Review of "Statistical Classification of Social Networks"

Eunseong Jang(22201806)

May 8, 2020

Contents

1. Introduction

- Social networks have been of research interest for a long time
- Some properties of networks can reveal the relationships or differences between different networks
- This paper studied whether the statistical differences of five properties on networks sufficiently characterize a social network
- The five properties are degree centrality, betweenness centrality, closeness centrality, eigenvector centrality and clustering coefficient
- Authors also tested with real data to support their hypothesis that the sufficient properties are degree centrality and clustering coefficient

2. Background: Centralities and Clustering Coefficient

- This section explains the definition of four centralities and clustering coefficient
- Throughout this paper, a network graph is assumed to be un-directed and its edges are un-weighted
- The Degree centrality describes how many direct neighbors one node has in a network
- The adjacency matrix(M) of a network is defined as:

$$M_{ij} = \begin{cases} 1, & \text{if nodes } i \text{ and } j \text{ are connected} \\ 0, & \text{if if nodes } i \text{ and } j \text{ are not connected} \end{cases}$$

2. Background: Centralities and Clustering Coefficient

- The degree centrality describes how many direct neighbors one node has in a network

- The definition of degree centrality of node i in a network is:

$$D(i) = \sum_{j=1,\ j \neq i}^{N} \frac{M_{ij}}{N-1},$$

where N is the total number of nodes

- And the degree of node i is defined as:

$$d(i) = \sum_{j=1,\ j \neq i}^{N} M_{ij}$$

- Betweenness centrality describes how important a node is when considering how much information flows through it in a network

- The definition of betweenness centrality of node i in a network is:

$$B(i) = \sum_{j=1,\ j \neq k \neq i}^{N} \frac{\sigma_{jk}(i)}{\sigma_{jk}},$$

  where $\sigma_{jk}(i)$ is the number of geodesic paths between node j and node k that goes through node i, and σ_{jk} is the number of geodesic paths between nodes j and k

- Closeness centrality describes how a node could pass information to the other nodes
- The definition of closeness centrality of node i in a network is:

$$C(i) = \sum_{j=1,\ j\neq i}^{N} \frac{d_{ij}}{N-1},$$

where d_{ij} is geodesic distance between nodes i and j

2. Background: Centralities and Clustering Coefficient

- Eigenvector centrality not only counts the number of links one node has, but also counts how important is the node it connects to

- The definition of eigenvector centrality of node i in a network is:

$$E(i) = \frac{1}{\lambda} \sum_{j=1,\ j \neq i}^{N} E(i) M_{ij},$$

  where $E(i)$ is eigenvector centrality for node i and λ corresponds to the largest eigenvalue of the adjacency matrix M

- $E(i)$ can be solved from the definition above

2. Background: Centralities and Clustering Coefficient

- A clustering coefficient measures how structured the neighborhood of a node is in a network
- The definition of clustering coefficient of node i in a network is:

$$CC(i) = \frac{2|\{e_{jk}\}|}{d(i)(d(i)-1)}$$

- set $\{e_{jk}\}$ contains all the links existing between the neighbors that node i immediately connects to
- $|\{e_{jk}\}|$ represents the number of triads that includes node i

3. Markov Graph

- A network system described by its adjacent matrix M can be considered as a sequence of random variables of M_{ij}'s
- We can also think of it as a random Markov field which has an underlying dependence structure describing the conditional dependence between M_{ij}'s
- Such dependence structure is called a dependence graph $D = \{node_D, edge_D\}$ for the network $M = \{node_M, edge_M\}$
- According to Hammersley-Clifford theorem, the probability of a general network to show up is:

$$P(G) = z^{-1} exp[\sum_{c \subseteq G} \alpha_c],$$

where z is the partition that normalizes $P(G)$ and α_c is a constant corresponding to a clique c in $\{D\}$

3. Markov Graph

- In a network system, two relationships which do not share a common node are conditionally independent of each other

- In this case, the cliques in D only correspond to triads and stars in M and M is called a $Markov\ Graph$

- Then, any homogeneous un-directed Markov graph has probability:

$$P(G) = z^{-1} exp[\tau t + \sum_{i=1}^{N} \theta_i d(i)],$$

  where t is number of triangles in network G, $d(i)$ is the degree for node i, τ and θ_i are arbitrary constant corresponding to t and $d(i)$

- Accordingly, authors claimed that there are the only two crucial properties describing a network system; clustering coefficient(τt) and degree centrality($\sum_{i=1}^{N} \theta_i d(i)$)

4. Relation Between Centralities and Clustering Coefficient

- To validate the hypothesis that only two crucial properties describe a network system, authors checked the statistics(moments) of properties introduced in section 2 except degree centrality

- They repeated the simulation 100 times following process

  1) sample a graph which obeys a certain graphical degree sequence, coming from a power law distribution.

  2) uniformly sample 10^5 sub-graphs that obey this degree and observe how the four centralities and clustering coefficients vary

- For each time, different degree sequence is selected

4. Relation Between Centralities and Clustering Coefficient

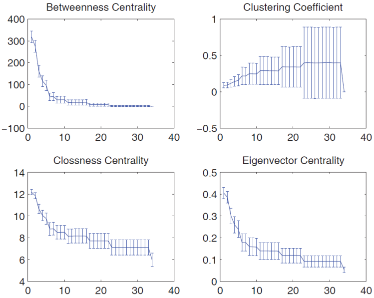

Figure: 1. Behavior of Betweenness Centrality, Closeness Centrality, Eigenvector Centrality and Clustering Coefficients

- From the results, we can see that for the same degree sequence, the variance of three centralities among sub-graphs remains small compared to that of clustering coefficient

4. Relation Between Centralities and Clustering Coefficient

Network Properties	STD/Mean
Betweenness Centrality	0.9101
Closeness Centrality	0.4705
Eigenvector Centrality	0.9178
Clustering Coefficients	2.6701

Figure: 2. STD/Mean of Network Properties

- The relative average standard derivation compared to the mean of the properties also shows that three centralities are less variable than clustering coefficient
- This indicates that when the distribution of degree centralities obeys a power law, the other centralities cannot provide any more information about the network

4. Relation Between Centralities and Clustering Coefficient

- Thus authors proposed that the characteristics of a network are determined by its clustering coefficient and its degree centrality
- To test the proposition, they build a model classifying different types of networks
- Their model uses statistics on degree centrality(D) and clustering coefficient(CC) up to fourth order as its variable set $\{A_i\}$:

$$\{A_i\} = \{mean(D), var(D), skewness(D), kurtosis(D),$$

$$mean(CC), var(CC), skewness(CC), knotisos(CC)\}$$

- By comparing the different sets $\{A_i\}$ of different networks, the model can figure out the class similarity of different networks

5. Network Classification

- The proposed model was tested using 800 sub-networks sampled from two giant networks
- One is a snap shot of the Internet at the level of autonomous systems measured by Mark Newman from data in July 22, 2006[3]
- The other one is a weighted network of co-authorships between scientists posting preprints on High-Energy Theory E-Print Archive between Jan 1, 1995 and Dec 31, 1999[3]
- All links between nodes are set to be un-weighted and natural, and the size of the sub-network is set to be from 50 nodes to 200 nodes
- There are 400 sub-networks for each type of the networks

- The mean of vector $\{A_i\}, i = 1, ..., n(A)$ is calculated for each data base:

  $\{\bar{A}_{i,INT}\}$ for the data from internet snap shot
  $\{\bar{A}_{i,HEP}\}$ for the data from the network of co-authorships among the scientists posting on High-Energy Theory E-Print Archive

- Then, the normalized distance from the $\{A_i\}$ of the tested sub-network to $\{\bar{A}_{ij}\}$ is calculated as:

$$ND_j = \sum_i \frac{|A_i - \bar{A}_{ij}|}{\bar{A}_{ij}}, \quad i = 1, ..., n(A) \text{ and } j = INT, HEP$$

5. Network Classification

- In order to show that the performance is best when $\{A_i\}$ is the statistics of degree centralities and clustering coefficients, $\{A_i\}$ is chosen to be in six cases:

  Case 1. Statistics of degree centralities and Clustering coefficients
  Case 2. Statistics of degree centralities
  Case 3. Statistics of clustering coefficients
  Case 4. Statistics of betweenness centralities
  Case 5. Statistics of closeness centralities
  Case 6. Statistics of eigenvector centralities

5. Network Classification

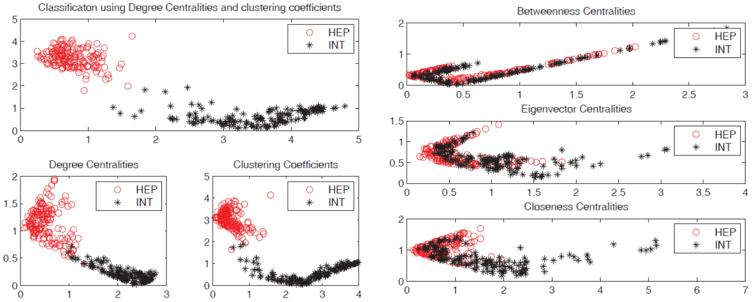

Figure: 3. Classification result of case 1 to 6. y axis represents ND_{HEP}, and x axis represents ND_{INT}

- From the results, we can see two distinct clusters in case 1 most clearly

6. Imitation of Paper's Work

- I repeated the proposed model using 800 sub-graphs sampled from same two giant networks:
  a snap shot of the Internet and co-authorships on High-Energy Theory E-Print Archive

- Each sub-graph was designed to have 50 to 200 nodes randomly

- To sample sub-graphs from two giant networks, snowball sampling was used, since it seemed that snowball sampling can catches centralized structures better than other methods

- To begin with, I plotted some samples from sub-graphs of each group to see if there is difference in graph structure between two groups

6. Imitation of Paper's Work

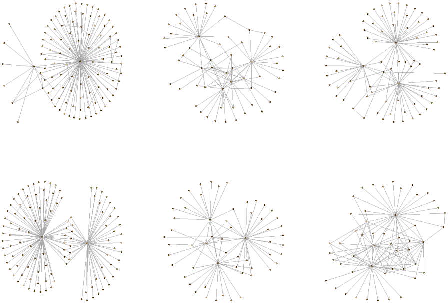

Figure: 4. Plots of Internet sub-graphs samples

Sub-graphs from Internet data tend to have few nodes with relatively much more neighbors than most nodes

6. Imitation of Paper's Work

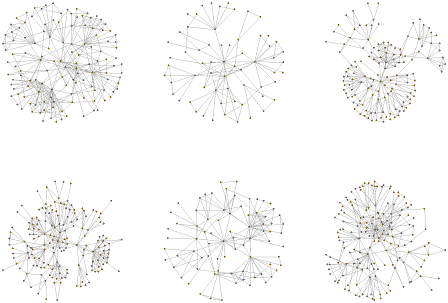

Figure: 5. Plots of co-authorships sub-graphs samples

Unlike previous plot, there are more than 1 or 2 nodes with high centrality
Next, analogous to paper's work, I calculated four statistics for all 6 cases, then plotted the sub-graphs' location on normalized distance space

6. Imitation of Paper's Work

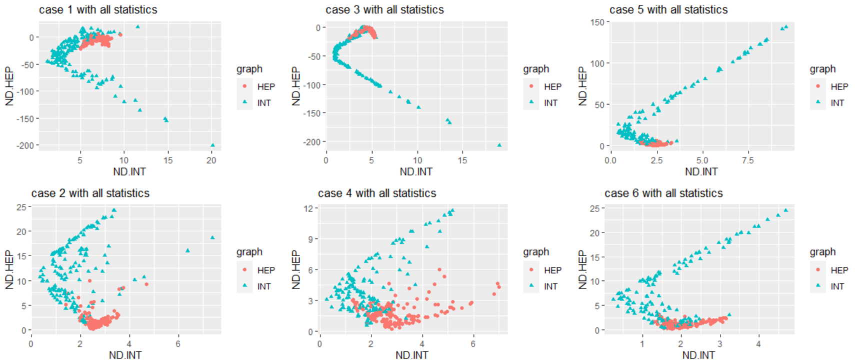

Figure: 6. The results with all statistics: INT for the sub-graphs from Internet data, HEP for the one from co-authorships data

Definitely two network group is separated most distinctly in case 1
However, there is some overlap zone between two groups and the domain of normalized distance is quite different from the paper(Figure 3)

6. Imitation of Paper's Work

Thus I re-plotted the sub-graphs' location while the normalized distances are calculated without kurtosis

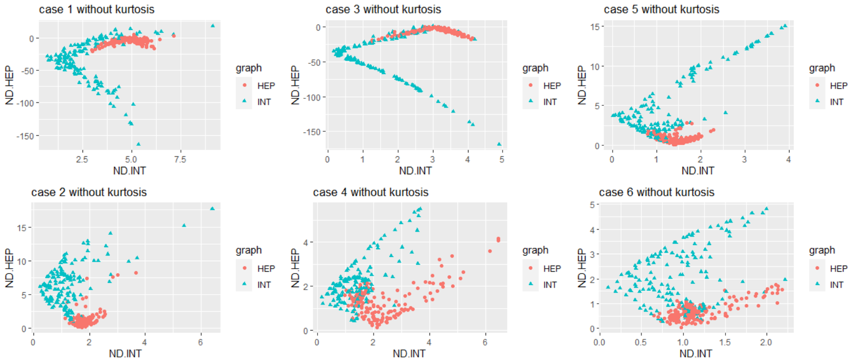

Figure: 7. The results without kurtosis

The problem is not fully resolved, so repeated the process again without skewness and kurtosis

6. Imitation of Paper's Work

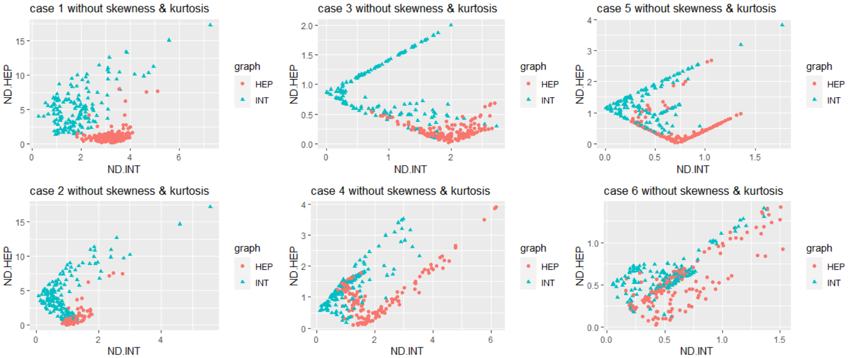

Figure: 8. The results without skewness & kurtosis

Now the result is similar with the paper's as shown in Figure 6.
Two group are clustered in distinct locations on above plot for case 1

References I

[1] Hamid Krim Tian Wang.
Statistical classification of social networks.
IEEE, 22(1):68–76, 2012.

[2]Eric D. Kolaczyk.
Statistical analysis of network data.
Springer-Verlag New York, 2009.

[3] M. E. J. Newman.
http://www-personal.umich.edu/~mejn/netdata.
website.