Winter 2021 MSBA 277

Team members : Cloris He, Joy Chen, Misha Khan, Ping-Chun Liu, Jialu Li

Social Network Team Assignment

1. Delete products that are not books from "*products*" and "*copurchase*" files. And then delete the books with salesrank>150,000 or salesrank = -1.

2. Create a variable named *in-degree*, to show how many "Source" products people who buy "Target" products buy; i.e. how many edges are *to* the focal product in "co-purchase" network.

```
> in_degree <- degree(g, mode='in')
> length(in_degree)
[1] 20684
```

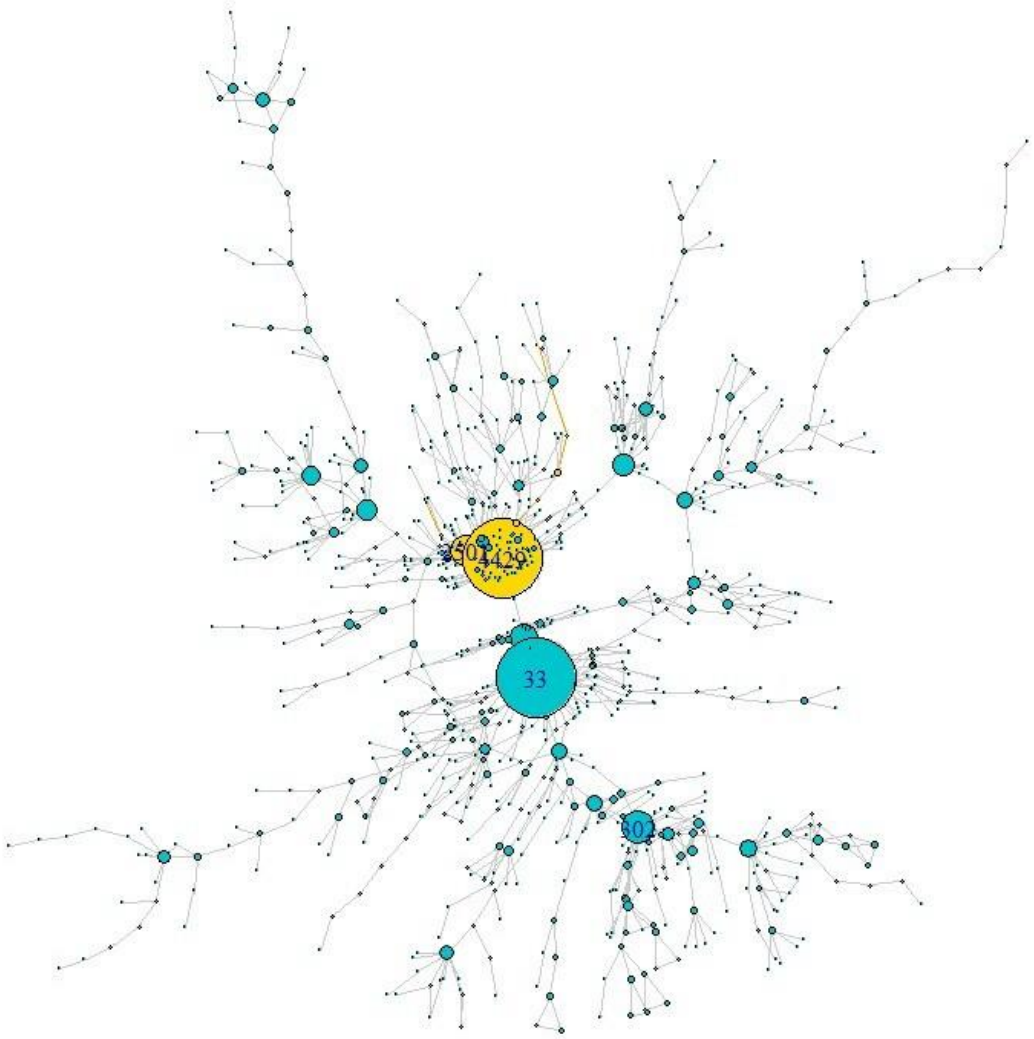
3. Create a variable named *out-degree*, to show how many "Target" products people who buy "Source" products also buy; i.e., how many edges are *from* the focal product in "co-purchase" network.

```
> out_degree <-degree(g, mode='out')
> length(out_degree)
[1] 20684
```

4. Pick up one of the products (in case there are multiple) with highest *degree* (*in-degree* + *out-degree*), and find its *subcomponent,* i.e., all the products that are connected to this focal product. From this point on, you will work only on this subcomponent.

We found out there are 2 products with the highest degree of 53: product 33 (Double Jeopardy (T*Witches, 6)) & product 4429 (Harley-Davidson Panheads, 1948-1965/M418) and we picked the product 33 (Double Jeopardy (T*Witches, 6)) for the following analysis based on its 904 subcomponents.

5. Visualize the subcomponent using iGraph, trying out different colors, node and edge sizes and layouts, so that the result is most appealing. Find the diameter, and color the nodes along the diameter. Provide your insights from the visualizations.



The social graph shows two main groups, one with Id 33 in the center: Double Jeopardy (T*Witches, 6), another one represented with Id 4429 in the center: Harley-Davidson Panheads, 1948-1965/M418.

Between 4429 and 33, there is a local bridge. It ties between two groups in a social graph that are the shortest route by which information might travel from those connected to one to those connected to the other. If the local bridge is removed, the distance between these two groups will increase. Also, the lack of the local bridge will significantly reduce the probability of co-purchasing behavior between the two groups and the frequency of products being bought.

The diameter, shown in yellow in the graph, is the longest path we can find among all the shortest distances of the vertices. The diameter is 9 and the nodes within this path are Id 37895, 27936, 21584, 10889, 11080, 14111, 4429, 2501, 3588, and 6676.

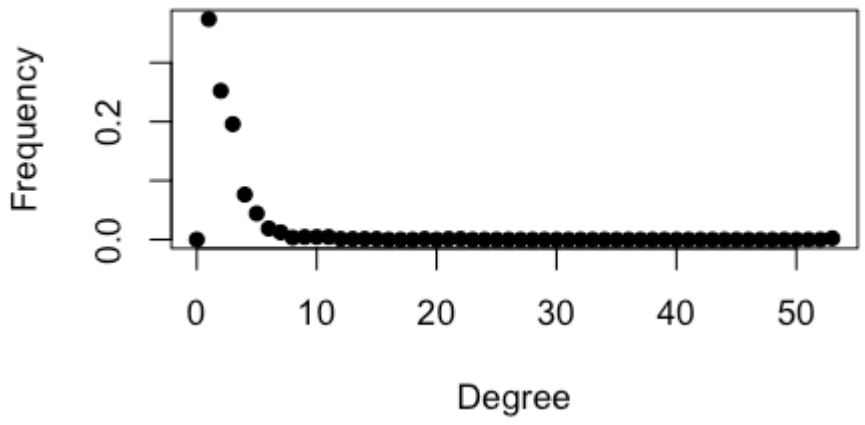
The size of the bubble is determined by how many connections they have with the other nodes. The larger the bubble, the more nodes link to it. And thus, from the graph, we can see that Id 33 and 4429 are the two biggest nodes that have the most connections. The smaller nodes spread on the edges of the network indicate fewer connections. The number of connections between nodes show how strong the relationship between the nodes are. The nodes clustered in the middle of the graph have a stronger relationship while the nodes that are spread around the border with long ties show a weaker relationship. And those products with long ties which only have 1-2 edges can be easily separated from the whole network.

6. Compute various statistics about this network (i.e., subcomponent), including degree distribution, density, and centrality (degree centrality, closeness centrality and between centrality), hub/authority scores, etc. Interpret your results.

- Degree Distribution

  Degree distribution is the probability distribution of these degrees over the whole network. In our subcomponent of product id 33, there are 904 nodes and more than 800 nodes have degrees less than 5. This shows that most products are only related to equal to or less than 5 other products.

  - 37.3% of all nodes have 1 edge.
  - 25.2% of all nodes have 2 edges.
  - 19.6% of all nodes have 3 edges.
  - 7.6% of all nodes have 4 edges.
  - 4.4% of all nodes have 5 edges.
  - And the above accounts for 94.1% of all nodes.

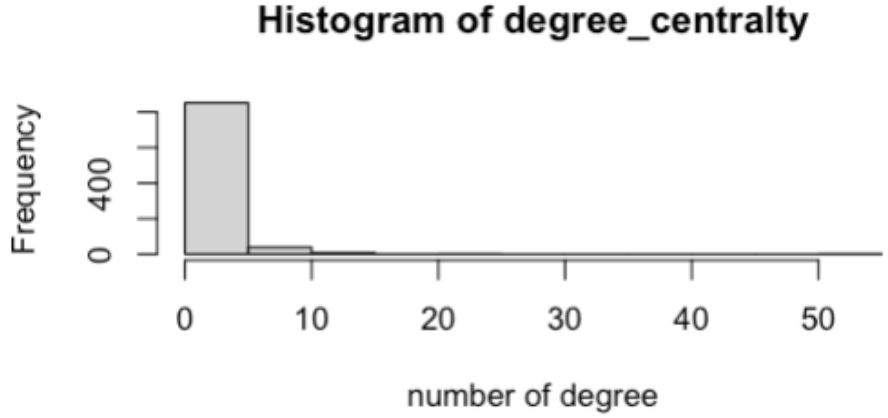


- Density

  Density measures the size of the network and gives us the probability that any edge exists given that it could exist. The edge density is 0.001436951. Because the value is fairly small, we can conclude that the network is dense.

- Degree

  The degree of a node in a network is the number of connections it has to other nodes. As mentioned in the degree distribution, over 94% of all nodes equal to or less than 5 edges.



- Closeness

  Closeness centrality shows how close a product is to all other products. In the co-purchase network, a product with higher closeness centrality is more likely to stimulate the purchase of other products quickly. Product id 33 has the highest closeness score in the subcomponent network, indicating that this product has the shortest overall distance to all other products in the network.

- Betweenness

  Betweenness centrality refers to the number of times that one product acts as the shortest bridge between the other two products. The more times a product acts as a "betweenness", the more centrality it becomes. Product id 2501 has the highest betweenness which means this product influences the flow around a whole system.

- Hub Scores

  Hub scores represent a product that points to other products which are the outgoing links. We found out that product id 195144 has the highest hub score which shows that it’s the product with most outgoing links.

- Authority Scores

  Authority scores represent a product that is linked by other products which are the incoming links. We found out that product id 33 has the highest authority score which means that this product has the most incoming links. And the following products have an

authority score of 0 which shows that they don't have any incoming links and no other products are pointing to them.

```
> auth[which(auth==0)]
   626   2423   2501   4429   7325   7544   8439  14950  18771  26080  26154  43813  69374 131572
     0      0      0      0      0      0      0      0      0      0      0      0      0      0
135616 148505 150979 159473 195420 213156 256680   6676   2563   2145   6096  21153  14827  25643
     0      0      0      0      0      0      0      0      0      0      0      0      0      0
 38866  66939  55978  12184  69915  56052  84076  37738
     0      0      0      0      0      0      0      0
```

7. Create a group of variables containing the information of neighbors that "point to" focal products. The variables include:

a. Neighbors' mean rating (nghb_mn_rating),

b. Neighbors' mean salesrank (nghb_mn_salesrank),

c. Neighbors' mean number of reviews (nghb_mn_review_cnt),

Refer to the code to see the full dataframe.

	id	nghb_mn_rating	nghb_mn_salesrank	nghb_mn_review_cnt
1	33	4.103774	82153.26	21.075472
2	77	4.666667	41744.00	4.000000
3	78	4.500000	73179.00	157.818182
4	130	4.500000	19415.00	6.000000
5	148	0.000000	46701.00	0.000000
6	187	4.500000	133546.67	3.666667
7	193	4.050000	59470.60	75.700000
8	224	3.250000	79068.00	167.500000
9	302	3.750000	73671.86	16.409091
10	321	4.750000	77093.50	3.500000
11	322	4.666667	93553.67	4.000000
12	422	4.000000	43866.50	8.000000
13	448	2.625000	64688.75	14.625000

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.

```
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 ***
downloads          2.457e-02  1.879e-04   130.759   <2e-16 ***
rating            -7.061e-03  1.098e-04   -64.314   <2e-16 ***
closeness         -1.789e+01  7.874e+00    -2.272   0.0231 *
betweenness       -7.349e-04  1.111e-05   -66.157   <2e-16 ***
hub_score1         2.452e-01  8.593e-04   285.400   <2e-16 ***
authority_score1   1.895e-01  4.754e-03    39.861   <2e-16 ***
in_degree_sub      2.801e-03  6.819e-05    41.069   <2e-16 ***
out_degree_sub     5.646e-02  2.057e-04   274.476   <2e-16 ***
nghb_mn_rating    -9.723e-03  1.253e-04   -77.613   <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 ***
---
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
  (386 observations deleted due to missingness)
AIC: 15321778
```

From the above chart, we can see that all the variables' p-values are less than alpha (assuming alpha = 0.05). Therefore, we can conclude that all the variables are significant to our model. Additionally, since the Poisson Regression coefficients are given on log scale, we need to convert them back in order to interpret appropriately.

Salesrank = exp (1.119e+01- 2.868e-02*review_cnt + 2.457e-02*downloads -7.061e-03*rating - 1.789e+01*closeness - 7.349e-04*betweenness + 2.452e-01*hub_score1 + 1.895e-01*authority_score1 + 2.801e-03*in_degree_sub + 5.646e-02*out_degree_sub -9.723e-03*nghb_mn_rating + 2.057e-07*nghb_mn_salesrank +7.386e-04*nghb_mn_review_cnt)

Meaning that:

Increasing of 1 unit of	**This will happen***
review_cnt	Salesrank will decrease by 0.9717321 units

downloads	Salesrank will increase by 1.024869 units
rating	Salesrank will decrease by 0.9929642 units
closeness	Salesrank will decrease by 1.704536 × 10^-8 units
betweenness	Salesrank will decrease by 0.9992654 units
hub_score1	Salesrank will increase by 1.277931 units
authority_score1	Salesrank will increase by 1.208656 units
in_degree_sub	Salesrank will increase by 1.002805 units
out_degree_sub	Salesrank will increase by 1.058080 units
nghb_mn_rating	Salesrank will decrease by 0.9903241 units
nghb_mn_salesrank	Salesrank will increase by 1 units
nghb_mn_review_cnt	Salesrank will increase by 1.000739 units

*The exponentiated coefficients are shown in the table below

Salesrank represents the rankings of book sales and therefore, the lower the rank number, the better the sales. With the increasing of downloads, hub_score1, authority_score1, in_degree_sub, out_degree_sub, nghb_mn_salesrank, nghb_mn_review_cnt, Salesrank will also increase but it means less sales of the books. And the increase of Review_cnt, rating, closeness, betweenness, nghb_mn_rating will lead to decrease of Salesrank which means more sales of the books and thus, potentially more revenue to the company.

	exp(poisson$coefficients)
(Intercept)	7.254152e+04
review_cnt	9.717321e-01
downloads	1.024869e+00
rating	9.929642e-01
closeness	1.704536e-08
betweenness	9.992654e-01
hub_score1	1.277931e+00
authority_score1	1.208656e+00
in_degree_sub	1.002805e+00
out_degree_sub	1.058080e+00
nghb_mn_rating	9.903241e-01
nghb_mn_salesrank	1.000000e+00
nghb_mn_review_cnt	1.000739e+00