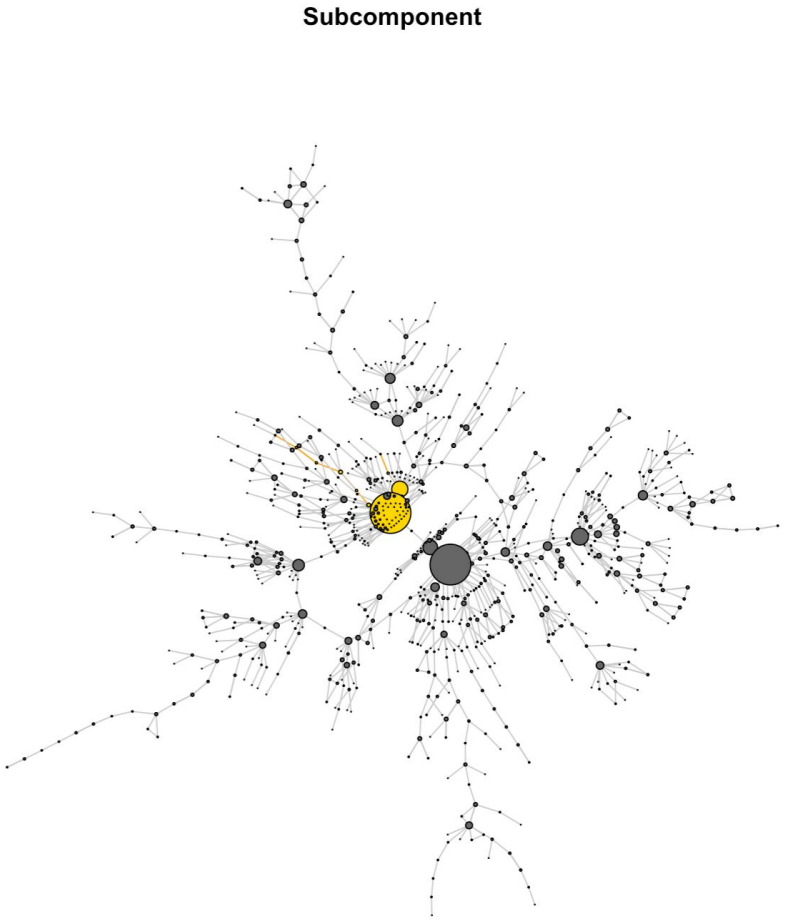
Poonam Ahuja, Sammi Chang, Isha Chaudhari, Sujaya Darke, Song Han

Cust & Social BANA277 Social Network Group Assignment

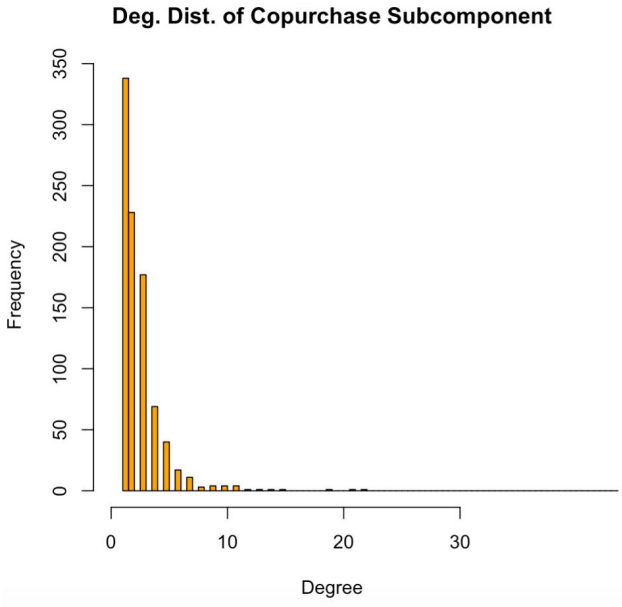
For question 1-3, please refer to the attached R file.

4. There were two nodes (33 and 4429) with 53 degrees, which was the highest number of degrees, and we chose node 33 to do our analysis.

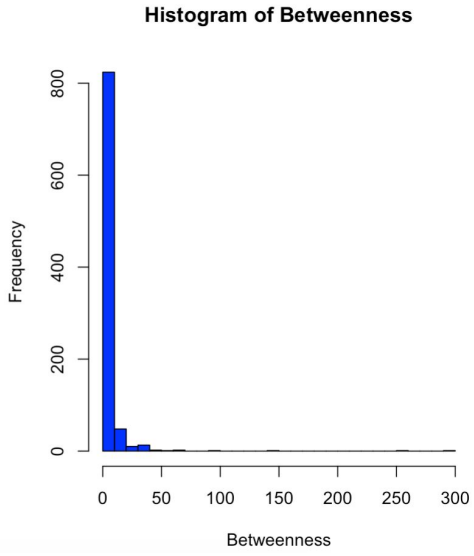
5. Interpretation of Subcomponent Graph: This is the subcomponent graph for book id = 33 with 904 vertices. The nodes are coded based on degree size: if there are more connections to the node (if it has more degrees), the node will be bigger. If the node is less connected and thus has less connections to other vertices, it is smaller. There are many nodes that are closer to the edge and thus are less connected to the network. Interestingly, the graph may show that most books are separated into two groups. This suggests that, among books connected to book id = 33, there are two clusters, and these clusters are only connected by a few nodes. There are two nodes that are the same size, and this agrees with our computation that there are two nodes in the raw data with max degree=53. The diameter is the longest path that traverses the graph between two nodes. The diameter has a length of 9, and the nodes are: 37895, 27936, 21584, 10889, 11080, 14111, 4429, 2501, 3588, 6676. Here, the diameter is yellow.



6. Interpretation of various statistics about this network: The **degree distribution** is quite clustered around 1-3 degrees. Most nodes have a degree of less than 10, and there are a few outliers that are greater than 15. **Density** is the proportion of realized edges over all potential edges in the network. The **edge density** is 0.0014369, which is small, so our network is pretty connected and considered quite dense. **Centrality** refers to centrality of the nodes in the network; since 53 is the highest degree, this is also the highest centrality. **Closeness** is defined as the inverse of the total distance of a node v to all others. Here, most values of closeness are small, so the nodes are pretty close together.



Betweenness measures the strength of a brokerage or bridge. It roughly measures the number of shortest paths going through a vertex. In our subgraph, we see that most nodes have a betweenness of 0, or less than 10, and there are a few “brokers” with a higher score that help to connect the graph further. This is visualized in the histogram below.



Node 33 has an **authority** score of 1 and a **hub** score of 0. This shows that node 33 is one of the top two influential nodes partially from having a high number of incoming links which we can see from the authority score. The hub and authority scores are at two extremes which we can see in the table below.

id	salesrank	review_cnt	downloads	rating	hub	betweenness	authority	closeness
33	97166	4	4	5.0	0.000000e+00	0.0	1.000000e+00	1.612383e-04

For question 7, please refer to the attached R file.

8. Interpretation of Poisson regression: Closeness is the least significant variable with a p-value of .0231, but it is still significant under an alpha level of 0.05. All other variables are significant at an alpha level of 0.0001, so this model predicts sales rank well. Median residuals are low, at -7.61, so the model explains variation in the data well.

Note that a lower sales rank means higher sales for the book, so a low y value is most advantageous. Thus, to find the variables that have a positive effect for book sales, we distinguish coefficients that have negative signs, namely, nghb_mn_salesrank, review_cnt, rating, betweenness and closeness. As these variables climb, sales rank decreases. On the other hand, high values in variables: in_degree, out_degree, downloads, authority_score, hub_score, nghb_mean_review_cnt and nghb_mean_salesrank are disadvantageous to the book's sales rank. As these variables increase, sales rank also increases. For example, as review count increases by 1, sales rank decreases by -.02868 when all other variables remain constant, showing that having more reviews improves the sales of a book.

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-363.25  -160.45    -7.61   122.01   519.58

Coefficients:
                    Estimate Std. Error   z value Pr(>|z|)
(Intercept)         1.119e+01  1.108e-03 10096.697   <2e-16 ***
review_cnt         -2.868e-02  1.877e-04  -152.749   <2e-16 ***
rating             -7.061e-03  1.098e-04   -64.314   <2e-16 ***
hub                 2.452e-01  8.593e-04   285.400   <2e-16 ***
betweenness        -7.349e-04  1.111e-05   -66.157   <2e-16 ***
downloads           2.457e-02  1.879e-04   130.759   <2e-16 ***
authority           1.895e-01  4.754e-03    39.861   <2e-16 ***
closeness          -1.789e+01  7.874e+00    -2.272   0.0231 *
in_degree           2.801e-03  6.819e-05    41.069   <2e-16 ***
out_degree          5.646e-02  2.057e-04   274.476   <2e-16 ***
nghb_mn_salesrank   2.057e-07  4.498e-09    45.733   <2e-16 ***
nghb_mn_review_cnt  7.386e-04  1.969e-06   375.165   <2e-16 ***
nghb_mn_rating     -9.723e-03  1.253e-04   -77.613   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 16968896  on 517  degrees of freedom
Residual deviance: 15315200  on 505  degrees of freedom
```