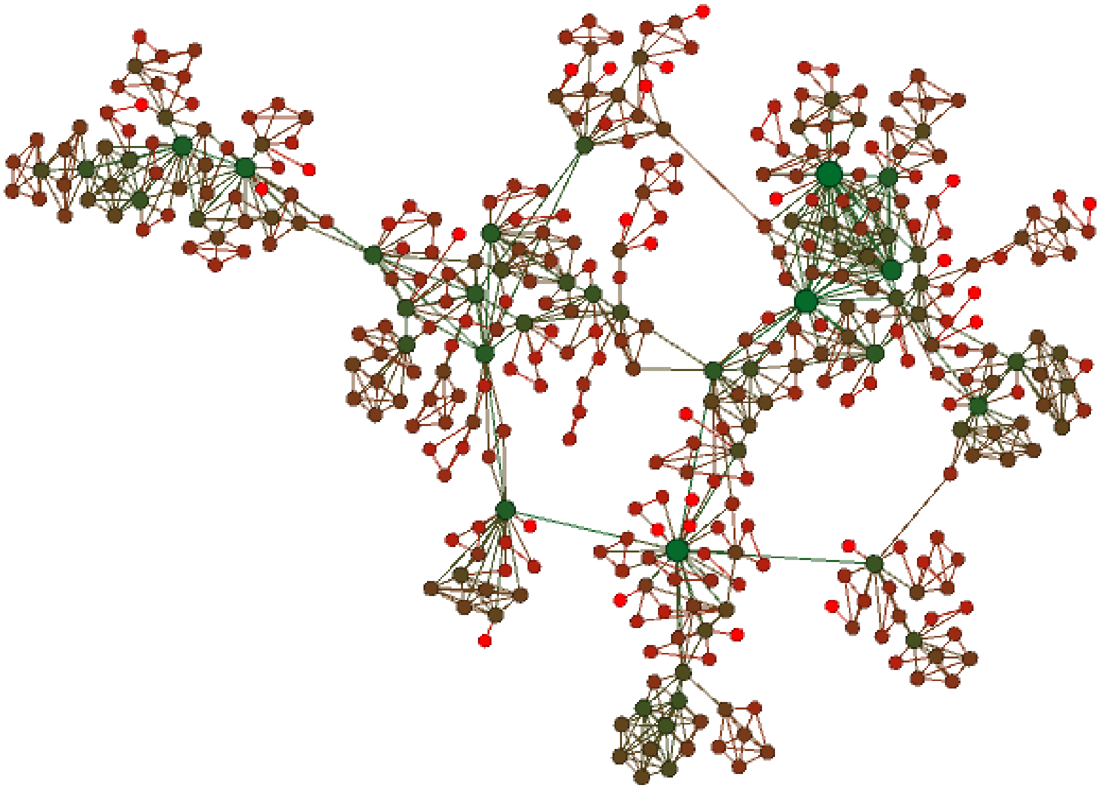
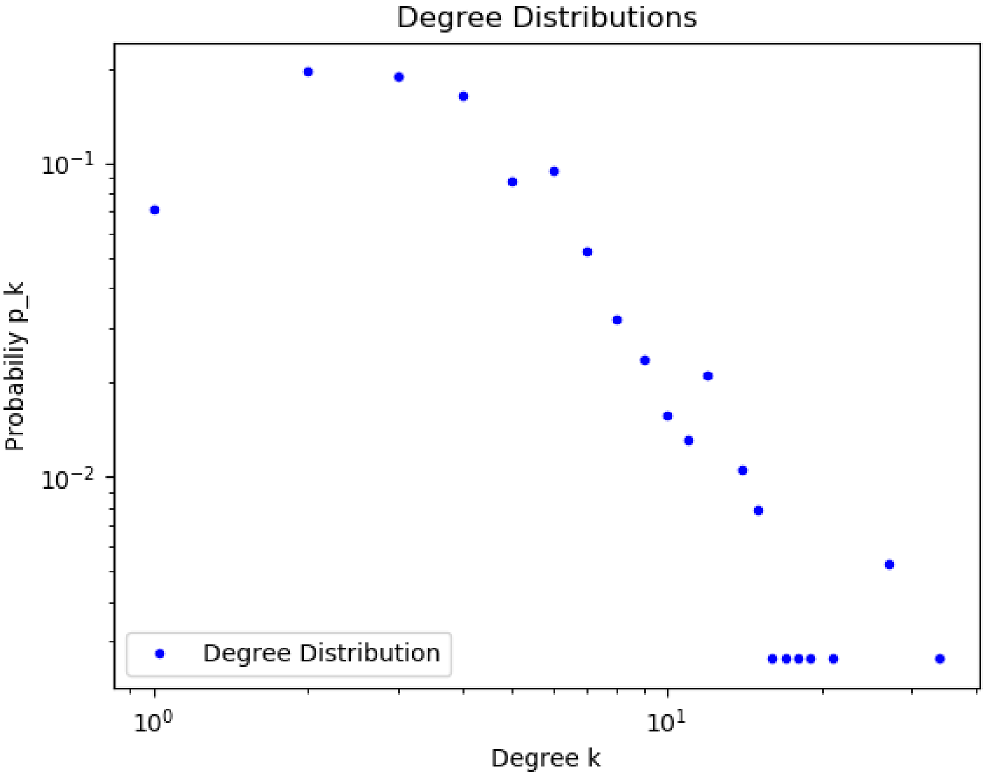
1a.



Larger nodes have a higher degree. Green nodes have high degree, light red nodes have low degree

1b.
Average shortest Path: 6.04186734794
Average clustering coefficient: 0.741230614293
Degree distribution:

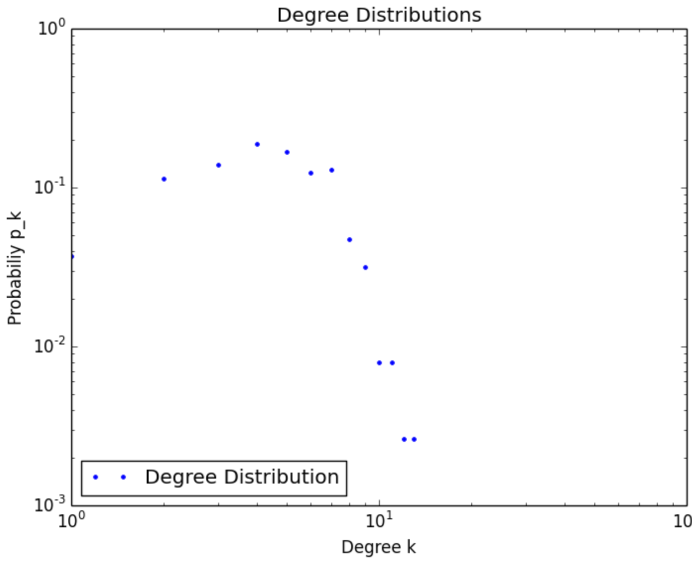


1c. Repository: homework-5-terrylu_kinaanpatel

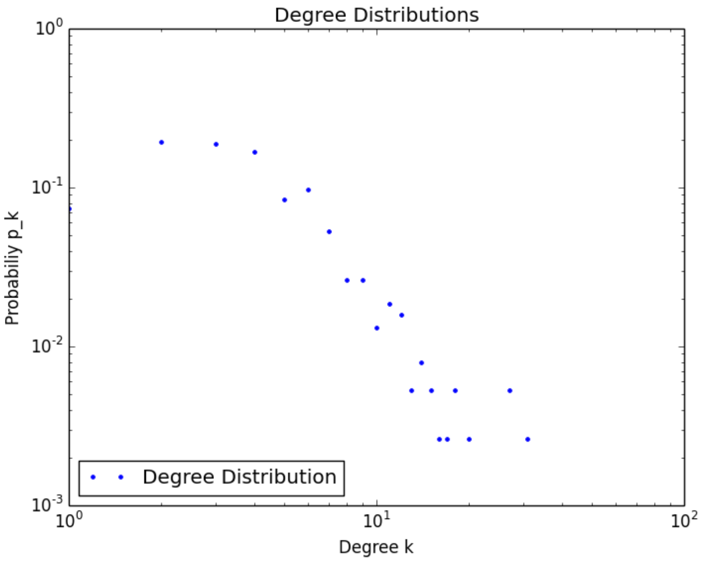
1d.

	Average path length	Average clustering coeffcient
Part b	6.04186734794	0.741230614293
ER	3.94220379445	0.0213711383896
SW	7.96138543368	0.397625329815
Config	3.78595859335	0.0247152744429
BA model	3.59714369477	0.0600561973557

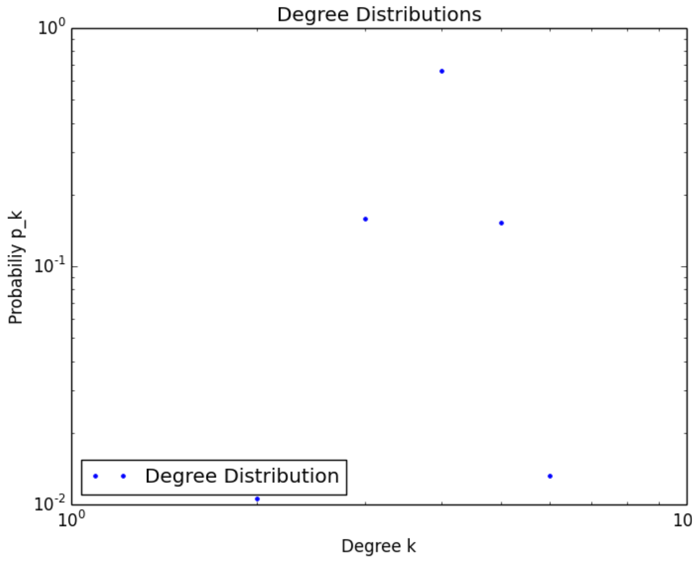
ER:



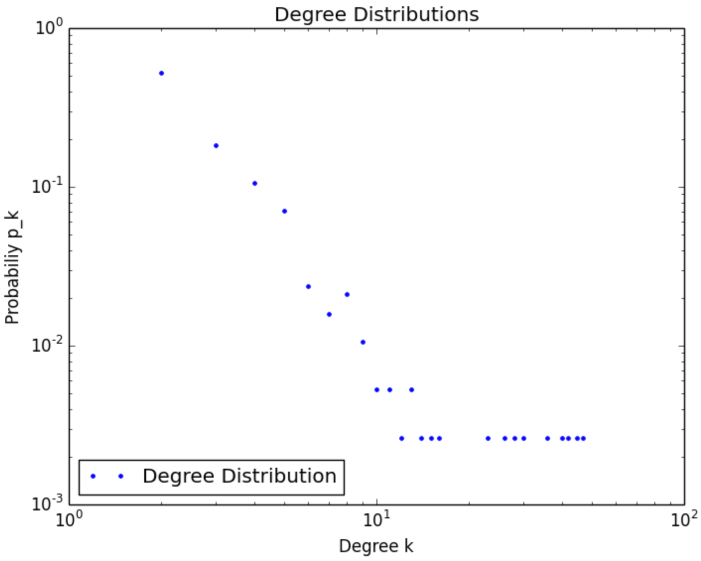
Configuration Model:



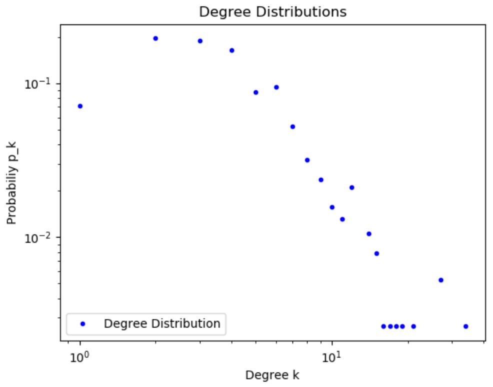
SW:



BA-graph:



Real world:



The small world model best represents the average path length and clustering coefficient in the real world model, however the degree distribution of the configuration model best matches the co-authorship network. There is no clear best model because the degree distribution would point to a model that has very different statistics than the real world model.

1e.

Small-world phenomenon: The network shows a little bit of the small world phenomenon because the distances between random people isn't that large (average of just over 6), but this is a small network, so it is hard to determine if the small-world phenomenon would hold if the network was expanded.

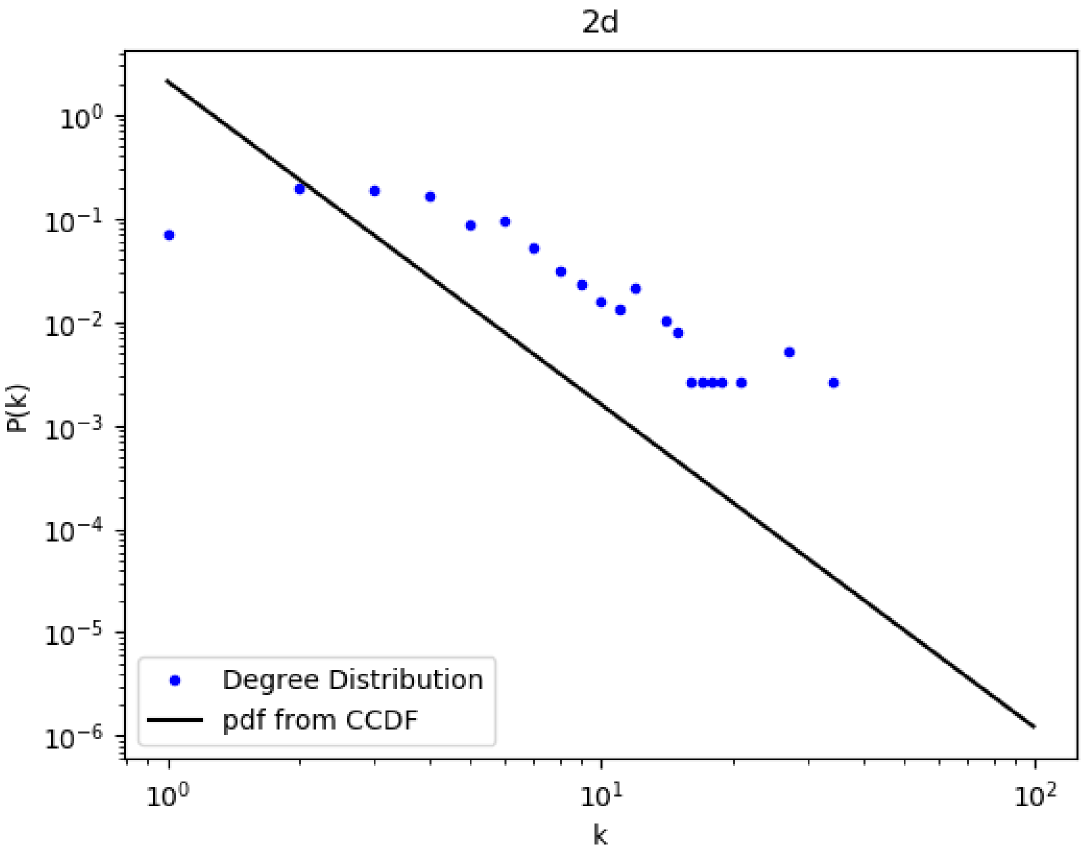
Hubs: There are definitely hubs throughout the co-authorship network. These can be seen in part A as the large green nodes that have many neighbors.

Rich-get-richer phenomenon: The rich get richer phenomenon is not fully present because the network does not display a true powerlaw degree distribution because there are not as many as expected nodes with degree 1. If the network had not been altered to only have the giant component, however, there were many more nodes with very small degrees that could have completed the powerlaw distribution which indicates the rich-get-richer effects.

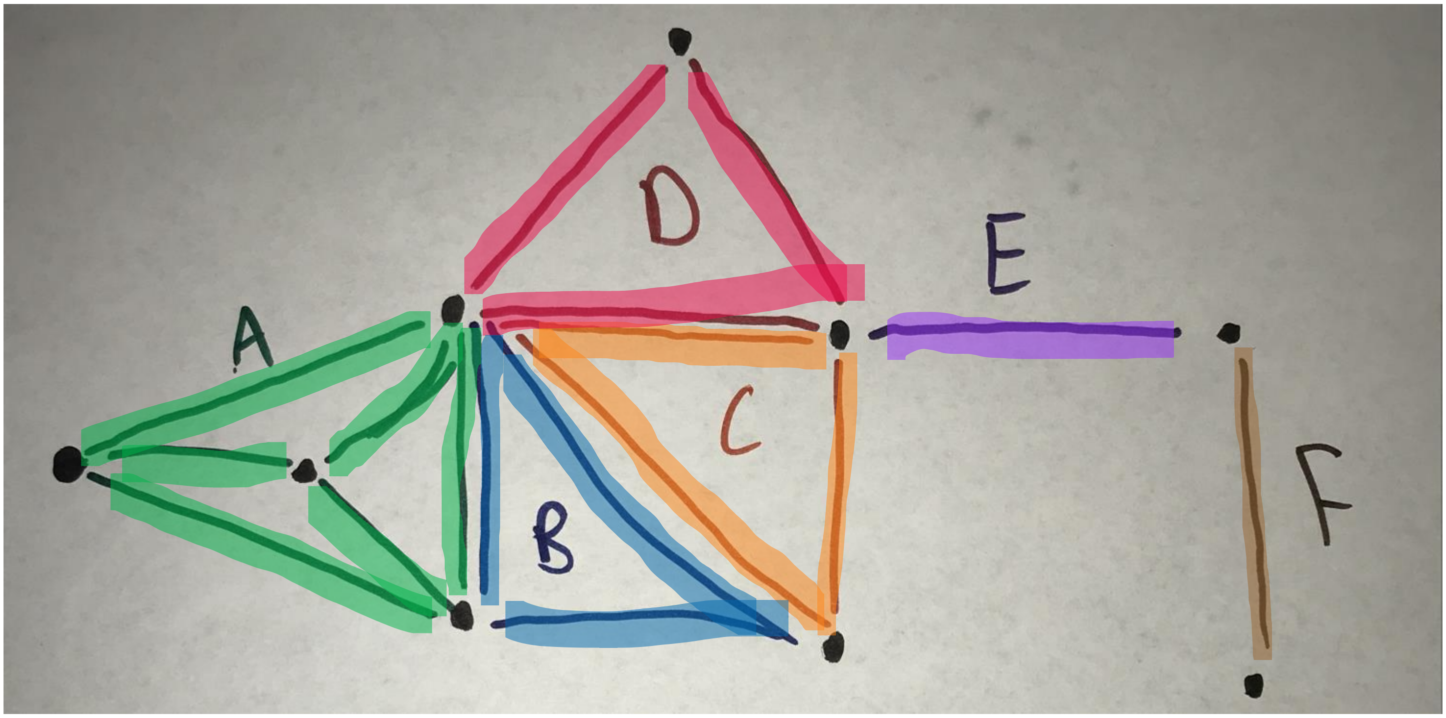
Local structure: This group does show local structure because there are many smaller groups (usually centered around one or two hubs) that are then connected.

1f [BONUS]

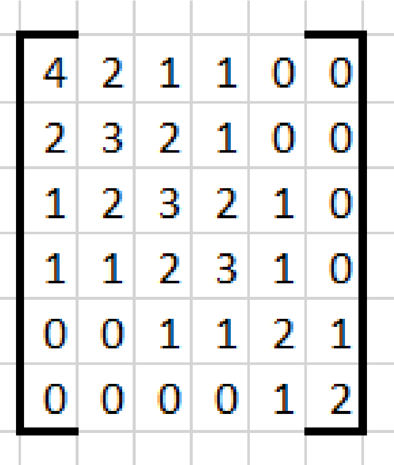
Using a CCDF, alpha = 3.12415997995



2a.



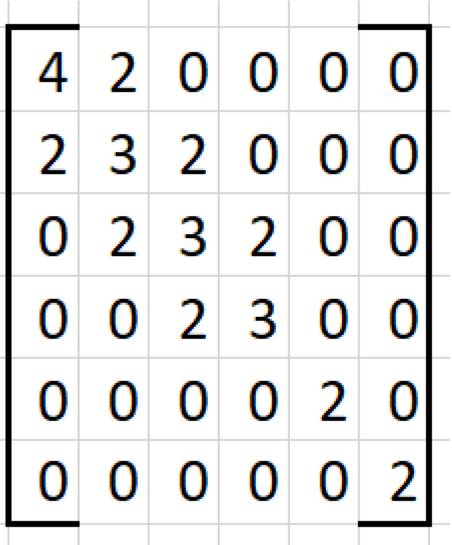
2b.



$$\begin{bmatrix} 4 & 2 & 1 & 1 & 0 & 0 \\ 2 & 3 & 2 & 1 & 0 & 0 \\ 1 & 2 & 3 & 2 & 1 & 0 \\ 1 & 1 & 2 & 3 & 1 & 0 \\ 0 & 0 & 1 & 1 & 2 & 1 \\ 0 & 0 & 0 & 0 & 1 & 2 \end{bmatrix}$$

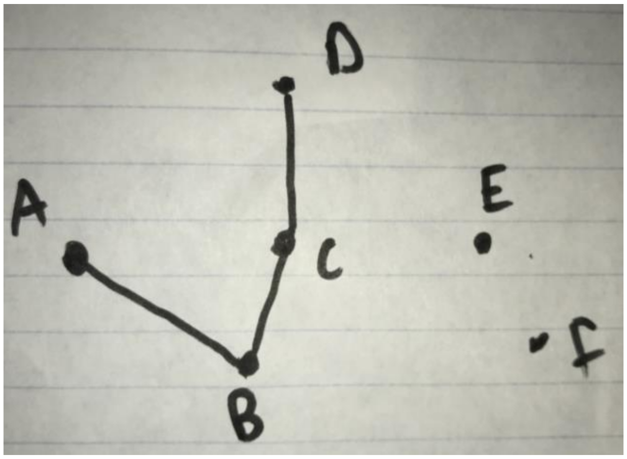
2c.

Threshold matrix

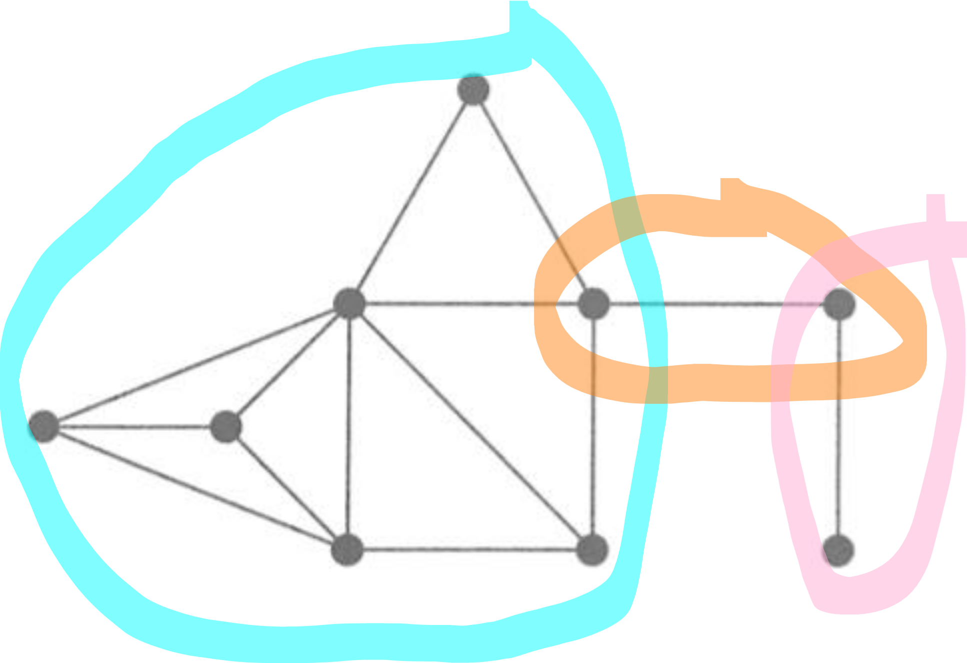


$$\begin{bmatrix} 4 & 2 & 0 & 0 & 0 & 0 \\ 2 & 3 & 2 & 0 & 0 & 0 \\ 0 & 2 & 3 & 2 & 0 & 0 \\ 0 & 0 & 2 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2 \end{bmatrix}$$

Resulting communities:



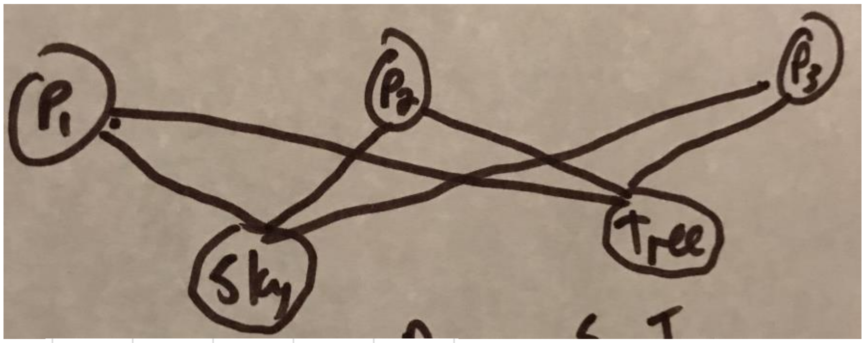
Therefore ABCD is a community, and E is a community, and f is a community



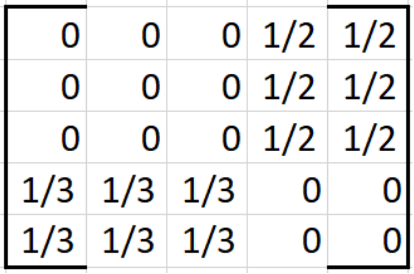
3a.

Picture 3 seems more similar to picture 1 because they have overlap in all of their items where as only half of the items for picture 1 overlap with picture 2.

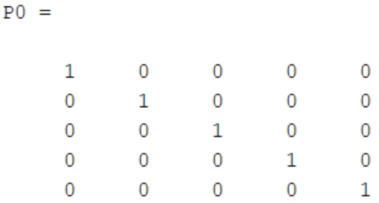
3b.



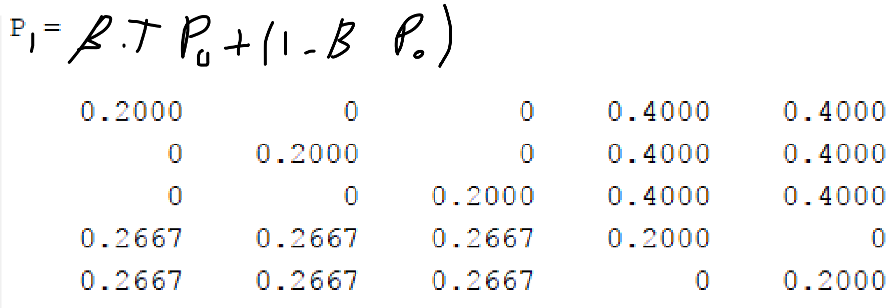
T =



0	0	0	1/2	1/2
0	0	0	1/2	1/2
0	0	0	1/2	1/2
1/3	1/3	1/3	0	0
1/3	1/3	1/3	0	0

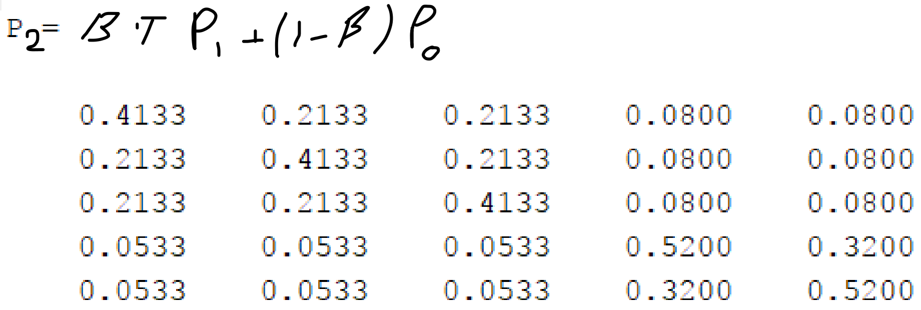


```
P0 =

     1     0     0     0     0
     0     1     0     0     0
     0     0     1     0     0
     0     0     0     1     0
     0     0     0     0     1
```



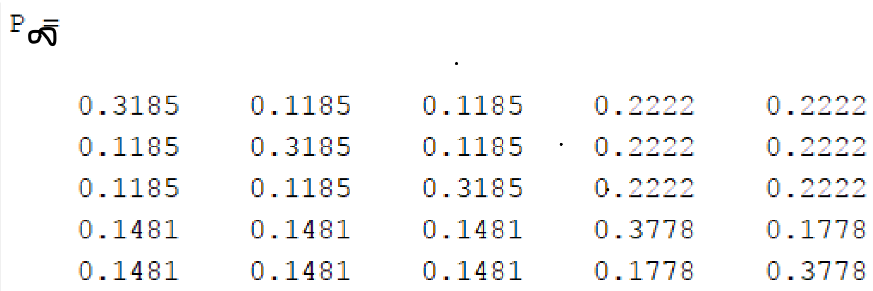
$P_1 = \beta \cdot T\ P_0 + (1 - \beta\ P_0)$

```
    0.2000         0         0    0.4000    0.4000
         0    0.2000         0    0.4000    0.4000
         0         0    0.2000    0.4000    0.4000
    0.2667    0.2667    0.2667    0.2000         0
    0.2667    0.2667    0.2667         0    0.2000
```



$P_2 = \beta \cdot T\ P_1 + (1 - \beta)\ P_0$

```
    0.4133    0.2133    0.2133    0.0800    0.0800
    0.2133    0.4133    0.2133    0.0800    0.0800
    0.2133    0.2133    0.4133    0.0800    0.0800
    0.0533    0.0533    0.0533    0.5200    0.3200
    0.0533    0.0533    0.0533    0.3200    0.5200
```

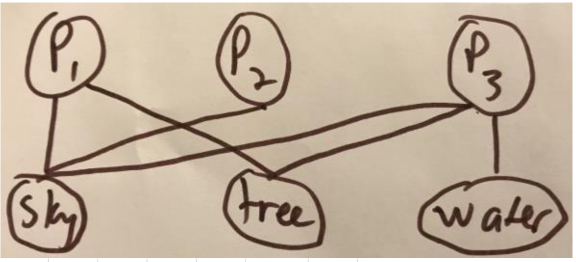


$P_\infty =$

```
    0.3185    0.1185    0.1185    0.2222    0.2222
    0.1185    0.3185    0.1185    0.2222    0.2222
    0.1185    0.1185    0.3185    0.2222    0.2222
    0.1481    0.1481    0.1481    0.3778    0.1778
    0.1481    0.1481    0.1481    0.1778    0.3778
```

Picture 2 and picture 3 are equally similar to picture 1 according to the first row of the P_{infinity} matrix

3c.



$$T = \begin{bmatrix} 0 & 0 & 0 & 1/2 & 1/2 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$

```
P0 =

     1     0     0     0     0     0
     0     1     0     0     0     0
     0     0     1     0     0     0
     0     0     0     1     0     0
     0     0     0     0     1     0
     0     0     0     0     0     1
```

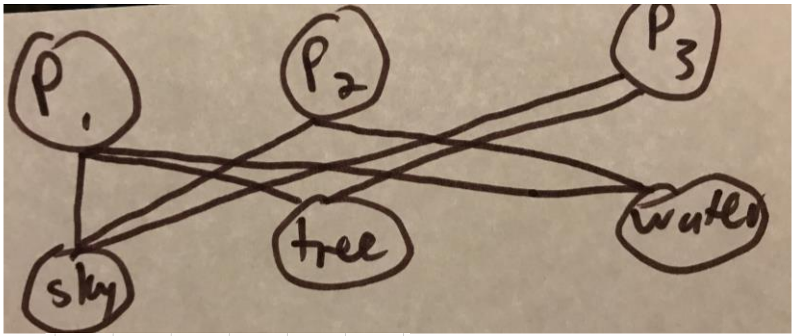
```
P1 =

    0.2000         0         0    0.4000    0.4000         0
         0    0.2000         0    0.8000         0         0
         0         0    0.2000    0.2667    0.2667    0.2667
    0.2667    0.2667    0.2667    0.2000         0         0
    0.4000         0    0.4000         0    0.2000         0
         0         0    0.8000         0         0    0.2000
```

```
P2=

    0.4667    0.1067    0.2667    0.0800    0.0800         0
    0.2133    0.4133    0.2133    0.1600         0         0
    0.1778    0.0711    0.5911    0.0533    0.0533    0.0533
    0.0533    0.0533    0.0533    0.5911    0.1778    0.0711
    0.0800         0    0.0800    0.2667    0.4667    0.1067
         0         0    0.1600    0.2133    0.2133    0.4133
```

```
P∞=

    0.3302    0.0597    0.1656    0.2240    0.1762    0.0442
    0.1195    0.2842    0.1519    0.3156    0.0883    0.0405
    0.1104    0.0506    0.3945    0.1899    0.1494    0.1052
    0.1494    0.1052    0.1899    0.3945    0.1104    0.0506
    0.1762    0.0442    0.2240    0.1656    0.3302    0.0597
    0.0883    0.0405    0.3156    0.1519    0.1195    0.2842
```

According to the first row of $P_{infinity}$, picture 1 is most like picture 3.

3d.



$$T = \begin{bmatrix} 0 & 0 & 0 & 1/3 & 1/3 & 1/3 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1/2 & 1/2 & 0 \\ 1/3 & 1/3 & 1/3 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 & 0 & 0 \end{bmatrix}$$

```
P0 =

     1     0     0     0     0     0
     0     1     0     0     0     0
     0     0     1     0     0     0
     0     0     0     1     0     0
     0     0     0     0     1     0
     0     0     0     0     0     1

P1 =

    0.2000         0         0    0.2667    0.2667    0.2667
         0    0.2000         0    0.4000         0    0.4000
         0         0    0.2000    0.4000    0.4000         0
    0.2667    0.2667    0.2667    0.2000         0         0
    0.4000         0    0.4000         0    0.2000         0
    0.4000    0.4000         0         0         0    0.2000

P2 =

    0.4844    0.1778    0.1778    0.0533    0.0533    0.0533
    0.2667    0.4667    0.1067    0.0800         0    0.0800
    0.2667    0.1067    0.4667    0.0800    0.0800         0
    0.0533    0.0533    0.0533    0.4844    0.1778    0.1778
    0.0800         0    0.0800    0.2667    0.4667    0.1067
    0.0800    0.0800         0    0.2667    0.1067    0.4667

P∞ =

    0.3544    0.1006    0.1006    0.1750    0.1347    0.1347
    0.1508    0.3214    0.0833    0.2021    0.0735    0.1688
    0.1508    0.0833    0.3214    0.2021    0.1688    0.0735
    0.1750    0.1347    0.1347    0.3544    0.1006    0.1006
    0.2021    0.0735    0.1688    0.1508    0.3214    0.0833
    0.2021    0.1688    0.0735    0.1508    0.0833    0.3214
```

Picture 2 and picture 3 are equally similar to picture 1 according to the first row of the $P_{infinity}$ matrix

4a.

https://github.com/MarionWashU/homework-5-terrylu_kinaanpatel

Bonus portion

4b.

	Percent of epidemics	Mean percent of people infected in epidemics	CHI square for likelihood of epidemics	P value for likelihood of epidemics	U stat for proportion of people infected	P value for proportion of people infected
ER Graph	82%	91.4%	1.835	0.176	199	$6*10^{-24}$
BA graph	73%	82.2%				

i)
Because the p value for the chi square estimate of likelihood to get epidemics is so high, it is insignificant to conclude the difference in susceptibility to epidemics.

ii)
because p value for U test is close to zero, we are confident to conclude that ER graph has larger final percentage infected if an epidemic occurs.

iii)
The likelihood of becoming an epidemic depends heavily on node degrees and which node is infected, and the probability of infection. Since the starting infected node is random in part b, we are unable to conclude which graph leads to an epidemic. The random ER graph leads to a higher proportion of people infected because of the small world phenomenon causing random nodes to be relatively close to each other.

4c

	Mean percent of people infected in epidemics	U stat for proportion of people infected	P value for proportion of people infected
ER Graph rand	90.8%	3324	0.326
ER graph High node	91.0%		

Part c : so the relative increase of percentage of people infected in epidemic compared to part b is. 1) for ER graph, the increase is 91% - 90.8% = 0.2%. , but it has a p value of 0.326, much larger than 0.10, so the result is not significant. 2) The relative increase for BA model graph is 81.864% - 81.003% = 0.861%. And the p value is about 0.0986, very close to 0.10. The result is statistically significant.

	Mean percent of people infected in epidemics	U stat for proportion of people infected	P value for proportion of people infected
BA Graph rand	81.003%	3258	0.0985998680957
BA graph High node	81.864%		

As a result, below is our answer for q4 part c

i) BA model graph seems to be more impacted by the targeting of the highest degree node.

ii) The nodes in BA model graph , compared to nodes in ER model graphs, have several nodes with extremely high degrees, serving as hubs. So if we infect hub nodes in the BA model graph, clearly it is much more susceptible to epidemic.

4d.

	Percent of epidemics	Mean percent of people infected in epidemics	CHI square for likelihood of epidemics	P value for likelihood of epidemics	U stat for proportion of people infected	P value for proportion of people infected
ER Graph	85%	91.1%	2.04	0.153	25.5	$9.4*10^{-28}$
Jazz graph	76%	78.6%				

	Percent of epidemics	Mean percent of people infected in epidemics	CHI square for likelihood of epidemics	P value for likelihood of epidemics	U stat for proportion of people infected	P value for proportion of people infected
BA Graph	66%	82.1%	1.97	0.16	1179.5	$2.7*10^{-8}$
Jazz graph	76%	78.6%				

The real network seems to be less susceptible than BA graphs, but more susceptible than ER graphs. When an epidemic occurs, the real jazz network also has fewer infected people, on average, than the synthetic graphs

Community structure plays an important role because while in a random graph everyone is very susceptible to catch an infection, in real-life graphs, those within a sub-group due to local structure, are more likely to get infected, but often the infection will not spread to all sub-groups because there are fewer connections between groups compared to in groups.