Amazon Product Co-Purchased Network

Explore the characteristics of the network and discover the most influential determinants in co-purchased item sales

Name: Xintian Shen, Ruixi Wang, Xinyi Chen, Xin Zhao

Course Code: BIA 658

CONTENT

Introduction

The emerge and mature of e-commerce brings obvious and giant changes to our daily lives – in the ways we work, we live and spend - changing our purchase behavior and revealing co-purchased product that are profoundly connected. What's less clear, however, is whether these new behaviors can be influenced or even dominated by certain marketing strategies of the brand or recommended system of website.

To assist platform and consulting company to optimize the recommended system and marketing strategies on co-purchase sales, this paper will mainly focus on exploring the characteristics of the e-consumer co-purchase behavior and discovering the most influential determinants in co-purchased item sales. We decide to leverage network analysis on products from one representative platform of a certain period. Amazon, the most successful and biggest e-commerce platform, comes into our mind. Its varied products and large order provide us sufficient and economically valuable data. We hereby collect Amazon product co-purchasing network metadata - collected in summer 2006 - from Stanford education as our dataset. The original dataset includes over five hundred thousand different products from four categories. After initial data preparation and visualization analysis, the metadata network looks quite sparse. Therefore, we attempt to scope the scaled-down network which includes sales star and their neighbors only.

Overall, our study shows that tags, categories influence customer co-purchase behavior and closeness centrality has negative correlation with sales. Despite of the unavoidable shortages such as sampling error and too limited categories, main conclusions we conclude here have huge potentials on the recommended system design and the product page design.

Dataset Explanation

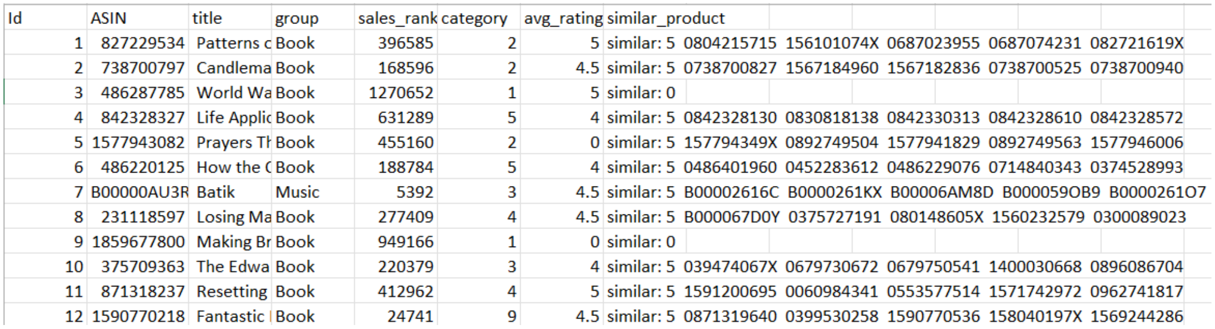
1. Amazon Product Metadata

The original dataset we chose is Amazon product co-purchasing network metadata and reviews from summer 2006 (Stanford). The data was collected by crawling Amazon website and contains product metadata and review information about 548,552 different products (Books, music CDs, DVDs and VHS video tapes).[1] For each product the following information of five dimensions is available:

1) Id: A series of numbers that help track each product in a database.
2) ASIN: ASIN stands for Amazon Standard Identifier number. Each product's co-purchased product is also shown in this format.
3) Title: The name of the product.
4) Group: The dataset has 4 product groups: Books, DVDs, videos, and Music.
5) Sales rank: The arrangement of items' sale in each category in ascending order.
6) List of similar products: the product that gets co-purchased with the current product. The maximum number of co-purchased products is 5, and it is stored in ASIN format.
7) Detailed product categorization: the specific subject name and corresponding id of each product's detailed category.
8) Product reviews: It includes time the customer makes review, the customer Id, the rating, the number of votes, the number of people that found the review helpful.
9) Average rating: The average rating of customers reviews.

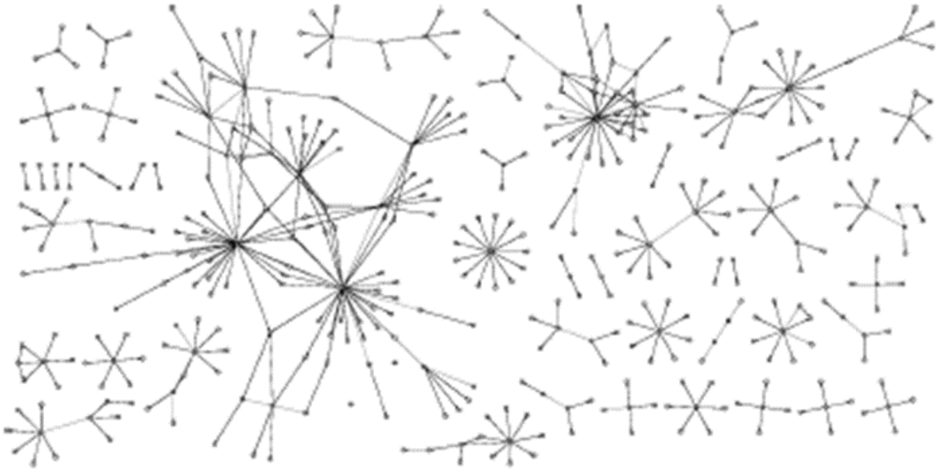
2. Data Preparation

The dataset we collect is initially in text format. To start with data preprocessing, we use regular expressions to extract information that is explained above except for customer reviews. We want to primarily focus on the product and its co-purchased products. Accordingly, we use average ratings to roughly represent customers' reviews. The extracted information is stored in dataframe and CSV format. The table below is the quick view of our data:



Id	ASIN	title	group	sales_rank	category	avg_rating	similar_product					
1	827229534	Patterns c	Book	396585	2	5	similar: 5	0804215715	156101074X	0687023955	0687074231	082721619X
2	738700797	Candlema	Book	168596	2	4.5	similar: 5	0738700827	1567184960	1567182836	0738700525	0738700940
3	486287785	World Wa	Book	1270652	1	5	similar: 0					
4	842328327	Life Applic	Book	631289	5	4	similar: 5	0842328130	0830818138	0842330313	0842328610	0842328572
5	1577943082	Prayers Th	Book	455160	2	0	similar: 5	157794349X	0892749504	1577941829	0892749563	1577946006
6	486220125	How the C	Book	188784	5	4	similar: 5	0486401960	0452283612	0486229076	0714840343	0374528993
7	B00000AU3R	Batik	Music	5392	3	4.5	similar: 5	B00002616C	B0000261KX	B00006AM8D	B000059OB9	B0000261O7
8	231118597	Losing Ma	Book	277409	4	4.5	similar: 5	B000067D0Y	0375727191	080148605X	1560232579	0300089023
9	1859677800	Making Br	Book	949166	1	0	similar: 0					
10	375709363	The Edwa	Book	220379	3	4	similar: 5	039474067X	0679730672	0679750541	1400030668	0896086704
11	871318237	Resetting	Book	412962	4	5	similar: 5	1591200695	0060984341	0553577514	1571742972	0962741817
12	1590770218	Fantastic	Book	24741	9	4.5	similar: 5	0871319640	0399530258	1590770536	158040197X	1569244286

The next step is to split the dataset into two files, one includes ASIN, title, group, sales rank, category, average rating as product attributes. We use ASIN as node Id. Another file includes the source column and target column as an edge list. The source column includes ASIN for all products in attributes file, and the target column is the products' co-purchased products. I split the similar product column into product and its co-purchased product pair. Thus, we can generate an amazon co-purchase network with edge list and attributes. The full network is shown below:



Graph source: The Dynamics of Viral Marketing

The full network is enormous and sparse. Accordingly, it caused R to fail loading attributes. What’s more, it is difficult to analyze a very sparse network. The transitivity score, centrality, and other network analyzing tools are extremely low and so these measurements lose their function to help analyze the network. Therefore, our team decides to shrink the entire data population to a smaller one for analyzing purposes. Firstly, we have decided to exclude products that degree below 100 and keep products degree above 100 and their related products. Now that we only have attributes for products in the source column, we further filter our dataset to only keep source products that are also in the target column. The benefit of this is that we can perform network analysis with a smaller and concentrated network. The downside of this is that we are taking a sampling risk for our result. It is because we are taking samples to evaluate the entire data population. The difference between sample and population can lead to a false result of a model.

3. Variable Explanations

The information of each node includes target node, category, and sales rank. Target node aimed to create a link between the source product and the target product. Category divides each product into one of four groups in Books, music CDs, DVDs and VHS video tapes.

Co-purchased Product Network Analysis

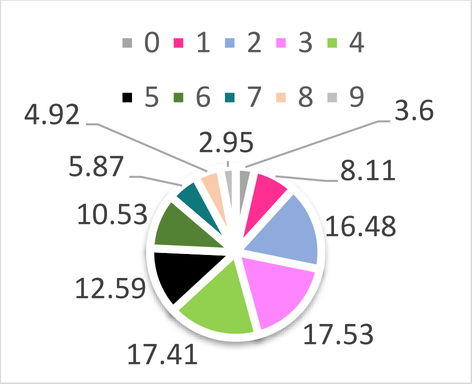
1. Network Analysis

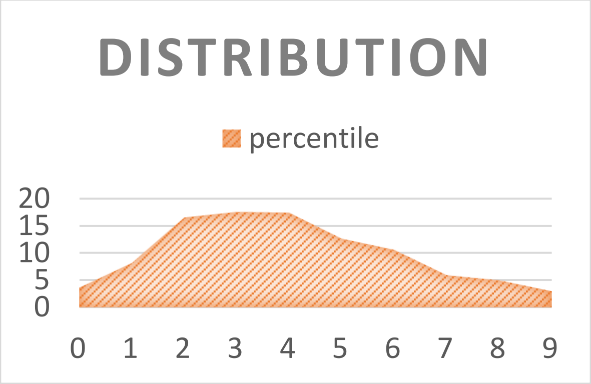
The selected undirected network includes all links and related nodes of source products with over 100 co-purchased items. This scoped network has 15,706 nodes and 38,291 edges. The node here is amazon product, and the edge is the amazon co-purchased product-project link: That is, if any customer purchases any product in the dataset statistics and meanwhile co-purchases another product in the same order, there will be an edge connecting two products.

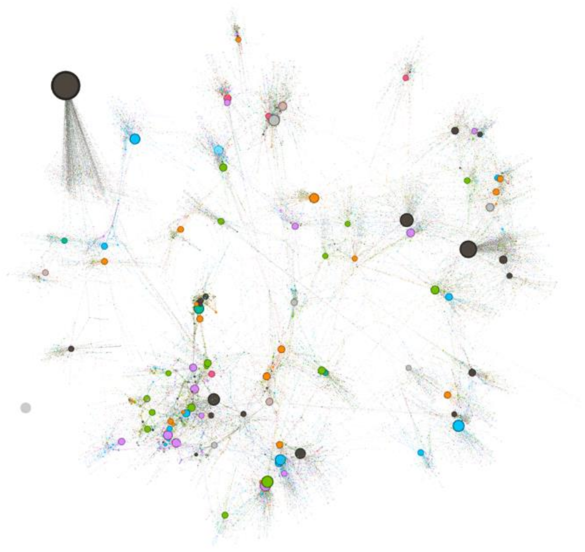
Graph density	Transitivity	Centralization Degree	Centralization Betweenness	Centralization Evcent	Mean Distance
0.0003104725	0.05244823	0.03464655	0.2528876	0.9981292	8.525939
Quite Sparse	Not bad	Not bad	Bridge-like nodes	Centralized	Local hub, No global hub

The above fact table shows the statistical score of the network under different measurements. The 0.0003 of graph density score indicates this network is still quite sparse. The transitivity and centralization degree score are relatively not bad if we consider the graph density, and we cannot conclude if the network is clustered based on those three dimensions. The centralization betweenness score of 0.2528876, indicates that there must be bridge-like nodes to connect different communities. The reason why centralization eigenvector score is almost 1 is that the network is formed by some influential nodes (source products with over 100 co-purchased items) together with its neighbors. In other word, most nodes know influential neighbor. The network are formed by many communities. So, the network is formed to be highly centralized based on the eigenvector dimension. In general, from the betweenness and eigenvector score we know the network is centralized which is because of the hubs we selected. Combine those three results we can conclude that due to the powerful bridge-like and centralized nodes, this network is quite likely to be a small-world. The mean distance of this network, however, is quite long as over 8, which indicates that even though the network is formed by the influential nodes and their neighbors within each community, there is no "super star" node that can connect most of other nodes.

Then we use Gephi to and explore the characteristics of the scoped network in three different ways. We also leverage pie chart and distribution chart to do basic analysis. We firstly color by number of category labels and size by node degree. Those charts tell that product with two to six labels are most likely to be co-purchased in our network. That makes sense because 0 or only 1 label makes customers hard to search for certain item by category label, while labels more than 6 usually describes a more general item and there are more competitors and some of those competitors are more likely to be more specific preferred items for the customer. Most of interesting finding here is that even though black nodes count for 12.59% of the network, almost all high degree nodes are black. To sum up, this visualization concludes that the items with five labels are most frequently co-purchased, and labels are not as more as better-ideal interval is two to six.

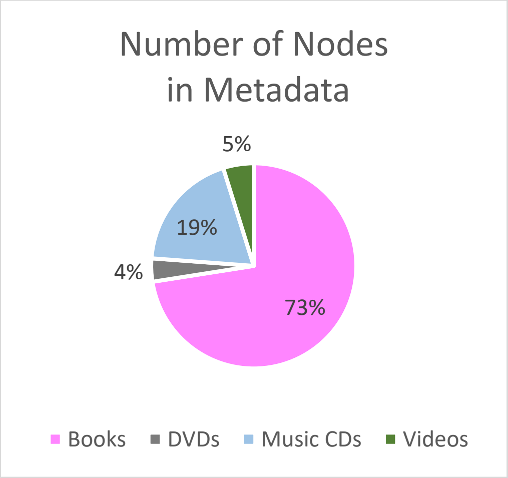


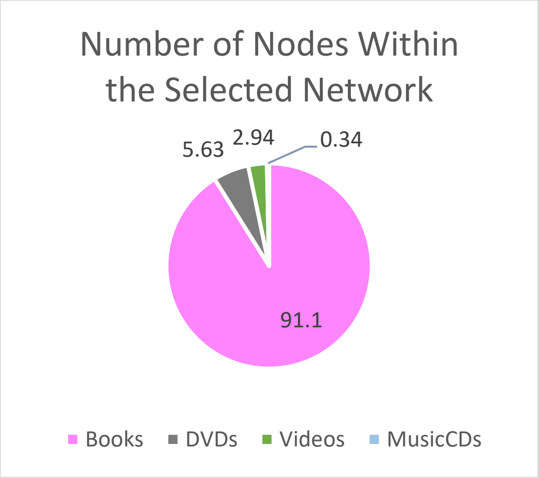


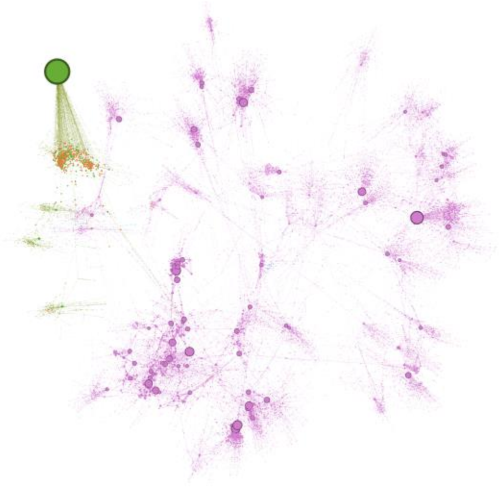


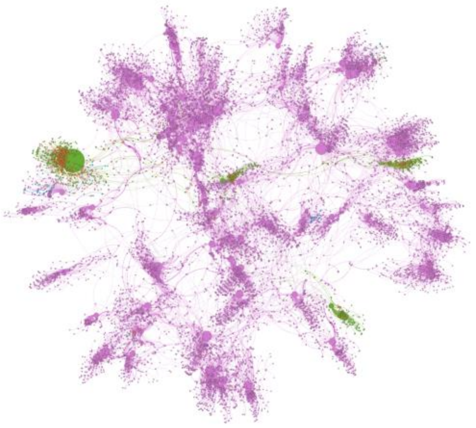
Secondly, we color by group and size by eigenvector value. And finally we color by group and size by degree. Since those two visualizations share the similar patterns. We would introduce them together.

In those two graphs, 91.1% nodes here are purple (Books) while in the original dataset books count for 73% only. That’s not only because most high degree nodes are purple but also because purple nodes tend to connect purple nodes only. Therefore, the first two findings in this visualization are that Books are most popular categories that customers tend to co-purchase and Amazon book customers tends to co-purchase products within same category. Then we realize that Music CDs counts for 19% in metadata but now only 0.34% in the scoped network. There're also no visible big blue nodes. Hereby we conclude that the least popular co-purchased category is Music CDs. DVDs and Videos in our network count for similar percentile as the dataset. The reason is that there is a community of the highest degree node: the green one (DVD Laura), it connects not only other DVDs but also Videos. This community contributes the DVDs and Videos popularity. So further we'll introduce the 1.0-degree ego-centric network to analyze the characteristics of this community has that brings it so successful. There is a pattern that the orange nodes connect not only orange but also green nodes. That is, customers who buy DVDs prefers to co-purchase not only other DVDs but also Video products. However, they barely co-purchase Music CDs and Books. So the last finding here is DVDs customers also tend to become Video customers.





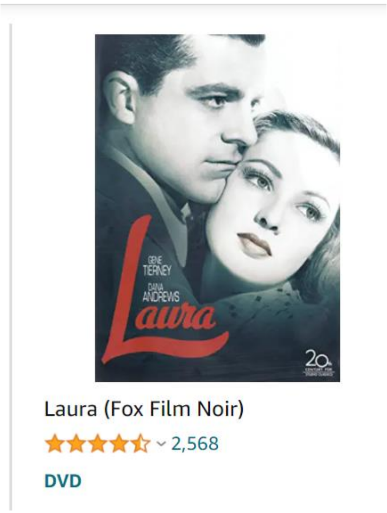




To sum up, the four interesting findings in the scoped network are:

1) Books are most popular categories that customers tend to co- purchase
2) Amazon book customers tends to co-purchase products within same category
3) The least popular co-purchased category is Music CDs
4) DVDs customers also tend to become Video customers

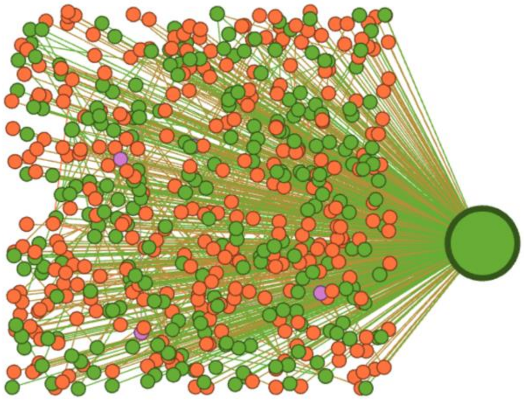
2. Highest Degree Node Analysis



Laura (Fox Film Noir)

★★★★☆ ⌄ 2,568

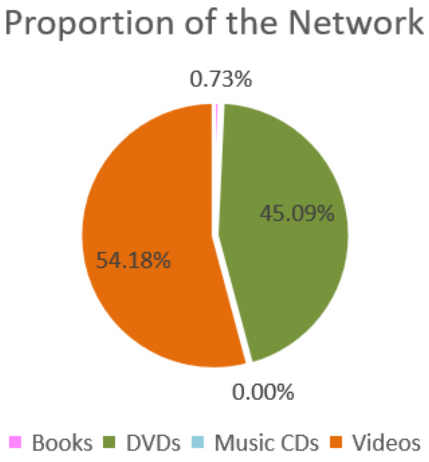
DVD



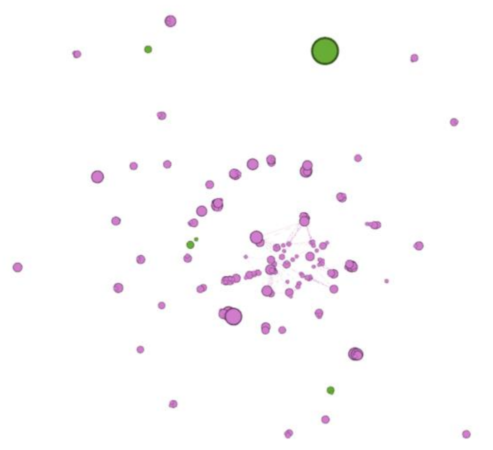
Then we focus on the DVD product Laura and its neighbors, the community of the highest degree in the selected network. This community contributes the DVDs and Videos popularity. Here we'll introduce the 1.0-degree ego-centric network to analyze the characteristics of this community has that brings it so successful. We leverage Gephi to do exploratory analysis.

This graph was grouped by Book (purple node), DVD (green node), Video (orange node), Music and it spread outward from a central ego (the biggest green node) to a series of alters. If we scoping to this 1.0-degree ego-centric network of the highest degree node, a product belonging to DVDs, we find that Laura in DVD group has the highest degree score and customer who purchase DVD co-purchase not only DVD products but also Video. However, they barely co-purchase Book and Music.

Besides, we can get the proportion of this network in Gephi. We find that the proportion of Video nodes is 54.18%, the proportion of DVD nodes is 45.09% and the proportion of Book nodes is 0.73%. Therefore, Video nodes and DVD nodes occupy the largest part of this network.



Thus, from the visualization in our analysis, we assume that this graph shows the large number of nodes about Video and DVD since the Video and DVD are all digital image in the same category and they highly have correlation between them, customer who like to purchase the DVD items are more likely to purchase Video products. However, the nodes of Music and Book are barely seen because Music and Book are not closely related to DVD group, customer are less likely to purchase Music and Book.



This graph shows that there's no popular items that around the biggest green node, this node connect with other unpopular products. So we can conclude that Laura DVD is the most popular DVD item at that time and even in all products. However, if we put this sub-network into the whole selected network for analysis, it's obvious that the DVD Luara has no significant neighbors. Customers who purchase Laura co-purchase those unpopular products at the same time.

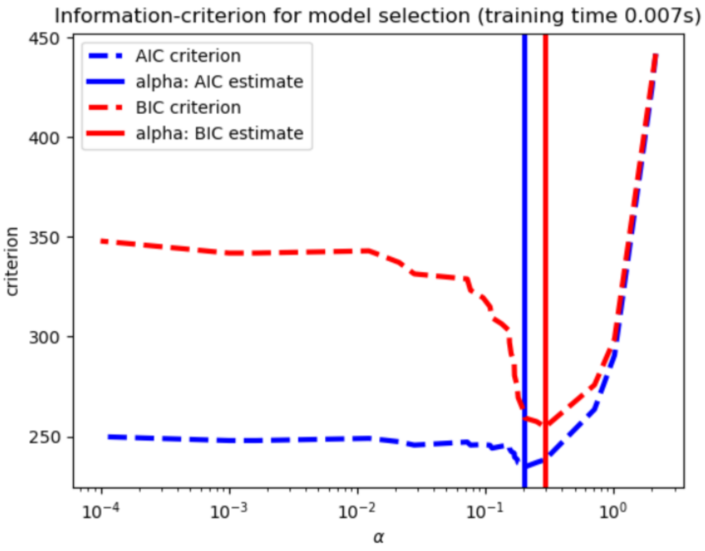
What's more, Based on centrality analysis in two standards: Degree centrality and Eigenvector centrality. We find that in degree centrality, this network shows highly centralized in degree centrality and hub Laura connects everyone else. In eigenvector centrality, this network is not centralized in eigenvector centrality and Laura knows no influential neighbors.

Therefore, even though the Laura has such high degree, it is unable to work as a bridge to connect other significant hubs. So what we can conclude here is "A local but giant hub in the scoped network is surprisingly not a significant bridge".

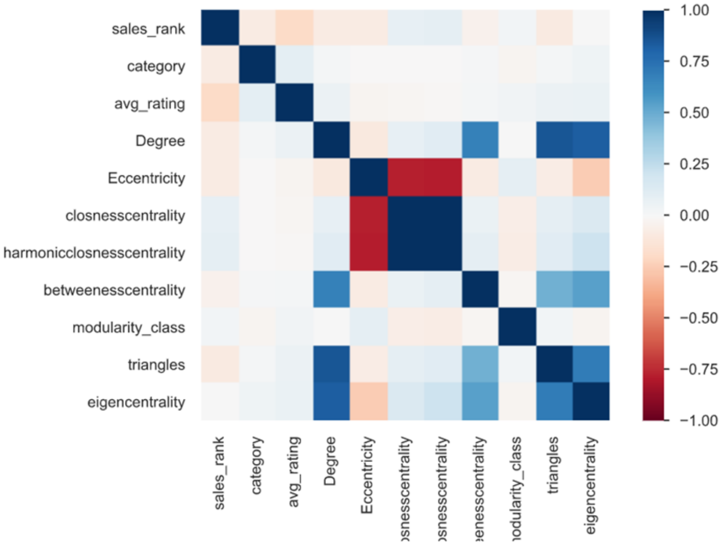
3. Regression Analysis

The highest degree node analysis demonstrates that it has a number of co-purchased products, but it has a low eigenvector centrality score. It is reasonable to assume that the more co-purchased products a product has, the higher sales it will be. So, what other factors can affect product sales? To answer this question, we build a linear regression model to evaluate the relationship between factors and sales.

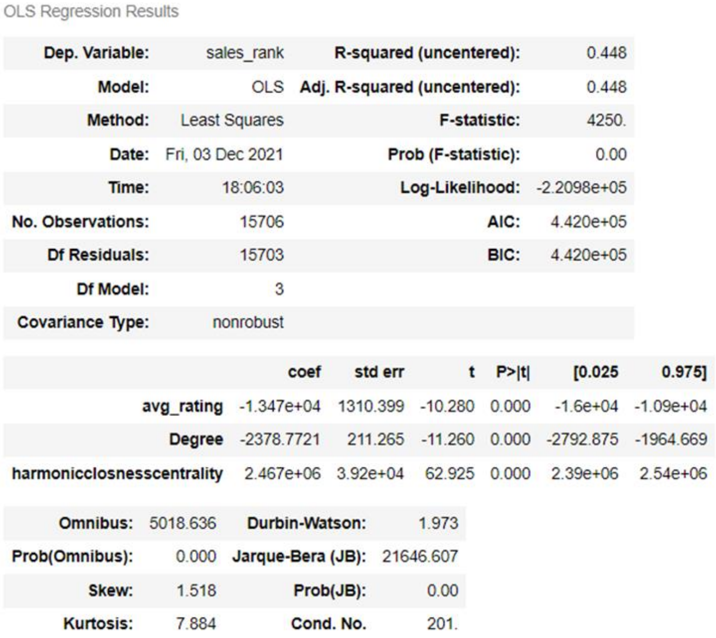
By utilizing Gephi, we export an attributes file that also includes eccentricity, closeness centrality, betweenness centrality, modularity class, triangles, and eigencentrality. Thus, we have all variables and factors to build a regression model. To select which factors should be included in the regression model, we use LassoLarsIC for model selection. Information-criterion based model selection is very fast, but it relies on a proper estimation of degrees of freedom. The figure below is the visualiztion of the LassoLarsIC:



The coefficient of LassoLarsIC algorithms suggests that the regression model should include avg rating, degree, eccentricity, closeness centrality, triangles, and eigencentrality. However, the correlation heat map shows that the degree is highly correlated with eccentricity, triangles, and eigencentrality. So, I exclude those factors that highly correlated with degree and only include avg rating, degree, and closnesscentrality to our model.



Below is the model result:

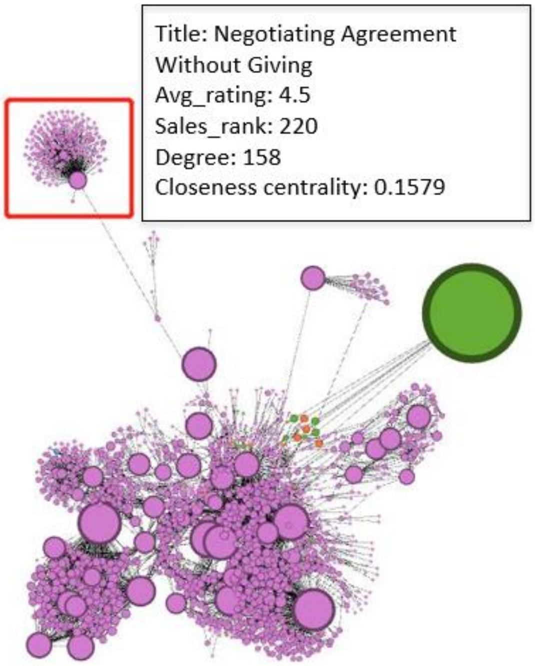

OLS Regression Results

Dep. Variable:	sales_rank	**R-squared (uncentered):**	0.448
Model:	OLS	**Adj. R-squared (uncentered):**	0.448
Method:	Least Squares	**F-statistic:**	4250.
Date:	Fri, 03 Dec 2021	**Prob (F-statistic):**	0.00
Time:	18:06:03	**Log-Likelihood:**	-2.2098e+05
No. Observations:	15706	**AIC:**	4.420e+05
Df Residuals:	15703	**BIC:**	4.420e+05
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
avg_rating	-1.347e+04	1310.399	-10.280	0.000	-1.6e+04	-1.09e+04
Degree	-2378.7721	211.265	-11.260	0.000	-2792.875	-1964.669
harmonicclosnesscentrality	2.467e+06	3.92e+04	62.925	0.000	2.39e+06	2.54e+06

Omnibus:	5018.636	**Durbin-Watson:**	1.973
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	21646.607
Skew:	1.518	**Prob(JB):**	0.00
Kurtosis:	7.884	**Cond. No.**	201.

The R squared represent that the percentage of target variable can be explained by X. For our model, the R square is 0.448. It’s quite low but it is reasonable because there are lots of factors that will influence sales and those are not in our dataset. Those factors can be product quality, brand awareness, competitors’ price, and so on. But we still can conclude some business insights with network data. The average customers rating and degree which represent the number of the co-purchased products has a negative coefficient. It is because that we use sales rank as target variable. The lower sales rank means higher sales. So, the negative coefficient means that the variable is negatively correlated to sales rank and so it is positively correlated with sales. However, we find that the closnesscentrality is positively correlated with sales rank, which means the higher closenesscentrality, the lower sales. So, how come a product has higher degrees but a lower closenesscentrality? The network below shows an example of a product that has high degree but low closenesscentrality.



Title: Negotiating Agreement Without Giving
Avg_rating: 4.5
Sales_rank: 220
Degree: 158
Closeness centrality: 0.1579

The size of the node is ranked by degree. The color of the node is category. The example product is a book called negotiating agreement without giving, the avg rating is 4.5, the sales rank is 220 out of 500k products. The degree is 158, and closenesscentrality is 0.1579. The degree of the product shows that it has 158 co-purchased products, but its co-purchased products' degree is very low. It means that the co-purchased products are barely co-purchased with other products. Accordingly, this product has a high degree with a low closenesscentrality. Thus, based on this finding, we can suggest amazon for an ads campaign. Amazon can partner with sellers for commercial premium, by giving products search recommendations with similar products but with a lower degree. That could help increase product sales.

Result Conclusion and Recommending Solution

1. Finding

From the analysis above, we find that the co-purchased network derived by the dataset has some characteristics under:

1.Products with 2 to 6 labels are most likely to have a higher connection in the co-purchased network, and it is probably because too many tags will make the products seem to have no specialized and practical purpose. At the meantime, 5 labels are most suitable for product to have high degree.

2.Music CDs have limited co-purchased ties. It seems people always buy music CDs especially; it is hard for music CDs to become a co-purchased product or bring other items to sale.

3.From LassoLarsIC, we find that there are three variables which include average rating, degree and closeness centrality have high co-relationship with sales. From the regression model, we can see the high sales products have high average rating, high degree and low closeness centrality.

2. Limitation

There are also some limitations in our project: The original network from the dataset is too sparse. In order to be convenient to study, we need to screen out nodes of which edges are lower than 100 and we only keep a small part of the data. And it will lead to sampling errors.

Reference

1. J. Leskovec, L. Adamic and B. Adamic. The Dynamics of Viral Marketing. ACM Transactions on the Web (ACM TWEB), 1(1), 2007.