Network Theory: Properties of Social and Fictional Networks

Math 485, Spring 2018

Christina West, Taylor Martins, Yihe Hao

Abstract:

Network Theory has been widely used and broadly researched to analyze various fields of networks. The main goal of this project is to utilize different patterns of network theory, and compare multiple calculated coefficients in order to conclude whether there's any similarities in between each other. Furthermore, network trees and degree distribution plot have also been used in order to aid visualization of the networking system. At the end, our result shows that similarities do exist in between different field of networks.

Introduction:

In this project, a methodology is created and utilized in order to determine the similarities and differences between different types of networks. This methodology is generated from the numerous tools provided by network theory for investigating the qualities of different networks. The parameters are calculated using data regarding the nodes and edges of the different networks. Real-world networks present different values for these parameters than fictitious networks. Analyzing these parameters in tabulated format allows for a simple determination of the similarities and differences of the presented networks. In this study, the particular networks analyzed were from *Star Wars*, *Harry Potter*, and congress members' Twitter profiles. The people in each of these networks were organized as the node data and their relationships with

each other comprised the edge data. Compiling this data with code in RStudio and using the iGraph package allows for the parameters desired to be determined and organized for analysis.

Methodology:

Data. Data was collected from github. There were data sets from three different networks consisting of csv files of nodes and edges. Two networks, Harry Potter characters and Star Wars Episode IV characters, were fictional while the third network, Congress Twitter accounts, was social. Data were analyzed using the igraph and poweRlaw packages for R.

Coefficient Computation. Csv files of nodes and edges were converted into a "graph" data frame using igraph. This data frame consists of vertices and edges. From there, we could determine the length of the vertices and edges to find out the number of nodes and edges. Degree distributions were calculated using the commands degree, degree_distribution, and poweRlaw's compare_distributions, which compared power law and exponential distributions for goodness of fit based on maximum likelihood ratios for the generated distributions from the network data. The minimum tail observations for the generation of these distributions were estimated using a bootstrapping algorithm developed by poweRlaw to minimize the Kolmogorov-Smirnov statistic. Clustering coefficient of the graphs were calculating using the transitivity command in igraph. Assortativity was calculated using igraph's assortativity_degree.

Assortativity. In the igraph package, documentation is included about how each coefficient referred to in this paper is calculated. For assortativity, it gives a measure of the homophyly of the vertices in the graph. This coefficient is calculated in slightly different ways

depending on if the graph is undirected, directed, weighted, or unweighted. The equation used when the graphs are undirected is given below. The measure of assortativity shows if there is some sort of preference for the vertices in the graph to attach in a similar way. For the networks that we are concerned with, this coefficient shows if the vertices of high degree associate with each other or with other vertices that have a low degree. When the assortativity coefficient calculated is positive, it means that the vertices associate with others that have like characteristics. When the assortativity coefficient calculated is negative, it means that vertices associate with other vertices that are different.

$$a = \frac{\sum(jk(e(j,k) - q(j)q(k)), j, k)}{\sigma(q)^2}$$

Clustering Coefficient. The clustering coefficient is found for each vertex in the graph. The igraph package looks at each vertex individually and considers all other vertices in the neighborhood. It then looks at the edges that are connecting the vertices and returns a fraction denoting how connected the particular vertex is based off the total number of possible connections that this vertex could have. In order to give a measure for the graph as a whole, the clustering coefficient is averaged across all the vertices in the graph. When this number is close to one, it means that the vertices in the graph are highly clustered. So, out of all the total connections that each vertex could make with the vertices in the neighborhood, almost all the connections are made. When the number is closer to zero, it means that each vertex is making the minimum number of connections possible with each vertex in the neighborhood. So, this measure gives a good idea of the connectedness of the graph when averaged over all the vertices.

$$c = \frac{|e_{jk} : v_j, v_k \epsilon N_i, e_{jk} \epsilon E)}{k_i(k_i - 1)}$$

Betweenness Centrality. The betweenness centrality is another measure that is calculated for one vertex at a time and then averaged across all the vertices in the graph. It checks the shortest path between two different nodes and then sees how many of those shortest paths go through a particular node. If a node has a high betweenness centrality it means that to go between particular nodes in the graph, the shortest path will be through some particular node. Concerning stories, this is typically going to be large when considering the effect of the main character. Often times, many periphery characters are unrelated to each other and the shortest way to connect them in the graph is to go through the main character who encounters most of the other people in the story. This case shows when the betweenness centrality of the graph will be very large. Otherwise when characters exist mainly in their own groups, one would expect the betweenness centrality to be lower.

$$r = \sum (\frac{\sigma_{st}(v)}{\sigma_{st}})$$

Giant Component. The giant component of the graph is found by the igraph package by first separating the graph into components. This is done by utilizing a random walk algorithm. The package utilizes a method that jumps into the graph at a random location and traverses the edges in order to detect different components. The components are then color coded in the graph that is generated. In order to calculate the giant component, the size of the largest component is calculated and divided by the total number of nodes in the network. This tells what percentage the largest component in the network covers compared to the size of the entire network. When

this number is close to one, it means that most of the nodes in the network are closely connected to each other. There would be a minimal number of vertices lying in the periphery regions. When the number is closer to zero it means that many of the nodes are in their own disjoint communities.

$$g = \frac{max(components)}{N}$$

Random Networks. Random networks were generated using igraph's sample_gnm, a random network generator which generates a simple random network of a specified number of nodes and edges with constant probability of degree. The random networks were matched to the number of nodes and edges of each of the Star Wars, Harry Potter, and Congress Twitter networks, for comparison of characteristic path length and clustering coefficient, to determine whether or not each network was small world.

Results:

The star wars network was small world, disassortative, had a power law distribution, and had a giant component larger than 90 percent of its nodes, all characteristics of fictional networks. The Harry Potter network was small world, disassortative and had a giant component larger than 90 percent of its nodes, and exhibited an exponential degree distribution. The Congress Twitter network was not small world, was disassortative, exhibited exponential degree distribution, and had a giant component larger than 90 percent. All coefficient calculations are included in Table 1.

	Star Wars Episode IV	Harry Potter	Congress Twitter Network	Random Networks (Star Wars | HP | Congress)
Number of nodes	22	65	517	22 | 65 | 517
Diameter	3	4	3	3 | 3 | 2
Characteristic path length	1.9095	2.0284	1.66	1.8381 | 1.7654 | 1.5357
Clustering Coefficient	0.5375	0.4134	0.57	0.2492 | 0.2468 | 0.4629
Degree Distribution	power law (R = 0.1697)	exponential (R = -1.0166)	exponential (R = -1.541)	N/A
Giant Component	0.9545	1.0	0.99	1.0 | 1.0 | 1.0
betweenness centrality	15.4932	Error	168.88	9.7727 | 24.4923 | 168.882
Assortativity	-0.1934	-0.2069	-0.1027231	0.0730 | -0.0792 | -0.0107

Table 1. Network coefficients calculated using igraph for each of 3 networks. The last column represents random networks that have the same number of nodes and edges as each of the 3 networks, generated with constant probability.

Discussion:

The Harry Potter network exhibited exponential degree distribution, which is unusual for a fictional network. One explanation for this finding is that the degree distribution was calculated using an algorithm in the poweRlaw package which calculated a minimum value for the probability of degree, and only looked at data in the tail of the distribution above this minimum value. For the Harry Potter network, only 18 of the 65 nodes were above this cutoff value, so essentially, the distribution was calculated with a small sample size and was prone to larger statistical errors.

The Star Wars network, though exhibiting characteristics consistent with fictional networks, was also calculated using a small sample size. The total number of nodes in the graph was 22 and the algorithm for calculating the distribution had a minimum cutoff value that excluded all but 9 of the nodes. Still, we conclude that the degree distribution was a power law because the R statistic was positive, indicating that it is more likely that the data came from a power law distribution than from an exponential distribution.

The Congress Twitter network had characteristics that were mostly inconsistent with social networks, namely, it was not small world, it was disassortative, and it had a giant component which was larger than 90 percent. This is somewhat unsurprising, as the types of social networks which have classically been studied do not include social media networks where connection between two nodes is more instant and artificial. The types of social networks that were cited in Paidrag, Carron, and Kenna (2012) are collaboration networks, such as two actors appearing in the same movie together, or two co-authors on a paper. One can speculate that nodes in these types of networks are less connected overall because it takes significantly more time, talent, and happenstance to become connected in this way, whereas Twitter connections are

virtual, instantaneous, and easier to obtain. Moreover, according to Su, Sharma & Goel (2016), the development of Twitter's "Friend of a friend" feature in 2010 has led to an increased popularity for accounts which are already popular.

In terms of our Congress Twitter network, this means that nodes which have high degree are likely to experience an increase in degree, and all nodes, regardless of degree, are likely to increase their connectivity with nodes of higher degree. This connection between nodes with high degree and other nodes in the network, regardless of degree, could lead to disassortativity in the network overall, as low-degree nodes form connections with high degree nodes. This could also lead to an increase in the giant component, as nodes with low (or no) degree are recommended by Twitter to connect with nodes of high degree, who act as a bridge into the larger network. The reason the Congress Twitter network was not small world is because it's clustering coefficient was roughly the same as the clustering coefficient for a random network of the same size. The random network used for comparison had the same number of nodes (517) and edges (61766) as the Congress Twitter network. It could be the massive number of connections that caused the random network to have a clustering coefficient which was higher than for the corresponding Harry Potter and Star Wars random networks.

Conclusion:

A network tree has been used in order to aid visualization of the networking system. Figure 1 shows the network connection in between Star-Wars and Harry Potter characters. Since Harry Potter has way more characters than Star-Wars, the graph seems to be much more

complicated than the network tree for Star-Wars. The results of our model indicate interesting trends within the data.

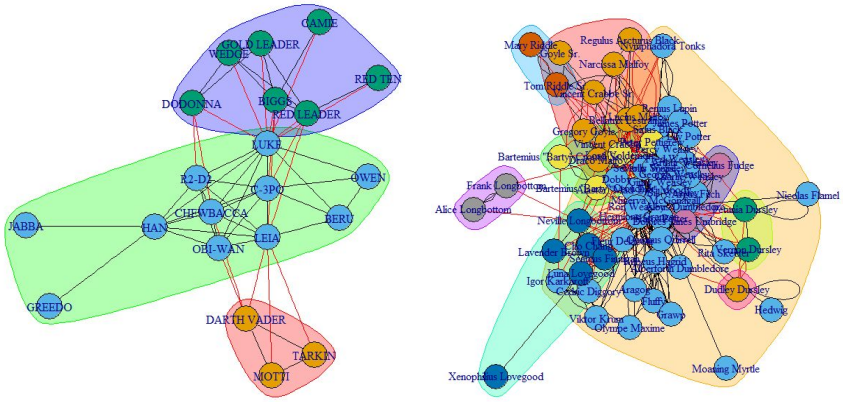

Figure 1. Network connection graph for Star-Wars & Harry Potter [3]

As we have discussed in the previous section, figure 2 shows the degree distribution for both the Congress Twitter network as well as the Harry Potter network.

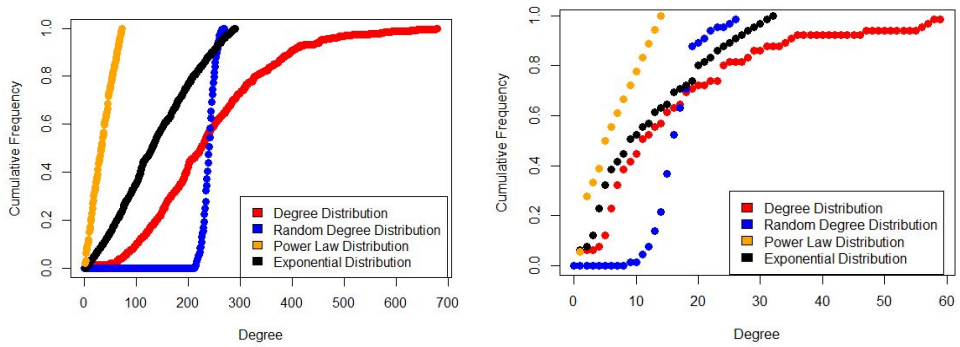

Figure 2. Dynamics of Degree Distribution (Congress Twitter [left] & Harry Potter [right])

Furthermore, by examining the calculated coefficients mentioned in the methodology section, we are able to conclude that there is similarity between fictional and social networks. This conclusion has been drawn based on degree distribution plot, clustering coefficient, giant component, betweenness centrality and assortativity.

References:

1. Colin S. Gillespie (2015). Fitting Heavy Tailed Distributions: The poweRlaw Package. Journal of Statistical Software, 64(2), 1-16. URL http://www.jstatsoft.org/v64/i02/.

2. D.(2014,November5).Dpmartin42/Networks.From https:// github.com/dpmartin42/Networks/tree/master/Harrypotter/data

3. Gábor Csárdi, Tamás Nepusz: The igraph software package for complex network research. InterJournal Complex Systems, 1695, 2006.

4. P.(2016, January 20). ablobarbera/data-science-workshop. Retrieved from https://github.com/pablobarbera/data science-workshop/blob/master/sna/data

5. Padraig Mac Carron and Ralph Kenna, Universal Properties of Mythological Networks, EPL, 99 (2012) 28002, doi: 10.1209/0295-5075/99/28002

6. Su, J., Sharma, A., & Goel, S. (2016, April). The effect of recommendations on network structure. In Proceedings of the 25th international conference on World Wide Web (pp. 1157-1167). International World Wide Web Conferences Steering Committee.