Poisson regression is run to predict salesrank of all the books based on predictors such as review count, rating, mean_rating (Neighbors' mean rating), mean_reviewcount (Neighbors' mean review count), mean_salesrank (Neighbors' mean salesrank), outdegr (out-going interactions of the focal product) and indegr(specific product as the point of focus).

Generalized Linear Model Regression Results

Dep. Variable:	salesrank	No. Observations:	171586
Model:	GLM	Df Residuals:	171578
Model Family:	Poisson	Df Model:	7
Link Function:	log	Scale:	1.0
Method:	IRLS	Log-Likelihood:	-3.4634e+10
Date:	Tue, 27 Feb 2018	Deviance:	6.9266e+10
Time:	15:28:15	Pearson chi2:	3.68e+13
No. Iterations:	8		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	13.8219	1.48e-05	9.31e+05	0.000	13.822	13.822
review_cnt	-0.0031	1.47e-07	-2.14e+04	0.000	-0.003	-0.003
rating	-0.1661	1.42e-06	-1.17e+05	0.000	-0.166	-0.166
mean_rating	-0.0009	2.29e-06	-371.804	0.000	-0.001	-0.001
mean_reviewcount	-2.909e-05	5.9e-08	-493.105	0.000	-2.92e-05	-2.9e-05
mean_salesrank	3.411e-10	7.81e-12	43.701	0.000	3.26e-10	3.56e-10
outdegr	0.0017	2.85e-06	587.450	0.000	0.002	0.002
indegr	0.0002	7.24e-07	319.661	0.000	0.000	0.000

- The coefficient for review_cnt is -0.0031. For a one-unit increase in review_cnt, we expect a 0.0031 decrease in the log-odds of salesrank, holding all other independent variables constant
- The coefficient for rating is -0.1161. For a one-unit increase in rating, we expect a 0.1161 decrease in the log-odds of salesrank, holding all other independent variables constant
- The coefficient for mean_ rating is -0.0009. For a one-unit increase in mean_ rating, we expect a 0.0009 decrease in the log-odds of salesrank, holding all other independent variables constant
- The coefficient for mean_reviewcount is -0.00002909. For a one-unit increase in mean_reviewcount, we expect a 0.00002909 decrease in the log-odds of salesrank, holding all other independent variables constant
- The coefficient for mean_salesrank 0.0000000003411. For a one-unit increase in mean_salesrank, we expect a 0.0000000003411 increase in the log-odds of salesrank, holding all other independent variables constant
- The coefficient for outdegr is 0.0017. For a one-unit increase in outdegr, we expect a 0.0031 increase in the log-odds of salesrank, holding all other independent variables constant
- The coefficient for indegr is 0.0002. For a one-unit increase in indegr, we expect a 0.0031 increase in the log-odds of salesrank, holding all other independent variables constant

Density of the network= 0.0000184 * 100%
= 0.00184%

```
edges = copurchase.shape[0]
nodes = len(set(np.append(copurchase.Source,copurchase.Target)))
density = edges/(nodes*(nodes-1))
print (density)
```

```
1.841229067865028e-05
```

Very low density indicates the connections in the network is unsubstantial.

```
print (np.exp(model.params)-1)
```

```
Intercept           1.006406e+06
review_cnt         -3.142164e-03
rating             -1.530755e-01
mean_rating        -8.500934e-04
mean_reviewcount   -2.909107e-05
mean_salesrank      3.411031e-10
outdegr             1.675324e-03
indegr              2.313136e-04
dtype: float64
```

Insights:

It is to be noted that, **"Lower salesrank means higher sales"**

1. Increase in number of reviews, lead to decrease in salesrank (0.314%). Therefore, products with more number of reviews will have higher sales, and hence book publishers in Amazon must improve their review count.

2. Increase in product rating, lead to decrease in salesrank (15.3%). Therefore, products with higher rating will have higher sales, and hence book publishers in Amazon must improve their product ratings in order to have higher sales.

3. Increase in mean neighbor rating, lead to decrease in salesrank (0.085%), translating to higher sales. Therefore, a network of products with higher average ratings has higher sales compared to network with lower average ratings even though it is occurring at negligible scale (0.085%).

4. Increase in mean neighbor review count, lead to decrease in salesrank (0.0029%), translating to higher sales. Therefore, a network of products with higher average review count has higher sales compared to network with lower average ratings even though it is occurring at negligible scale (0.0029%).

5. Out-degree represents the number of times the focal product interacts with others (or, in this case, how many "Target" products, people who buy "Source" product also buy). The increase in out-degree of a product causes the salesrank to increase by (0.167%) which is decreasing sales for the product. In a same network, the book will be of similar relevance and same category, thereby reducing its significance for customer as there will be many options to choose from, converting to lower number of sales (0.167%), however small it may be.

6. Behavior of all products in the network is based on their relation to the focal point of the "in-degree" product (or, in this case, how many "Source" products, people who buy "Target" products buy). The increase in in-degree for a product causes the salesrank to increase by (0.02313%) which is decreasing sales for the product by a paltry amount and can be overlooked.

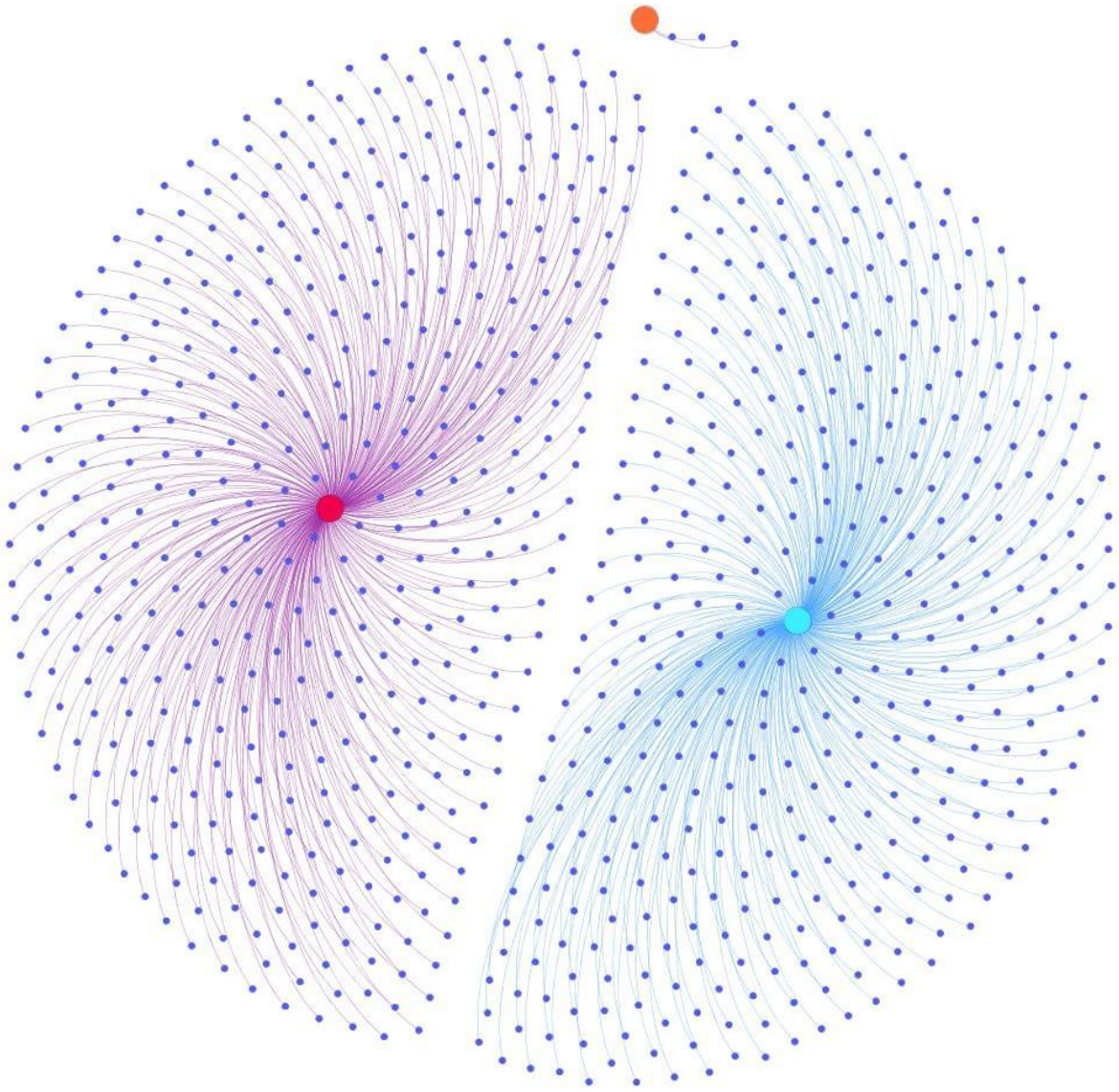
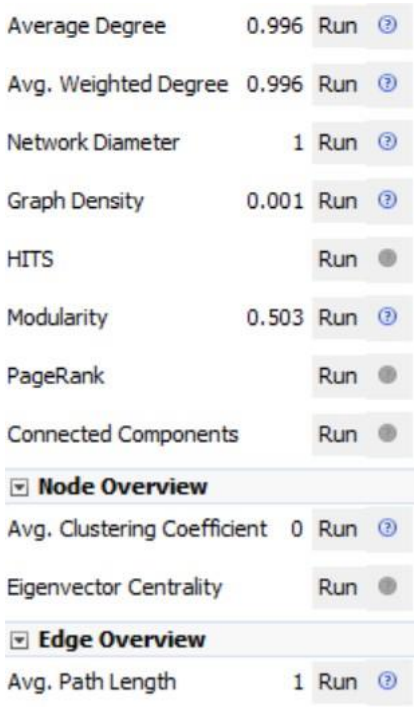


Fig: GEPHI Network Visualization

(i) the one with the highest in-degree
(ii) the one with the highest out-degree
(iii) the one with the highest degree (in-degree + out-degree)



1. Graph Density = 0.001;
   From the statistic, as the value is closer to zero the graph is sparse as observed and therefore a majority of available connections have not been completed as expected, which means a weak network.
2. Average Degree: 0.996;
   The degree of a node is the number of relation (edge) it has. The bigger the better. From the statistic, it can be interpreted that it has fair number of relations (edges).
3. Network Diameter = 1, Average Path Length = 1;
   which is a measure of the distance between the two most distant nodes in this network. If we have two networks with same network diameter, the one with the higher number of nodes is efficient. From the statistic, it can be interpreted that for one connection to reach the other, it is required to pass along the big node.
4. Modularity = 0.503;
   which is a measure of the strength of division of a network into groups or clusters. A high modularity score indicates sophisticated internal structure. From the network statistic, it can be interpreted that the network has intermediate internal structure in complexity.