# Generation of Custom Shops From An Amazon Co-Purchasing Network

Samuel Hamilton, Shirley Liu, Aria Wang, Joe Zhang

## Summary of the results of the preliminary analysis and data description

- What we did/discovered in our project proposals

Traditional marketing practices rely on demonstrating the value of a few profitable products to a large number of potential customers so as to increase revenue or for the ability to charge a greater margin on said products. However, E-commerce retailers such as Amazon are able to market and sell a larger number of products than traditional retailers do. For Amazon, niche products comprise a large portion of total revenue, and the volume of sales from these products are crucial for both Amazon and sellers on the platform. Therefore, connecting the right products to interested consumers is critical for revenue. For this project, we aimed to apply network analysis on the data about co-purchased Amazon products and help Amazon make better recommendations on what customers are likely to purchase next. Based on the co-purchasing product networks, we sought to develop a method to accurately connect Amazon customers to products they may be interested in.

- Where we found our data

The dataset for our project was a product co-purchasing network collected from Amazon on March 2, 2003. We acquired this data from the Stanford Large Network Dataset collection website (http://snap.stanford.edu/data/amazon0302.html). We integrated this dataset with an Amazon product metadata dataset collected in the summer of 2006 (http://snap.stanford.edu/data/amazon-meta.html). This dataset was also available from the Stanford Large Network Dataset collection website

- The variables we included in this project

In our original data, a node represents a product, and every edge indicates the frequent co-purchase relationships between 2 products. The Amazon metadata contains detailed information about each product such as the categories of the products and customer reviews (who gave reviews, the content of the reviews, the number of consumers who found that review helpful, and the aggregated review scores). In our analysis, we decided to include the product group, the product category within the product group (such as the book genres a book belongs to).

- The benefits and/or limitations of this data

A benefit of this data is that it is clean and straightforward. We did not spend much time on cleaning the data or extracting potentially useful information from the original data.

One limitation of this dataset is that it defines high co-purchasing as a binary variable. In the data, items are either frequently co-purchased or not. This binary classification masks the degree to which products are co-purchased, which could have been a highly informative statistic.

The other limitation of this data is that it covers a large number of unique products, which led to too many nodes in the network and thus posed challenges to our network analysis.

- Descriptive statistics about our data (e.g., # of nodes, edges, etc.)

The Amazon metadata product metadata dataset consists of 548,552 products, 1,788,725 product-product edges and 2,509,699 total product category memberships. There are 393,561 books, 19828 DVDs, 103144 Music CDs, and 26132 videos. The Amazon product co-purchasing network has 262,111 nodes and 1,234,877 edges. The largest WCC had all of the total edges and nodes, 262,111 nodes and 1,234,877 edges. The largest SCC had 241,761 nodes, and 1,131,217 edges. The average clustering coefficient was 0.4198. In addition, there were 717,719 triangles,

0.09339 of which were closed. The Diameter of the network was 92, with a 90% effective diameter of 11.

## The in-depth questions we are going to answer

- Questions are we attempting to answer

The goal of our project is to identify products that are commonly bought with any given product so that Amazon can recommend the correct products to consumers. In addition, we aim to find out whether the qualities of a product, such as its product group and product category, would affect its likelihood to be co-purchased with other products. Furthermore, for this project, we would like to learn whether there is in-degree or out-degree bias within the network of co-purchased Amazon products.

- What we hope to learn from this analysis

We hope to better understand how qualities of products affect their likelihood to be co-purchased. We are interested in learning whether there is an in-degree or out-degree bias within co-purchasing networks, in order to help Amazon determine whether to use co-purchasing networks to direct shoppers to potential purchases.

- Why are these questions important for Amazon?

These questions are important to Amazon because niche products comprise a large proportion of their total margin. However, since individually all of these products do not contribute highly to revenue, investing in advertising all of them separately is not cost efficient. Therefore, there is a critical need for Amazon to develop algorithms that automatically connect customers to products they are more likely to purchase, in order to increase the volume of these niche product sales at a low cost.

## Analysis

- Description of our analysis

We would like to test whether the groups or categories have an effect on the co-purchasing behavior. Initially, we had the assumption that if consumers buy a product in a certain group or category, it is likely for them to buy something else from the same or the complementary group. For example, if Joe buys coke, he probably has a tendency to buy another drink, like sprite, as well, because there may be a party waiting for him and he is just the one to prepare different kinds of drinks. However, there are also cases when people buy things from the complementary groups or categories as the original product. In the example of Joe's case, if he buys some bottled beer, it is likely that he will also buy an opener. Therefore, there exist two types of considerations when people make a purchase. Determining which dominates the other is one main purpose of the analysis. For this purpose, we used an ERGM model for the data and checked whether the coefficients for groups and categories are significant.

- What we did to conduct these analyses

First of all, as the product categories with too few products are not helpful for our analysis, we removed the categories with less than 2 products. The raw data originally contained 90 distinct categories, 15 of which have only 1 product, so we ended up with 75 distinct categories in the data after the removal.

Then we conducted a stratified sampling on the product category over the original network. The main purpose of the sampling process is to facilitate the ERGM model in the next step. We tried to fit the ERGM model on the whole dataset first, but the large original dataset with so many isolated nodes always led to the crash of the program. Therefore, to solve this issue, we implemented the stratified sampling with 0.05 sampling ratio over the data. The

sampling process drawed 3,584 different samples from the original data, and the distributions of product groups and categories remained unchanged.

At last, we fed the new dataset into an ERGM model and looked into the coefficients. The predictors and some hyperparameter settings are as follows.

```
model2 <- ergm(DataNetwork ~ edges
               # Structural patterns
               + mutual
               + gwidegree(1.06, fixed = T)
               + gwodegree(log(2), fixed = T)
               + dgwesp(log(2), type = "OTP", fixed = T)
               + dgwdsp(log(2), type = "RTP", fixed = T)
               # Node attribute effects
               + nodematch("Group")
               + nodematch("Categories"),
               nr.maxit = 100)
)
summary(model2)
```

- Why we chose these methods

To solve the project questions, we need to identify the effect of different node attributes on the relationships between different actors and on the presence of any particular patterns. ERGM is a perfect model for this job, as its model result gives the coefficient values and the variable significance for any predictors we are interested in.

- How these analyses help us to answer our questions

The ERGM model result shows the coefficient and significance of predictors. From the sign of the coefficients, we can learn whether the predictors have positive or negative effects on the probability of the occurrence of the target network. Meanwhile, the significance of the coefficients can tell us whether the effects are significant or not. For example, if the coefficient value for the product group or category is significantly different from zero, this co-purchasing data indicates that the product group or category affected the co-purchasing behaviors on

Amazon, whereas if the coefficients are not significant, it indicates that Amazon consumers are not more or less likely to buy different products together within the same group or category.

## Findings

- The results we uncovered with our analyses

The co-purchasing network demonstrates some in- and out-degree biases, but the product group and category are not the best predictors of the network. All the estimates are significant except for the product group and category (Table 1). In other words, products from the same group or category do not significantly impact consumers' decisions on co-purchasing.

- The numerical output and its interpretation in the context of the methodological definition of the concepts we are exploring

The Monte Carlo result (Table 1) shows that all the estimates are significant except for groups and categories, which indicates that people are not more or less likely to buy products from the same group or category. As for the goodness of fit (Table 2), none of the observed edges, indegree, outdegree, nodematch.group and nodematch.categories is significantly different from the simulated values compared to mutual, OTP and RTP. If we check the distribution of typical patterns in the network (visualized in plot 1, 2, 3), we can find that all of them follow long-tail distribution, which suggests that in general, the model is a fair one and we can accept the results of the model.

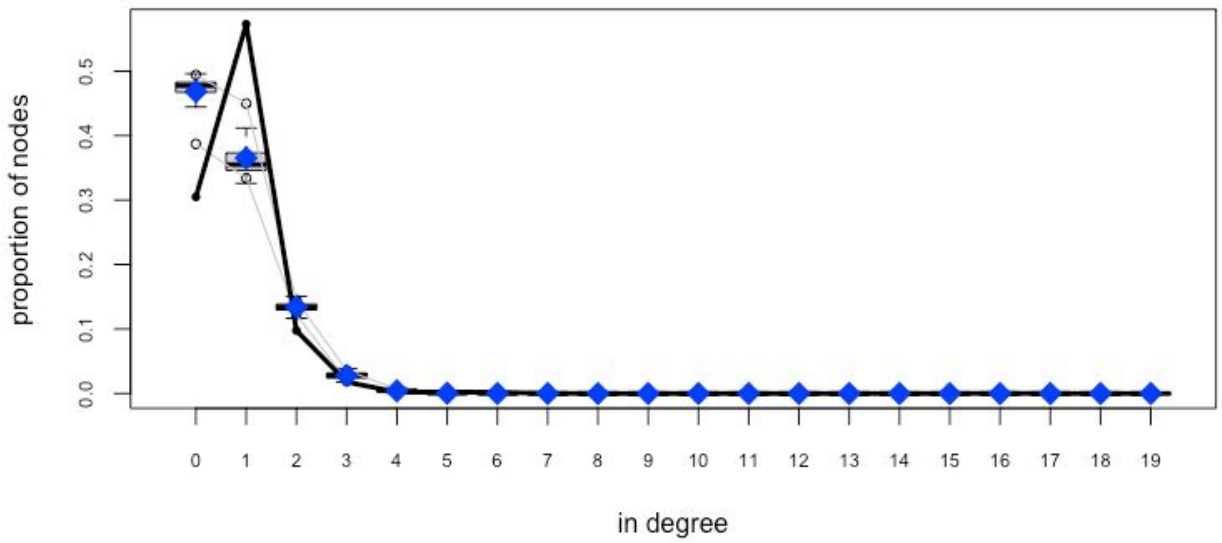
| | Estimate | Standard error | MCMC % | Z value | Probability |
|---|---|---|---|---|---|
| edges | -15.01333 | 0.24800 | 0 | -60.537 | <1e-04*** |
| mutual | 12.83730 | 0.41725 | 8 | 30.766 | <1e-04*** |
| gwideg.fixed | 1.71338 | 0.20363 | 0 | 8.414 | <1e-04*** |

| | | | | | |
|---|---|---|---|---|---|
| gwodeg.fixed | 5.74803 | 0.21441 | 0 | 26.809 | <1e-04*** |
| gwesp.OTP.fixed | 3.18092 | 0.35574 | 3 | 8.942 | <1e-04*** |
| gwdsp.RTP.fixed | 8.30550 | 0.58314 | 3 | 14.243 | <1e-04*** |
| nodematch.Group | 0.06144 | 0.07483 | 1 | 0.821 | 0.412 |
| nodematch.Categories | 0.11775 | 0.10740 | 1 | 1.096 | 0.273 |

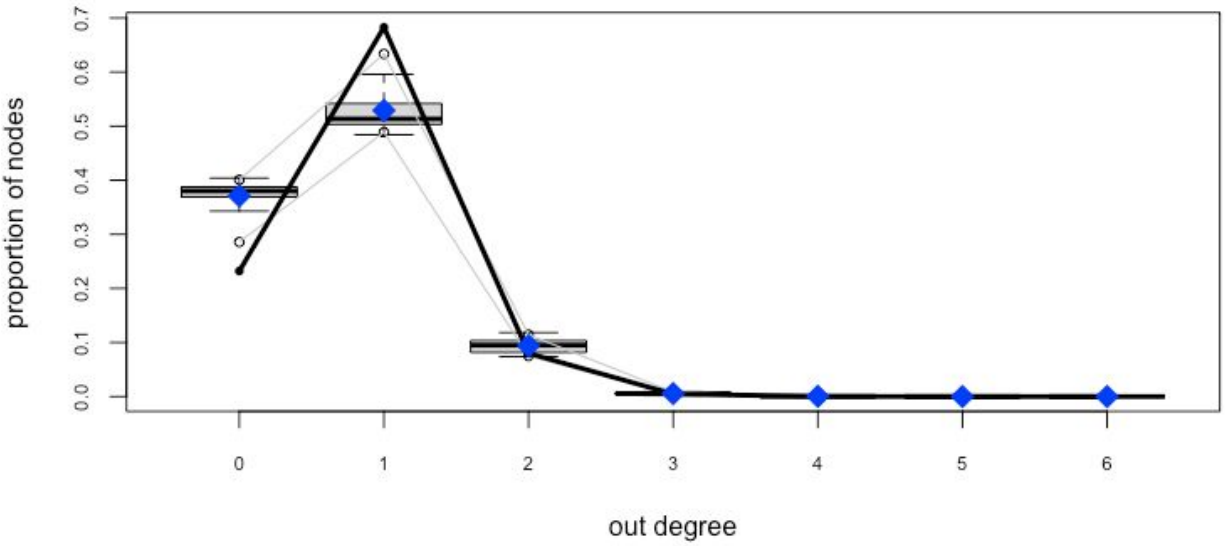
Table 1. ERGM Model Summary

| | obs | min | mean | max | p-value |
|---|---|---|---|---|---|
| edges | 3076 | 3014 | 3142 | 3266 | 0.22 |
| mutual | 849 | 849 | 931 | 1003 | 0.01 |
| gwideg.fixed | 2826 | 2755 | 2865 | 2966 | 0.47 |
| gwodeg.fixed | 2909 | 2850 | 2961 | 3062 | 0.17 |
| gwesp.OTP.fixed | 297 | 228 | 251 | 297 | 0.01 |
| gwdsp.RTP.fixed | 143 | 143 | 163 | 192 | 0.01 |
| nodematch.Group | 1792 | 1715 | 1797 | 1884 | 0.89 |
| nodematch.Categories | 372 | 344 | 381 | 417 | 0.61 |

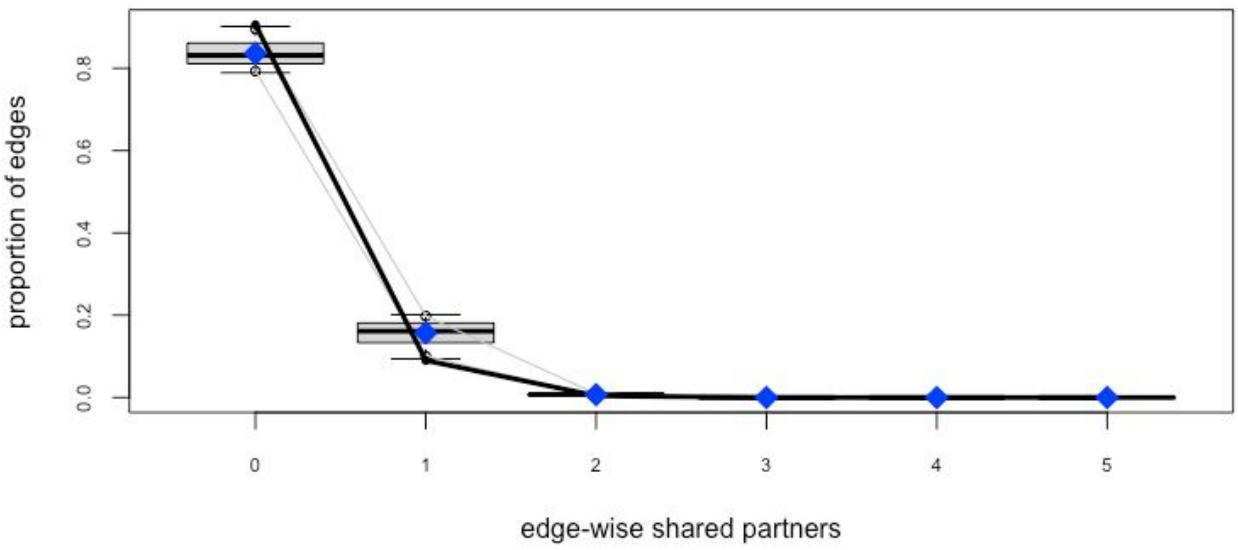
Table 2. Goodness-of-Fit for Model Statistics



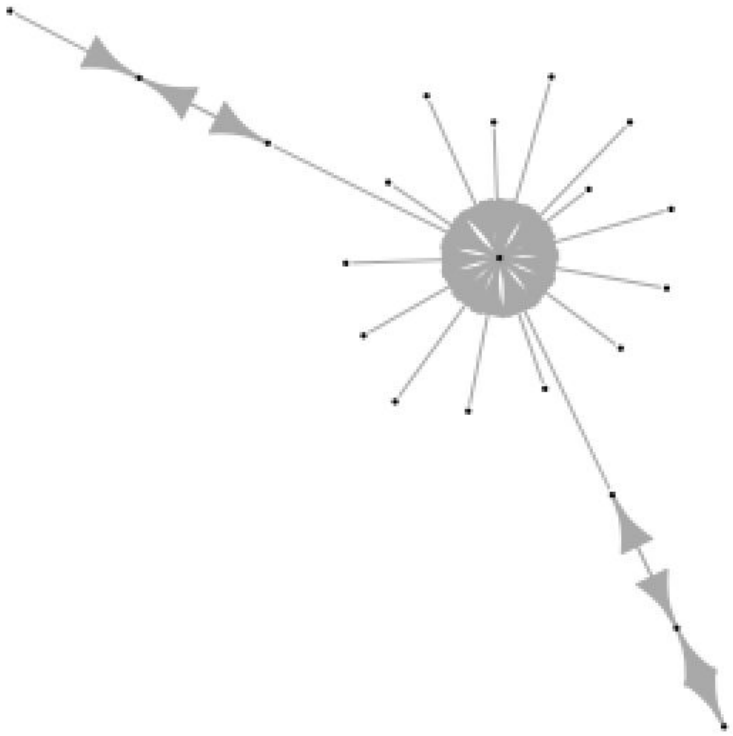

Plot 1. Goodness-of-Fit Indegree Distribution




Plot 2. Goodness-of-Fit Outdegree Distribution




Plot 3. Goodness-of-Fit Edgewise Shared Partners Distribution



Plot 4. Largest connected component within our data sample consists of just 24 points. Displays high in-degree bias.

## Implications

- The implications of our findings for the organization

From the results, we can conclude that whether products are within the same group/category does not impact whether they are often purchased together. The product category/group does not play a significant role in the co-purchasing likelihood.

- More specifically, our recommendations for the organization

When Amazon aims to connect consumers to the right items and increase the sales of more products, the recommendation can be based less on the group or category information of the products.

## Reflection

- Were we able to answer the questions you wanted to ask? Did the results surprise us?

We were able to answer the questions that we asked. The initial question we wanted to address was whether consumers tend to purchase products from the same category or different categories. The result from our analysis surprised us in that the product category turned out to be an unimportant factor for co-purchasing behaviors. As we talked about in the “Analysis” section, there are 2 different possible scenarios when consumers determine their co-purchases, and we hoped the model results would support one of the two scenario. However, our result from the ERGM model shows that both of the 2 scenarios could be possible. None of the 2 cases is more common than the other in the data, and it actually does not matter to the co-purchasing likelihood that whether the products are of the same category or not.