APRIL, 2021

WRITTEN ANALYSIS

NETWORK

Shynuie

PROBLEM 1

SUGGESTING SIMILAR PAPERS

A citation network is a directed network where the vertices are academic papers and there is a directed edge from paper A to paper B if paper A cites paper B in its bibliography. **Google Scholar** performs automated citation indexing and has a useful feature that allows users to find similar papers. In the following, we analyze two approaches for measuring similarity between papers.

Two papers are said to be cocited if they are both cited by the same third paper. The edge weights in the cocitation network correspond to the number of cocitations. In this part, we will discover how to compute the (weighted) adjacency matrix of the cocitation network from the adjacency matrix of the citation network.

PART (C) : (2)

HOW DOES THE TIME COMPLEXITY OF YOUR SOLUTION INVOLVING MATRIX MULTIPLICATION IN PART (A) COMPARE TO YOUR FRIEND'S ALGORITHM?

By friend's algorithm, we have n rows in the adjacency matrix A. And we only need to check with $\binom{n}{2}$ different pair nodes in each row. Therefore the complexity for friend's algorithm will be n*n*(n-1)/2, which is O($(n^3 - n^2)/2$). And the complexity of our solution using matrix multiplication will be only O(n^2) because it only involves matrix multiplication. Thus, the computation using solution involving matrix multiplication will be much easier.

PART (D) : (3)

BIBLIOGRAPHIC COUPLING AND COCITATION CAN BOTH BE TAKEN AS AN INDICATOR THAT PAPERS DEAL WITH RELATED MATERIAL. HOWEVER, THEY CAN IN PRACTICE GIVE NOTICEABLY DIFFERENT RESULTS. WHY? WHICH MEASURE IS MORE APPROPRIATE AS AN INDICATOR FOR SIMILARITY BETWEEN PAPERS?

Co-citation would be weighted if both paper cite the same paper. High value in co-citation indicates that there are lots of papers cited by both these two paper. In other word, these two papers involve many same theorem or technique. And bibliographic coupling would be weighted if both paper are cited by the same paper. High value in bibliographic coupling indicates that there are lots of papers citing with both these two paper. In other words, these two papers might be both important, general or even relative to each for some academic filed. However, before determining which measure is more appropriate as an indicator for similarity, we'd better define the similarity of two papers first.

If we say that two papers are similar as they are both important in lots of science topics or fields, then bibliographic might be a proper measure for the similarity.

If we say two papers are similar as they involve lots of same theorem or technique, then co-citation would be a proper measure for the similarity.

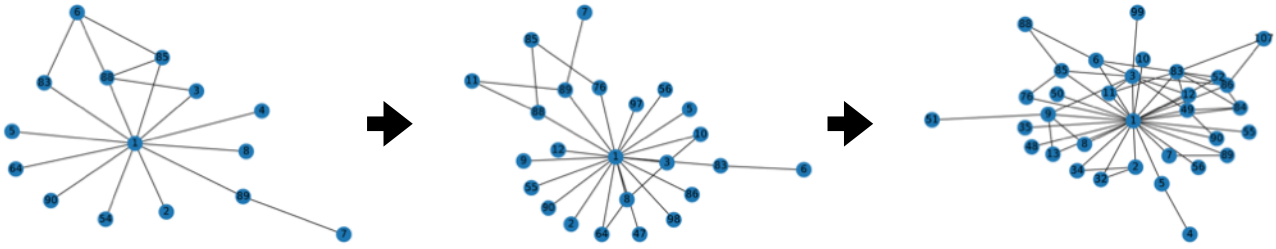
PROBLEM 2

INVESTIGATING A TIME-VARYING CRIMINAL NETWORK

The CAVIAR investigation lasted two years and ran from 1994 to 1996. The operation brought together investigation units of the Montréal police and the Royal Canadian Mounted Police of Canada. During this two year period, 11 wiretap warrants, valid for a period of about two months each, were obtained.

PART (C) : (2)

OBSERVE THE PLOT YOU MADE IN PART (A) QUESTION 1. THE NUMBER OF NODES INCREASES SHARPLY OVER THE FIRST FEW PHASES THEN LEVELS OUT. COMMENT ON WHAT YOU THINK MAY BE CAUSING THIS EFFECT. BASED ON YOUR ANSWER, SHOULD YOU ADJUST YOUR CONCLUSIONS IN PART (B) QUESTION 5?



Phase 1 Phase 2 Phase 3

The number of nodes in the network indicates how big this drug business is. Therefore without arresting any suspects in the first three phases, the business grows really fast. And then police start arresting in phase 4 and cause that the growth of network, or we can say the business, is not as obvious as growths in first 3 phases are. I think the main purpose is that the suspect police arrested is some nodes that somehow temporally highly central to the network.

However, in question 5, we are interested in which players consistently remained active and central throughout most of the phases and which didn't. To make analysis basing on this purpose, I think it is not necessary to adjust the conclusion since we are looking for players consistently remained active and central throughout most of the phases.

PART (D) : (5)

IN THE CONTEXT OF CRIMINAL NETWORKS, WHAT WOULD EACH OF THESE METRICS TEACH YOU ABOUT THE IMPORTANCE OF AN ACTOR'S ROLE IN THE TRAFFIC? IN YOUR OWN WORDS, COULD YOU EXPLAIN THE LIMITATIONS OF DEGREE CENTRALITY? IN YOUR OPINION, WHICH ONE WOULD BE MOST RELEVANT TO IDENTIFY WHO IS RUNNING THE ILLEGAL ACTIVITIES OF THE GROUP? PLEASE JUSTIFY.

The explanation for three centrality measurements is as followed:

Degree centrality-

This measurement will indicate the importance of a node basing on how many other nodes it can contact. Higher degree centrality in this criminal network indicates that this member contacts and interacts lots of other members. However, this measurement can't capture the cascade effect. For example, a member charge in sales could have as many contact member as a real core member has. But the network won't be highly affected if we remove the sales man since it's easy to appoint other member to take this job. So this might be the drawback of the degree centrality.

Betweenness centrality-

This measurement will indicate the importance of a node basing on how many shortest paths between any other pair nodes will pass through this node. With higher proportion, number of shortest paths passing through versus number total shortest paths, it indicates that this node is quite important for the connection of other pair nodes. If we remove node with high betweenness centrality, it will have higher chance to break large fraction of the connections between other nodes.

Eigenvector centrality-

This measurement will indicate the importance of a node basing on the importance of its neighbor nodes. Some one contacts lots of member in the network might be important. However another member contact with this member will be weighted with the importance of this member. For example, a CEO in the company might need to only contact with other chief officers in different apartment. But other chief officers might need to contact with lots of staff, clients or vendors respectively. If we consider the degree of each nodes only, the CEO might not be an important node since it only contact with other chief officers. But if we use eigenvector centrality, the importance of other chief officers will propagate to CEO and the centrality of him will be high enough to indicate the real importance of this position.

With explanation of different centralities discussed above, if our purpose is to identify the core member "who is running the whole illegal group", I will prefer choosing eigenvector centrality measurement as my indicator. However, if we pursue to break the network efficiently, then I will choose to arrest the member with higher betweenness centrality first although this might not bring the network a permanent damage.

PART (E) : (3)

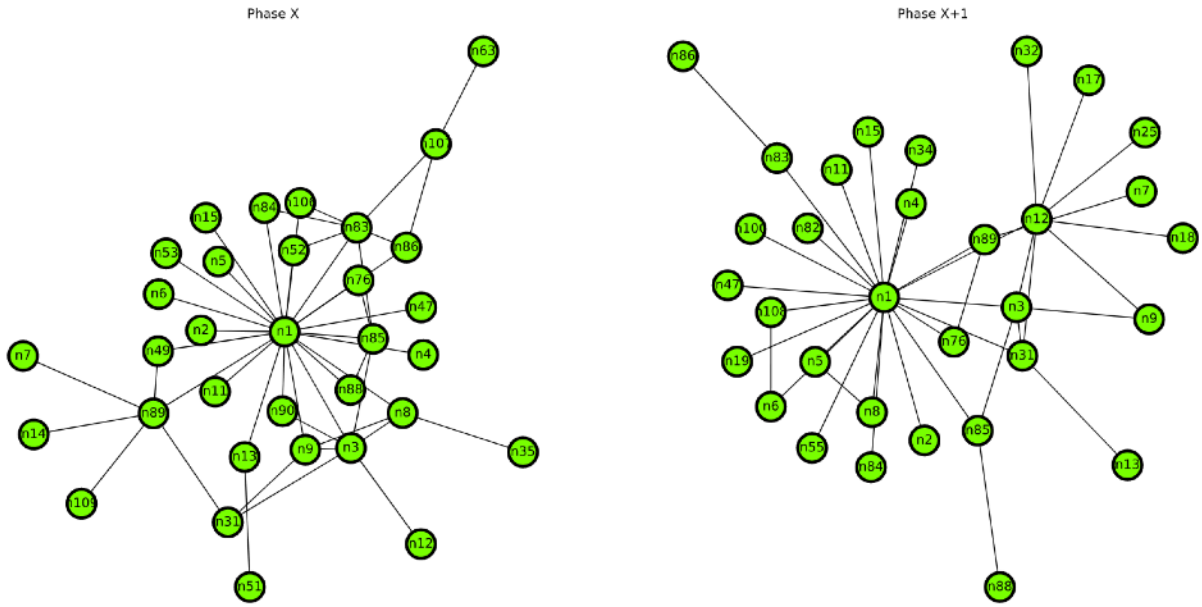
IN REAL LIFE, THE POLICE NEED TO EFFECTIVELY USE ALL THE INFORMATION THEY HAVE GATHERED, TO IDENTIFY WHO IS RESPONSIBLE FOR RUNNING THE ILLEGAL ACTIVITIES OF THE GROUP. ARMED WITH A QUALITATIVE UNDERSTANDING OF THE CENTRALITY METRICS FROM PART (D) AND THE QUANTITATIVE ANALYSIS FROM PART PART (B) QUESTION 5, INTEGRATE AND INTERPRET THE INFORMATION YOU HAVE TO IDENTIFY WHICH PLAYERS WERE MOST CENTRAL (OR IMPORTANT) TO THE OPERATION.

According the result from betweenness centrality, we may find that player 1, 12 and 3 are the players having first three highest centrality. Which means that most of the paths in the network will pass through these three players. According to the roles list provided from the police, these three respectively are mastermind of the network, principal importer and principal lieutenant. It's obvious that their absence will bring huge damage to the network. However, if we want to identify who is responsible for running the illegal activities of the group, I will prefer using eigenvector centrality to identify the players connect to most important players. According to the result, the eigenvector centrality

indicated that player 1, 3 and 85 are the first three players contacting most important players. According to the roles list provided from the police, these three respectively are mastermind of the network, principal lieutenant and accountant. And these three roles are most important to the whole criminal network as they are the roles closest to the mastermind of the network among all phases.

PART (F) QUESTION 2 : (3)

THE CHANGE IN THE NETWORK FROM PHASE X TO X+1 COINCIDES WITH A MAJOR EVENT THAT TOOK PLACE DURING THE ACTUAL INVESTIGATION. IDENTIFY THE EVENT AND EXPLAIN HOW THE CHANGE IN CENTRALITY RANKINGS AND VISUAL PATTERNS, OBSERVED IN THE NETWORK PLOTS ABOVE, RELATES TO SAID EVENT.



By comparing with patterns corresponds to each phase, we can know that these two patterns correspond to phase 4 and phase 5 respectively. According to the seizures information, we know that there's one seizure taking place at the end of phase 4. Through cross phases indexes checking, I found that there are 11 players absent and 10 new players participating in the network in phase 5.

Then we want to view the difference in centrality caused by the seizure. Following tables show the information of betweenness centrality and eigenvalue centrality respectively.

Both tables are sorted by values in Phase 4.

When we look to the table of betweenness centrality, we may find that centrality of 4 remained members decrease to 0. The proportion of nonzero-centrality-nodes in phase 4 is 33% (11/33), and the proportion in phase 5 is 22% (7/32). We could take this decrease as a sign that some horizontal connections between members has been destroyed or abandoned on purpose. Maybe it is a result of high level member change.

And when we look to the table of eigenvector centrality, we can find that rank high in phase 4 will get decrease in phase 5 except node 1. And rank low in phase 4 will get increase in phase 5.

BETWEENNESS CENTRALITY

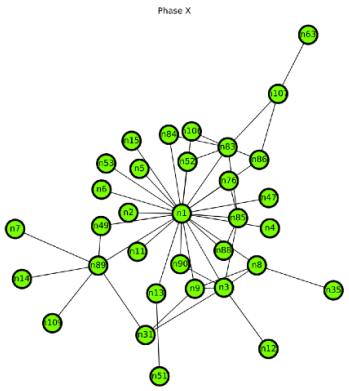
ID	Difference	Phase 4	Phase 5	Rank 4	Rank 5
1	0.0446	0.8393	0.8839	1.0	1.0
89	-0.1317	0.1962	0.0645	2.0	4.5
3	-0.0464	0.0904	0.0441	3.0	7.0
83	-0.0151	0.0796	0.0645	4.0	4.5
13	-0.0625	0.0625	0.0000	6.0	20.0
107	NA	0.0625	NA	6.0	NA
8	-0.0625	0.0625	0.0000	6.0	20.0
86	-0.0474	0.0474	0.0000	8.0	20.0
85	0.0480	0.0165	0.0645	9.0	4.5
9	-0.0148	0.0148	0.0000	10.0	20.0
31	0.0524	0.0121	0.0645	11.0	4.5
51	NA	0.0000	NA	22.5	NA
49	NA	0.0000	NA	22.5	NA
35	NA	0.0000	NA	22.5	NA
52	NA	0.0000	NA	22.5	NA
109	NA	0.0000	NA	22.5	NA
53	NA	0.0000	NA	22.5	NA
106	NA	0.0000	NA	22.5	NA
63	NA	0.0000	NA	22.5	NA
14	NA	0.0000	NA	22.5	NA
12	0.2699	0.0000	0.2699	22.5	2.0
15	0.0000	0.0000	0.0000	22.5	20.0
2	0.0000	0.0000	0.0000	22.5	20.0
4	0.0000	0.0000	0.0000	22.5	20.0
47	0.0000	0.0000	0.0000	22.5	20.0
6	0.0000	0.0000	0.0000	22.5	20.0
11	0.0000	0.0000	0.0000	22.5	20.0
7	0.0000	0.0000	0.0000	22.5	20.0
76	0.0000	0.0000	0.0000	22.5	20.0
84	0.0000	0.0000	0.0000	22.5	20.0
88	0.0000	0.0000	0.0000	22.5	20.0
5	0.0000	0.0000	0.0000	22.5	20.0
90	NA	0.0000	NA	22.5	NA
100	NA	NA	0.0000	NA	20.0
108	NA	NA	0.0000	NA	20.0
17	NA	NA	0.0000	NA	20.0
18	NA	NA	0.0000	NA	20.0
19	NA	NA	0.0000	NA	20.0
25	NA	NA	0.0000	NA	20.0
32	NA	NA	0.0000	NA	20.0
34	NA	NA	0.0000	NA	20.0
55	NA	NA	0.0000	NA	20.0
82	NA	NA	0.0000	NA	20.0

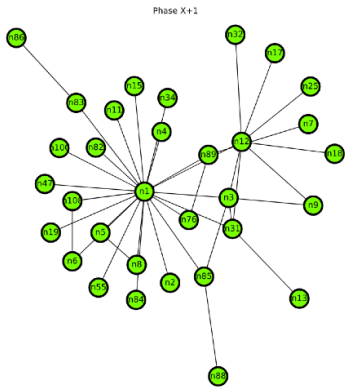
EIGENVECTOR CENTRALITY

ID	Difference	Phase 4	Phase 5	Rank 4	Rank 5
1	0.0298	0.6104	0.6402	1.0	1.0
3	0.0030	0.2726	0.2757	2.0	3.0
83	-0.1443	0.2710	0.1267	3.0	12.0
85	-0.0704	0.2517	0.1813	4.0	5.0
9	-0.0997	0.2059	0.1062	5.0	24.0
8	-0.0436	0.1945	0.1509	6.0	9.5
86	-0.1421	0.1662	0.0242	7.0	32.0
89	-0.0061	0.1629	0.1568	8.0	6.0
90	NA	0.1530	NA	9.0	NA
52	NA	0.1527	NA	11.0	NA
106	NA	0.1527	NA	11.0	NA
84	-0.0306	0.1527	0.1221	11.0	18.0
88	-0.1148	0.1493	0.0346	13.5	30.0
76	0.0027	0.1493	0.1520	13.5	7.0
49	NA	0.1340	NA	15.0	NA
31	0.1258	0.1111	0.2369	16.0	4.0
13	-0.0638	0.1090	0.0452	17.0	29.0
47	0.0164	0.1057	0.1221	21.5	18.0
4	0.0164	0.1057	0.1221	21.5	18.0
15	0.0164	0.1057	0.1221	21.5	18.0
11	0.0164	0.1057	0.1221	21.5	18.0
2	0.0164	0.1057	0.1221	21.5	18.0
5	0.0451	0.1057	0.1509	21.5	9.5
6	0.0451	0.1057	0.1509	21.5	9.5
53	NA	0.1057	NA	21.5	NA
107	NA	0.0781	NA	26.0	NA
12	0.2337	0.0472	0.2810	27.0	2.0
35	NA	0.0337	NA	28.0	NA
7	0.0017	0.0282	0.0299	30.0	31.0
109	NA	0.0282	NA	30.0	NA
14	NA	0.0282	NA	30.0	NA
51	NA	0.0189	NA	32.0	NA
63	NA	0.0135	NA	33.0	NA
100	NA	NA	0.1221	NA	18.0
108	NA	NA	0.1509	NA	9.5
17	NA	NA	0.0536	NA	26.5
18	NA	NA	0.0536	NA	26.5
19	NA	NA	0.1221	NA	18.0
25	NA	NA	0.0536	NA	26.5
32	NA	NA	0.0536	NA	26.5
34	NA	NA	0.1221	NA	18.0
55	NA	NA	0.1221	NA	18.0
82	NA	NA	0.1221	NA	18.0

PART (G) : (4)

WHILE CENTRALITY HELPS EXPLAIN THE EVOLUTION OF EVERY PLAYER'S ROLE INDIVIDUALLY, WE NEED TO EXPLORE THE *GLOBAL* TRENDS AND INCIDENTS IN THE STORY IN ORDER TO UNDERSTAND THE BEHAVIOR OF THE CRIMINAL ENTERPRISE. DESCRIBE THE COARSE PATTERN(S) YOU OBSERVE AS THE NETWORK EVOLVES THROUGH THE PHASES. DOES THE NETWORK EVOLUTION REFLECT THE BACKGROUND STORY?





NETWORK INFO

Phase	Network diameter	Avg degree	Avg path lenth	Edge density	Transitivity	Homophily
Phase 1	4.00	2.40	2.03	0.17	0.11	-0.65
Phase 2	4.00	2.33	2.18	0.10	0.05	-0.57
Phase 3	4.00	3.39	2.13	0.11	0.16	-0.43
Phase 4	5.00	2.91	2.39	0.09	0.14	-0.44
Phase 5	4.00	2.44	2.36	0.08	0.09	-0.51

By looking to the coarse pattern, we may observe that the edges in the phase 5 network kindly become sparse comparing to previous phase. It seems that the arresting did bring damage to the network although number of nodes didn't sharply decrease. And if we observe more deeply, we may find out that the number of triangles, factor of clustering coefficient, has been decreased also. This match my assumption about decrease of horizontal connection between nodes.

Table above shows some measures describing global network information. By looking to the rows indicating for phase 4 and phase 5 respectively, we can find 2 obvious change, average degree and transitivity , between these 2 phases. The average degree describes how many contacts a member has on average. And the transitivity describes the clustering status of the network. Both these two measures decreased a lot.

With these results, I have an assumption for the background story. Since the police started to arrest some members of the group, the criminal enterprise got cautious and trying to recruit new blood. The decrease of number of edges is not because of the decrease of number of nodes since number of nodes only decreased for 1 only. The

main reasons might be two: First, they have to be cautious to police, therefore they have to avoid unnecessary contacting to prevent wiretap. Second, since they've recruited someone new to the group, they can't trust the new members and have to avoid them contacting to important role for temporary. Therefore the number of edges decreased as result, and so did the transitivity.

PART (H) : (2)

ARE THERE OTHER ACTORS THAT PLAY AN IMPORTANT ROLE BUT ARE NOT ON THE LIST OF INVESTIGATION (I.E., ACTORS WHO ARE NOT AMONG THE 23 LISTED ABOVE) ? LIST THEM, AND EXPLAIN WHY THEY ARE IMPORTANT.?

AVERAGE BETWEENNESS CENTRALITY TOP 40

Rank	ID	Centrality	Rank	ID	Centrality
1	1	0.65505	21	31	0.00696
2	12	0.16756	22	13	0.00568
3	3	0.12940	23	107	0.00568
4	76	0.08379	24	5	0.00568
5	87	0.06133	25	2	0.00512
6	41	0.05037	26	11	0.00511
7	89	0.04795	27	58	0.00431
8	14	0.03267	28	6	0.00403
9	83	0.03178	29	71	0.00322
10	82	0.02920	30	78	0.00235
11	85	0.02373	31	27	0.00218
12	79	0.02194	32	30	0.00181
13	37	0.01595	33	19	0.00102
14	88	0.01244	34	49	0.00068
15	8	0.00917	35	20	0.00055
16	7	0.00798	36	73	0.00044
17	96	0.00786	37	24	0.00033
18	9	0.00748	38	84	0.00006
19	86	0.00704	39	102	0.00006
20	22	0.00698	40	90	0.00000

CORE MEMBERS 23

ID	Centrality
1	0.65505
12	0.16756
3	0.12940
76	0.08379
87	0.06133
89	0.04795
83	0.03178
82	0.02920
85	0.02373
88	0.01244
8	0.00917
96	0.00786
86	0.00704
5	0.00568
11	0.00511
6	0.00403
84	0.00006
106	0.00000
77	0.00000
17	0.00000
80	0.00000
33	0.00000
16	0.00000

It's hard to not talk about the centrality when we discussed which nodes should be important. And this time I will focus on two measurements, one for betweenness centrality (BC) and one for eigenvector centrality (EC).

First, let's check the average BC of all phases for each node. The data of beside tables correspond to top 40 average BC and average BC of 23 listed members respectively. We might find out that some BC of 23 members are even 0. So I colored all the members not in listed 23 but nonzero in top 40 BC list since they all might be important. They are n:41, 14, 79, 37, 9, 22, 31, 13, 107, 2, 58, 71, 78, 27, 30, 19, 49, 20, 73, 24, 102.

Then, let's check the average EC of all phases for each node. Still, I've generated two tables correspond to top 40 average EC and average EC of 23 listed members respectively. It's surprised that all top-40 centralities are larger than the minimum average centrality of listed members. So again, I colored all the members not in listed 23 but nonzero in top 40 EC list since they all might be important. They are n: 2, 9, 37, 4, 41, 19, 90, 13, 81, 14, 15, 55, 34, 27, 31, 47, 49, 64, 7, 20, 52.

Further more, if we look to these two list respected to BC and EC, we can find nodes that are listed by both lists. These nodes represent members with both high BC and EC but not noticed and listed. They are n: 41, 14, 37, 7, 9, 31, 13, 2, 27, 19, 49, 20 and 90.

CORE MEMBERS

ID	Centrality
1	0.5464
3	0.2981
85	0.1906
76	0.1659
83	0.1535
8	0.1524
12	0.1419
87	0.1411
82	0.1001
6	0.0973
11	0.0927
88	0.0864
5	0.0855
84	0.0819
89	0.0784
86	0.0582
77	0.0406
17	0.0291
96	0.0268
106	0.0139
16	0.0138
80	0.0009
33	0.0009

AVERAGE EIGENVECTOR CENTRALITY TOP 40

Rank	ID	Centrality	Rank	ID	Centrality
1	1	0.5464	21	19	0.0619
2	3	0.2981	22	90	0.0617
3	85	0.1906	23	86	0.0582
4	76	0.1659	24	13	0.0521
5	83	0.1535	25	81	0.0520
6	8	0.1524	26	14	0.0517
7	12	0.1419	27	15	0.0478
8	87	0.1411	28	55	0.0419
9	2	0.1143	29	77	0.0406
10	9	0.1007	30	34	0.0401
11	82	0.1001	31	27	0.0371
12	6	0.0973	32	31	0.0354
13	11	0.0927	33	47	0.0339
14	88	0.0864	34	49	0.0329
15	5	0.0855	35	64	0.0325
16	84	0.0819	36	7	0.0315
17	89	0.0784	37	20	0.0295
18	37	0.0710	38	17	0.0291
19	4	0.0709	39	52	0.0288
20	41	0.0639	40	96	0.0268

PART (I) : (2)

WHAT ARE THE ADVANTAGES OF LOOKING AT THE DIRECTED VERSION VS. UNDIRECTED VERSION OF THE CRIMINAL NETWORK?

I think one of the answers of this question was already referred in problem 1. When we define importance of a node, a directed graph provides us an option to choose in-degree or out-degree as consideration. Members who contact others a lot and members who are contacted a lot might be two quite different roles in this criminal enterprise. The

difference between this two kinds of roles can be easily identified with directed network but not with undirected network.

PART (J) : (4)

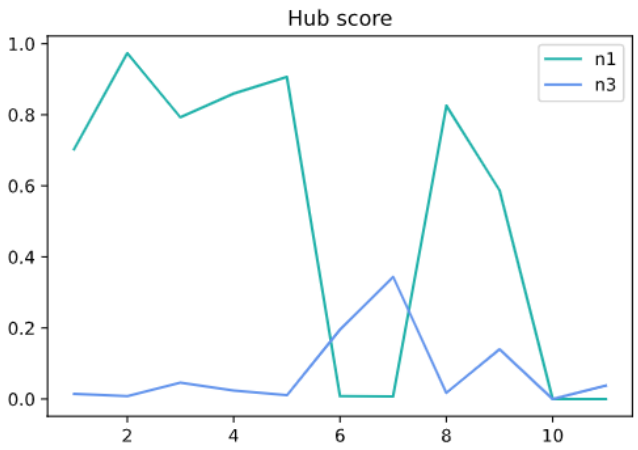
RECALL THE DEFINITION OF HUBS AND AUTHORITIES. COMPUTE THE HUB AND AUTHORITY SCORE OF EACH ACTOR, AND FOR EACH PHASE. USING THIS, WHAT RELEVANT OBSERVATIONS CAN YOU MAKE ON HOW THE RELATIONSHIP BETWEEN N1 AND N3 EVOLVES OVER THE PHASES. CAN YOU MAKE COMPARISONS TO YOUR RESULTS IN PART (G)?

HUBS SCORE

ID	Phase 1	Phase 2	Phase 3	Phase 4	Phase 5	Phase 6	Phase 7	Phase 8	Phase 9	Phase 10	Phase 11
1	0.7031	0.9730	0.7931	0.8598	0.9065	0.0080	0.0068	0.8259	0.5879	0.0000	0.0001
3	0.0144	0.0076	0.0463	0.0240	0.0105	0.1953	0.3433	0.0174	0.1395	0.0000	0.0379

AUTHORITIES SCORE

ID	Phase 1	Phase 2	Phase 3	Phase 4	Phase 5	Phase 6	Phase 7	Phase 8	Phase 9	Phase 10	Phase 11
1	0.0118	0.0003	0.0032	0.0022	0.0006	0.8054	0.7274	0.0020	0.0162	0.0000	0.0000
3	0.1357	0.3367	0.1496	0.2755	0.3236	0.0321	0.0069	0.4672	0.0675	0.0000	0.0000

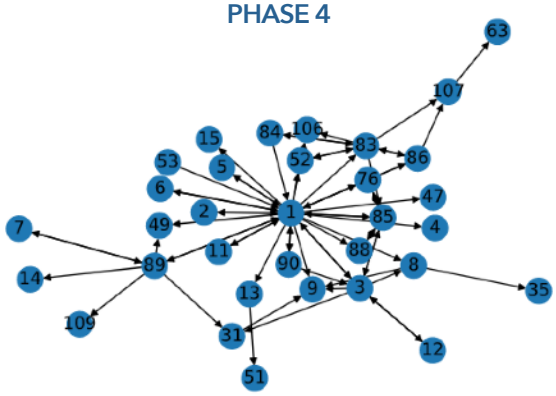


The hubs and authorities score of each n1 and n2 node are presented in the tables above. We may find out that as score of n1 increased the score of n3 would decrease, and vice versa, as score of n1 decreased the score of n3 would increased as result. These two nodes somehow kindly offset to each other in score performance which corresponds to their roles. This can be easily observed if we look to the line charts beside.

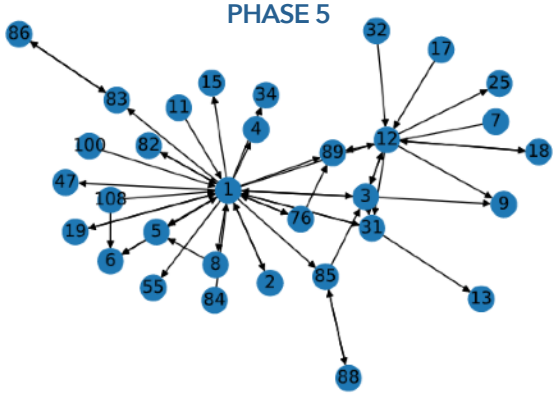
To compare with the results in part (g), we can focus on the scores in phase 4 and phase 5. The hub score of n1 increased and became the second highest score over all phases. Meanwhile the authority score of n3 also increased and became the third

highest score over all phases. Which means that n1 contacted most of important authorities by himself instead of assigning other member doing this. With increasing hub score of n1, the authority score of n3, who are pointed by n1, also became larger.

PHASE 4



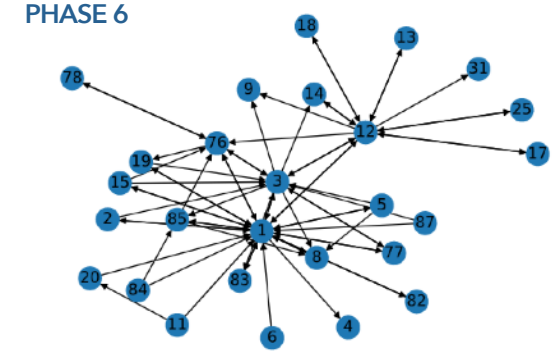
PHASE 5



The relationship between n1 and n3 can be observed also well by looking the maps beside corresponding directed graph. We may find that the shortest path length between these two nodes are 1, and the direction is double way. These features show that these two member are closed and they have well communication since they can contact each other initiatively.

There are other interesting phenomenons showed in the score table. The hub score of n1 dropped sharply in phase 6, meanwhile the authority score of n1 increased also sharply. This might indicates that the role n1 played in phase 6 somehow transformed. So I also made the graph map of phase 6 and degree table as below.

PHASE 6



When we look to the map, we can find out that the linkages of n3 increased obviously from phase 5 to phase 6. But it's hard to tell the difference of n1 if we only look to the map. So let's move on to the degree table.

In the degree table, the change of n1 is much more obvious. We can find out that the in-degree increase a lot and out degree decrease a lot in phase 6. This really explains the change performed by hubs and authorities score.

DEGREE TABLE

Member	Degree type	Phase 4	Phase 5	Phase 6
N1	In-degree	11	11	17
	Out-degree	21	18	13
N3	In-degree	4	3	9
	Out-degree	6	4	9

PROBLEM 3

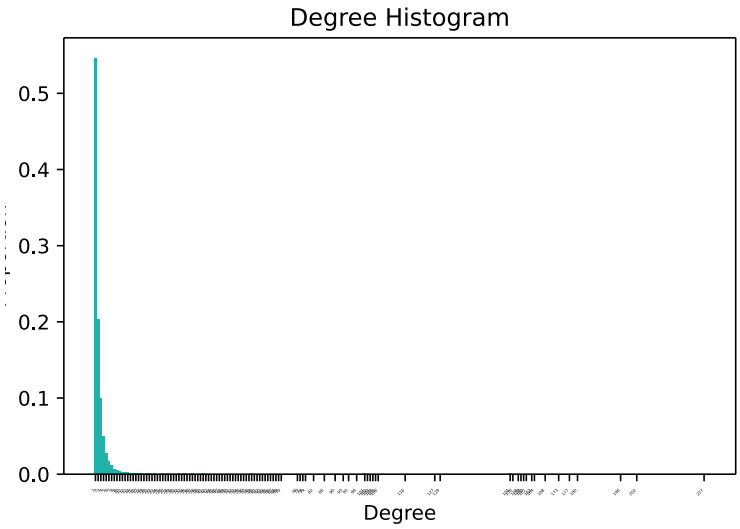
CO-OFFENDING NETWORK

The data for this problem set consists of individuals who were arrested in Quebec between 2003 and 2010. Some of the individuals have always acted solo, and have been arrested alone throughout their criminal career. Others co-offended with other individuals, and have been arrested in groups. The goal of this problem set is to construct and analyze the co-offender network. The nodes in the network are the offenders, and two offenders share a (possibly weighted) edge whenever they are arrested for the same crime event.

PART (G) : (3)

PLOT THE DEGREE DISTRIBUTION (OR AN APPROXIMATION OF IT IF NEEDED) OF G. COMMENT ON THE SHAPE OF THE DISTRIBUTION. COULD THIS GRAPH HAVE COME FROM AN ERDOS-RENYI MODEL? WHY MIGHT THE DEGREE DISTRIBUTION HAVE THIS SHAPE?

Before plotting, I've made the scale on y-axis, representing the numbers of nodes corresponding to each degree, transform into the proportion by dividing by number of total nodes. And following figure is the result.



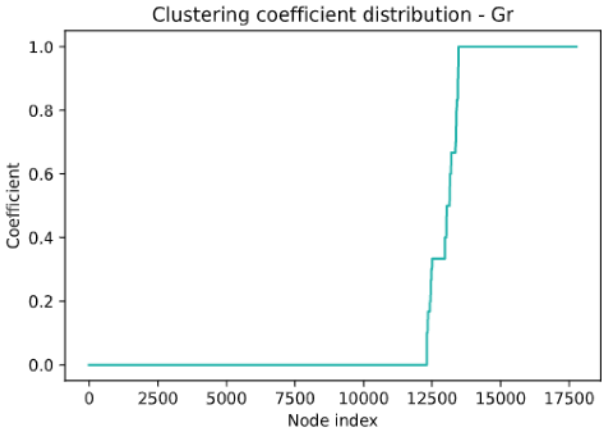
I don't think this graph came from an Erdos-Renyi model. The main reason is that the degree distribution generating from Erdos-Renyi model is not a power-law distribution. Which means that the probability of occurring nodes with high degree approximately equals to 0. However, in the beside graph,

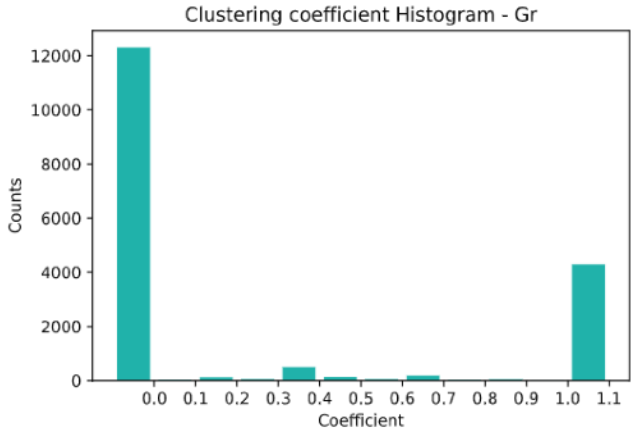
we can find that there are some nodes with pretty high degree. Therefore, the graph is probably not coming from an Erdos-Renyi model.

Most degree distribution coming from real data will conform the power-law distribution which allows distribution having a long tail. In real world, we know that the number of offenders involving a co-offending case could be 2 for general case, liked robbery or stealing, or could be huge for some financial criminal case, liked scam group or financial criminal enterprise. So the degree distribution of co-offending will have high height in the head but still a long tail.

PART (M) : (4)

PLOT THE DISTRIBUTION OF CLUSTERING COEFFICIENTS FOR EACH NODE FOR *Gr* AND *Gnr*. WHAT SHAPE DO THE PLOTS MAKE? WHAT DOES THIS TELL YOU ABOUT THE BEHAVIOR OF THE ACTORS?

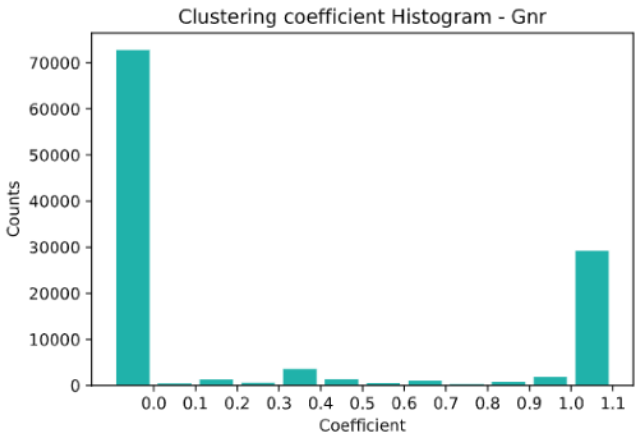




First, let's view the distribution of clustering coefficients for each node for graph Gr. The left figure above, is made with nodes sorted by their coefficients. We can observe that the line begin from 0 for a large distance, and become cliffy in a short segment. Then it become horizontal again with a second largest distance when coefficient equalling to 1.0. So we expect that the clustering coefficients of most of the nodes in Gr will be 0.0 or 1.0. Then I made a histogram to perform the distribution. The leftest interval indicates number of nodes that its clustering coefficient equals to 0. And the rightest interval indicates number of nodes that its clustering coefficient equals to 1.0. And the interval between these two indicates that the number of nodes that its clustering coefficient is in interval: (X, X+1). And we can find that most of the coefficients are in 2 sides. Only few coefficients appear in other intervals.

We all know that the clustering coefficient is the proportion of number of closed triplets versus number of closed and open triplets. In this case, open triplets only exists when 3 offenders offends the law twice above and two of these 3 didn't have cooperation. Therefore, a intermediated clustering coefficient only occurs when an actor has offends the law twice above and work with at least 2 totally un-connected groups or person. Thank God that such criminal masters are minorities. And this might be one reason that explains why most clustering coefficients are 0 or 1.

Then, continue with viewing the distribution of clustering coefficients for each node for graph Gnr.



The results are similar to the results of graph Gr. However, the proportion of intermediated clustering coefficients seems exceeding the one in graph Gr. The result of this is quite reasonable. If an actor choose not to work with same co-offender twice, when he makes another offending he must choose someone or some group else as his co-offenders. As a result, there's higher chance that his clustering coefficient is not equal to one or zero as long as one of actor in second group hasn't co-offend with any actors in previous group.

PART (N) : (4)

PICK A CENTRALITY MEASURE (DEGREE, EIGENVECTOR, BETWEENNESS, ETC) AND COMPUTE THE SCORES FOR THE TOP (LARGEST) COMPONENT OF *Gr* AND *Gnr*. COMPARE THE DISTRIBUTION OF THE CENTRALITY ACROSS NODES (FOR EXAMPLE, WITH SUMMARY STATISTICS AND/OR A HISTOGRAM). EXAMINE THE NUMBER OF CRIMES COMMITTED BY THE MOST CENTRAL ACTOR IN THE REPEAT OFFENDER GRAPH, DOES THIS SUPPORT YOUR CONCLUSIONS?.

Let's start with Graph Gr. Following are the tables describing betweenness centralities and eigenvector centralities for each node.

BETWEENNESS CENTRALITY - *Gr*

OffenderID	EigenvectorC	# of Crimes	OffenderID	EigenvectorC	# of Crimes
610924	0.5931	46	592064	0.0000	12
605832	0.5736	29	584555	0.0000	4
592205	0.4572	24	618650	0.0000	4
596719	0.3788	24	633944	0.0000	28
628662	0.2644	38	636382	0.0000	4
596946	0.2292	36	612729	0.0000	6
643287	0.2226	42	604085	0.0000	4
627701	0.2216	37	624186	0.0000	13
608039	0.2127	27	633584	0.0000	3
626681	0.1624	31	555332	0.0000	8
548976	0.1197	18	636666	0.0000	20
596056	0.1188	33	610898	0.0000	20
630371	0.1141	41	563283	0.0000	13
609728	0.0909	67	598961	0.0000	2
590256	0.0894	38	597408	0.0000	6
639523	0.0457	30	604892	0.0000	4
593515	0.0440	55	621772	0.0000	2
627921	0.0327	10	635391	0.0000	3
507801	0.0308	21	213763	0.0000	11
607935	0.0308	15	561479	0.0000	7
605573	0.0308	9	606328	0.0000	10
610612	0.0308	42	611611	0.0000	7
591077	0.0270	37	631037	0.0000	2
624870	0.0151	11	582160	0.0000	8
634941	0.0151	17	650691	0.0000	2
648679	0.0149	13	596406	0.0000	19
582830	0.0034	19	559931	0.0000	9
643988	0.0005	24	583987	0.0000	3
632643	0.0000	11	611781	0.0000	23
616582	0.0000	2	642507	0.0000	12
604843	0.0000	53	598630	0.0000	7
623782	0.0000	7	595269	0.0000	28
652027	0.0000	18	601732	0.0000	8

EIGENVECTOR CENTRALITY - *Gr*

OffenderID	EigenvectorC	# of Crimes	OffenderID	EigenvectorC	# of Crimes
596946	0.4056	36	631037	0.0232	2
610924	0.4042	46	605573	0.0214	9
626681	0.3699	31	611611	0.0206	7
608039	0.3280	27	616582	0.0176	2
593515	0.2777	55	618650	0.0176	4
627701	0.2454	37	624186	0.0176	13
630371	0.2205	41	596719	0.0101	24
591077	0.2114	37	584555	0.0077	4
582830	0.1843	19	607935	0.0074	15
643988	0.1457	24	596406	0.0072	19
632643	0.1148	11	582160	0.0066	8
627921	0.1118	10	598630	0.0066	7
596056	0.0993	33	595269	0.0056	28
605832	0.0990	29	604085	0.0055	4
623782	0.0981	7	555332	0.0053	8
590256	0.0969	38	213763	0.0053	11
636382	0.0933	4	559931	0.0044	9
635391	0.0933	3	628662	0.0036	38
604892	0.0880	4	643287	0.0031	42
609728	0.0849	67	636666	0.0028	20
610898	0.0840	20	592064	0.0026	12
563283	0.0837	13	642507	0.0026	12
598961	0.0837	2	639523	0.0024	30
583987	0.0837	3	648679	0.0023	13
604843	0.0679	53	624870	0.0023	11
612729	0.0575	6	634941	0.0023	17
633944	0.0457	28	597408	0.0021	6
592205	0.0272	24	621772	0.0015	2
610612	0.0271	42	606328	0.0014	10
507801	0.0265	21	652027	0.0010	18
561479	0.0262	7	650691	0.0010	2
611781	0.0256	23	601732	0.0006	8
548976	0.0254	18	633584	0.0005	3

If we only look to the tables, no matter table of betweenness centrality or eigenvector centrality, we may find that there are lots of nodes of high number of crimes performing a high centrality. How ever it's still hard to make conclusion that number of crimes of a offender does relate to the centrality. Therefore I measured the correlation coefficient in each centrality measurement. Besides are the correlation matrix of these two measurement.

BETWEENNESS CENTRALITY CORRELATION MATRIX - *Gr*

Factor	Centrality	# of Crimes
Centrality	1.0000	0.4704
# of Crimes	0.4704	1.0000

EIGENVECTOR CENTRALITY CORRELATION MATRIX - *Gr*

Factor	Centrality	# of Crimes
Centrality	1.0000	0.4817
# of Crimes	0.4817	1.0000

From the correlation tables, we know that correlation coefficients of centrality versus number of crimes are all near to 0.5 for both betweenness centrality and eigenvector centrality. With these results, we may say that number of crimes and centrality are moderately positive correlated.

And let's move on to the Gnr graph but this time I choose to present the correlation matrix only because the number of nodes is quite large. And also I will measure both betweenness centralities and eigenvector centralities. However, with 12,806 nodes, it really really took a while to compute the betweenness centralities. Besides are the correlation matrix corresponding to BC and EC respectively.

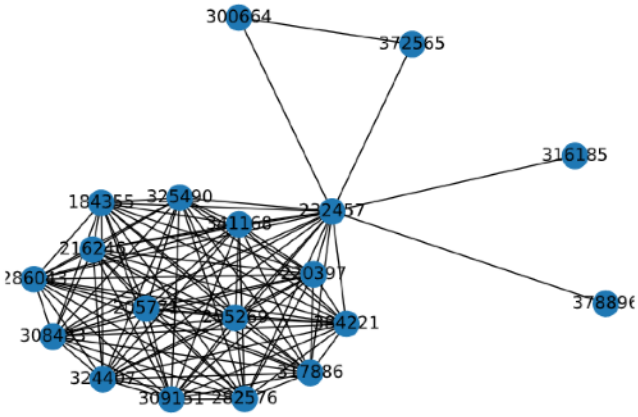
BETWEENNESS CENTRALITY CORRELATION MATRIX - *Gnr*

Factor	Centrality	# of Crimes
Centrality	1.0000	0.2350
# of Crimes	0.2350	1.0000

EIGENVECTOR CENTRALITY CORRELATION MATRIX - *Gnr*

Factor	Centrality	# of Crimes
Centrality	1.0000	-0.0599
# of Crimes	-0.0599	1.0000

It's surprised that the results of graph Gnr are quite different to results of graph Gr. We can find that the correlations between number of crimes of an offender and his centralities of both betweenness and eigenvector are really low. To discuss with this result, I think it's better to draw out a graph to illustrate. However, when I chose the node that offended with most times, I found that he somehow usually offended solely. And he offended with one co-offender for only one time. I think there probably are some offenders in Gnr corresponding to this type of offender, and their centralities will be really low not surprisingly. This type of offender might also help us to understand that



why there are quite lots number of nodes having a zero clustering coefficient. Then I continue with choosing the offender who offended with second most times and draw the subgraph with his neighbored nodes.

I think the graph beside is quite suitable for illustrating the centralities issue especially for the eigenvector one. When one node in graph Gnr cooperate with one big group. Each node of this group will have be full connected to others in this group. Therefore, if we measure the eigenvector centralities of them, the centralities of nodes in big group will be large, much higher than nodes not in big group and equal to each other in big group. As a result, with adjustment generating from largest lambda value, all the centralities of nodes in this graph will become quite small. Therefore, the eigenvector centralities in graph Gnr will definitely be closed to zero and uncorrelated with the number of crimes.

ARE THESE OFFENDERS POLITICALLY CORRECT?

OPEN-CLOSED PROJECT

In these years, political correctness becoming more and more important in the whole world. More and more people aware that it's not appropriate to label someone with self gender awareness, race, nationality…etc. Political correctness culture might have brought us to a world which is more civilized. But it also bring us some new cultural conflict. Regardless the dispute caused by political correctness culture, we have to admit that it did somehow change the world. However, I'm curious about, did it make some change to the criminal network?

PART (A) - DEFINE THE PROBLEM:

Since the data we have contains few information about the offenders, the topic we can discuss with are restrict to only physical gender and age range. Therefore I choose the physical gender as our discussion topic since the age range is not specific enough. So how do we define that a network is political correct with physical gender or not? Is an increasing of proportion of gender as minorities in the network matching this idea?

I think this answer might be somehow worthy. But the proportion of gender in the data has a drawback that it can't tell us whether someone of minor gender are invited into a group consisting members with different gender. I think this might be an advantage of network analysis. The network analysis can tell us some structural information about the data. Therefore it's worth to do the network analysis on this topic.

I think the homophily is an appropriate measurement for this issue. If we label the node with its gender, we can measure the homophily score of the network. The homophily

score should be high if most nodes of same type are linked to each other. And since we are concerning about the political correctness about gender, instead of pursuing high homophily score, a homophily score closed to 0 will correspond to the purpose of this issue properly. With homophily score closed to 0, it indicates that gender is not a factor for connecting nodes in the network. So, in this project I will focus on the performance of homophily score.

PART (B) - ESTABLISH THE PROCESS:

Before starting the data processing and analysis, it's better to establish the process and make sure all these processes are logical and reasonable for our purpose. My process is as followed:

First, as same as process in previous problem, we have to reassign indexes to replace the offenders' identifiers and crimes' identifier for generating adjacency matrix with proper shape.

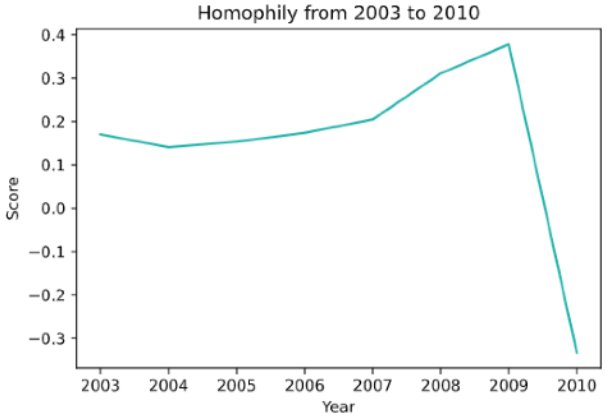
Second, I'm curious about the difference between performance of homophily scores in different year, so I will make 8 graphs each with nodes corresponds to its crime year, between 2003 to 2010. Before making this, I should split the data frame into 8s according crime year. This part won't be too hard.

Third, make the unweighted and undirected graph. Since we are discussing whether an offender of minority is hard to be recruited into a criminal group, the co-offending times of a node is not that important to us. Therefore it's not necessary to generate graph with weights and direction.

Fourth, add on the gender labels to each nodes in each graph. Then measure the homophily scores of each graphs and make simple analysis.

PART (C) - RESULTS:

Besides are the table and chart describing the homophily score over the years. Instead of decreasing, the scores surprisingly increased in most of periods. Dose this mean that offenders prefer to work with co-



HOMOPHILY

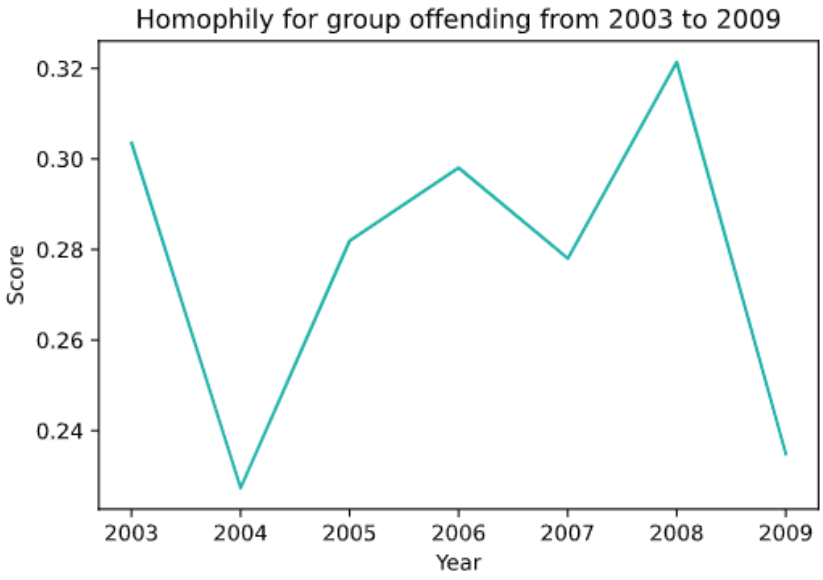
Year	Score
2003	0.1704
2004	0.1404
2005	0.1531
2006	0.1737
2007	0.2045
2008	0.3108
2009	0.3781
2010	-0.3333

offenders of same gender year after year? There's one more interesting phenomenon. As we can see that the line drops sharply in 2010, it's hard to be not curious about this. However, the reason of this is quite simple. The samples of 2010 we have are only 7. Therefore the performance in this year can be ignored when we make analysis.

Then I reconsider with the process trying to find something that I didn't aware before. And yes, there's one thing that I did ignore. When we discuss topic about social culture, especially the interaction between people, we should only concern about the event consisting 2 or more people. Which means that the criminal cases involving only 1 offender should be removed from the data in case that it will affect our result. So I processed the data again and did the rest of process. The results are described by following chart and table. But this time, I took off the result of year 2010.

HOMOPHIL FOR GROUP

Year	Score
2003	0.3035
2004	0.2274
2005	0.2819
2006	0.2980
2007	0.2780
2008	0.3213
2009	0.2349



I have to say, without obvious trend ,these results are more closed to real life.

PART (D) - SUMMARY:

Although we can't find out some obvious trend across years, there is still one thing that we can point out. The scores of homophily are non-negative and around 0.22 to 0.32. This indicate that offender potentially prefer co-offend with offenders of same gender. Gender might not be a major factor since the scores is not closed enough to 1 or -1, but some of offenders did co-offend with offender of same gender more often.

However, there are lots of potential reasons may explain for the results. Fo example, most offenders in the data are male, with proportion 428884 male to 110709 female. Therefore

nodes are more likely linked to the same gender as result and raise the score of homophily. Secondly, remember that the data only consists offenders who's been arrested. This might involves issue of law of survivors. There might be a possibility that criminal groups consisting members of different genders are more outstanding and harder to be arrested. Maybe the offenders in the data are arrested just because that they forgot to invite Julia Roberts or Gal Gadot, who knows. In my opinion, I think the political correctness might not be an issue in the criminal activity. The main factor of choosing criminal partner might be much simpler than we think. They are just more closed than others.