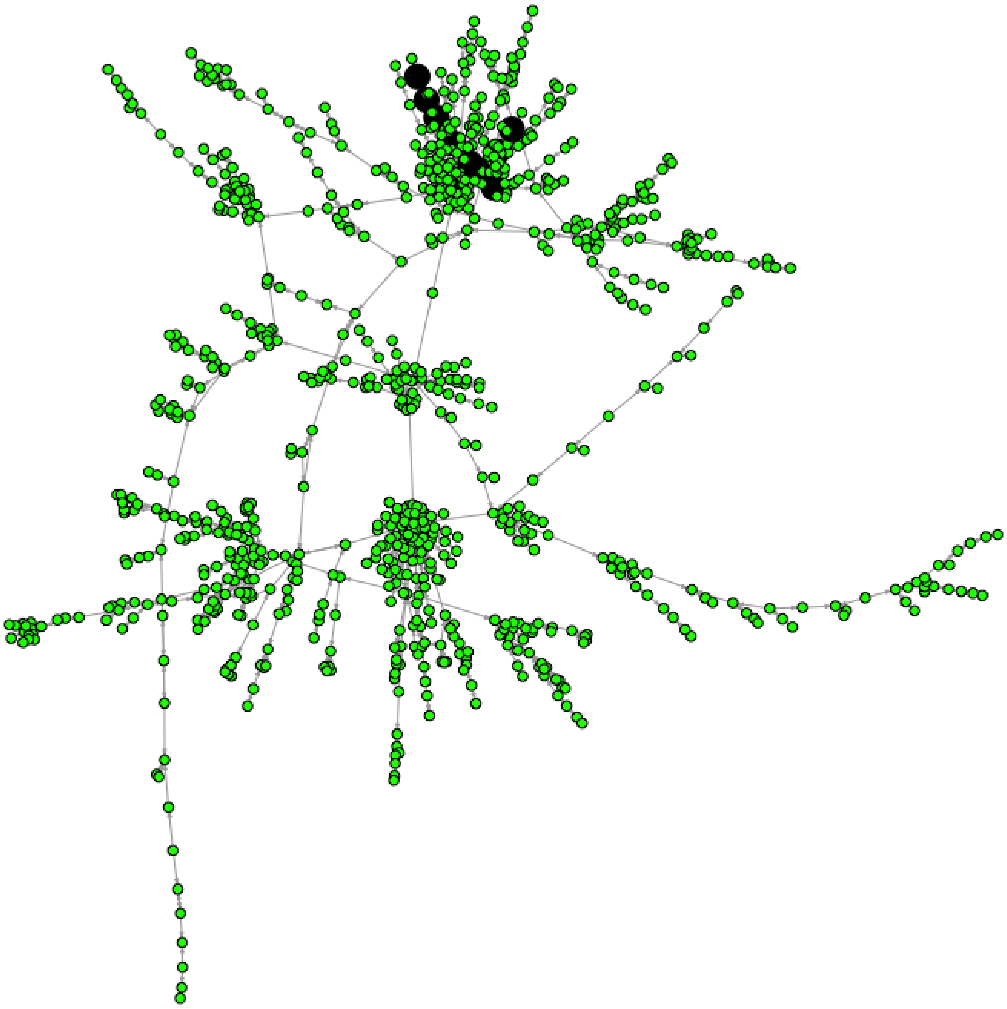
Social Network Team Assignment 1_Team14_SecA

Garrett Chaffey, Chih Han Chi, Sammi Chueh, Pratheek Kumar



5. Provide your insights from the visualizations of the subcomponent using iGraph

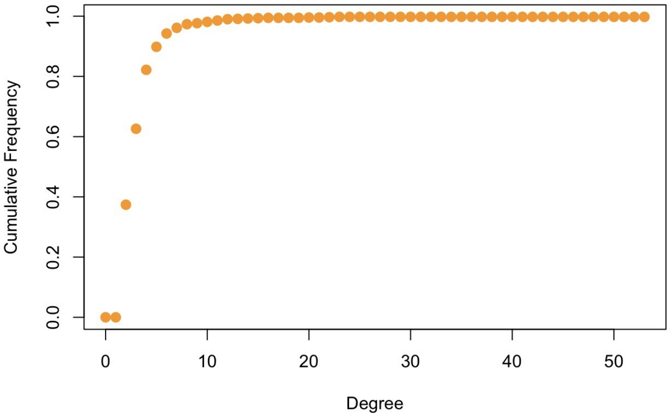
The graph shows 904 nodes, which means there are 903 nodes related to book whose id = 33. There are a few (three specifically) nodes scattered throughout the graph which act as bridges between the largest clusters of the graph representing all of the nodes. Additionally the largest clusters with the most connections seem to reside in the middle of the graph (along the vertical axis), alongside other clustered but not necessarily well connected branches that branch off of the central area. These branches consist of nodes that are connected to each other through other nodes but which do not necessarily have connections to all of the nodes themselves (E.g. A -> B -> C and A is connected to C through B but not connected to C directly). In this case, we will focus more on the directed side, due to our data and subcomponent have direction. The diameter of directed nodes is 9. The undirected diameter is 42 nodes, while the directed diameter is 10 vertices. The directed diameter consists of these vertices: 37895, 27936, 21584, 10889, 11080, 14111, 4429, 2501, 3588, 6676.

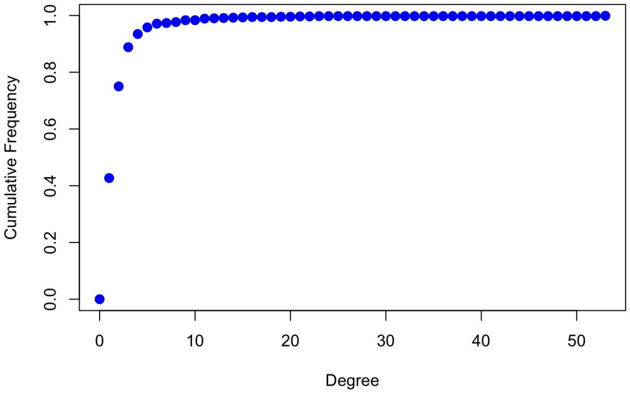
(directed graph, dark blue dots represent nodes which are part of the diameter)

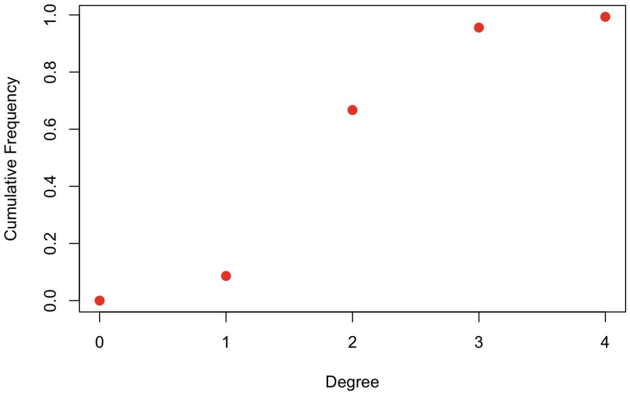
6. Interpret your results after computing various statistics about this network: degree distribution, density, closeness, betweenness, hub/authority score

Degree Distribution-Centrality:

The degree of a node refers to the number of ties associated with a node. “Deg1” measures all the ties going in and out. “Deg2” measures all the ties going in, and” Deg3” measures all the ties going out. In our case, deg1 gives an output of books ids and their corresponding total number of links going to and from the focal product in the network. From those graphs which show the degree distribution of Deg1, Deg2, and Deg3, we can see that the distribution of Deg3 has a big difference compared to Deg1 and Deg2. For Deg1 and Deg2, the degree are mainly in 0 to 5 and have cumulative frequency to 100% at 53. However, in Deg3, its degree is mainly on 2 and get cumulative frequency to 100% at 4. These phenomena can also be known when we see the centralization. Through those graphs, we can anticipate that the centralization of Deg3 will be the smallest and Deg2 will be the largest. The truth is as we thought, the centralization of Deg3 is 0.003 and the centralization of Deg2 is 0.057.


Cumulative Frequency
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
50
Degree


Cumulative Frequency
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
50
Degree


Cumulative Frequency
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
4
Degree

(the graphs shown above represent deg1, deg2, and deg3 respectively)

Density -

The density is 0.001436951, which is a very low number, which indicates that the network is quite dense. Density is computed as a ratio representing the total edges present in a graph compared to the maximum number of edges that the graph can contain, the low number shows that the nodes are gathered together in a clump as opposed to being scattered over a larger "area".

Closeness –

Our graph had a closeness ranging from a minimum of approximately 0.000043 (belonging to book 123953 aka Ratchet and Clank: Prima's Official Strategy Guide) to approximately 0.000161(belonging to book 33 aka Double Jeopardy (T*Witches, 6)). Book number 123953 (ie Ratchet and Clank) has a low closeness, meaning that it has a low centrality due to a large distance to other nodes. In contrast book 33 (or Double Jeopardy) has a higher closeness. This means that it has a high centrality due to short distances to other nodes in the network. Closeness centrality is 0.11.

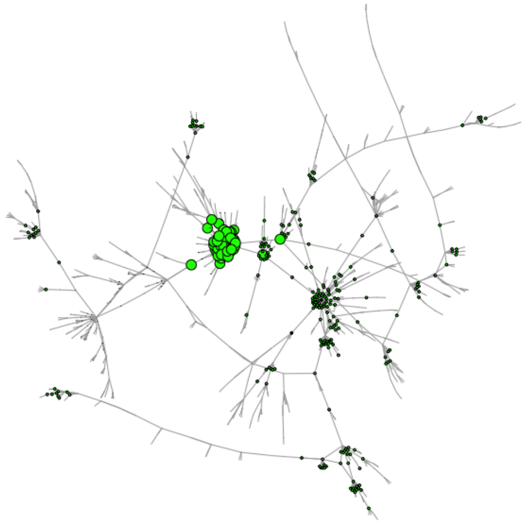
Betweenness –

The node with the highest betweenness score stands at 298 and represents the book The Narcissistic Family : Diagnosis and Treatment. This node acts as a "bridge" between a large number of other nodes. If this node were to be removed from the network a large number of other nodes would see their distance from each other increase. Between centrality is 0.00036.

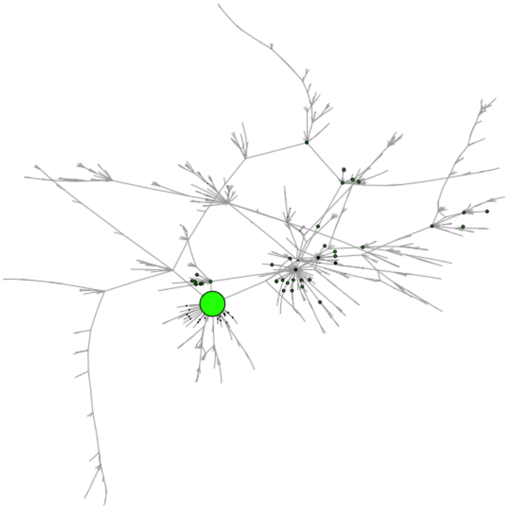
Hub/authority scores -

Hub refers to outgoing nodes, and authority refers to incoming nodes. In this case, the max hub score is 1, which is the main hub, whose id = 33. And the max authority score is also 1. This tells us the main authority whose id = 195144.

Hubs

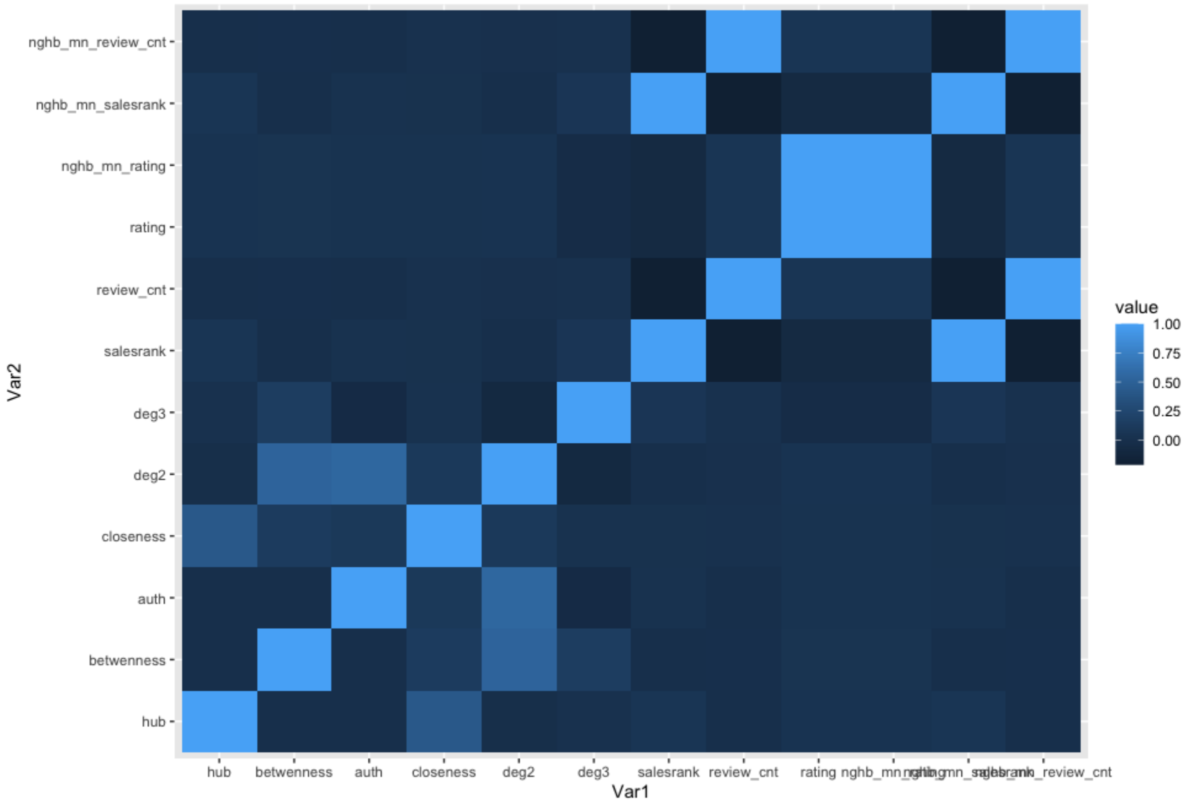


Authorities



8. Include the variables (taking logs where necessary) created in Parts 2-6 above into the "products" information and fit a Poisson regression to predict *salesrank* of all the books in this subcomponent using products' own information and their neighbor's information. Provide an interpretation of your results.

When examining the model results all of the variables seem to be significant. We noticed that running a model with both neighbor statistics and the non-neighbor statistic (ie ngbh_product_reviews and product_reviews) that we were running into issues relating to colinearity which we resolved by running two models, one with the original variables sans neighbor variables and another model with the neighbor variables sans "original" values. Returning to our original point, when looking at model statistics associated with our review count, betweenness, and closeness variables, all of these variables have some of the most impactful numbers which we think makes sense. Logically speaking, a book with a higher rating coefficient will have a lower sales rank value as a higher rated book with more reviews will also likely sell at a higher rate. When we look to buy books we want to not waste money by buying a product we will not like so we use reviews to try and get an idea of whether the book in question aligns with what we want. Other people think similarly, thus it makes sense that a book with a higher number of reviews will be more likely to be bought and will thus have a lower sales rank. Betweenness and closeness also make sense as important factors for determining sales rank as well because people tend to search for books using comparisons to other books they have read. Searches following this pattern tend to feature search results of books that are similar in some way to the book being used as a touchstone (which is reliant on the closeness score of a book compared to its neighbors) and when people buy a book that is part of a series they will buy other books in the series as well thus we can expect to see the sales rank for these books increase as well.



(in the heatmap above we can see the correlations between the different variables we created, note the high/perfect correlation between nghb variables and their non nghb counterparts).