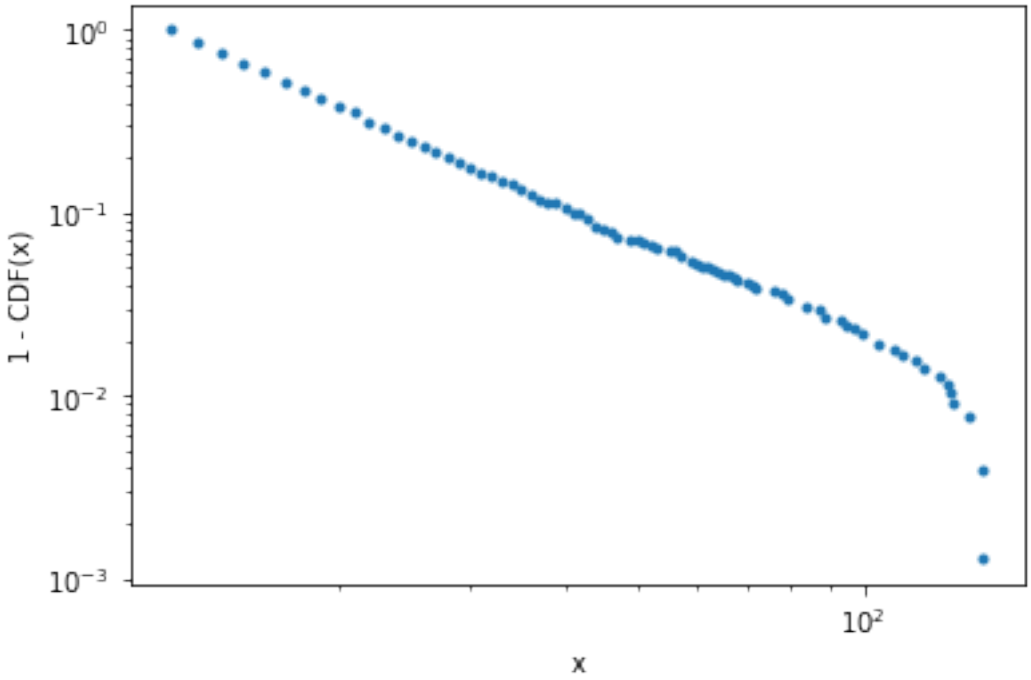

Figure 4: Degree distribution of Barabasi-Albert network

Figure 4 shows distribution of degrees for Barabasi-Albert network. It is scale free as it should be.

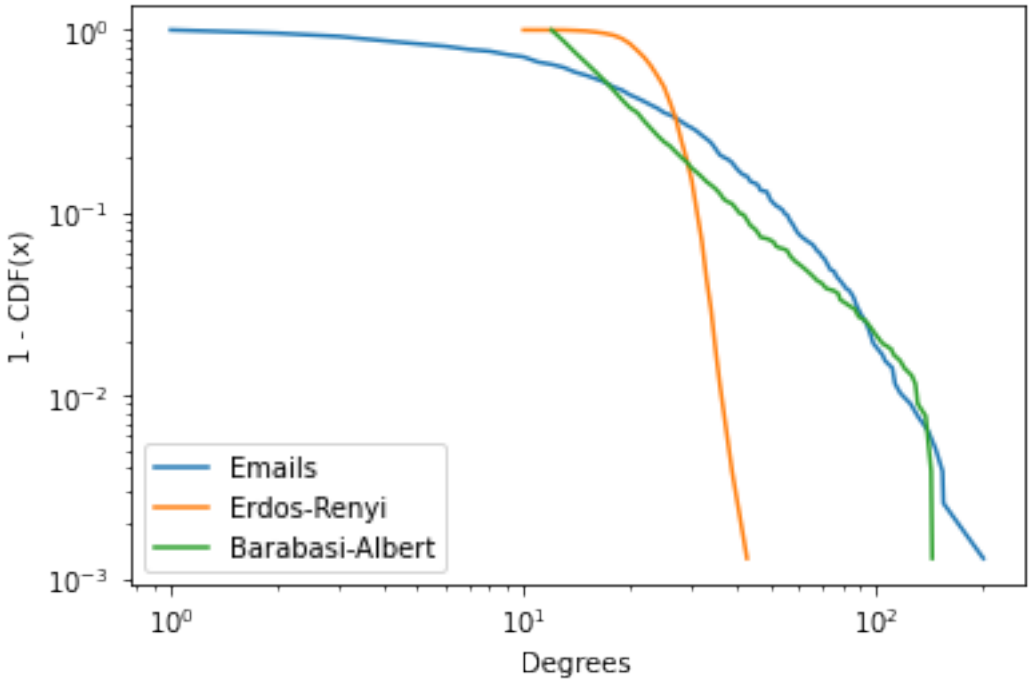

Figure 5: Degree distribution of all three networks

Finally figure 5 shows all three distributions in one figure. Because of algorithms Erdos-Renyi network's and Barabasi-Albert network's smallest degree is about 10 when in Email network smallest degree is 1. This figure shows clearly that Email network and Erdos-Renyi network are not same even if their distribution shapes resembled each other.

2.3 Centrality values

This section exploits some centrality properties of three networks.

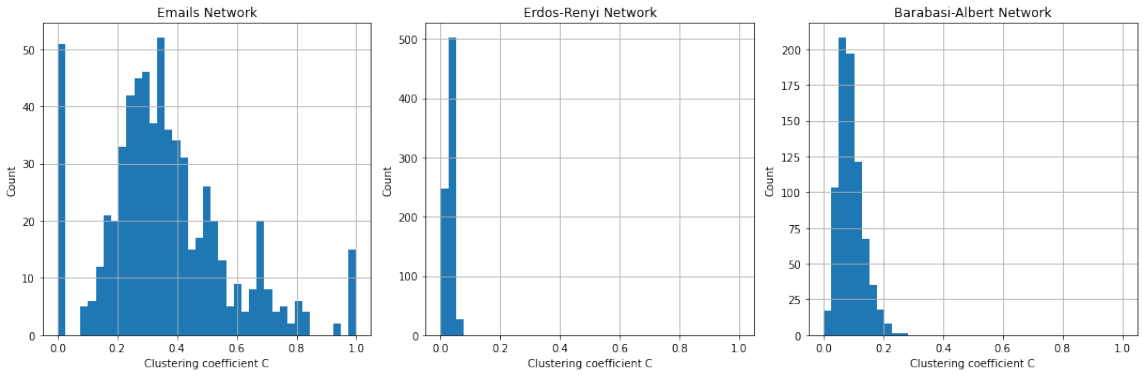

Figure 6: Clustering Coefficients

Figure 6 shows clustering coefficient centralities. As we can see, distributions of centrality coefficients are very different to each other even if they have a same amount of nodes and links. Erdos-Renyi network has very low centrality and its variance is also very small. This means that nodes of Erdos-Renyi network doesn't form communities and links are pretty random.

Centrality coefficients of Barabasi-Albert network are more spread but they are still pretty below, most coefficients are below 0.2. This network might contain some communities.

Centrality coefficients of Email network gets all values from zero to one. Variance and mean are a lot higher than other two. My interpretation to this is that Email network indeed contains communities where groups knows almost all other members of groups. There is also a lot of zero centrality nodes. These nodes in Email network might be mailing lists or some kind of contact person in university.

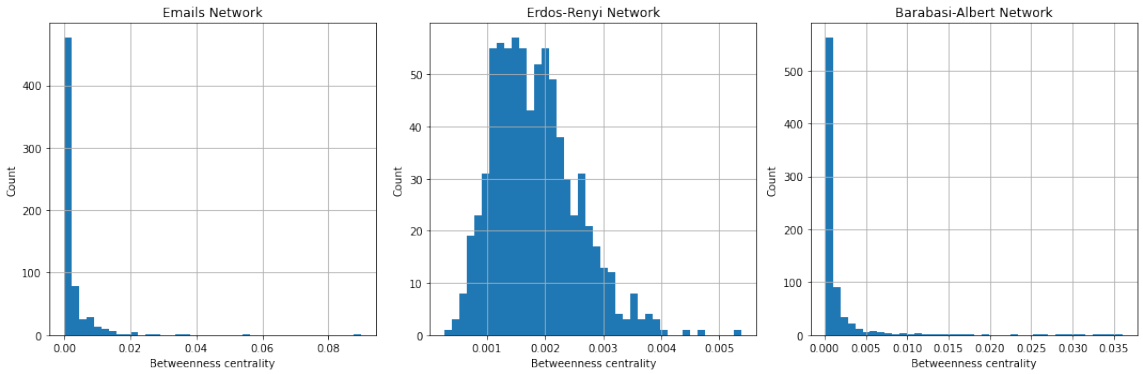

Figure 7: Betweeness centrality

Figure 7 shows betweeness centrality points. The higher score the more

important node is in network. There is not any important nodes in Erdos-Renyi network.

Betweenness scores for Email network and Barabasi-Albert network are similiar. The distribution seems to be same but Email network might have more important nodes.

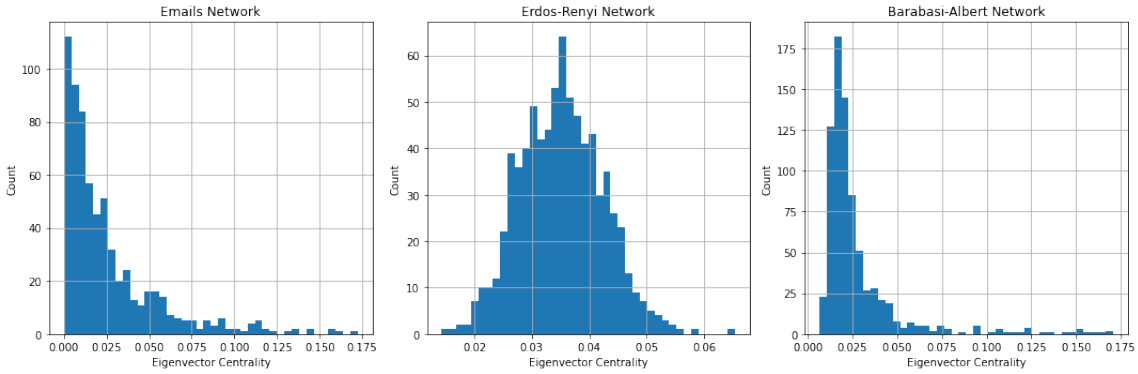

Figure 8: Eigenvector centrality

Figure 8 shows eigenvector centrality scores. Differences seems to be similiar in comparison to betweenness scores. Email network have more zero eigenvector centrality nodes which are leafs. Barabasi-Albert network have less lead nodes otherwise distributions are pretty similiar.

3 Community discovery

In this part I explore if some community algorithms finds communities which includes real departments from Email network. In this part I examined three different algorithms. They were Demon,K-Clique and Infomap.

There is not strong communities in email network but there are several weak communities. These departments are: 0, 1, 2, 4, 5, 7, 8, 10, 13, 14, 16, 17, 19, 20, 21, 25, 28, 30 and 35. So 19 out of 42 departments are considered as weak community. I think that this result means that Email network really contains communities and they are not just random.

In further analysis I was only interested found communities contained real department community. For this purpose I computed proportion of most common department in given community and I also computed proportion of most common department of real department(number of most common department nodes in given community/size of department in whole network). For further analysis I took communities where most common department has at least 50% proportion of community.

3.1 Demon

This algorithm finds communities which overlaps to each other, so algorithm might produce several communities which are close to each other. Algorithm needs only one parameter: epsilon. By adjusting epsilon demon algorithm produces different number of communities. A small epsilon produced few huge(over 600 nodes) communities, so I chose epsilon = 0.65 which procured 59 communities. Community sizes were from 20 to 500.

Following method found 7 real departments. They are listed in next table. Group of found nodes are union of several groups of same departments.

Table 2: Found departments by demon algorithm

Department	Found nodes	size of department	proportion
0	32	49	65.31%
1	27	65	41.54%
4	47	109	43.12%
7	34	51	66.67%
14	74	92	80.43%
17	25	35	71.43%
37	12	15	80.00%

Only department 37 isn't weak community, other six departments are weak communities.

3.2 K-Clique

This algorithm doesn't find overlapping communities and there will be less communities. There is only parameter: minimum size of community. In this case I set 8 as minimum community size. This produced 15 communities and most contained 10-20 nodes.

Following method found 10 real departments out of 15 communities. They are listed in next table. Some groups of found nodes are union of several groups of same departments.

Table 3: Found departments by K-clique algorithm

Department	Found nodes	size of department	proportion
0	11	49	22.45%
1	13	65	20.00%
4	18	109	16.51%
5	5	18	27.78%
7	23	51	45.10%
14	16	92	17.39%
16	12	25	48.0%
17	21	35	60.00%
19	16	29	55.17%
20	7	14	50.0%

K-Clique algorithm found real communities but pretty small ones. All found real departments are also weak communities. This algorithm found 10 out of 19 weak communities, at least part of it.

3.3 Infomap

Infomap algorithm doesn't take any parameters. So, it just computed stuff and gave 14 communities. Size of communities was pretty same with k-clique. Infomap found 11 departments out of 14 communities, so infomap possibly is best of three used algorithms to find real departments. The table below shows results.

Table 4: Found departments by Infomap algorithm

Department	Found nodes	size of department	proportion
1	32	65	49.23%
2	6	10	60.0%
4	6	109	5.50%
5	14	18	77.78%
8	12	19	63.16%
9	5	32	15.62%
14	74	92	80.43%
16	12	25	76.00%
17	24	35	68.57%
19	20	29	68.97%
37	11	15	73.33%

Departments 9 and 37 aren't weak communities and rest are. Seems that infomap found most of nodes of each department. Proportions are mostly 60-80 %.

3.4 Community discovery for Erdos-Renyi and Barabasi-Albert network

3.4.1 Erdos-Renyi

Demon algorithm found almost 600 communites with sizes 5-10 with same parameters. K-Clique algorithm didn't find any communities.

3.4.2 Barabasi-Albert

Demon algorithm found 300 communities with sizes 5-300 with same parameters. K-Clique algorithm found only one 40-node community.

3.4.3 Conclusion

I think results above means that both networks don't have real communities. Result of centrality coefficients earlier supports this result.

4 Spreading phenomena

I used EoN(Epidemics of Networks) package for spreading phenomena analysis. This package is explained here: https://arxiv.org/pdf/2001.02436.pdf. I ran every simulation 20 times and averaged them to get smoother trajectories.

4.1 SIS

I tested spreading phenomena with six different tau-values which mean how likely node infects another node. Next three figures shows S and I curves with all six tau values and three used networks.

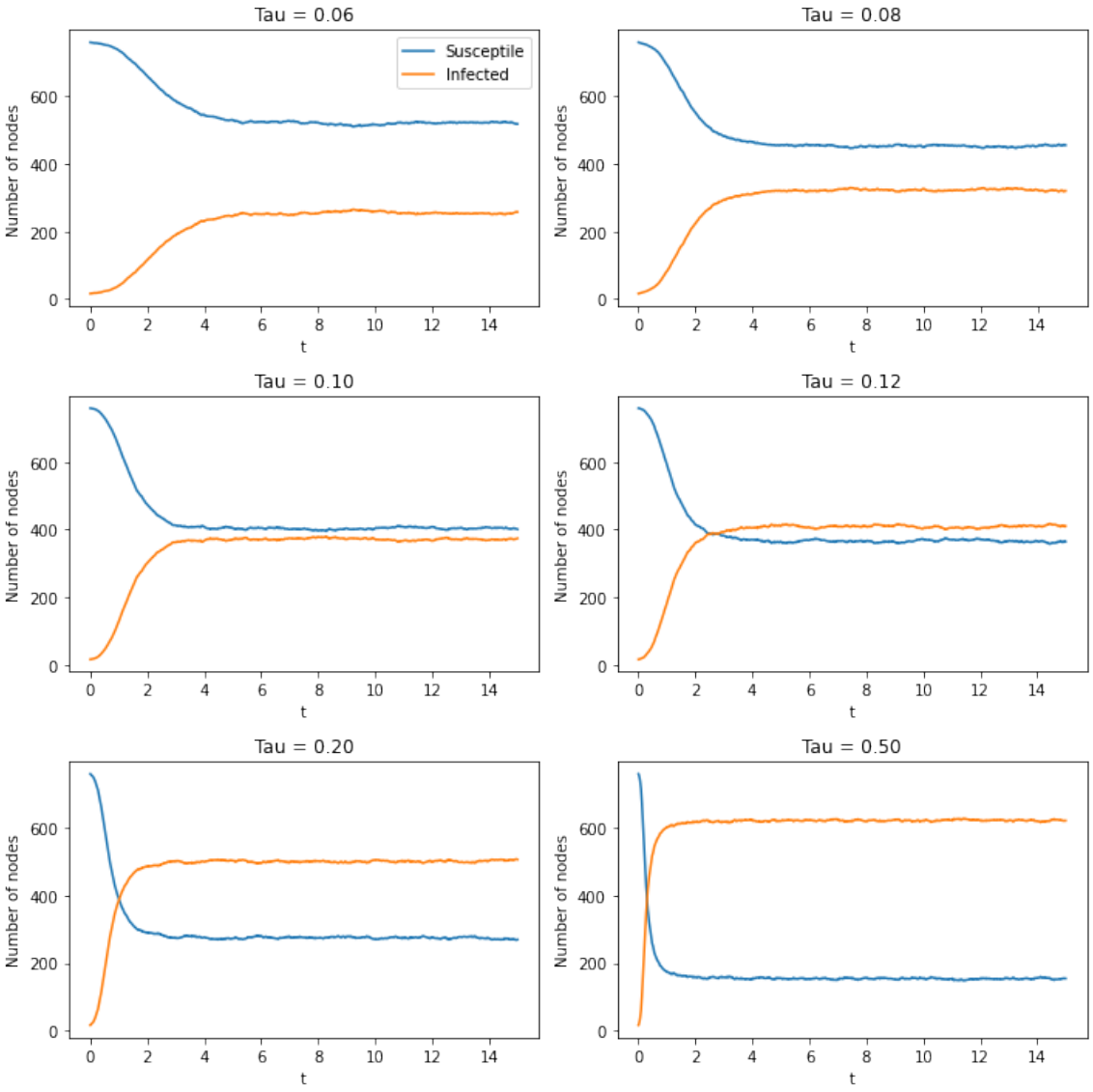

Figure 9: SIS spreading with Email network

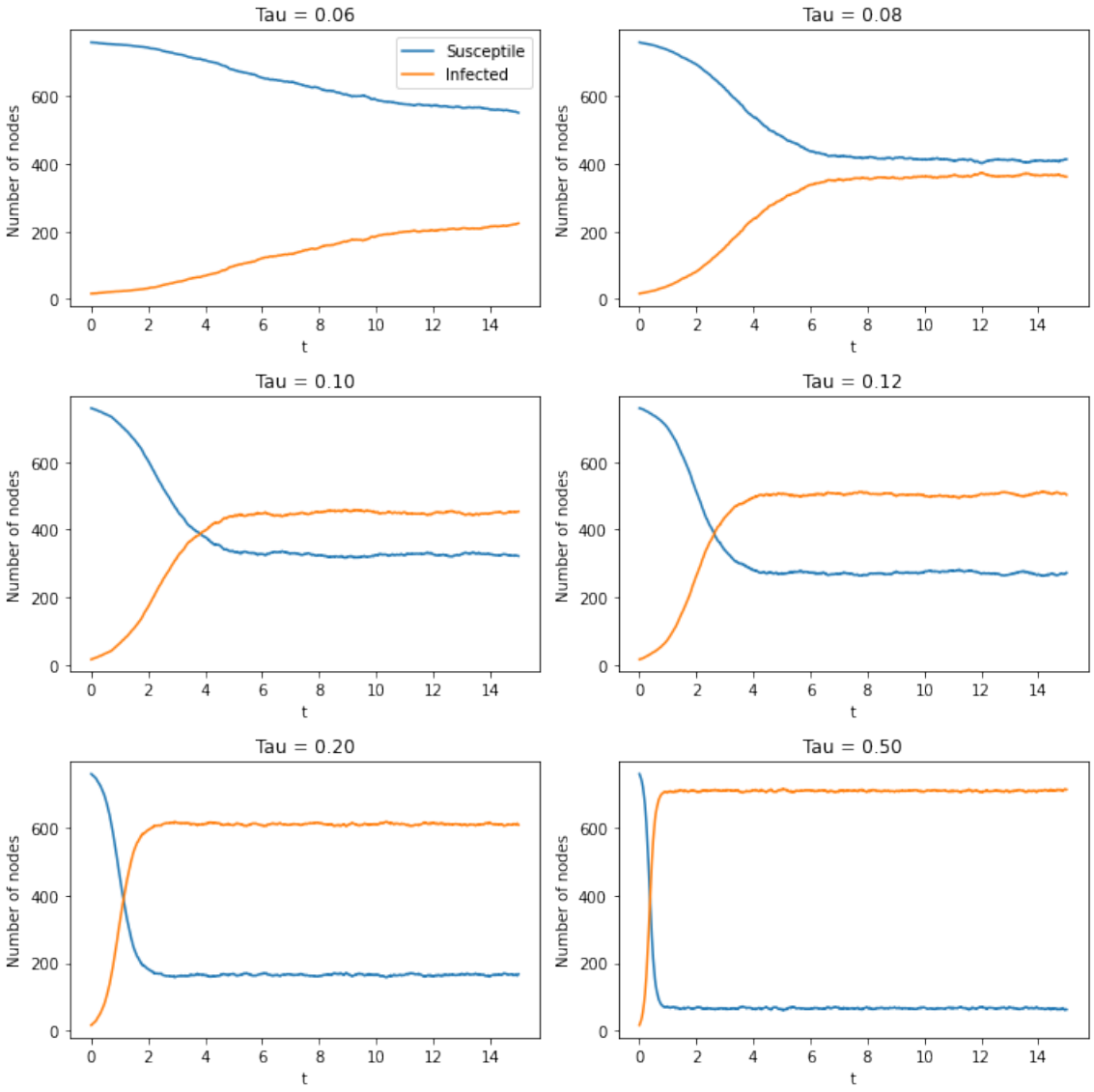

Figure 10: SIS spreading with Erdos-Renyi network

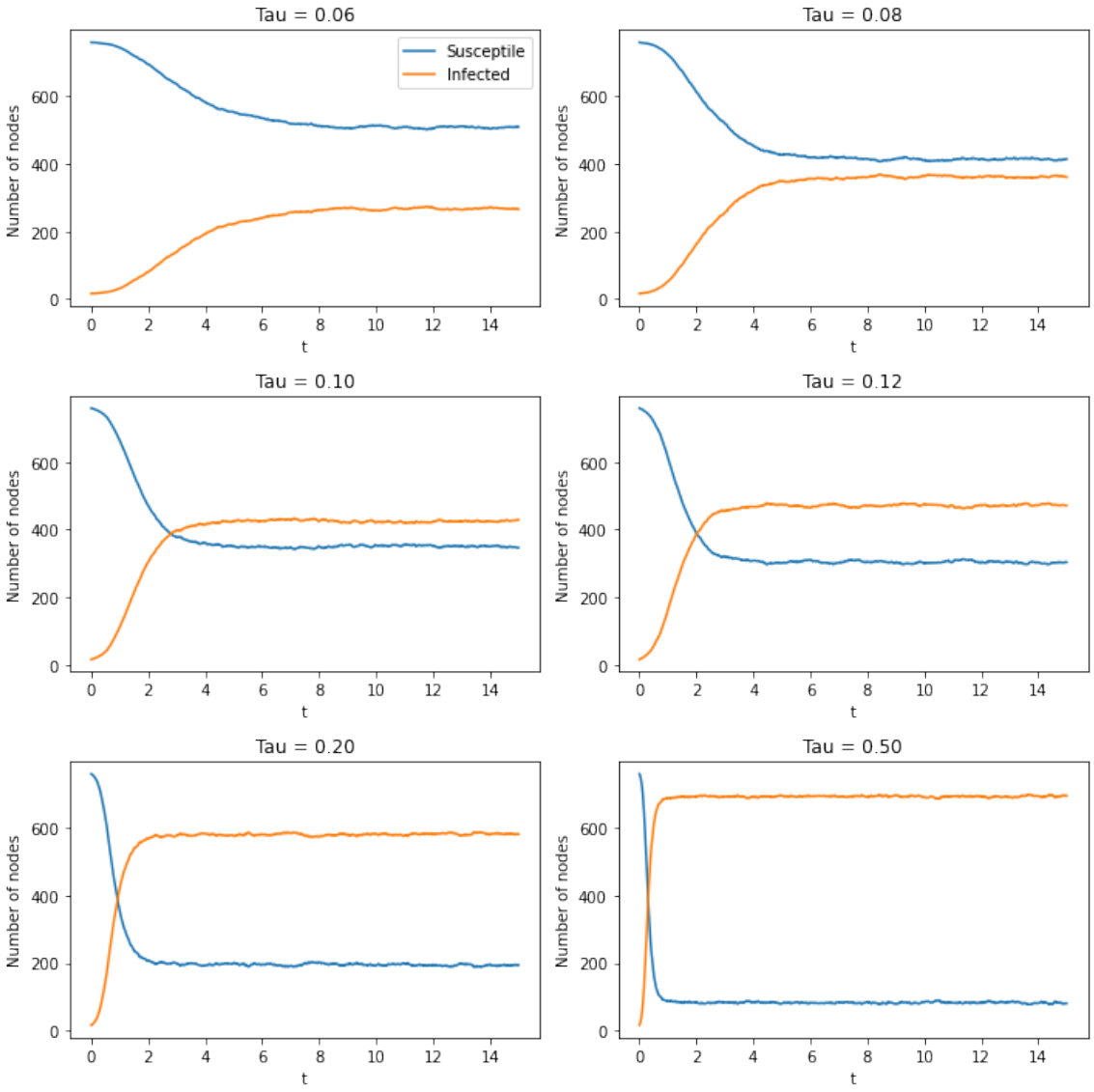

Figure 11: SIS spreading with Barabasi-Albert network

As figures 9,10 and 11 show, increasing tau value accelerates spreading which is not surprising. Interestingly equilibrium value for infected state increases when tau value increases.

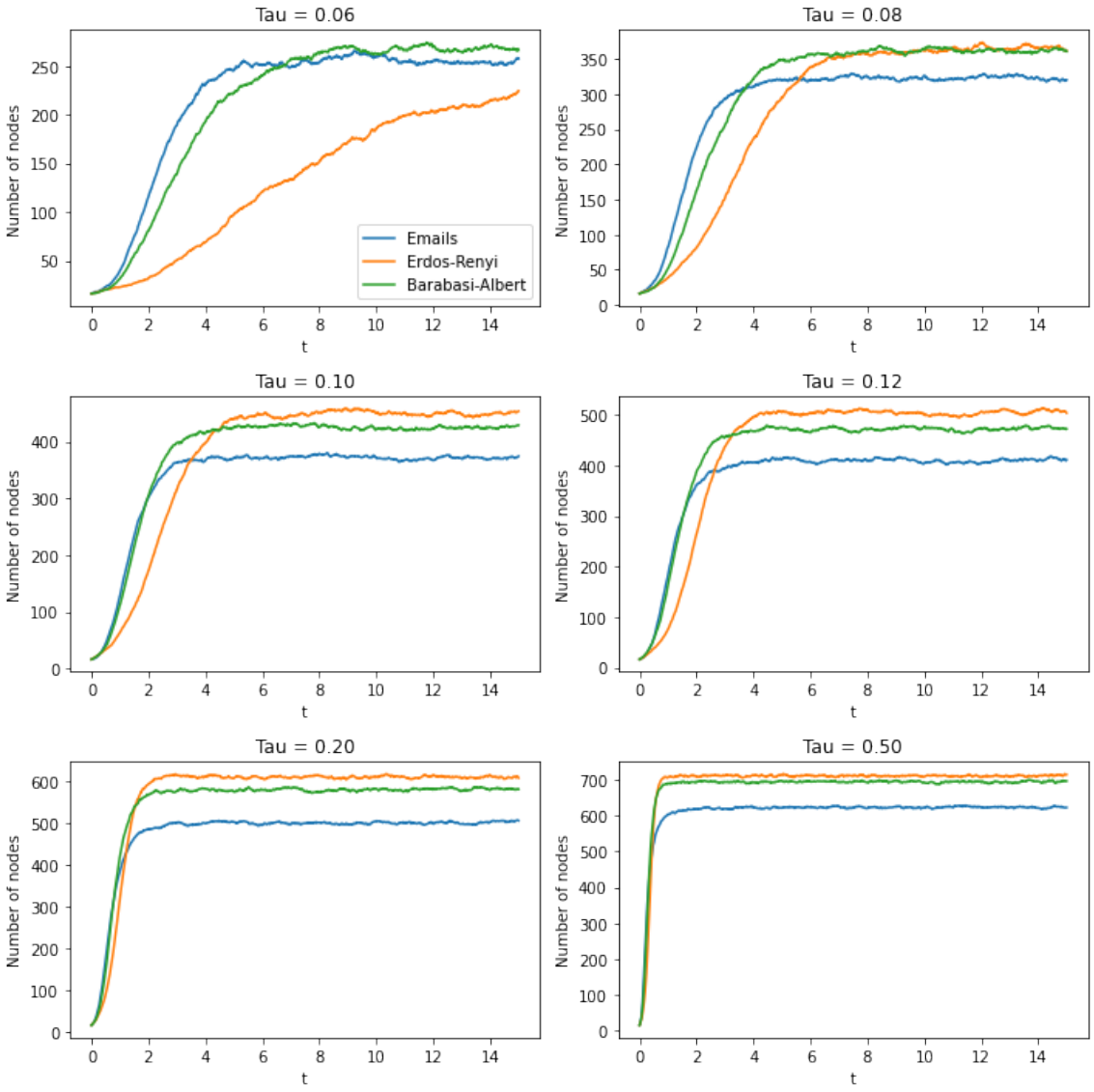

Figure 12: SIS spreading with all three networks,Infected

Figure 12 compares all three networks using infected values. Now there is some differences among used networks. There is similar behavior in curves even if tau value changes. In the beginning Email and Barabasi-Albert network spreads a lot faster than Erdos-Renyi network. However, Erdos-Renyi network gets more nodes infected than other two networks in the end. So, infections are increasing very fast in the beginning and then stops increasing with Email and Barabasi-Albert network. Spreading stops earlier in Email network than

for other networks. This might happen due to community structure.

4.2 SIR

I tested spreading phenomena with six different tau-values which mean how likely node infects another node. Next three figures shows S,I and R curves with all six tau values and three used networks.

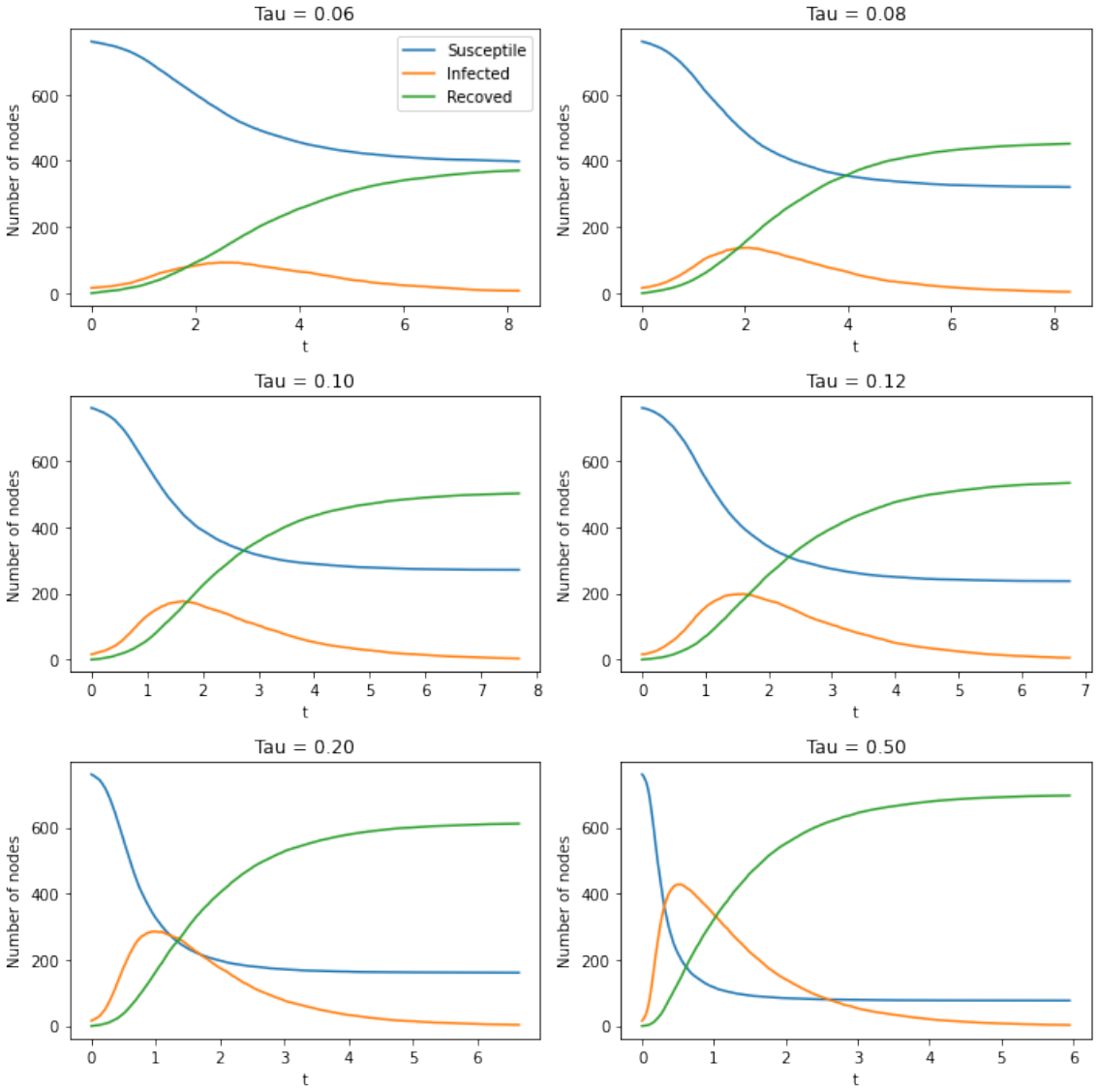

Figure 13: SIR spreading Email network

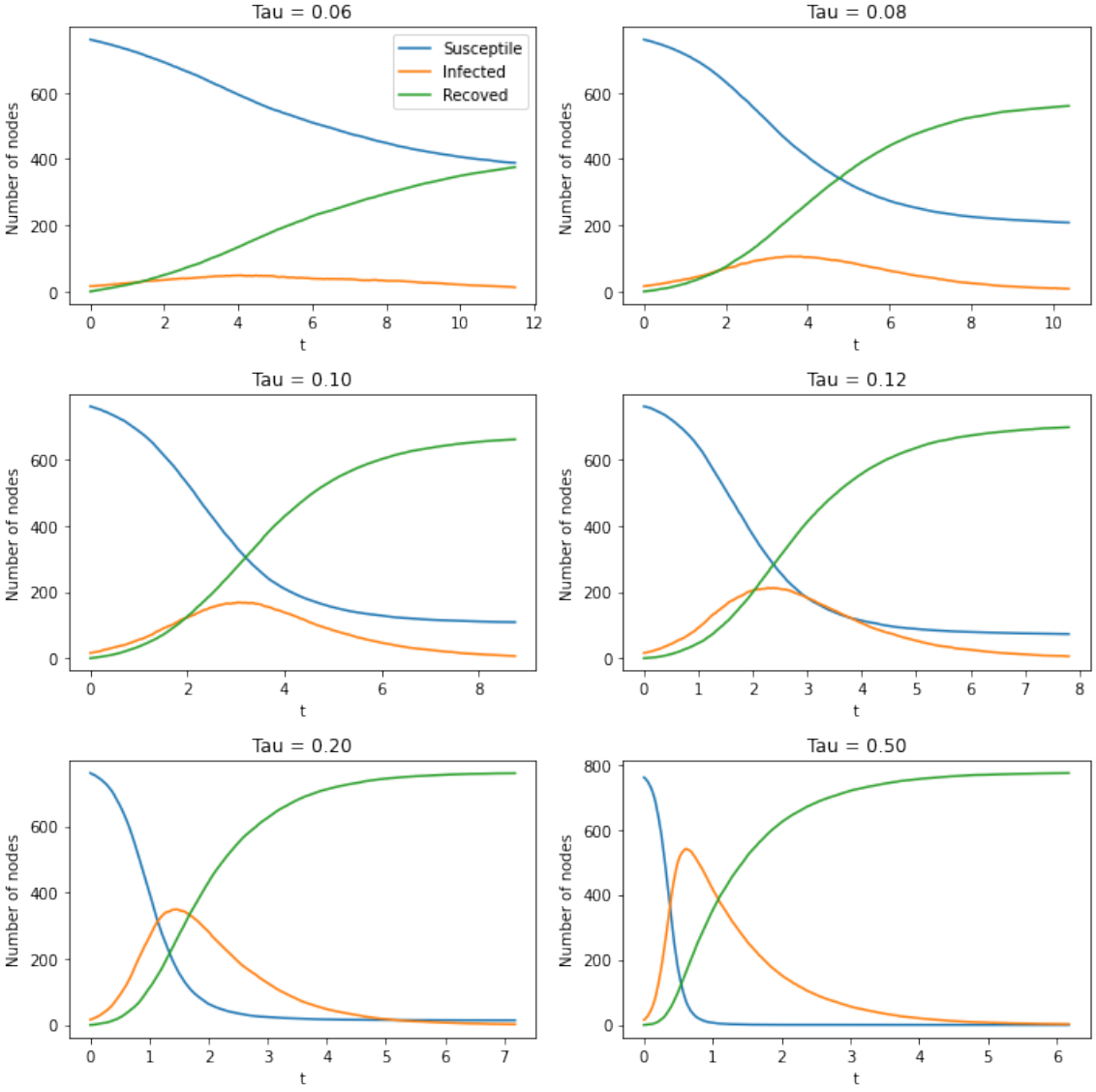

Figure 14: SIR spreading Erods-Renyi network

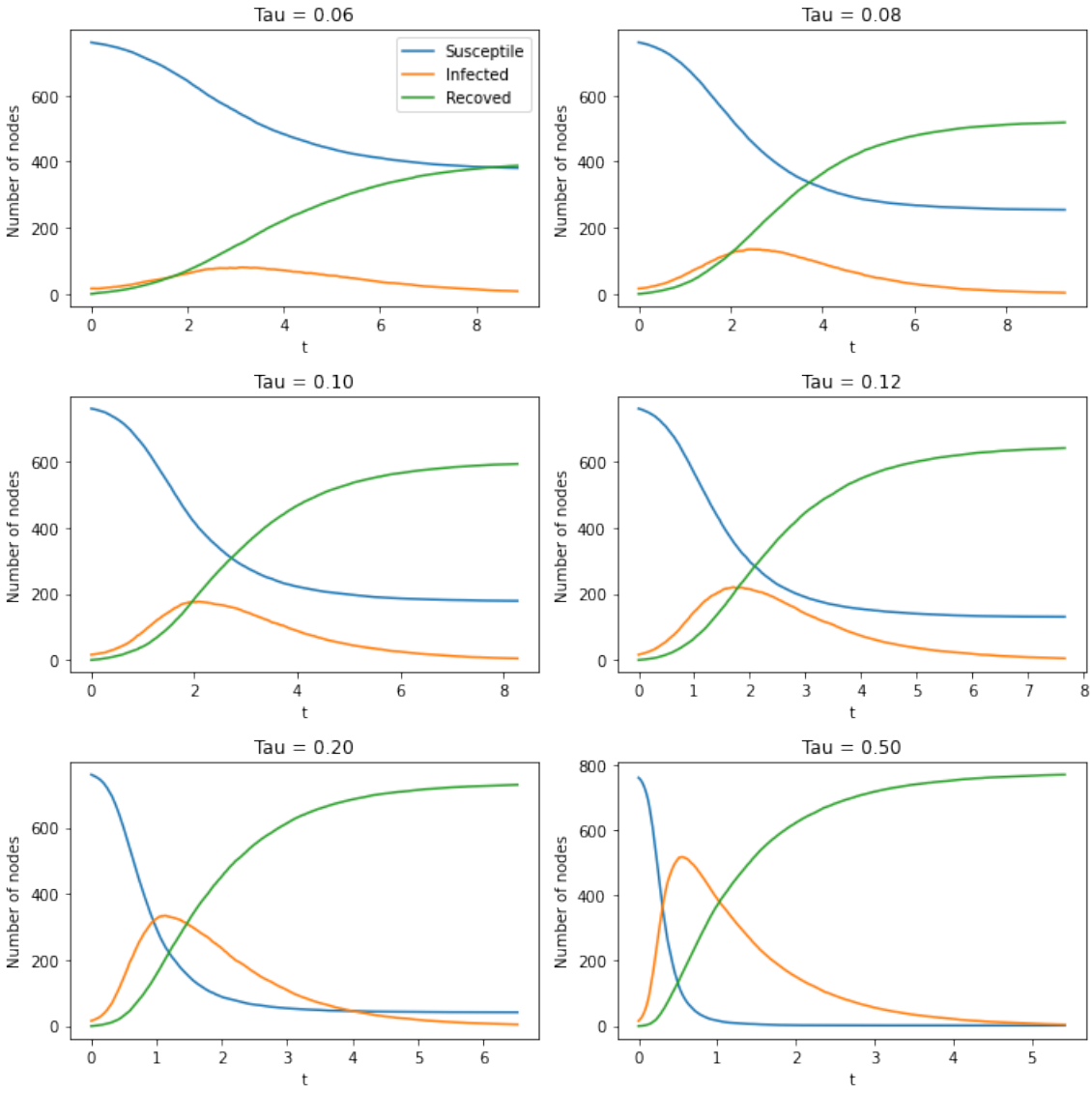

Figure 15: SIR spreading Barabasi-Albert network

Figures 16 and 17 compares networks with different tau-values. Figure 16 shows infected-curves and figure 17 recoved values. As in SIS simulation Email and Barabasi-Albert networks resembles each other and spreading happens very fast in the beginning. Also spreading in the Email network dies faster than Barabasi-Albert network.

Figure 16 shows same properties than SIS simulation. Erdos-Renyi network gets most nodes infected, Barabasi-Albert network second most and Email

network least nodes get infected. In case where tau = 0.5 Erdos-Renyi and Barabasi-Albert network gets same amount of nodes infected, I think this happens because all nodes in network have infected.

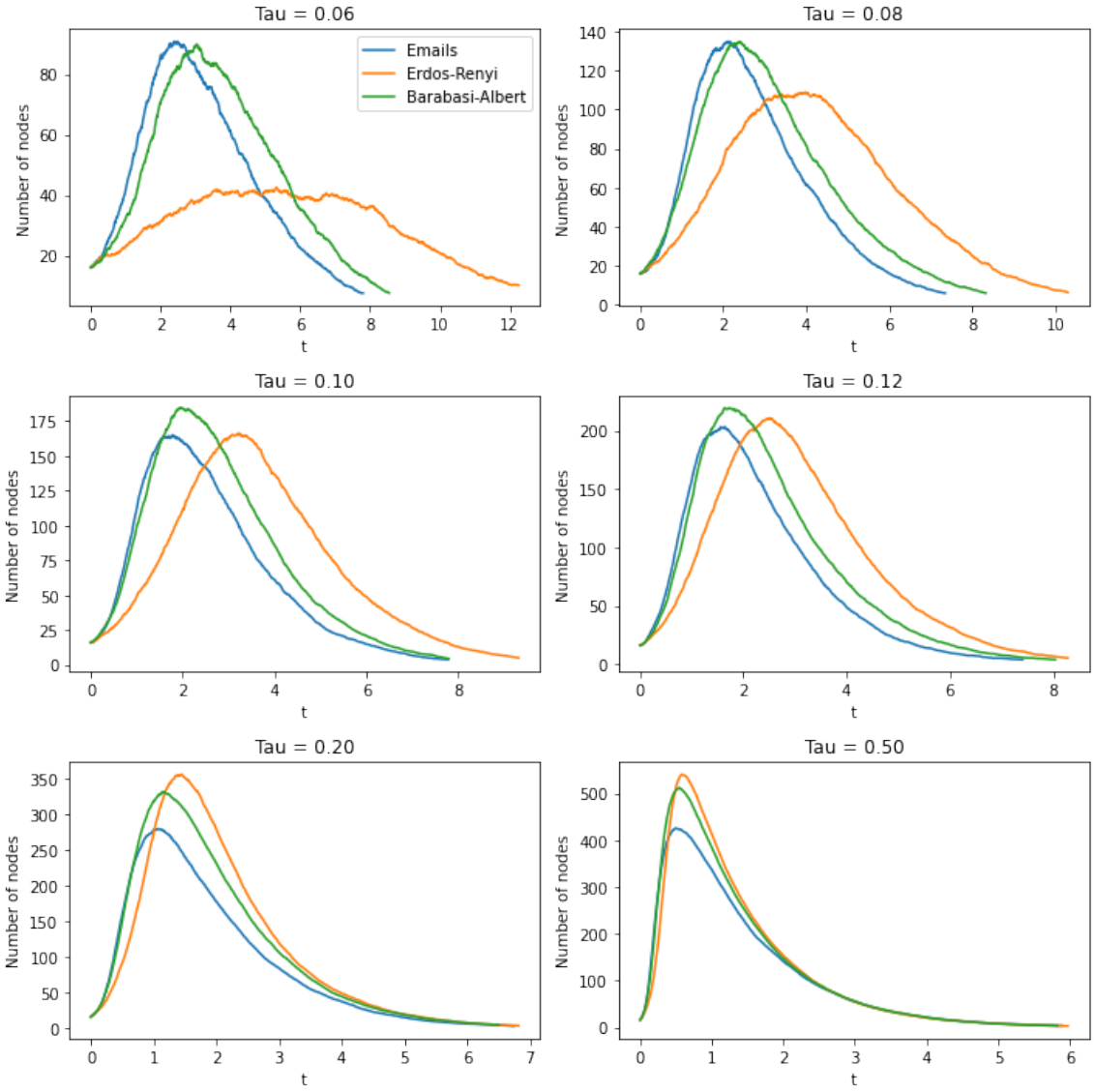

Figure 16: SIR spreading with all three networks,Infected

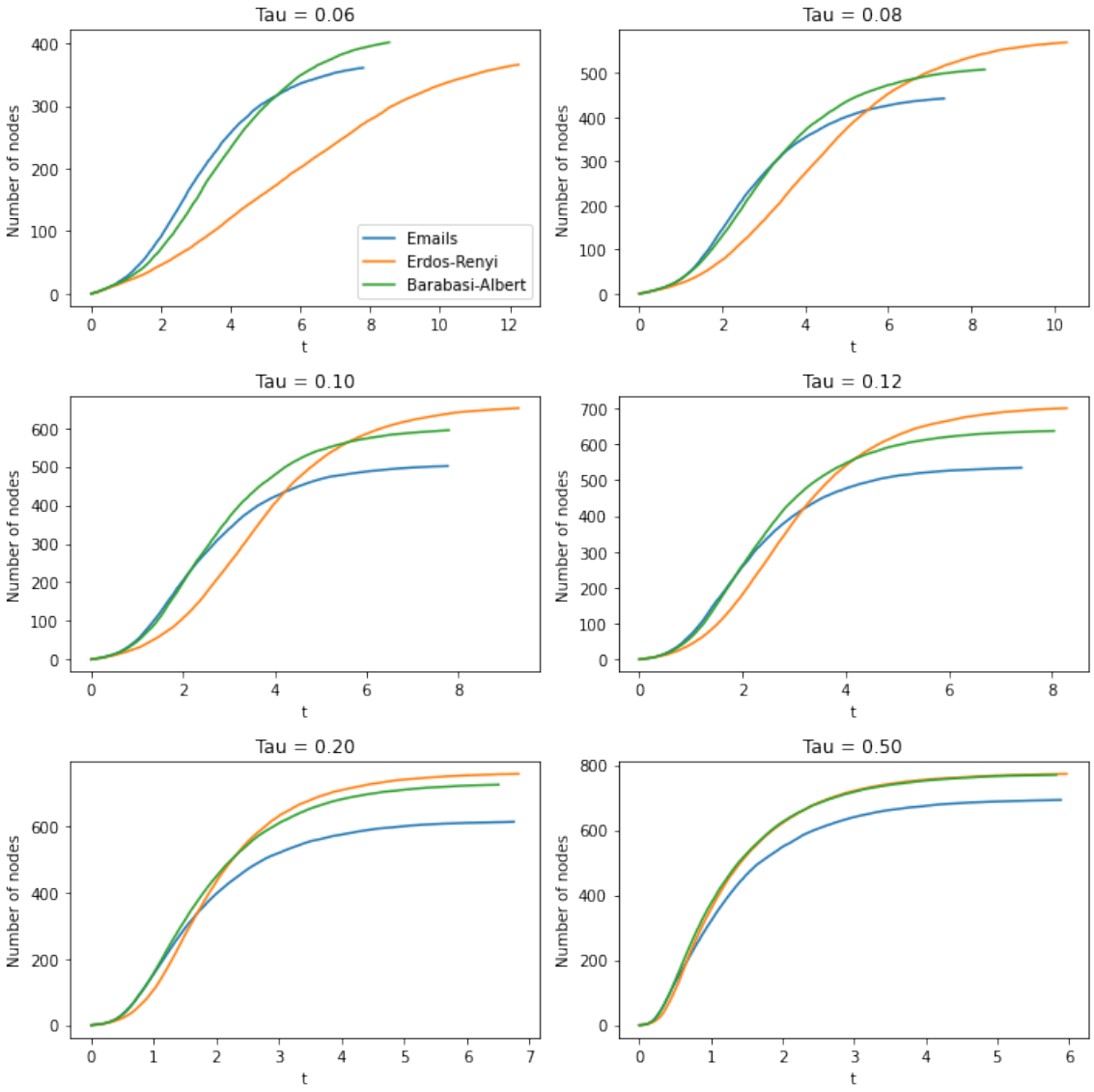

Figure 17: SIR spreading with all three networks,Recoved

I also tested different rho values, rho value indicates how many nodes get infected in the beginning. Figure 18 shows that the bigger rho value leads to more nodes to get infected in all three networks. Seems that Email network gets less nodes infected than other networks, I didn't find other differences.

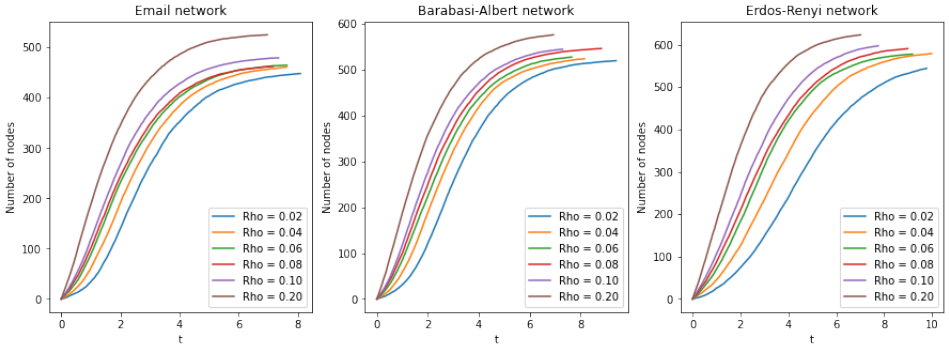

Figure 18: SIR spreading with all three networks and different rho values,Recoved