Assignment of EE 4389 - Group 11
Modeling and Data Analysis in Complex Networks

Nivedita Prasad – 4712099
Chitra Balasubramanian – 4742907

March 12, 2018

Contents

1 A – Compute the following network properties for Graph G.

1. **What is the number of nodes N, the number of links L, the link density p, the average degree E[D] and the degree variance Var[D]?**
   - **Number of nodes :** 327
   - **Number of Links. :** 5818
   - **Link Density p :** 0.1091537
   - **Average degree E[D] :** 35.5841
   - **Degree variance Var[D] :** 182.7406
2. **Plot the degree distribution. Which network model, Erdos-Renyi (ER) random graphs or scale-free networks, could better model this network? Why?**

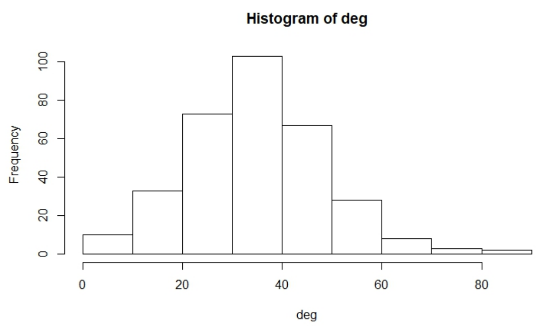

Figure 1: Degree distribution

The degree distribution follow Gaussian distribution hence Erdos-Renyi graph is better for this model.

3. **What is the degree correlation (assortativity) D? What is its physical meaning?** *Physical Meaning :* The preference of network's nodes to attach itself to other nodes that are similar to it .It is the correlation between the degrees of the node. *Degree Assortativity :* 0.03317577

4. **What is the clustering coefficient C?** *Clustering Coefficient* is 0.5035048

5. **What is the average hop-count E[H] of the shortest paths between all node pairs? What is the diameter Hmax?** *Average hop-count E[H]* is 2.159434. *Diameter Hmax* is 4. The nodes in the *Diameter* are 5, 69, 155, 278, 24.

6. **Has this network the small-world property? Justify your conclusion quantitatively.** The clustering coefficient is less than 0.5 hence it does not follow small world property.

7. **What is the largest eigenvalue (spectral radius) 1 of the adjacency matrix?** *Largest Eigenvalue:* 41.23161

8. **What is the second smallest eigenvalue N1 of the Laplacian matrix (algebraic connectivity)?** *Algebraic Connectivity :* 1.930049

Table 1: Metrics and Values

Metrics	**Value**
Number of nodes	327
Number of Links	5818
Link Density p	0.1091537
Average degree E[D]	35.5841
Degree Variance	182.7406
Degree Assortativity	0.03317577
Clustering Coefficient	0.5035048
Average hop-count E[H]	2.159434
Diameter Hmax	4
Largest Eigenvalue	41.23161
Algebraic Connectivity	1.930049

2 B. Information spreading on a temporal network

9. **Taking all the N iterations into count, plot the average number of infected nodes E[I(t)] together with its error bar (standard deviation sq.root of Var[I(t)]) as a function of the time step t.**

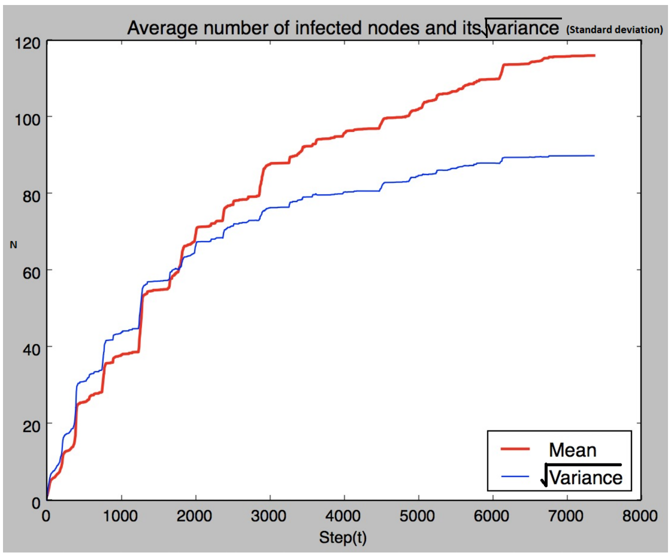

Figure 2: Average number of infested nodes and it's standard deviation

Figure 2, is a plot of the number of infested nodes (N) vs the time steps (t) . It can be seen from the graph that the distance between the mean curve and the Standard deviation curve is very minimal, showing that the data lies mostly around the mean and one standard deviation away. It can also be observed that as the number of infested nodes increase over time the standard deviation moves lower than the corresponding mean. This separation is observed when the infested nodes hit 80 at time step around 2500.

10. **How influential a node is as a seed node could be partially rejected?**
Given the total number of nodes to be 327, we traverse every single node starting from node 1 up to node 327 till 80% of the network is infested on every iteration. This is done on iteration for all 327 nodes. The time stamps are recorded for each iteration and then arranged in ascending order to compute the shortest time taken. We noted that at t=5321, 80% of the nodes were infested and the seed node is node=202. When the seed node considered is node 202, the time taken to infest 80% of the total number of nodes is 5321.

11. **Explore which nodal level network feature could well suggest the nodal influence discussed in 2. Compute the degree and clustering coefficient of each node in the aggregated network G and rank the importance of the nodes according to these two centrality metrics respectively. Plot rRD(f) and rRC(f) as a function of f where f = 0:05; 0:1; 0:15; :::; 0:5. Which metric, the degree or the clustering coefficient could better predict the influence of the nodes? Why?**

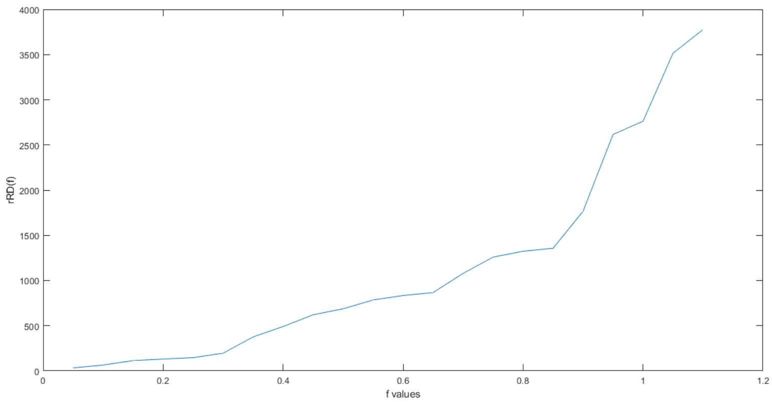

Figure 3: A graph of f values vs rRD(f)

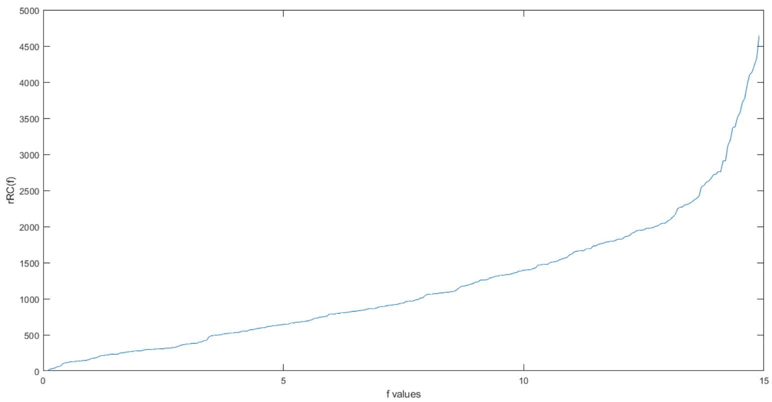

Figure 4: A graph of f values vs rRC(f)

Degree centrality is defined as the number of neighbors a given node has. The highly influential nodes in the network have a higher degree centrality than the other nodes in the same network. Transitivity or clustering coefficient is known as "the extent to which the relation that relates two nodes in a network that are connected by an edge". From Figure 3 and 4 it is quite evident that the clustering coefficient helps predict the influence of a given node in the network as it is more linear. It also measures a higher variance than degree centrality.

12. **Propose another two nodal/centrality features that could possibly well predict nodes' influence.Compare the two features you proposed and the two features proposed in question 11): which feature better/badly reflects how influential a node is and why?**
The other two temporal metrics that can be considered are closeness centrality and betweenness centrality. Closeness centrality can be regarded as a measure of how long it will take to spread information from the seed node to all the other nodes sequentially in a given network. "The betweenness centrality for each vertex is the number of these shortest paths that pass through the vertex".Higher the betweenness centrality stronger the connection in the network. Figure 5 shows a plot of f values and it's rRB(f)(closeness centrality).It can be seen that this graph is more linear compared to figure 4.Hence proving that it is a better measure to measure the influence of the network.

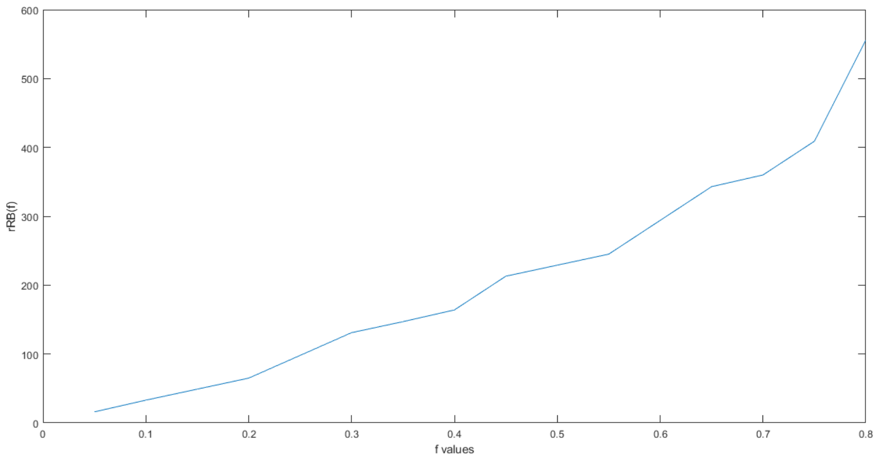

Figure 5: A graph of f values vs rRB(f)

13. **Reflection on B: What can be improved in the aforementioned analysis method to discover which nodal feature could better reflect how influential a node is as a seed node?**

A better model or a better approach can be suggested instead of basic methods of understanding the spread of the infection. Famous models like the SI(susceptible - infected networks) or the SIS(susceptible-infection-susceptible) models would better help understand the susceptibility and infestation of nodes. These work on the principle of decay or derivation and also on the rate at which nodes are getting infested. This model can work better as it has decay rates and information/infection spread rates, which can better assess the network.

3 C. Influence of temporal network features on information spreading.

14. **Construct the following three temporal networks. Gdata, G2, G3** *Gdata :* Network constructed with Highschool dataset . *G2 :* Network constructed with Highschool dataset's time variable shuffled without replacement. *G3 :* Network constructed with only the uniques values of High school dataset that is resulting in 5818 unique links between nodes without considering the time stamp column. Highschool dataset's time variable shuffled with replacement for each of the time value ranging from 1 to 7353. For each time value, the links between nodes are picked from the graph G(Graph used in Part A questions)'s nodes . This is repeated for 1000 times to get a proper shuffle of nodes for each time stamp.

15. **Plot the average number of infected nodes E[I(t)] and the standard deviation sq.root Var[I(t)] as a function of the time step t for Gdata, G2 and G3 respectively. Compare and rank the information spreading performance (e.g. prevalence or speed of the spread) on these three temporal networks. Interpret/explain your observation. For example, which temporal network features could possibly explain the different spreading performance on these temporal networks?**

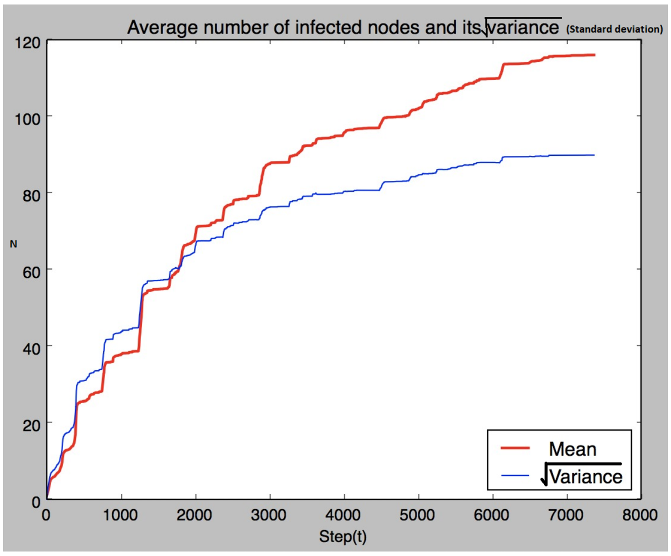

Figure 6: Average number of infested nodes and it's standard deviation for G-data