Module 4: Network Analysis

IDS.131

Problem Set 4

Yash Dixit

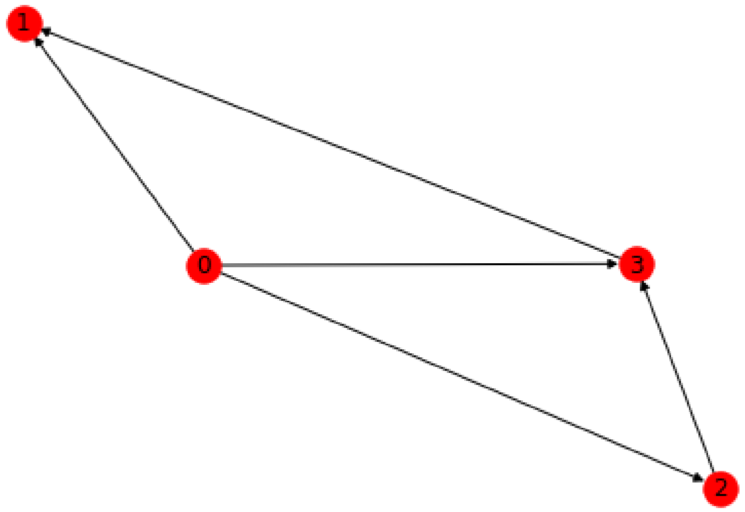
Collaborators:
Karan Bhuwalka
Benny Ng

Problem 4.1: Suggesting similar papers

A is the citation matrix with a_{ij} as the entries (i is the paper which cites, j is the paper which is being cited) and $a_{ij} = 1$ if there exists a citation as per the above directionality

(a) Co-citation network

- $C = AA^T$ represents the adjacency matrix for the co-citation network.
- Illustration



```
Array ([[3, 0, 1, 1],
        [0, 0, 0, 0],
        [1, 0, 1, 0],
        [1, 0, 0, 1]])
```

(b) Bibliographic coupling

- $B = AA^T$ represents the adjacency matrix for the bibliographic coupling network.

(c) Comparison

- It is obvious to infer that both metrics give different results and convey different aspects of the network.

- They exhibit different temporal behavior

  - The co-citation matrix can evolve over time and contains more entries, as more people continue to cite both the papers. In a field there can be some highly decorated works that are similar and cited by a lot of further research. The bibliographic coupling on the other hand is fixed because it refers to past research, and has a smaller body of work that can be cited.

    - Thus, picking the co-citation matrix as an indicator of similarity might make more sense. If a lot of papers in the future keep citing two papers among a growing body of work, it is highly likely they are similar. Meanwhile a bibliographic matrix could just identify some of the main works in the field that a lot of papers (including the two under consideration) cite.

- However, more generally, both can be useful metrics to determine similarity and a combination of the results of the two matrices should be seen (maybe a weighted sum or a index-wise product). The specific nature of the index should be contextual and decided based on the set of disciplines that the paper operates within.

Problem 4.2: Investigating a time-varying criminal network

(a) Centrality measures of the network for all listed phases

- Snippets of the code used to generate the network

```
phase = pd.read_csv("phase11.csv", sep = ',')

sp = phase.to_sparse(fill_value=0)

p = phase.iloc[0:,1:]
cols = p.columns.get_values()
clist = cols.tolist()
p.index = clist

#G = nx.from_numpy_matrix(p.values)
G = nx.DiGraph(p.values)
G = nx.relabel_nodes(G, dict(enumerate(p.columns)))
```

- Sample network visualizations

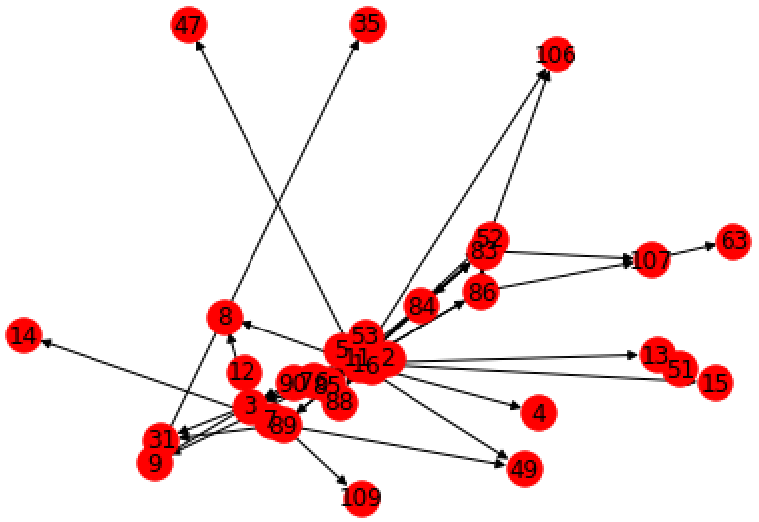


Fig: Phase 4

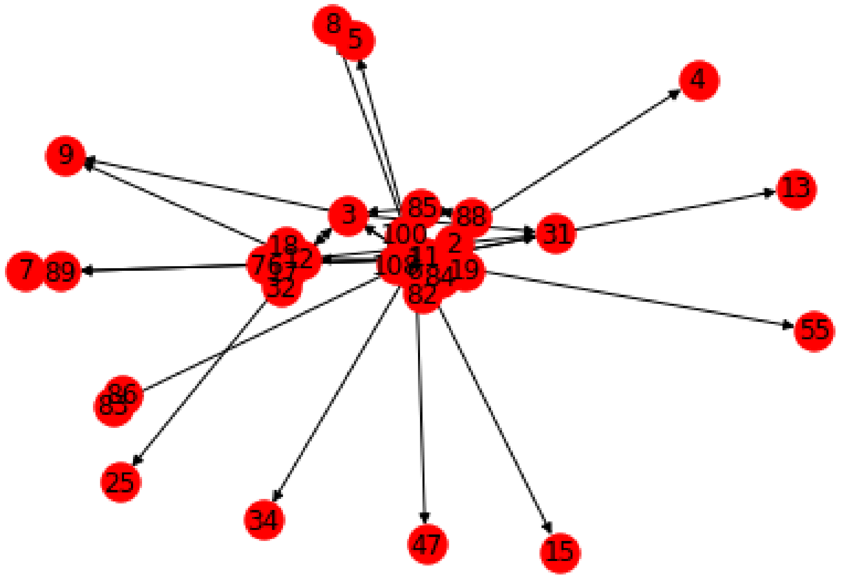
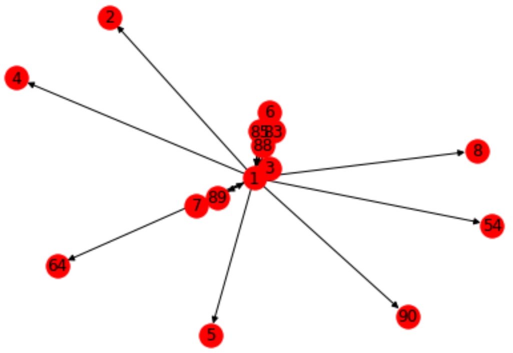


Fig: Phase 5

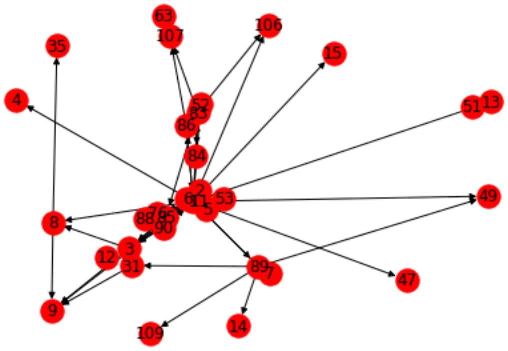
- Centrality measures for phases 1, 4, 5 and 11
  inc: incentrality | ouc: outcentrality | btc: betweenness | elgc, ergc: left, right eigen cen

Phase 1



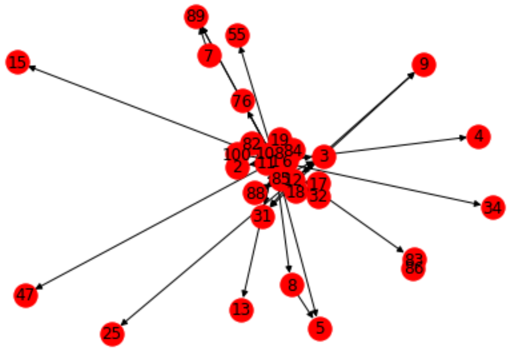
#	inc	ouc	btc	elgc	ergc
1	0.357143	0.785714	0.387363	4.70E-01	4.70E-01
3	0.142857	0.142857	0	3.35E-01	3.35E-01
5	0.071429	0	0	1.63E-01	1.63E-01
6	0.214286	0.071429	0.002747	2.89E-01	2.89E-01
8	0.071429	0	0	1.63E-01	1.63E-01
11	NaN	NaN	NaN	NaN	NaN
12	NaN	NaN	NaN	NaN	NaN
16	NaN	NaN	NaN	NaN	NaN
17	NaN	NaN	NaN	NaN	NaN
33	NaN	NaN	NaN	NaN	NaN
76	NaN	NaN	NaN	NaN	NaN
77	NaN	NaN	NaN	NaN	NaN
80	NaN	NaN	NaN	NaN	NaN
82	NaN	NaN	NaN	NaN	NaN
83	0	0.142857	0	2.21E-12	2.21E-12
84	NaN	NaN	NaN	NaN	NaN
85	0.142857	0.214286	0.008242	3.35E-01	3.35E-01
86	NaN	NaN	NaN	NaN	NaN
87	NaN	NaN	NaN	NaN	NaN
88	0.285714	0.285714	0.085165	4.96E-01	4.96E-01
89	0.142857	0.142857	0.098901	1.85E-01	1.85E-01
96	NaN	NaN	NaN	NaN	NaN
106	NaN	NaN	NaN	NaN	NaN

Phase 4



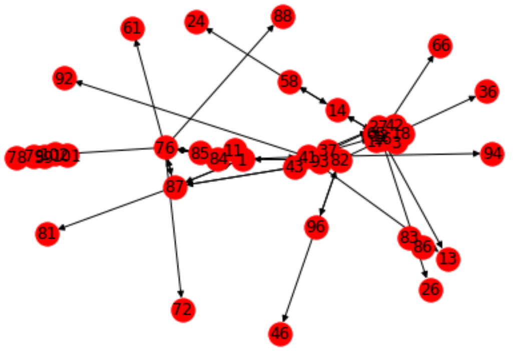
#	inc	ouc	btc	elgc	ergc
1	0.34375	0.65625	0.428931	0.442196	0.442196
3	0.125	0.1875	0.064516	0.275366	0.275366
5	0.03125	0.03125	0	0.119841	0.119841
6	0.03125	0.03125	0	0.119841	0.119841
8	0.0625	0.0625	0.018145	0.194469	0.194469
11	0.03125	0.03125	0	0.119841	0.119841
12	0.03125	0.03125	0	0.074627	0.074627
16	NaN	NaN	NaN	NaN	NaN
17	NaN	NaN	NaN	NaN	NaN
33	NaN	NaN	NaN	NaN	NaN
76	0.0625	0.03125	0	0.222663	0.222663
77	NaN	NaN	NaN	NaN	NaN
80	NaN	NaN	NaN	NaN	NaN
82	NaN	NaN	NaN	NaN	NaN
83	0.125	0.1875	0.06502	0.237031	0.237031
84	0.03125	0.0625	0.012265	0.06424	0.06424
85	0.15625	0.125	0.086358	0.379397	0.379397
86	0.0625	0.09375	0.026378	0.184081	0.184081
87	NaN	NaN	NaN	NaN	NaN
88	0.0625	0.03125	0	0.222663	0.222663
89	0.0625	0.1875	0.082661	0.129341	0.129341
96	NaN	NaN	NaN	NaN	NaN
106	0.0625	0	0	0.184081	0.184081

Phase 5



#	inc	ouc	btc	elgc	ergc
1	0.354839	0.580645	0.372043	4.70E-01	4.70E-01
3	0.096774	0.129032	0.091935	3.06E-01	3.06E-01
5	0.064516	0	0	2.05E-01	2.05E-01
6	0.064516	0.032258	0	1.55E-01	1.55E-01
8	0.032258	0.032258	0	1.55E-01	1.55E-01
11	0	0.032258	0	1.57E-10	1.57E-10
12	0.16129	0.16129	0.112366	2.86E-01	2.86E-01
16	NaN	NaN	NaN	NaN	NaN
17	0	0.032258	0	1.57E-10	1.57E-10
33	NaN	NaN	NaN	NaN	NaN
76	0.032258	0.064516	0	1.55E-01	1.55E-01
77	NaN	NaN	NaN	NaN	NaN
80	NaN	NaN	NaN	NaN	NaN
82	0.032258	0.032258	0	1.55E-01	1.55E-01
83	0.064516	0.032258	0.019355	1.73E-01	1.73E-01
84	0	0.032258	0	1.57E-10	1.57E-10
85	0.064516	0.064516	0.041935	1.73E-01	1.73E-01
86	0.032258	0.032258	0	5.70E-02	5.70E-02
87	NaN	NaN	NaN	NaN	NaN
88	0.032258	0.032258	0	5.70E-02	5.70E-02
89	0.096774	0	0	2.05E-01	2.05E-01
96	NaN	NaN	NaN	NaN	NaN
106	NaN	NaN	NaN	NaN	NaN

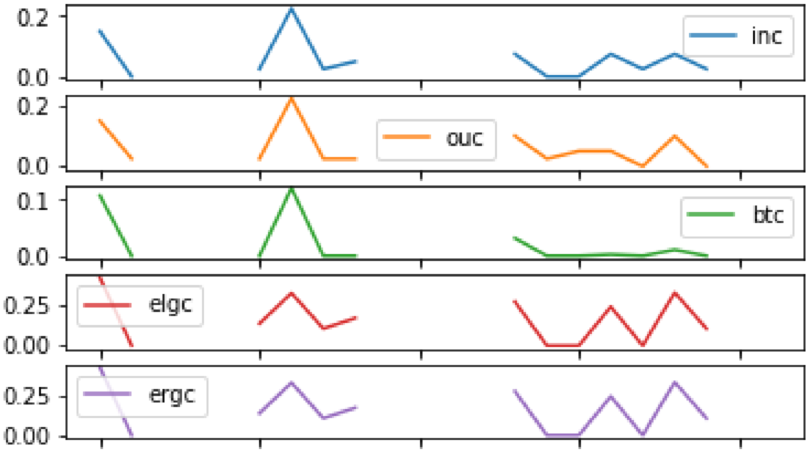
Phase 11

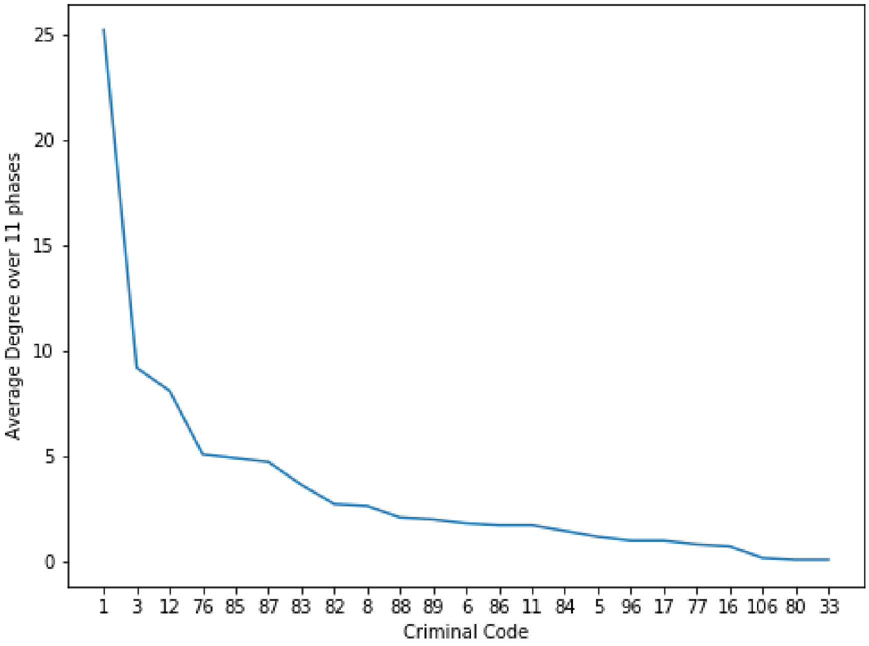


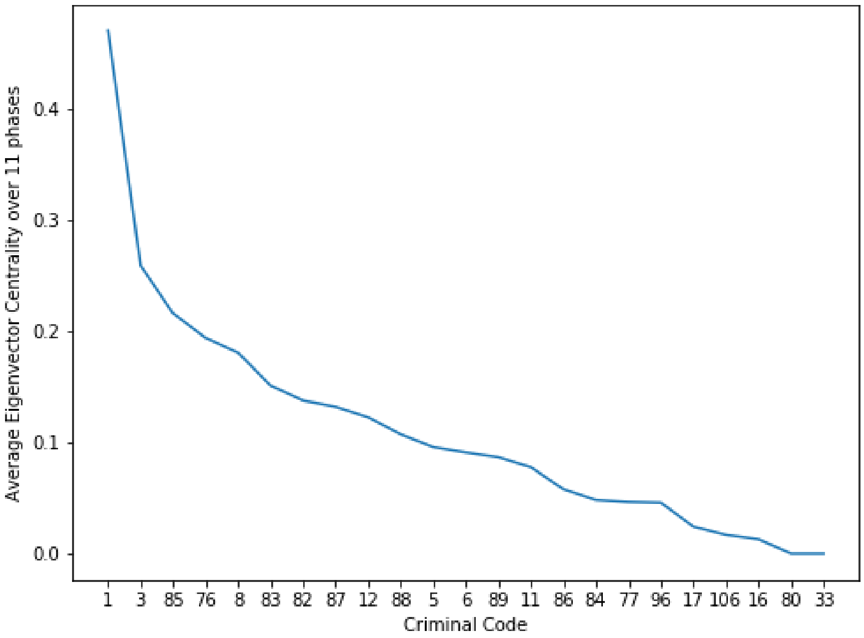
#	inc	ouc	btc	elgc	ergc
1	0.15	0.15	0.107372	4.26E-01	4.26E-01
3	0	0.025	0	4.56E-44	4.56E-44
5	NaN	NaN	NaN	NaN	NaN
6	NaN	NaN	NaN	NaN	NaN
8	NaN	NaN	NaN	NaN	NaN
11	0.025	0.025	0	1.38E-01	1.38E-01
12	0.225	0.225	0.120726	3.30E-01	3.30E-01
16	0.025	0.025	0	1.07E-01	1.07E-01
17	0.05	0.025	0	1.73E-01	1.73E-01
33	NaN	NaN	NaN	NaN	NaN
76	0.075	0.175	0.032372	3.25E-01	3.25E-01
77	NaN	NaN	NaN	NaN	NaN
80	NaN	NaN	NaN	NaN	NaN
82	0.075	0.1	0.03141	2.75E-01	2.75E-01
83	0	0.025	0	4.56E-44	4.56E-44
84	0	0.05	0	4.56E-44	4.56E-44
85	0.075	0.05	0.002564	2.44E-01	2.44E-01
86	0.025	0	0	3.24E-42	3.24E-42
87	0.075	0.1	0.010897	3.33E-01	3.33E-01
88	0.025	0	0	1.05E-01	1.05E-01
89	NaN	NaN	NaN	NaN	NaN
96	0.025	0.05	0.004487	8.92E-02	8.92E-02
106	NaN	NaN	NaN	NaN	NaN

(b) Identifying central actors using centrality metrics

- Visualization of multiple metrics







- Summary of the centrality measures

Phase #	Important nodes inferred from centrality measures					Marginally relevant nodes
	inc	ouc	btc	elgc	ergc	
1	1,88,85	1,88,85	89	88,1,3	88,1,3	83,8,5
2	1,8,11,76,83,88,89	1,3,8,85,88,89	1,88,8,89	1,8,89	1,8,89	12,85,86
3	1,3,85	1,83,3	1,3,83	1,3,86,85	1,3,86,85	88,89,76
4	1,85,3,83	1,89,3,83	1,85,89	1,85	1,85	11,12
5	1,12,3,89	1,12,89	1,12,3	1,3,12	1,3,12	11,17,84
6	1,3,12	1,3,12	1,12,3	1,3,76	1,3,76	84,87
7	1,12	1,3	1,76	1,83,76	1,83,76	16,17
8	1,12,3	1,3,87	1,12,3	1,3	1,3	33,6,80
9	1,3,12,87,82	3,1,12,87,82	3,12,87	1,87,82,3	1,87,82,3	88,89
10	1,82	1,87	1,87,82	1,87,82	1,87,82	16,17,86
11	12,1	12,76	12	1,87,12	1,87,12	83,84,86,88

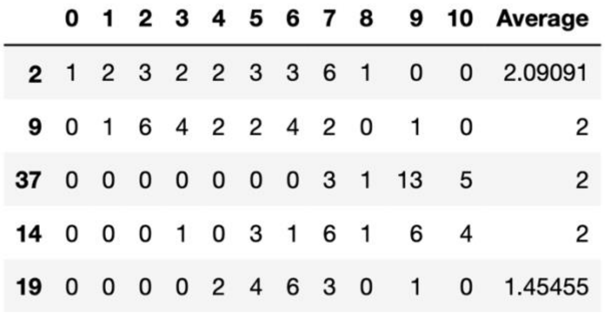
- Suspects absent from call logs

Phase #	Suspects absent from the phase call logs													
1	11	12	16	17	33	76	77	80	82	84	86	87	96	106
2	16	17	33	77	80	82	84	87	96	106				
3	16	17	33	77	80	82	87	96	106					
4	16	17	33	77	80	82	87	96						
5	16	33	77	80	87	96	106							
6	16	33	80	86	88	89	96	106						
7	33	80	82	84	86	89	96	106						
8	5	88	89	96	106									
9	5	33	77	80	84	86	106							
10	5	6	11	33	77	80	88	89	106					
11	5	6	8	33	77	80	89	106						

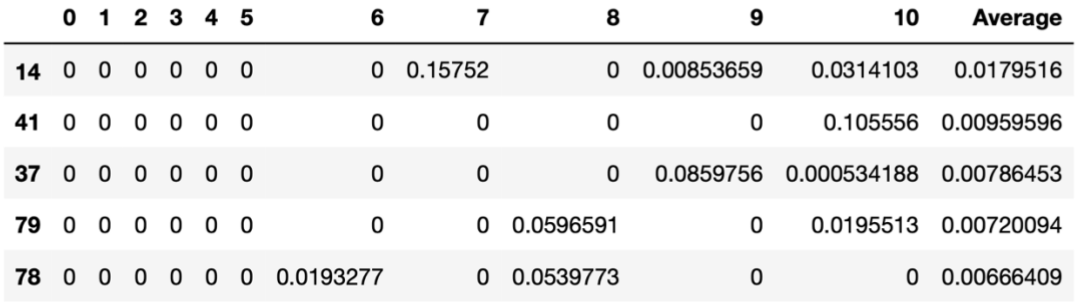
- Important nodes in the network (as per the chosen metrics)
    - Serero, Daniel (n1) : Mastermind of the network.
    - Pierre Perlini (n3) : Principal lieutenant of Serero, he executes his instructions.
    - Ernesto Morales (n12): Principal organizer of the cocaine import, intermediary between the Colombians and the Serero organization.

(c) Other actors who are important, but aren't being investigated

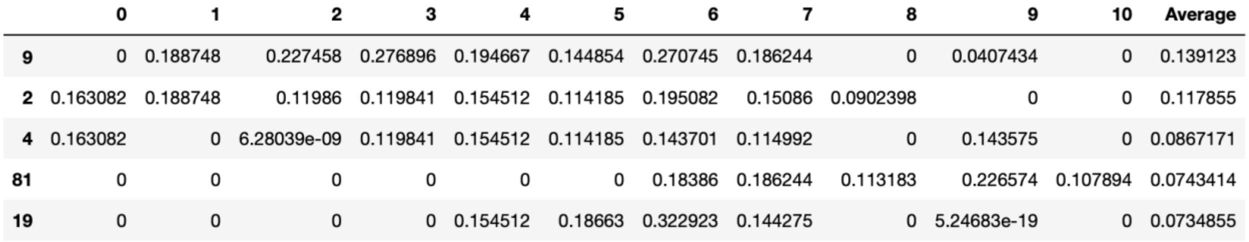
- Centrality measures for nodes other than those mentioned in the investigation (Degree, betweenness and eigenvalue centrality)



	0	1	2	3	4	5	6	7	8	9	10	Average
2	1	2	3	2	2	3	3	6	1	0	0	2.09091
9	0	1	6	4	2	2	4	2	0	1	0	2
37	0	0	0	0	0	0	0	3	1	13	5	2
14	0	0	0	1	0	3	1	6	1	6	4	2
19	0	0	0	0	2	4	6	3	0	1	0	1.45455



	0	1	2	3	4	5	6	7	8	9	10	Average
14	0	0	0	0	0	0	0	0.15752	0	0.00853659	0.0314103	0.0179516
41	0	0	0	0	0	0	0	0	0	0	0.105556	0.00959596
37	0	0	0	0	0	0	0	0	0	0.0859756	0.000534188	0.00786453
79	0	0	0	0	0	0	0	0	0.0596591	0	0.0195513	0.00720094
78	0	0	0	0	0	0	0.0193277	0	0.0539773	0	0	0.00666409



	0	1	2	3	4	5	6	7	8	9	10	Average
9	0	0.188748	0.227458	0.276896	0.194667	0.144854	0.270745	0.186244	0	0.0407434	0	0.139123
2	0.163082	0.188748	0.11986	0.119841	0.154512	0.114185	0.195082	0.15086	0.0902398	0	0	0.117855
4	0.163082	0	6.28039e-09	0.119841	0.154512	0.114185	0.143701	0.114992	0	0.143575	0	0.0867171
81	0	0	0	0	0	0	0.18386	0.186244	0.113183	0.226574	0.107894	0.0743414
19	0	0	0	0	0.154512	0.18663	0.322923	0.144275	0	5.24683e-19	0	0.0734855

- None of the actors listed above have high enough values for degree or centrality to compare to the top 5 suspects in the original investigation list.
- However, they do figure near the top 10 compared to the suspects.
- n2 and n9 figures high on both eigenvector centrality and degree. Comparatively both of them are lower according to betweenness suggesting that they may be important in isolated/separated networks within the larger network.
- n14 should be included in the suspects list because his/hers betweenness centrality would be 8th (averaged over phases). This means that many people are connected through n14.
- n9 is 7th by eigenvector centrality so he/she has connections who have influential connections - might be effective to include him/her in suspects to see who n9 is connected to.

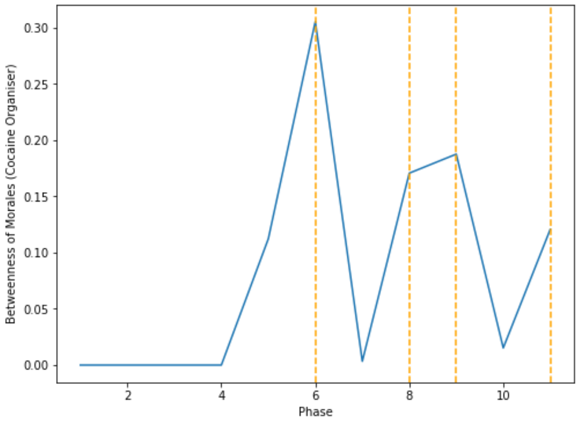
(d) Observable pattern from the evolution of the network

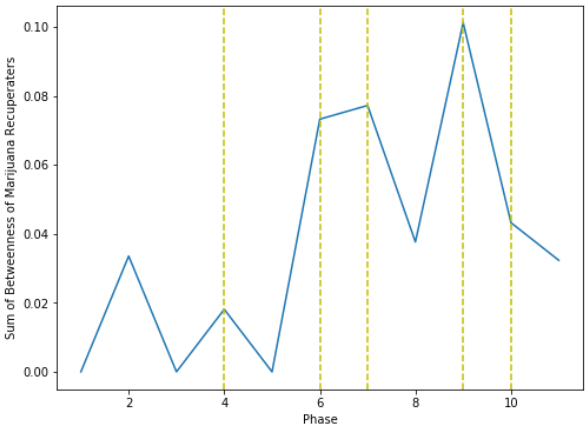
- Until Phase 4 i.e. the first seizure, we can see that the 'money movers' (83, 86, 88, 89) are quite active.
    - Alain (n83) and G ́erard (n86) Levy : Investors and transporters of money.
    - Wallace Lee (n85) : Takes care of financial affairs (accountant).
    - Lee Gilbert (n88): Trusted man of Wallace Lee (became an informer after the arrest).
- Furthermore, the cocaine segment of the network isn't much active until the first seizure leading to an insight that marijuana and cocaine are potential substitutes.

- Within the active nodes for the cocaine segment, we can see that 12 does the firefighting and takes action after heavy seizures
    - Ernesto Morales (n12): Principal organizer of the cocaine import, intermediary between the Colombians and the Serero organization.
- As the cumulative magnitude of the seizures increases, we see the 'transportation' guys becoming increasingly active, intuitively so
    - Patrick Lee (n87): Investor.
    - Salvatore Panetta (n82): Transport arrangements manager.
- Gabrielle Casale (n76; Charged with recuperating the marijuana) also occurs in the top 4 of all metrics. Gleeson (n5) and Quinzio (n8) alongwith Casale are in charge of marijuana but only Casale shows up in the top by these metrics- suggesting that he may be more important than them.
- Lee (n85) i.e. the accountant is important for finances. Interestingly, his eigenvector centrality is higher than other measures- he has friends who has friends, although he isn't directly connected as much
- Additional comment:
  It will be incredibly useful to pair the economics of marijuana / cocaine trade to further assess the impacts of the trade. We can get useful insights of supply elasticity and relative importance of both the drugs in the network

(e) Role of central actors and activity patterns

- To assess activity patterns, we can look at the 'total communication' index that represents the sum of all edges in the network.



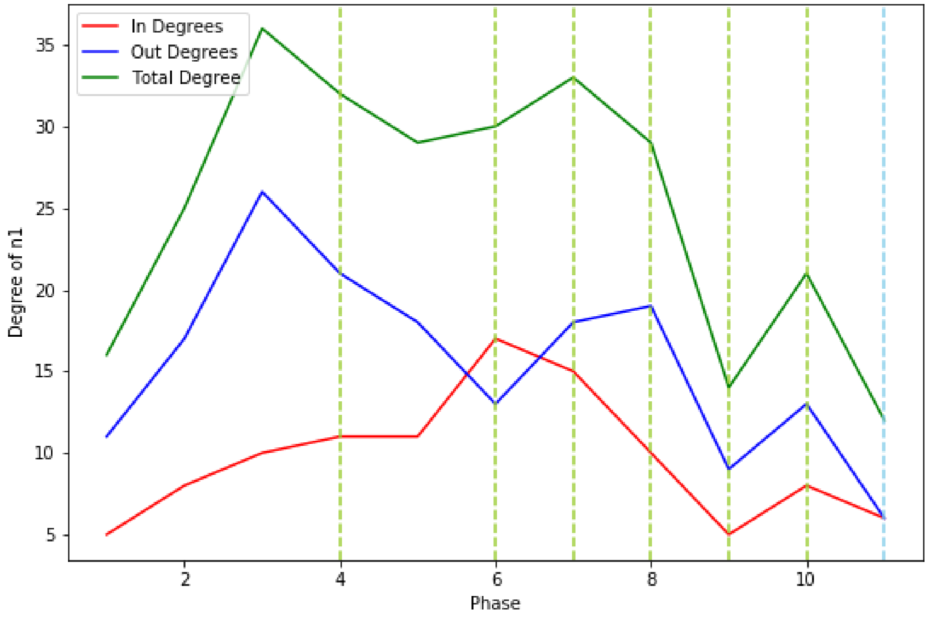


- We can see that the level of communication drops drastically after a major seizure. A plausible explanation could be that a seizure incentivizes the network to operate furtively and under the radar.

- With respect to Morales' activity (n12), we see that every time there is a seizure, there is a reduction in influence/communication through Morales. However, communication resumes after one more phase

(f) Analysis of communication for n1

- Visualization of the 'total communication' index for kingpin



- While Serero seems to have exhibited a decreased degree over time, the entire network had increasing degree. Maybe this means that n1 is isolating himself and more people are communicating in other parts of the network to keep business running. n1 had increasing degree until the first interception by the police.
- In the period after the first seizure of drugs, n1 has decreasing communications outward but more communication inward.
- This indicates a cautious communication strategy by Serero, wherein although he gets access to the important information of the network, the execution is heavily delegated to his lieutenant (n3). And this can be validated by looking at the 'total communication' index of n3 as illustrated in the preceding section

(g) Comments on the strategy adopted by the police

- The effect of the early seizures indicate that cocaine and marijuana are easily substitutable. Thus, it is interesting to note that, when only marijuana is seized, while marijuana betweenness reduces, cocaine betweenness goes up. This suggests that the criminal network changes the drug its dealing after one of the drugs gets seized by authorities.

- When both cocaine and marijuana is seized, betweenness for both cocaine and marijuana organizers goes down showing that there is less communication. The police didn't decrease drug trading overall but only displaced communication from one drug to the other every time they seized a particular drug. So different parts of the criminal network are more active based on whether there has been interception by the police.

- Given the unstable and reactive nature of communication post a seizure, we can plausibly conclude that the strategy would increase the likelihood of directing law enforcement towards the most central members of the network

Problem 4.3: Co-offending network

(a) – (d): General aspects of the network

- Snippets of the code (preliminary questions and creation of offender-offender matrix)

```
phase = pd.read_csv("Cooffending.csv", sep = ',')

sp = phase.to_sparse(fill_value=0)

of = sp['NoUnique']
ev = sp['SeqE']
m = sp['SeqE'].shape

rows = of.index.get_values()
rlist = rows.tolist()

ones = np.ones(m)

on = pd.DataFrame(ones, columns = ['Ones'])
on.set_index = rlist

m1 = pd.concat([of,ev,on],axis=1)

mat_coo = sparse.coo_matrix((m1['Ones'], (m1['NoUnique'], m1['SeqE'])))
mat_coo.toarray()

oc = mat_coo
```

```
cc = (oc)*(oc.transpose())

G = nx.DiGraph(cc, format='weighted_adjacency_matrix')
```

- Events / Offenders
  - Size of the cooffending matrix - 1280459 rows x 15 columns
  - No. of unique offenders: 539593
  - No. of unique events: 1164836
  - Events by year
    - [(2003, 122284),
    - (2004, 133705),
    - (2005, 188115),
    - (2006, 203381),
    - (2007, 214315),
    - (2008, 220791),
    - (2009, 197861),
    - (2010, 7)]

  - Event-event matrix - <1639824x1639824 sparse matrix of type '<class 'numpy.float64'>' with 12575236 stored elements in Compressed Sparse Row format>
  - Offender-offender matrix - <670537x670537 sparse matrix of type '<class 'numpy.float64'>' with 896419 stored elements in Compressed Sparse Row format>

- Notable inference about events and offenders
  - Event with the highest number of offenders: 27849
  - Number of offenders: 156

- Municipality with crime that had most offenders: [66023]

(e) Nodes / Solo offenders

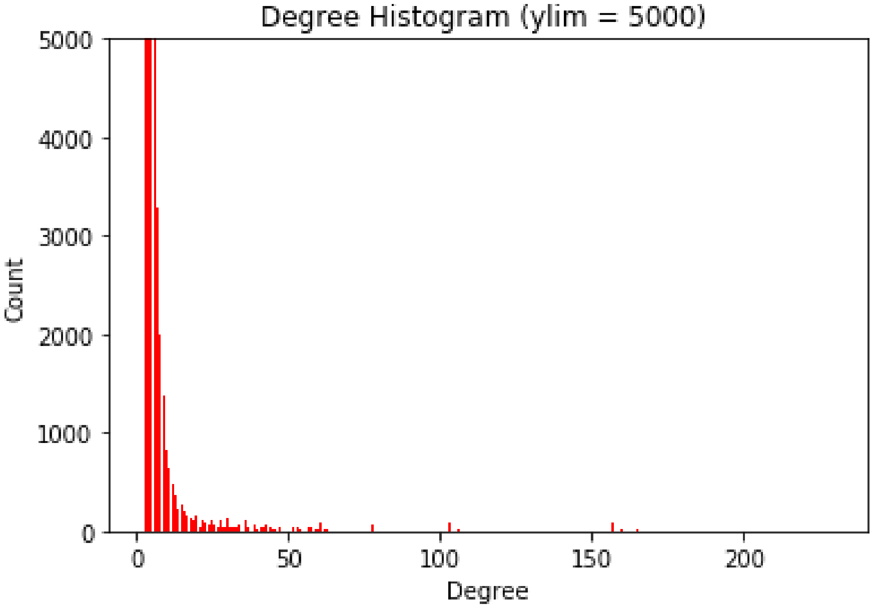
- No. of nodes including solo offenders: 539593
- No of solo offenders: 418434
- No of nodes (excluding solo offenders): 121159
- No of edges 178413
- Code (used for removing solo offenders)

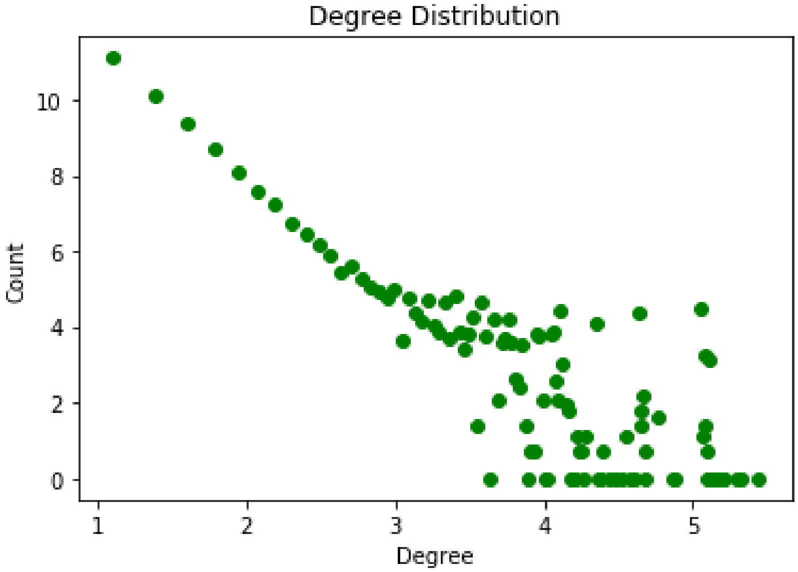
```
remove = [node for node,degree in G.degree() if degree <= 2]
```

```
G.remove_nodes_from(remove)
#if degree=3, two connected criminals
```

(f) – (g) Degree distribution and connected components

- Degree distribution

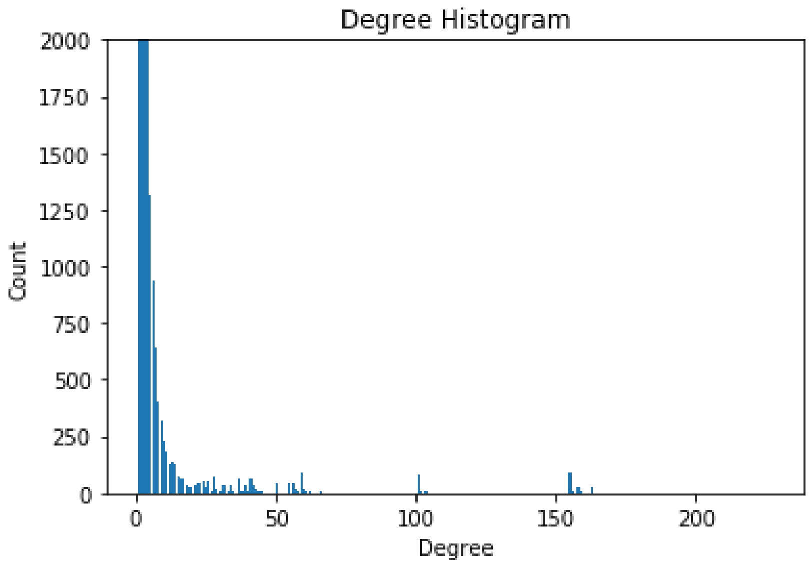


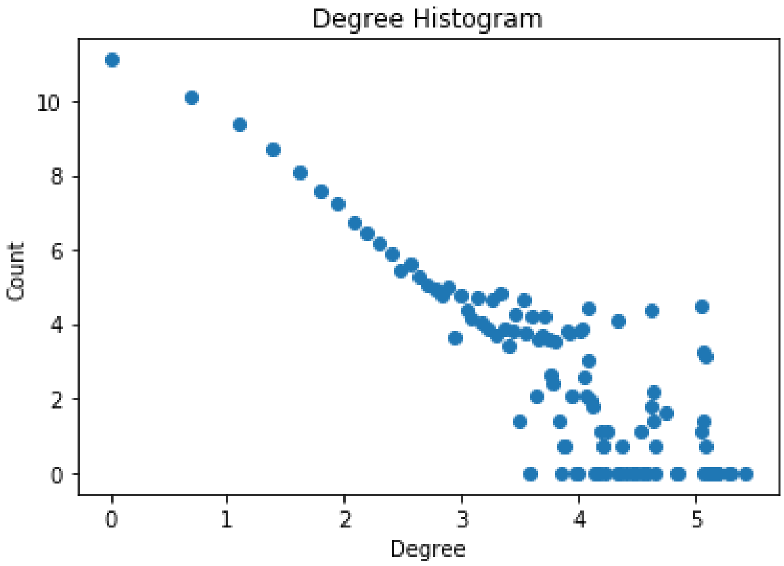


- Number of connected components: 36098

(h) – (j) Analysis of the largest connected component

- Distributions for the largest cc





- Network metrics for the largest connected component

```
print( 'Density of the component: ', nx.density(largest_cc))
Density of the component:  0.00035564548486577961
```

```
print( 'Clustering (unweighted): ', nx.average_clustering(largest_cc))
```

```
Clustering (unweighted):  0.4784259694538959

print( 'Clustering (weighted): ', nx.average_clustering(largest_cc, weight='weight'))
Clustering (weighted):  0.006007783646767325

#nx.diameter(largest_cc)
print ('Diameter:', 48)
Diameter: 48
```

(k) – (m) Further investigation

- Number of crimes committed by youngsters without an adult: 0

  The above result most likely indicates that the data has sieved out juvenile crimes

- The most common crimes are:

  'INTRO PAR EFFR. DANS RES.' (code: 21201, count: 1606, almost 10%)

  'INTRO EF. ETA. COM. PUBL.' (code: 21203, count: 1110)

  'VOL 5000$ - A L''ETALAGE' (code: 21405, count: 1074)

  This suggests that some crimes turn up a significantly greater number of times than others so very specific crimes are committed by youth and adult.

- The most common crime committed by youths is 'Breaking into a residence'. It would be interesting to think of why this is the most common crime.

- When Youths and Adults are both involved in a crime:
  Avg # Youths: 1.4771567649991202
  Avg # Adults: 2.8064629640490293

  This is an interesting result. This means most of the times multiple adults are there with fewer youth. This could be that seniors in a gang are "training" juniors. This could also be because an adult would not go with multiple juniors for risk of someone messing up.

- Interesting extensions for the co-offending network
  - Superimpose spatial clusters to identify vulnerable sections of the youth population to enable preemptive social interventions
  - Rank the nodes as per multiple centrality measures and selectively target those who have higher eigenvalue centralities, but lower degree / betweenness – this indicates that the big gangs are potentially recruiting new members