Social Network Analysis Report

Team XI - Jiayan Han, Ryan Chen, Kasandra Woo, Sean Jung

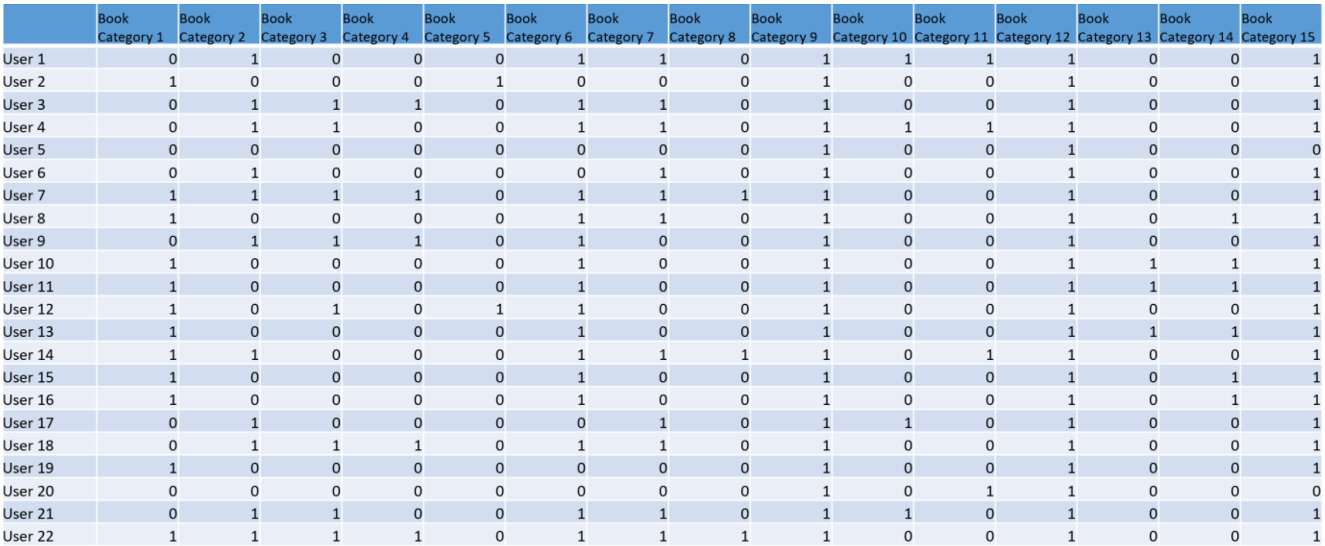
Introduction and Research Question

The Comscore data available via WRDS represents a vast sample of US internet users' internet browsing behavior, purchases, and demographics. Specifically, it is a subset of the opt-in panel data used to inform Media Metrix — Comscore's premier audience measurement product. One of the Comscore databases contains detailed information about the transaction history that spans prominent E-commerce websites such as eBay and Amazon for the year 2019.

The purpose of our analysis is to build a recommendation system specifically designed to recommend book products to the customers who are most likely to purchase a book in a particular category based on different algorithms such as Jaccard similarity, Cosine similarity, and R recommenderlab package. We seek to measure relationships and multiple types of dimensions, represent relationships between various types of objects, extend the individual, group, and network-level measurements for multiple types of objects at once, and finally understand the similarity or dissimilarity of different objects using multiple networks.

Background

Multiple group affiliations are a fundamental concept in defining individuals' social identity as more extensive patterns of similarities among collection individuals can be inferred. Affiliation networks look at collections or subsets of actors or subsets rather than ties between pairs of actors themselves. Looking at relationships through the lens of events shows us how actors create ties among events. Therefore, the affiliation matrix is an appropriate method to model the relationship between actors and events as a whole system.

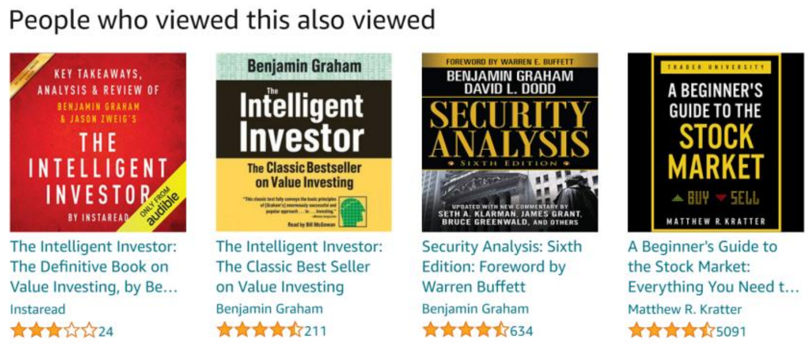


	Book Category 1	Book Category 2	Book Category 3	Book Category 4	Book Category 5	Book Category 6	Book Category 7	Book Category 8	Book Category 9	Book Category 10	Book Category 11	Book Category 12	Book Category 13	Book Category 14	Book Category 15
User 1	0	1	0	0	0	1	1	0	1	1	1	1	0	0	1
User 2	1	0	0	0	1	0	0	0	1	0	0	1	0	0	1
User 3	0	1	1	1	0	1	1	0	1	0	0	1	0	0	1
User 4	0	1	1	0	0	1	1	0	1	1	1	1	0	0	1
User 5	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0
User 6	0	1	0	0	0	0	1	0	1	0	0	1	0	0	1
User 7	1	1	1	1	0	1	1	1	1	0	0	1	0	0	1
User 8	1	0	0	0	0	1	1	0	1	0	0	1	0	1	1
User 9	0	1	1	1	0	1	0	0	1	0	0	1	0	0	1
User 10	1	0	0	0	0	1	0	0	1	0	0	1	1	1	1
User 11	1	0	0	0	0	1	0	0	1	0	0	1	1	1	1
User 12	1	0	1	0	1	1	0	0	1	0	0	1	0	0	1
User 13	1	0	0	0	0	1	0	0	1	0	0	1	1	1	1
User 14	1	1	0	0	0	1	1	1	1	0	1	1	0	0	1
User 15	1	0	0	0	0	1	0	0	1	0	0	1	0	1	1
User 16	1	0	0	0	0	1	0	0	1	0	0	1	0	1	1
User 17	0	1	0	0	0	0	1	0	1	1	0	1	0	0	1
User 18	0	1	1	1	0	1	1	0	1	0	0	1	0	0	1
User 19	1	0	0	0	0	0	0	0	1	0	0	1	0	0	1
User 20	0	0	0	0	0	0	0	0	1	0	1	1	0	0	0
User 21	0	1	1	0	0	1	1	0	1	1	0	1	0	0	1
User 22	1	1	1	1	0	1	1	1	1	0	0	1	0	0	1

This example affiliation matrix above illustrates some of each book category's purchasing behavior among users who have visited the Amazon.com website and purchased the book product. Each row indicates one type of actor (user), while columns indicate each book category.

For example, we can observe that both users 17 and 18 purchased the multiple items that belong in the same book categories (Book categories 2, 7, 9, 12 and 15), indicating the presence of homophily and consequently posing potential to make the use of the recommender system.

Calculating the similarity measures using affiliation matrix is the basic logic behind Amazon and Netflix's product recommendation algorithms. Traditionally, marketing campaigns primarily used their purchase data to recommend products to potential customers. However, numerous research and pre-existing use-case have been published to provide highly accurate systems incorporated to fine-tune product recommendation systems using 'neighbors' purchasing history.



People who viewed this also viewed

The Intelligent Investor: The Definitive Book on Value Investing, by Be…
Instaread
★★★☆☆24

The Intelligent Investor: The Classic Best Seller on Value Investing
Benjamin Graham
★★★★☆211

Security Analysis: Sixth Edition: Foreword by Warren Buffett
Benjamin Graham
★★★★☆634

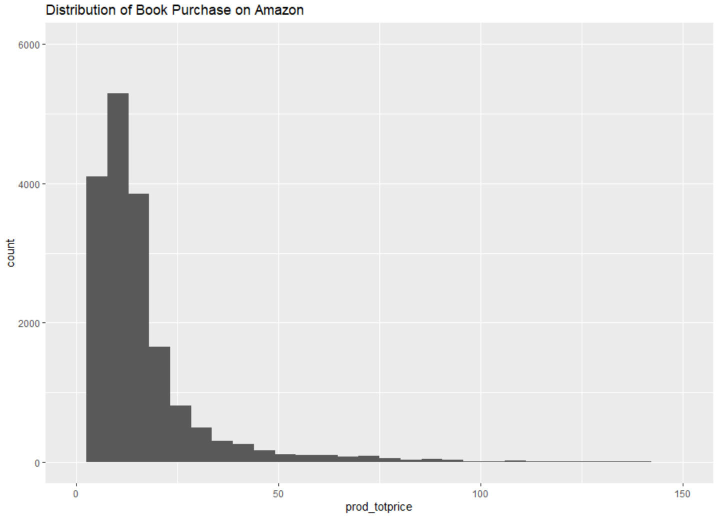
A Beginner's Guide to the Stock Market: Everything You Need t…
Matthew R. Kratter
★★★★☆5091

KEY TAKEAWAYS, ANALYSIS & REVIEW OF BENJAMIN GRAHAM & JASON ZWEIG'S
THE INTELLIGENT INVESTOR
BY INSTAREAD
ONLY FROM audible
Benjamin Graham
The Intelligent Investor
The Classic Bestseller on Value Investing
FOREWORD BY WARREN E. BUFFETT
BENJAMIN GRAHAM DAVID L. DODD
SECURITY ANALYSIS
SIXTH EDITION
UPDATED WITH NEW COMMENTARY BY SETH A. KLARMAN, JAMES GRANT, BRUCE GREENWALD, AND OTHERS
TRADER UNIVERSITY
A BEGINNER'S GUIDE TO THE STOCK MARKET
BUY SELL
MATTHEW R. KRATTER

Amazon's use-case recommendation system based on shared affiliation uses co-affiliations to define the similarity of customers. Then, the corresponding weight of the customer's purchasing decisions proximal to the target customer is calculated to predict whether the target customer will favor a particular product. Additionally, it is possible to train the algorithm to make better predictions over time to delve into a question such as '*What similar features of customer similarity are most salient for different kinds of purchasing decisions?*'

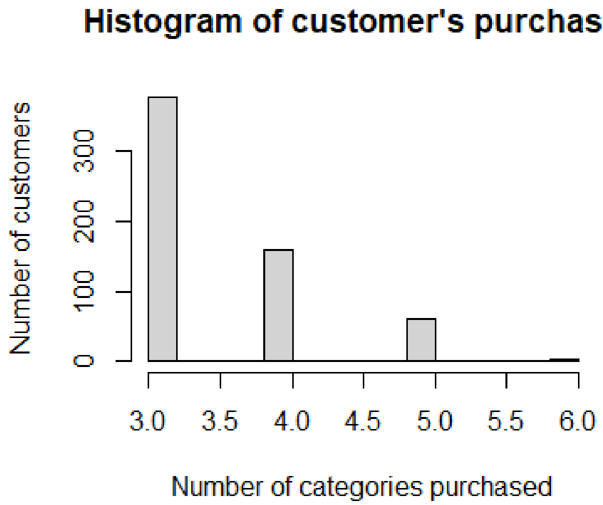
Data Description

Transactions 2019.csv contains information about different purchase categories. This dataset consists of approximately 150 of the largest e-commerce retailers in the United States. *machine_id* column contains information about unique machine identifiers of website users in the form of an integer value. *prod_category_id* column includes information on unique identifiers for the category of the product purchased in the form of an integer value.



The histogram indicates right skewness, meaning that some people purchased significantly higher amounts of books on Amazon. Most people, in contrast, spent less than 50 dollars on Amazon.com buying books.

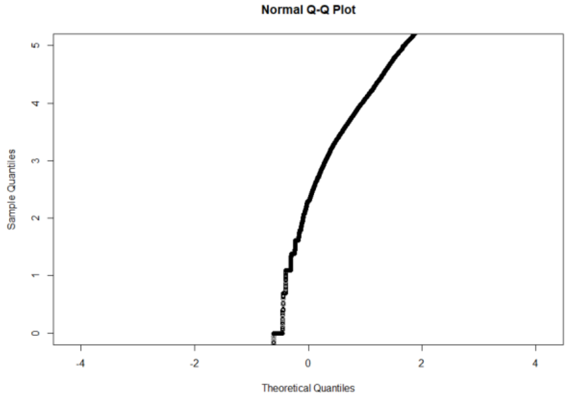
For customers who have made purchases in one or two categories of books, we are just going to recommend them to the most popular book types. For customers who have purchased at least three types of books, we will design a recommendation system for them to help them better locate the book types they are interested in.



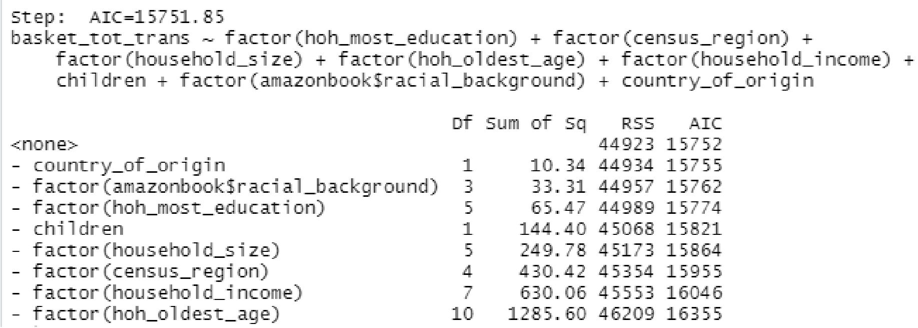
Regression Analysis

We were interested in learning about the relationship between demographic features in the dataset (e.g., education level, household income, etc.) and the number of book purchases on Amazon.

Log transformation was performed for the target variable to make the distribution more linear. This is how the Q-Q plot looks like after the transformation:



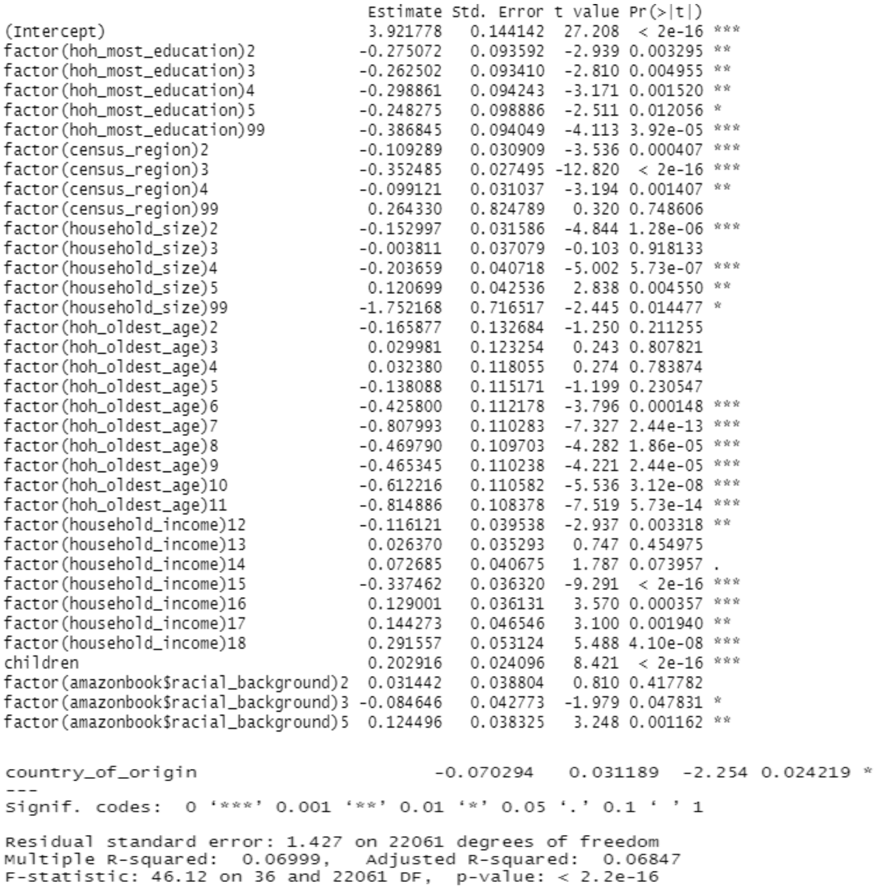
We then performed backwards elimination for variable selection, and the remaining variables are shown below. Connection speed was the only variable that got removed from our model.



```
Step:  AIC=15751.85
basket_tot_trans ~ factor(hoh_most_education) + factor(census_region) +
    factor(household_size) + factor(hoh_oldest_age) + factor(household_income) +
    children + factor(amazonbook$racial_background) + country_of_origin

                                       Df Sum of Sq   RSS   AIC
<none>                                              44923 15752
- country_of_origin                     1     10.34 44934 15755
- factor(amazonbook$racial_background)  3     33.31 44957 15762
- factor(hoh_most_education)            5     65.47 44989 15774
- children                              1    144.40 45068 15821
- factor(household_size)                5    249.78 45173 15864
- factor(census_region)                 4    430.42 45354 15955
- factor(household_income)              7    630.06 45553 16046
- factor(hoh_oldest_age)               10   1285.60 46209 16355
```

Finally, we ran our optimized regression model and discovered that even though it had a low R squared, most features had p values below 0.05, meaning they were correlated with the target variable in a statistically significant way.



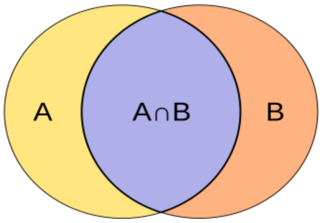
```
                                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)                           3.921778   0.144142  27.208  < 2e-16 ***
factor(hoh_most_education)2          -0.275072   0.093592  -2.939 0.003295 ** 
factor(hoh_most_education)3          -0.262502   0.093410  -2.810 0.004955 ** 
factor(hoh_most_education)4          -0.298861   0.094243  -3.171 0.001520 ** 
factor(hoh_most_education)5          -0.248275   0.098886  -2.511 0.012056 *  
factor(hoh_most_education)99         -0.386845   0.094049  -4.113 3.92e-05 ***
factor(census_region)2               -0.109289   0.030909  -3.536 0.000407 ***
factor(census_region)3               -0.352485   0.027495 -12.820  < 2e-16 ***
factor(census_region)4               -0.099121   0.031037  -3.194 0.001407 ** 
factor(census_region)99               0.264330   0.824789   0.320 0.748606    
factor(household_size)2              -0.152997   0.031586  -4.844 1.28e-06 ***
factor(household_size)3              -0.003811   0.037079  -0.103 0.918133    
factor(household_size)4              -0.203659   0.040718  -5.002 5.73e-07 ***
factor(household_size)5               0.120699   0.042536   2.838 0.004550 ** 
factor(household_size)99             -1.752168   0.716517  -2.445 0.014477 *  
factor(hoh_oldest_age)2              -0.165877   0.132684  -1.250 0.211255    
factor(hoh_oldest_age)3               0.029981   0.123254   0.243 0.807821    
factor(hoh_oldest_age)4               0.032380   0.118055   0.274 0.783874    
factor(hoh_oldest_age)5              -0.138088   0.115171  -1.199 0.230547    
factor(hoh_oldest_age)6              -0.425800   0.112178  -3.796 0.000148 ***
factor(hoh_oldest_age)7              -0.807993   0.110283  -7.327 2.44e-13 ***
factor(hoh_oldest_age)8              -0.469790   0.109703  -4.282 1.86e-05 ***
factor(hoh_oldest_age)9              -0.465345   0.110238  -4.221 2.44e-05 ***
factor(hoh_oldest_age)10             -0.612216   0.110582  -5.536 3.12e-08 ***
factor(hoh_oldest_age)11             -0.814886   0.108378  -7.519 5.73e-14 ***
factor(household_income)12           -0.116121   0.039538  -2.937 0.003318 ** 
factor(household_income)13            0.026370   0.035293   0.747 0.454975    
factor(household_income)14            0.072685   0.040675   1.787 0.073957 .  
factor(household_income)15           -0.337462   0.036320  -9.291  < 2e-16 ***
factor(household_income)16            0.129001   0.036131   3.570 0.000357 ***
factor(household_income)17            0.144273   0.046546   3.100 0.001940 ** 
factor(household_income)18            0.291557   0.053124   5.488 4.10e-08 ***
children                              0.202916   0.024096   8.421  < 2e-16 ***
factor(amazonbook$racial_background)2  0.031442   0.038804   0.810 0.417782    
factor(amazonbook$racial_background)3 -0.084646   0.042773  -1.979 0.047831 *  
factor(amazonbook$racial_background)5  0.124496   0.038325   3.248 0.001162 ** 

country_of_origin                     -0.070294   0.031189  -2.254 0.024219 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.427 on 22061 degrees of freedom
Multiple R-squared:  0.06999,	Adjusted R-squared:  0.06847 
F-statistic: 46.12 on 36 and 22061 DF,  p-value: < 2.2e-16
```

Demonstration of Jaccard Similarity

Jaccard Similarity

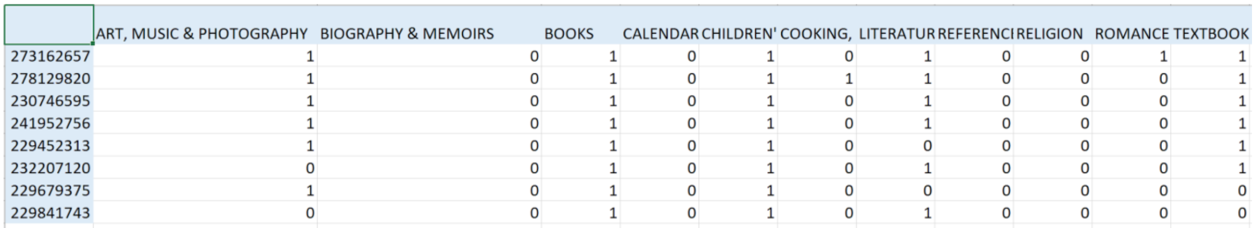


$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

The Jaccard similarity measures the similarity between objects of purely binary attributes known as similarity coefficients. It computes similarity as the proportion of shared membership to all membership held by either one of the nodes. For two nodes A and B, the number of concurrent instances is divided by the union of the events (these are unique events). The similarity measure is equal to 0 if no events are shared and is equal to 1 if all events are shared.

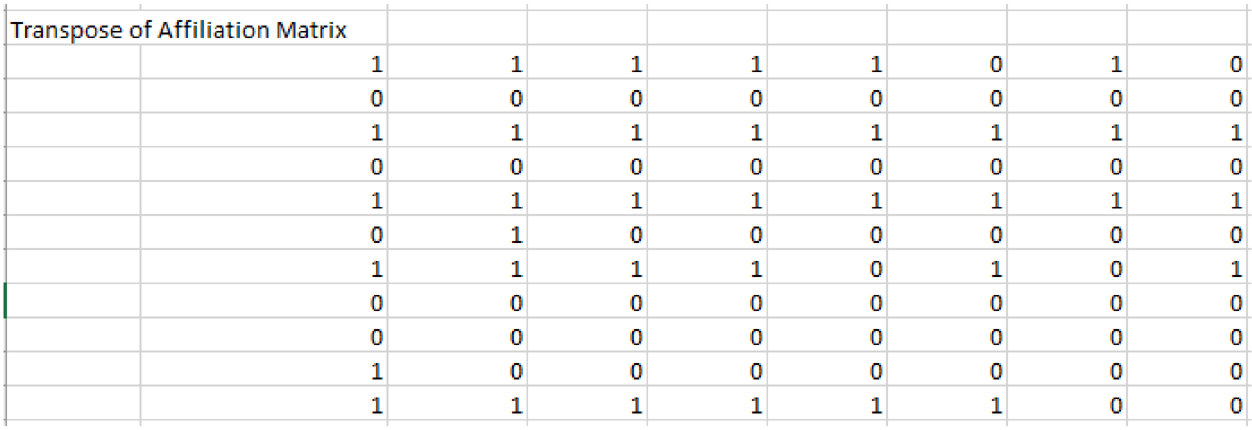
We were able to utilize the Comscore data to generate the Jaccard similarity measures. After doing so, the measures were used to make book category recommendations to consumers. To demonstrate how to use Jaccard similarity to make a recommendation system, we took a sample of eight consumers, two who had purchased six books, two that had purchased five books, two that had purchased four books, and two purchased three books. The recommendation system using Jaccard similarity would not be as good of a fit for users who purchased only one or two book types. We would use an alternative method to make recommendations. For those users, we would recommend the most popular genres.

Initially, we created an affiliation matrix that shows each consumer and the genres of the books they purchased. We used this information to create an affiliation matrix for the sample. This matrix shows the user ID, and if they purchased a book in the genre, that relationship is denoted by a 1, and if not, it is denoted by a 0.



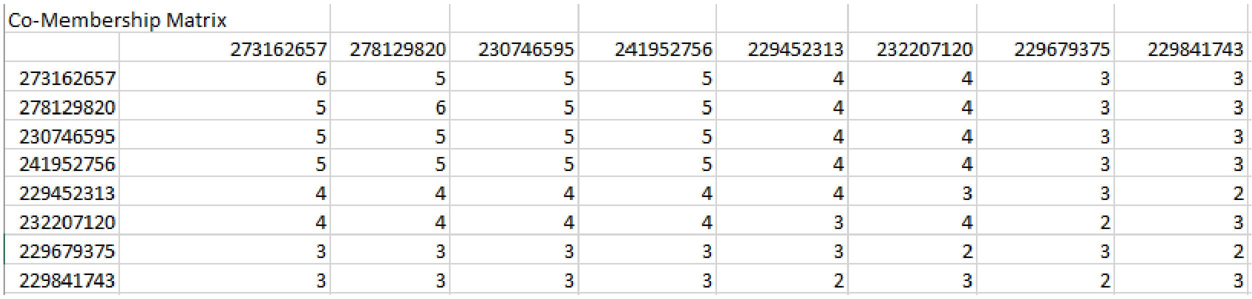
	ART, MUSIC & PHOTOGRAPHY	BIOGRAPHY & MEMOIRS	BOOKS	CALENDAR	CHILDREN'	COOKING,	LITERATUR	REFERENCI	RELIGION	ROMANCE	TEXTBOOK
273162657	1	0	1	0	1	0	1	0	0	1	1
278129820	1	0	1	0	1	1	1	0	0	0	1
230746595	1	0	1	0	1	0	1	0	0	0	1
241952756	1	0	1	0	1	0	1	0	0	0	1
229452313	1	0	1	0	1	0	0	0	0	0	1
232207120	0	0	1	0	1	0	1	0	0	0	1
229679375	1	0	1	0	1	0	0	0	0	0	0
229841743	0	0	1	0	1	0	1	0	0	0	0

In order to find the Jaccard similarity, it is necessary to know how many genres the users share. Therefore, creating a co-membership matrix is useful. You can use formula $X^N = AA'$ to find the co-membership matrix, where $X^N$is the co-membership matrix, $A$is the affiliation matrix, and $A'$is the transpose of the affiliation matrix.



Transpose of Affiliation Matrix								
	1	1	1	1	1	0	1	0
	0	0	0	0	0	0	0	0
	1	1	1	1	1	1	1	1
	0	0	0	0	0	0	0	0
	1	1	1	1	1	1	1	1
	0	1	0	0	0	0	0	0
	1	1	1	1	0	1	0	1
	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0
	1	0	0	0	0	0	0	0
	1	1	1	1	1	1	0	0

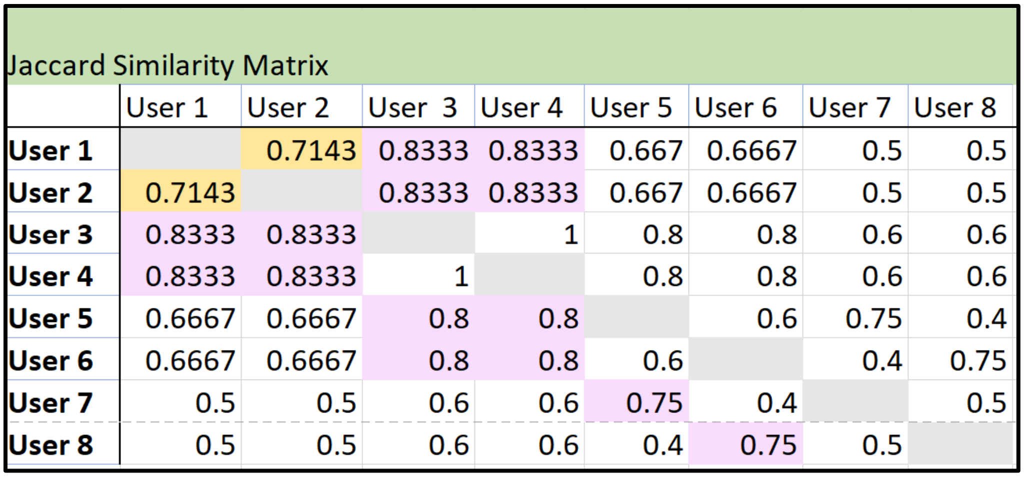
After finding the transpose of the affiliation matrix, you can use the formula stated above to create the co-membership matrix.



Co-Membership Matrix								
	273162657	278129820	230746595	241952756	229452313	232207120	229679375	229841743
273162657	6	5	5	5	4	4	3	3
278129820	5	6	5	5	4	4	3	3
230746595	5	5	5	5	4	4	3	3
241952756	5	5	5	5	4	4	3	3
229452313	4	4	4	4	4	3	3	2
232207120	4	4	4	4	3	4	2	3
229679375	3	3	3	3	3	2	3	2
229841743	3	3	3	3	2	3	2	3

This co-membership matrix tells us how many genres these users share with one another. The diagonal tells us how many genre categories for that user. This information can be applied to the

Jaccard similarity formula stated earlier in order to find the Jaccard similarity measure for each of these relationships. In order to best aggregate this information, we created a matrix that shows the Jaccard similarity between each of these users (I renamed the users to User 1, User 2, etc. instead of using the User IDs for readability)



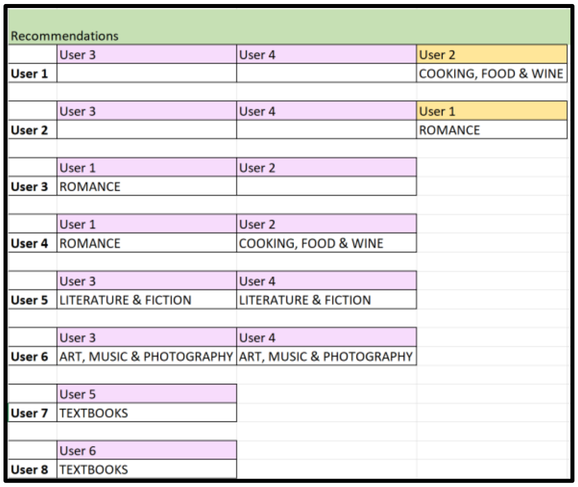
Jaccard Similarity Matrix

	User 1	User 2	User 3	User 4	User 5	User 6	User 7	User 8
User 1		0.7143	0.8333	0.8333	0.667	0.6667	0.5	0.5
User 2	0.7143		0.8333	0.8333	0.667	0.6667	0.5	0.5
User 3	0.8333	0.8333		1	0.8	0.8	0.6	0.6
User 4	0.8333	0.8333	1		0.8	0.8	0.6	0.6
User 5	0.6667	0.6667	0.8	0.8		0.6	0.75	0.4
User 6	0.6667	0.6667	0.8	0.8	0.6		0.4	0.75
User 7	0.5	0.5	0.6	0.6	0.75	0.4		0.5
User 8	0.5	0.5	0.6	0.6	0.4	0.75	0.5	

This shows us how similar these users are to each other in regards to the types of book genres they purchase. The higher the Jaccard similarity, the more similar they are. From there, we can compare each user to their most similar user and recommend book genres based on genres the user has not purchased, but their most similar user has purchased.

Illustration of Results

After finding the most similar users, you can use the affiliation matrix (created earlier) to compare which products were not purchased by the user when comparing them to their most similar user. For example, with User 4, their most similar users were users 1 and 2 (tied). We then compared what books User 4 purchased to the books User 1 purchased. User 1 purchased romance books while User 4 did not, and because they are similar, we would recommend romance books to User 4 as well.



Recommendations

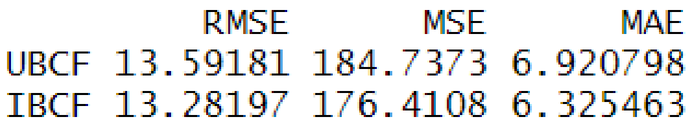
	User 3	User 4	User 2
User 1			COOKING, FOOD & WINE
	User 3	User 4	User 1
User 2			ROMANCE
	User 1	User 2	
User 3	ROMANCE		
	User 1	User 2	
User 4	ROMANCE	COOKING, FOOD & WINE	
	User 3	User 4	
User 5	LITERATURE & FICTION	LITERATURE & FICTION	
	User 3	User 4	
User 6	ART, MUSIC & PHOTOGRAPHY	ART, MUSIC & PHOTOGRAPHY	
	User 5		
User 7	TEXTBOOKS		
	User 6		
User 8	TEXTBOOKS		

You might notice that there are some blank spaces in the illustration above. This is because some users have purchased all genres their most similar user has purchased. Therefore, there are no new book genres to recommend. In those instances, we look at their next most similar user based on Jaccard similarity and make recommendations based on that user.

R Recommenderlab Analysis

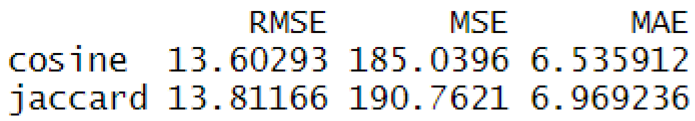
Another method is implemented by using the Recommenderlab package in R. Generally speaking, there are two types of recommendation algorithms. One is a content-based approach; the other one is collaborative filtering. Recommenderlab package focuses on the collaborative filtering method. The idea is that given rating or purchase history data by many users for many items, a prediction is made such that top-N lists of recommended items can be created from other similar user's ratings. The premise is that users who agreed on the rating for some items typically also agree on the rating for other items. In our case, both user-based (UBCF) and item-based collaborative filtering (IBCF) will be used to train the data. The input data will be the affiliation matrix that we showed earlier. To evaluate the two methods, we will split the data to train and test, and compare the value of MSE, RMSE, MAE.

As shown below, RMSE, MSE, and MAE of item-based collaborative filtering are smaller than that of UBCF, indicating that item-based CF generates more accurate predictions than user-based CF.



```
         RMSE      MSE      MAE
UBCF 13.59181 184.7373 6.920798
IBCF 13.28197 176.4108 6.325463
```

Based on the previous result, we then compared the Cosine and Jaccard similarity method. Cosine similarity has a lower RMSE, MSE, and MAE, which suggests that the Cosine similarity outperforms the Jaccard similarity method.



```
            RMSE      MSE      MAE
cosine  13.60293 185.0396 6.535912
jaccard 13.81166 190.7621 6.969236
```

Conclusions and Implications

Before we get to our analysis results, it is noteworthy to summarize what we have done and look at their implications. We started by visually inspecting the histogram of the number of book categories purchased in our network. We found that customers are most likely to purchase the three different categories. We then moved down to running a simple linear regression model to learn more about the relationship between demographic features and the number of books purchased on Amazon. We also performed backward elimination for variable selection and finally ran an optimized regression model. From the output of the results, we can conclude that several variables, such as racial background, education levels, household size, census region, household income, oldest age in the household, are related to the transaction amount attributed to the variable (total basket size).

While we think that we have performed sufficient analysis to our ability, we would like to propose additional avenues to extend our research to other topics. Firstly, we suggest building the recommendation system for the different categories in the dataset. We can use big data analytics tools such as Apache Spark Stream to automate this analysis as the data comes into the database in real-time. We can also improve our research by finding centrality measures for each customer who has purchased the particular category in Amazon.com, Inc. By identifying the edge-level attribute of each customer. We can build a more sound recommendation system that relies less on second-guessing but more on the evidence-based solution that guarantees the tangible return on investment when deployed into real life.

We can also identify each node's significance and relevance to other customers who have purchased different categories of product for each purchase. This will potentially inform us of the intrinsic/extrinsic motivation behind which customers purchase particular categories of the product, allowing us to adopt additional recommendation features that can make appropriate decisions to the customers at the right time. We can also perform further regression analysis on different demographic factors such as the customer's geographical location from the ZIP code to evaluate if there exists a correlation between the location and categories of items purchased.

Though we have not conducted any cost/benefit analysis to build a feasible expected value model around our recommendation system, we can also look at the correlation between the categories of item purchased as well as the amount of transaction of each purchase and their respective centrality measures to analyze trends if any and determine if some categories are overpriced/underpriced based on the behavioral analysis to their counterparts. Lastly, we also suggest that analyzing dataset spanning more than one year would be beneficial to see the longitudinal effect of attributes on the outcome variables and their impact on Amazon's bottom line over time, which could be crucial in establishing a corporate strategy that determines the primary focus based on Pareto principle (20 percent of item categories that yields 80 percent of market dominance and profit).

Reference

1. Michael Hahsler (2020). recommenderlab: Lab for Developing and Testing Recommender Algorithms. R package version 0.2-6. https://github.com/mhahsler/recommenderlab
2. Social Network Analytics class 6 lecture slides - Representing Multiple Types of Nodes