Amazon Co-Purchasing Ground-Truth Communities

**Kanika Mohpal, Sushaanth Srirangapathi, Vijitha Vasantha Kumar, Yi Zheng**

1. **Introduction**
   1. **Background**

For a long time, retail companies would like to use many techniques to extract information from their mass transactions data, which may help them in affinity positioning, cross-selling and so on. Some techniques such as Association Rules become an important part in data mining algorithms.

Nowadays more and more shopping happens online, but the heat of market-basket analysis is never off. On the contrary, it becomes easier to get co-purchasing data for making better analyses. Amazon is the largest Internet-based retailer in the world by total sales and market capitalization. Its net revenue reaches 107.1 billion dollars in 2015. On December 5th, Amazon took the wraps off Amazon Go, a real-world grocery store that comes with a twist: there’s no checkout process. You just grab the stuff you want and walk out; the order posts to your Amazon account afterwards. There is another cool scenario where the value of “Customer who bought this item also bought” shows up.

In the perspective of network, every pair of co-purchasing could be a link between two items and the total “Customer who bought this item also bought” system forms a big network. Taking advantage of some concepts and techniques in network analysis will help to exact patterns in the whole complex dataset and find key items that would be a sign of many other items purchasing. What’s more, it is a good way to provide visualization of the data than other techniques like association rules.

* 1. **Our goals**

1. Find patterns of Amazon co-purchase behavior

Some items are co-purchased more frequently than others, which form communities where edges appear with high concentration among the members of the community. It is reasonable to group the co-purchased items into different communities. Our first goal is to detect communities. It will make the decisions about cross-selling and affinity positioning more supportive (more frequent).

1. Select solid communities

Then it comes to our primitive goal of this project. There are several algorithms to detect communities in network analysis. Jaewom Yang and Jure Leskovec in their *Defining and Evaluating Network Communities based on Ground-truth* give 13 commonly used structural definitions of network communities and examine their sensitivity, robustness and performance in identifying the ground-truth. Our goal is to go farther in this and find a more confident measure by combining some initial measures.

1. Find relatively important items within the communities

After detecting communities, we’d like to get some insight within the community by elaborating our focus on finding key items which should be given more inventory and more exposure on the website.

* 1. **Data description (Subject to Kinika’s confirmation)**

Our dataset is from the publicly available Stanford Network Analysis Platform (SNAP). The description about this dataset is as follows, “Network was collected by crawling Amazon website. It is based on Customers Who Bought This Item Also Bought feature of the Amazon website. If a product i is frequently co-purchased with product j, the graph contains an undirected edge from i to j. Each product category provided by Amazon defines each ground-truth community.” Some detailed information about the dataset is in the following table 1-1.

|  |  |
| --- | --- |
| Table 1-1 Dataset statistics | |
| Nodes | 334863 |
| Edges | 925872 |
| Nodes in largest WCC | 334863 (1.000) |
| Edges in largest WCC | 925872 (1.000) |
| Nodes in largest SCC | 334863 (1.000) |
| Edges in largest SCC | 925872 (1.000) |
| Average clustering coefficient | 0.3967 |
| Number of triangles | 667129 |
| Fraction of closed triangles | 0.07925 |
| Diameter (longest shortest path) | 44 |
| 90-percentile effective diameter | 15 |

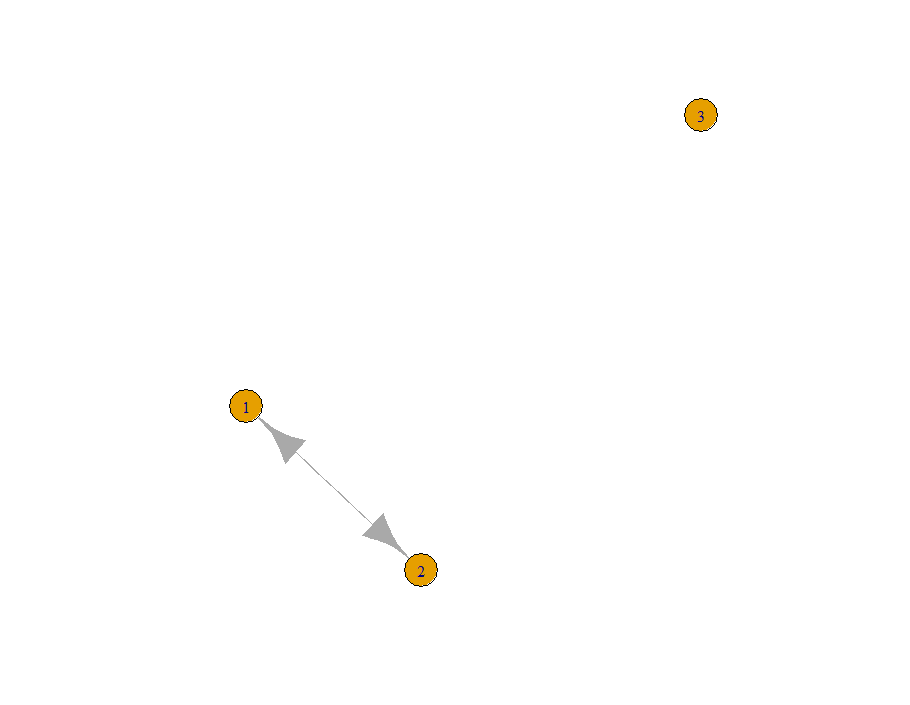
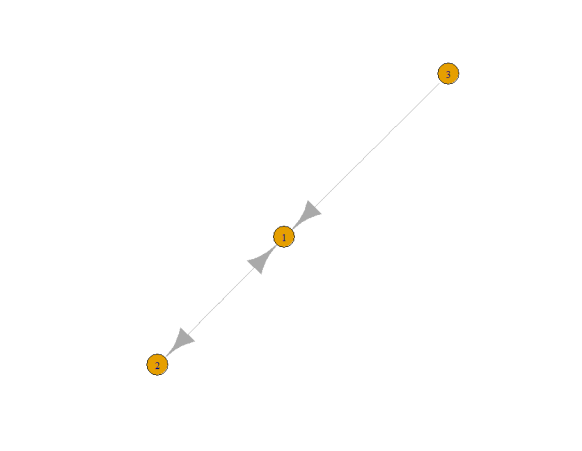
1. **Patterns of Amazon co-purchase**
   1. **Micro Pattern – Motif Analysis**

Understanding the interactions at a localized and micro scale may help in isolating the noise and point to interesting trends that can be extrapolated to the larger scale. Motifs have been used extensively to study inter-actions in biological networks. Milo et all in his article *Network Motifs: Simple Building Blocks of Complex Networks* defined “Network Motif” as patterns of inter-connections occurring in complex networks at numbers that are significantly higher than those in randomized networks. These motifs in our network will show the most frequent micro patterns of co-purchasing.

1. **3-node motifs**

Considering 3-node motifs, the first 4 most frequent motifs are as follows in Figure 2-1. From this we could see that there are many pairs of items in the network, which may be very obvious in our daily life, such as bread and butter. What is more interesting is that in many motifs, there is another node leads to one or both two nodes in the pair (as in MotifID 5 and MotifID 8), which might be not so obvious relationship between items. It means that this third node should be noticed more.

**Figure 2-1 3-node motifs**

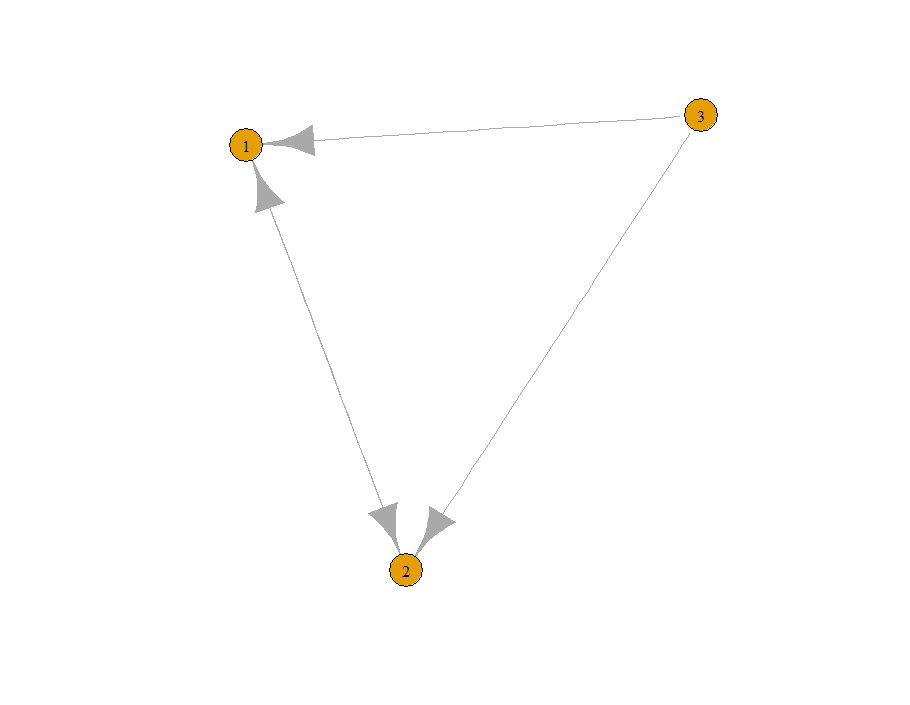
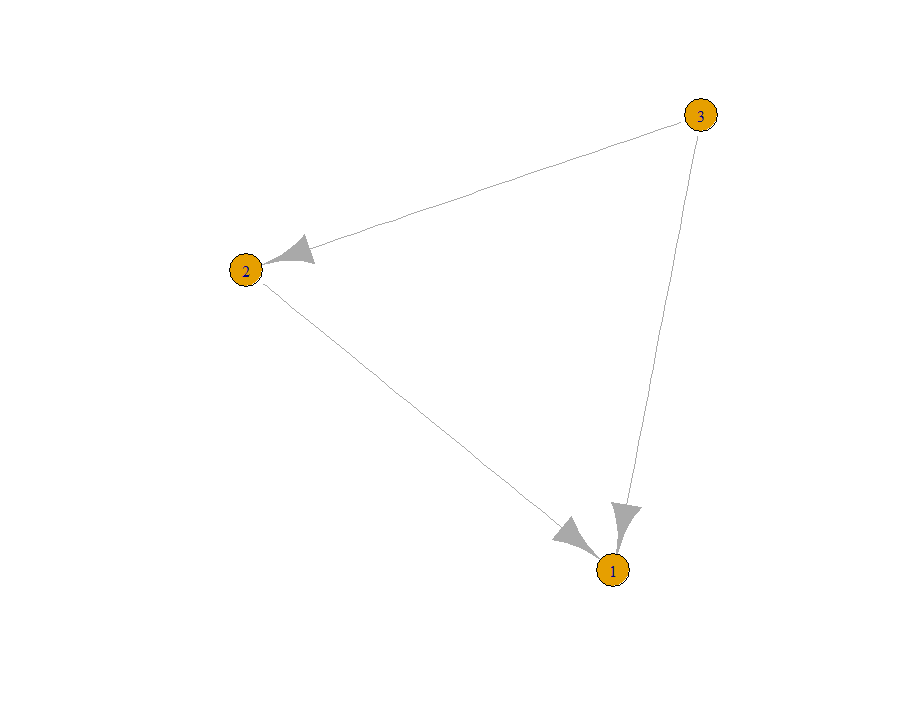
 

MotifID: 5

Count: 2483796

MotifID: 3

Count: 2723405



MotifID: 8

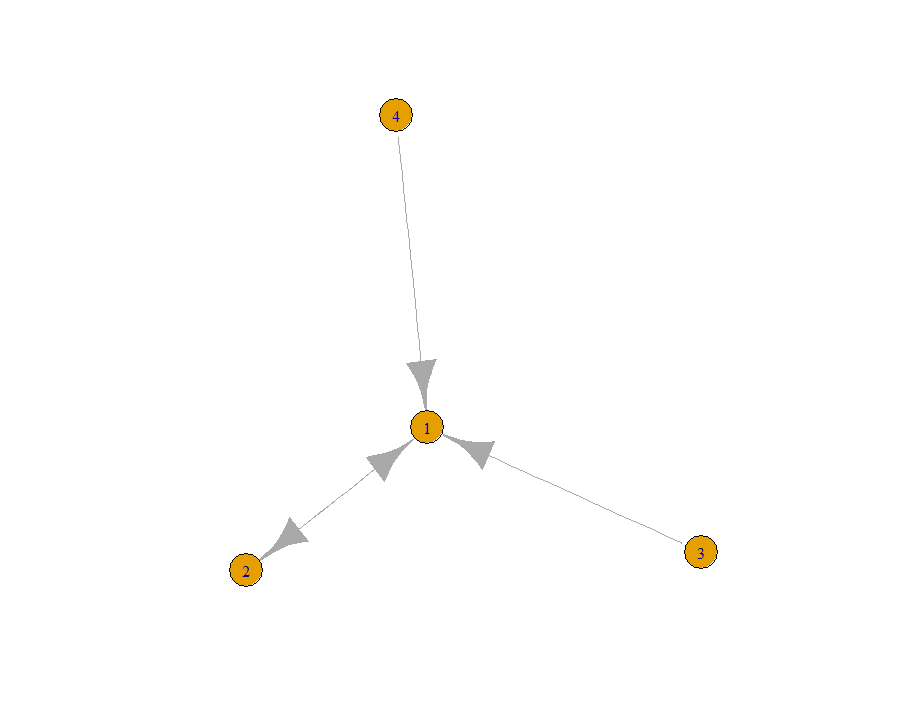
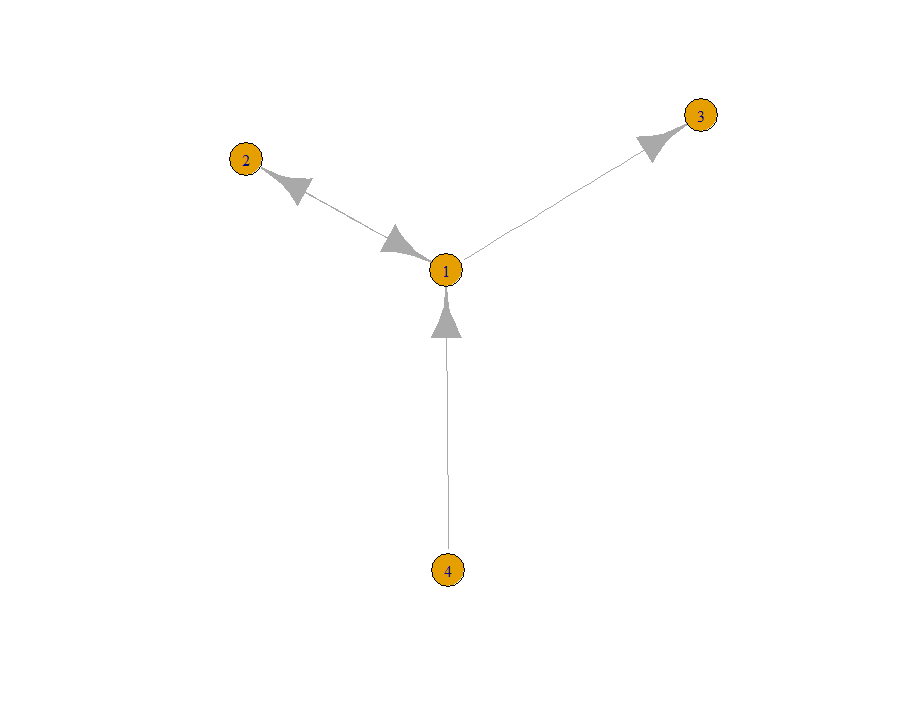
Count: 667129

MotifID: 7

Count: 2543598

1. **4-node motifs**

**Figure 2-2 4-node motifs**

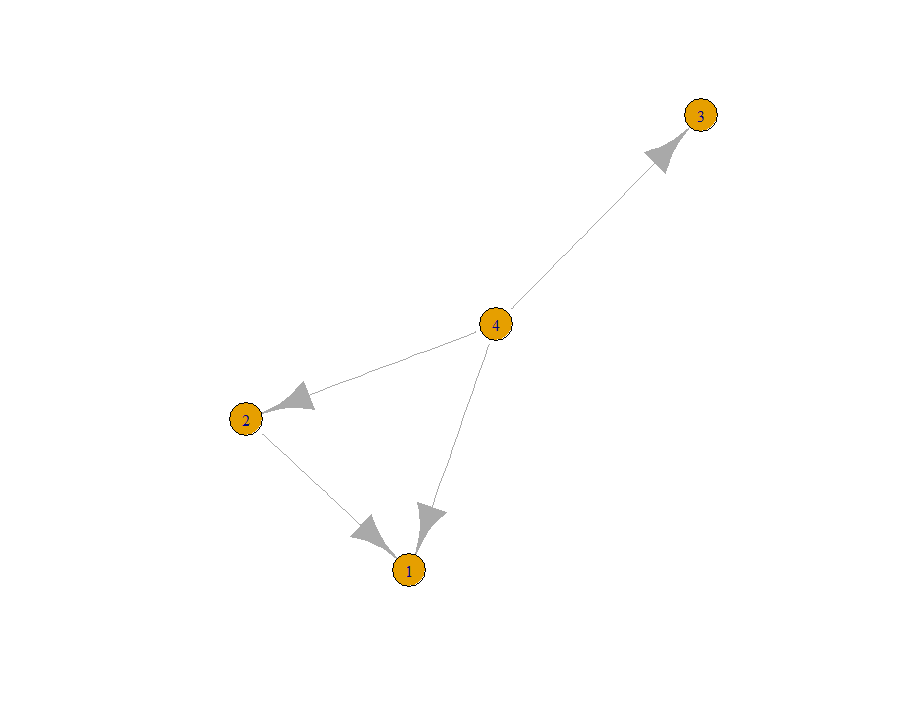
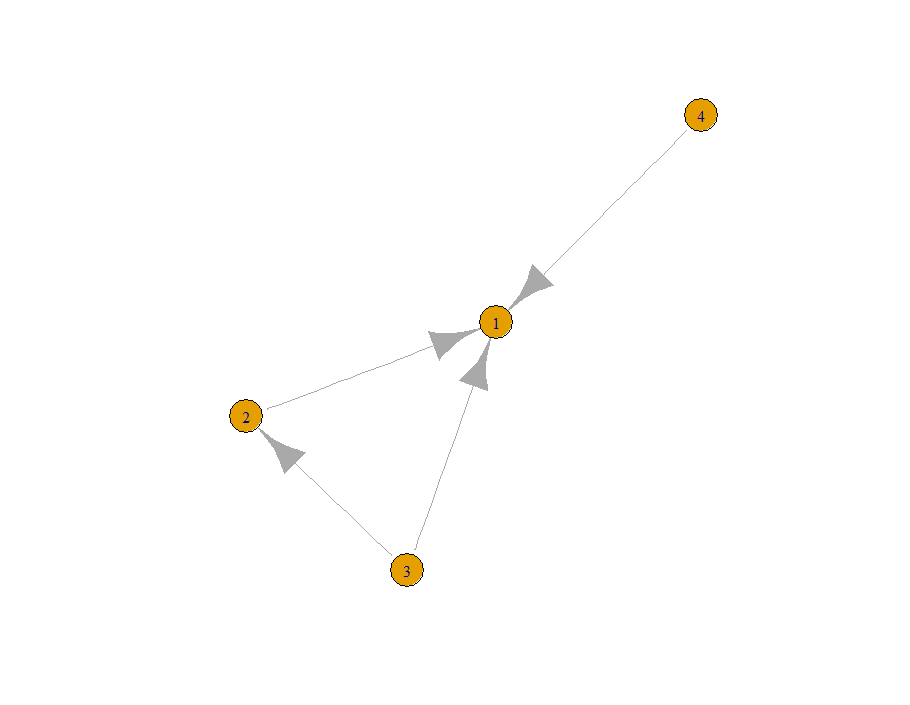
 

MotifID: 25

Count: 23048484

MotifID: 8

Count: 24443779

MotifID: 14

Count: 15695173

MotifID: 77

Count: 23029165

Figure 2-2 shows the first 4 most frequent motifs of 4 nodes (excluding what is the same in 3-node motifs). As in MotifID 8 and MotifID 14, we see there are some nodes in the center of co-purchasing micro patterns. They are “wild cards” in the network, which need more inventory and should give more positions. On the other hand, in MotifID 77 and MotifID 25, the node in the center is a leading item or an important tube to other items, which might be a good candidate to be held in promotion.

* 1. **Macro Pattern - Communities**