

IDS 564: Amazon Co-Purchasing Ground-Truth Communities

Sushaanth Srirangapathi

Kanika Mohpal

Vijitha Vasantha Kumar

Yi Zheng

1. Introduction

1.1 Background

For a long time, retail companies liked using many different techniques to extract information from their mass transactions data, which may help them with affinity positioning, cross-selling and so on. Some techniques such as Association Rules have 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 the network, every pair of co-purchased items is a link/edge between two items and the total "Customer who bought this item also bought" system forms a bigger 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 as a set would be co-purchased items. What's more, it is a good way to provide visualization of the data than other techniques like association rules.

1.2 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 (based on more frequent co-purchasing).

(2) Select solid communities

Then it comes to our primitive goal of this project. There are several algorithms to detect communities in network analysis. Jaewon 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 of the initial measures.

(3) 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.

1.3 Data description

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.

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

Table 1-1

1.4 Network Structure

The network shows the social phenomenon of ‘rich gets richer’ that is used to describe preferential attachment, as shown in Fig 1.4, of earlier nodes/existing product or service in Amazon network, that means these nodes tend to attract more links early on. Because of preferential attachment the product/service that acquire more connections than another one will increase its connectivity that is popularity, at higher rate, and this will increase as the network grows or more people buy or avail product/services from Amazon.

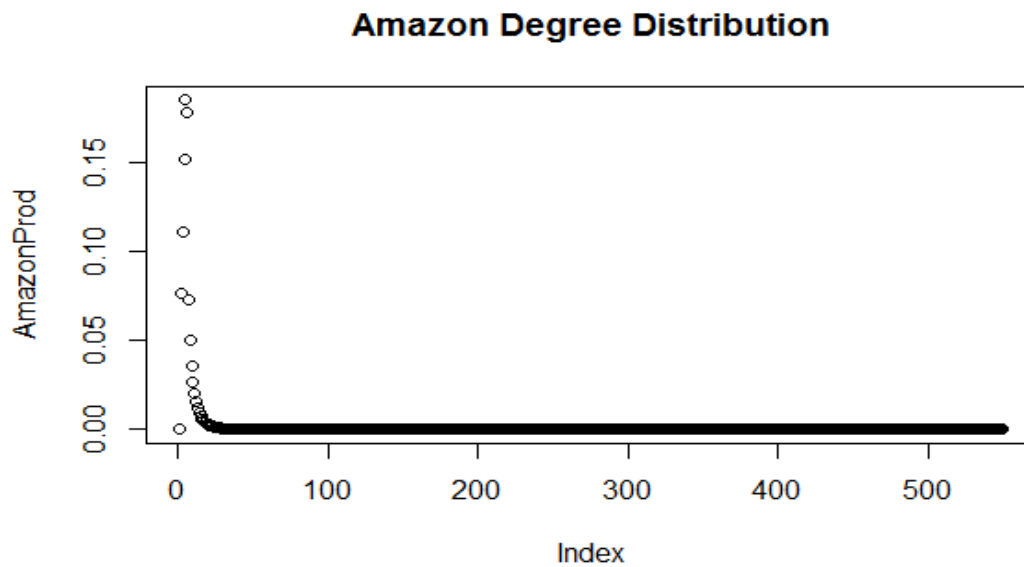


Fig 1.4

2. Patterns of Amazon co-purchase

2.1 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 interactions in biological networks. Milo et al in his article *Network Motifs: Simple Building Blocks of Complex Networks* defined “Network Motif” as patterns of interconnections 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.

(a) 3-node motifs

Considering 3-node motifs, the first 4 most frequent motifs are as follows in Figure 2.1.1. 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.

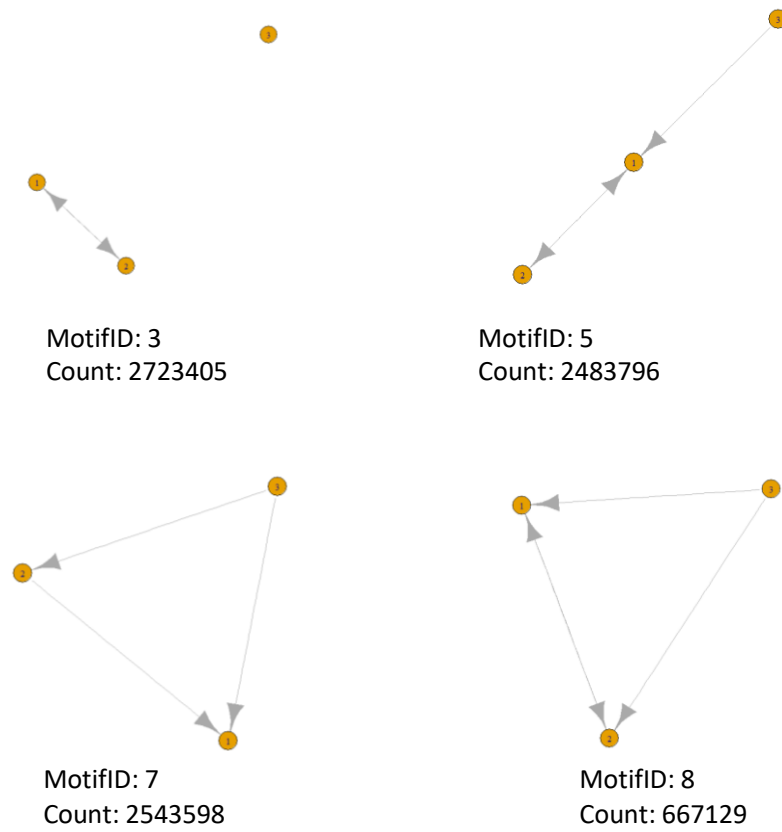


Figure 2.1.1 3-node motifs

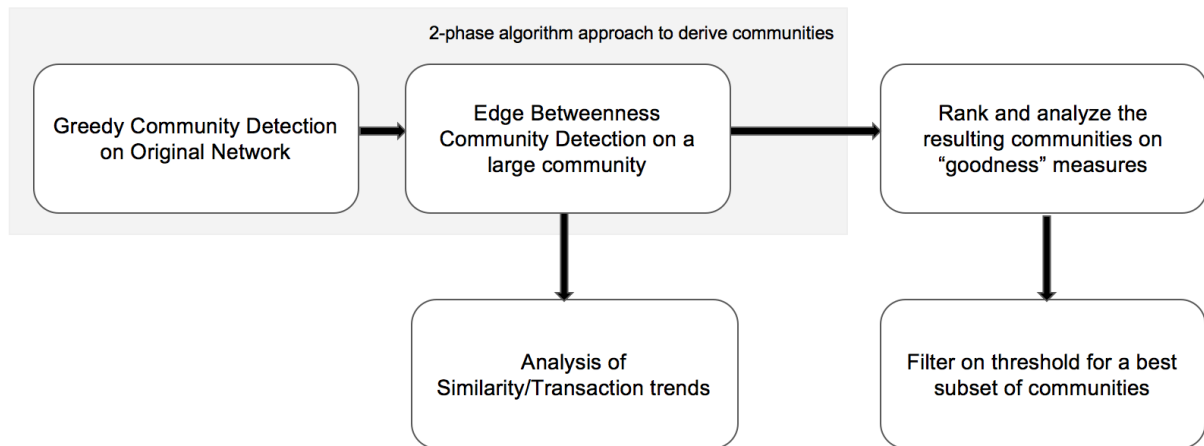
(b) 4-node motifs



Figure 2.1.2 4-node motifs

Figure 2.1.2 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.

2.2 Macro Pattern – Communities Analysis



Flow diagram 2.2.1

2-phase algorithm approach:

We follow a two-phase algorithm, as shown in Flow diagram 2.2.1, that extracts all possible candidates first, and then optimizes the community hierarchy.

1st Phase – Fast greedy community detection:

On our large dataset of ~350,000 nodes and ~1M edges, the first approximation we deploy is a fast and greedy community detection for purposes of speed based on locally optimal modularity scores offered by the fast greedy community detection algorithm.

We then follow this by a series of algorithms learnt in class to find the effectiveness from each approach –

- Edge betweenness community detection algorithm
- Fast Greedy community detection algorithm
- Walktrap community detection algorithm

2nd Phase – Edge Betweenness:

For purposes of detailed analysis, the second algorithm applied on the output of the Fast Greedy algorithm is the Edge Betweenness Community Detection algorithm. We now proceed to score and rank the resulting communities of the algorithm.

Measures of Community Goodness:

There are certain measures which symbolize a well-defined community. A community is deemed 'good' if it is compact and well connected internally while being well distanced from the rest of the network.

Inducing four of the resulting communities into a graph in Gephi, we obtained in Graph 2.2.1



Graph 2.2.2

- The size of the vertex is based on the **degree** of the node
- The color of the vertex is based on the **modularity**

Visually, we find that these measures influence/measure the goodness of a community. What are certain quantities that inculcate degree in their measure? Will they be an appropriate guide in measuring the goodness of a community? With this in mind, we have come up with 5 quantities that measure the goodness.

Scoring function based on External Connectivity – Between Communities:

We will rely on the 2nd algorithm to take care of distancing communities from each other. For this we will use the *Modularity* measure. Modularity is the quality of division of a network into communities.

$$\text{Score} = 1/\text{Modularity}$$

Scoring function based on Internal Connectivity – Within a Community:

Transitivity/Clustering Coefficient

This is the ratio of the triangles and the connected triples of the graph. Connected triangles or triples are perhaps the strongest measure of connectedness in a graph. The more triples there are, the better connected the graph is.

$$\text{Score} = \text{Transitivity}$$

Density

Having looked at triangles (3 nodes), we now look at edges (2 nodes) within the graph. The more edges there are the better connected the graph is. For this, we find the ratio of the number of edges in the graph and the number of all possible edges. This would help penalize some of the sparse and large communities (~40-50 nodes).

$$\text{Score} = \text{Density}$$

Average Path Length

Average of shortest path of all possible pairs of network nodes. The lower this value the highly connected our community is.

$$\text{Score} = 1/\text{Average path length}$$

Diameter

This is the longest shortest path in the graph. This measure is not skewed, unlike the average

path length and hence we have considered this separately as a measure.

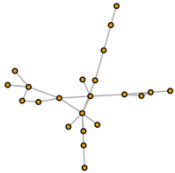
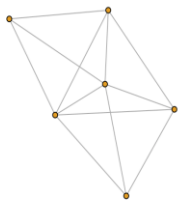
$$\text{Score} = 1/\text{Diameter}$$

Some of the community scoring functions are correlated with each other to a certain extent.

Mathematical formula of scoring function:

$$\text{Score} = \frac{1}{\text{Transitivity}} + \frac{1}{(1 - \text{Density})} + \frac{0.375 * \text{Average Path Length} + 0.625 * \text{Diameter}}{\text{Transitivity} * \text{Density} * \text{Diameter}}$$

In the Table 2.2.3 we compare 2 communities and see how the individual measures impact the overall score. The community 27 on the right is a well-defined community by just looking at it visually. We will see if our results indicate the same, in Table 2.2.3

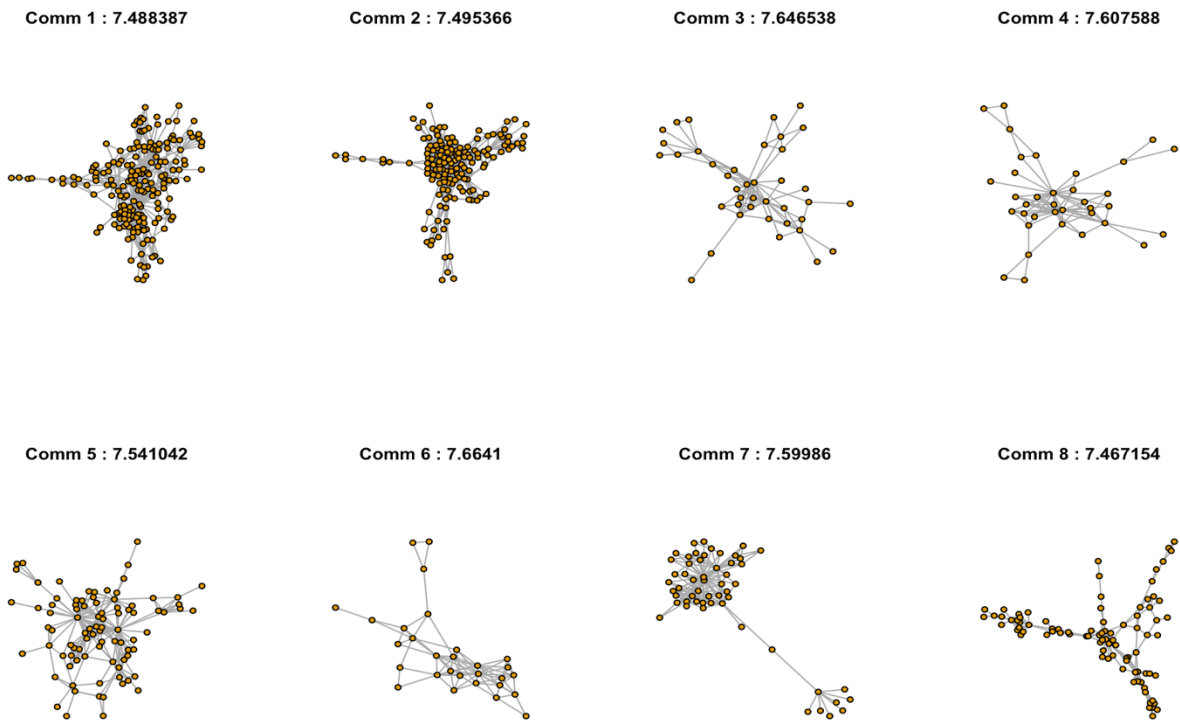
	14	27
Goodness Measure		
Transitivity/Clustering coeff.	0.1154	0.7894
Density	0.1039	0.8000
Average Path Length	3.4500	1.2000
Diameter	8.0000	2.0000

Modularity	0.8640	0.8640
Score	7.4980	8.5020

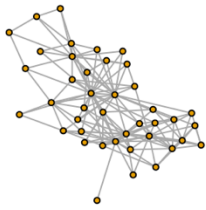
Table 2.2.3

Community 27 on the right has a much higher score than the one on the left as expected.

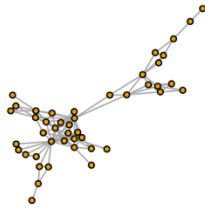
Applying the scores on all the output communities, we get the scores, as mentioned in Fig 2.2.4



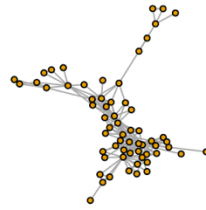
Comm 9 : 7.71849



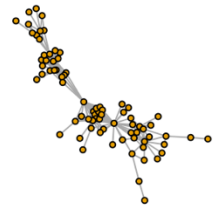
Comm 10 : 7.52402



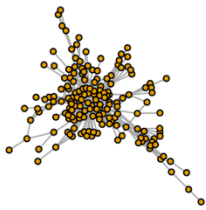
Comm 11 : 7.547265



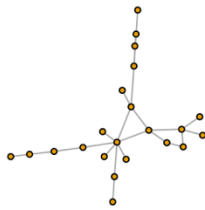
Comm 12 : 7.521097



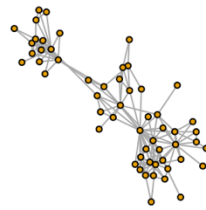
Comm 13 : 7.485349



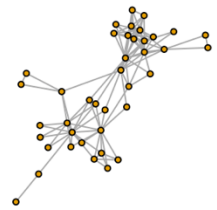
Comm 14 : 7.498081



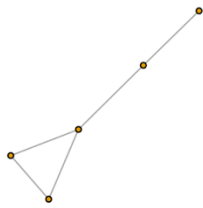
Comm 15 : 7.603508



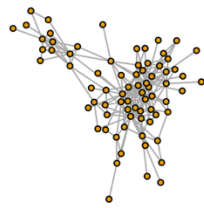
Comm 16 : 7.631884



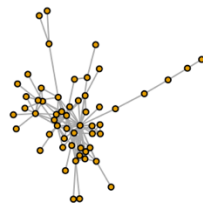
Comm 17 : 7.968338



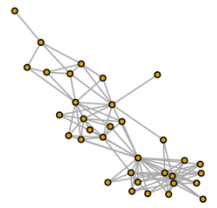
Comm 18 : 7.583303



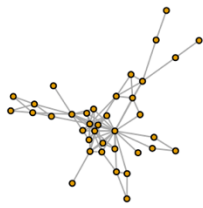
Comm 19 : 7.534367



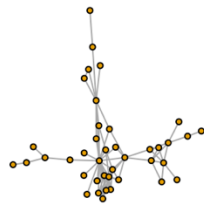
Comm 20 : 7.633517



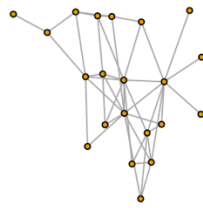
Comm 21 : 7.617618



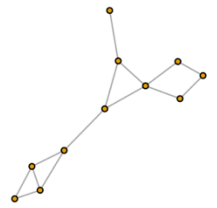
Comm 22 : 7.526915



Comm 23 : 7.665808



Comm 24 : 7.627451



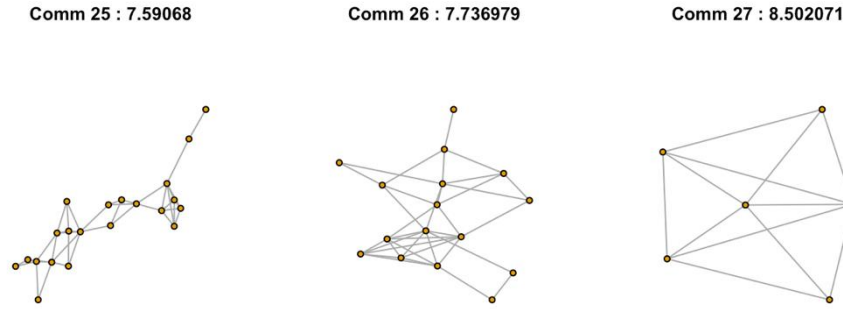


Fig 2.2.4 Scores of the Edge Betweenness Algorithm Communities

Results of Fast Greedy algorithm: Generates 25 communities with the following statistics:

Average score of 7.003106

Max score of 8.666735

Min score of 6.800011

Results of the Walktrap algorithm: Generates 84 communities with the following statistics:

Average score of 6.637386

Max score of 8.051979

Min score of NaN (communities with 2 nodes and a single edge between them)

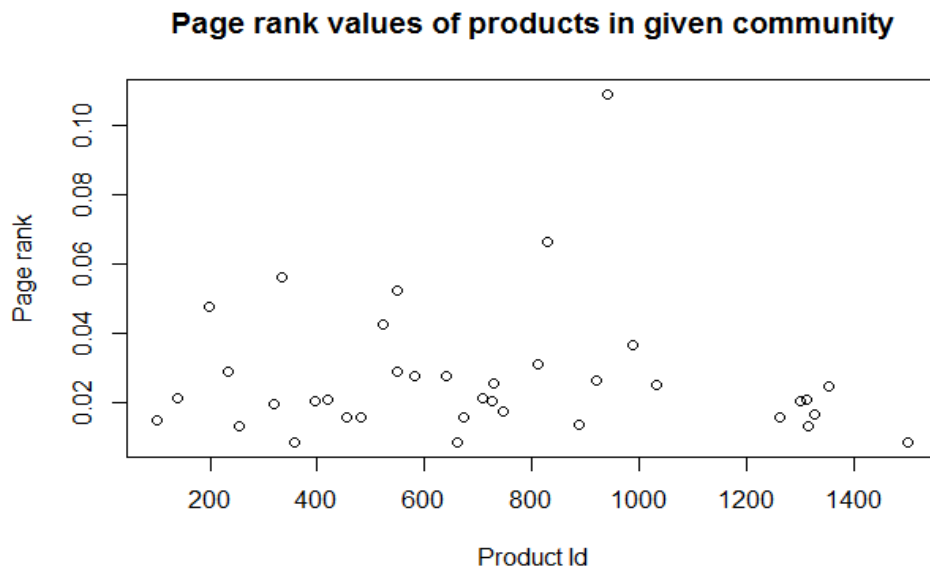
Conclusion:

In large datasets, the standard community detection algorithms focus more on the between community measures to detect communities - Fast Greedy (modularity), Edge Betweenness (edge betweenness), Walktrap (short random walks). We propose taking into account a score on the goodness that measures characteristics within the generated communities to provide a best subset of communities for consideration to the user. Rank the communities on the descending order of their scores and filter on a threshold value to get a list of ground truth

communities.

Now that we have identified communities that are almost pure. Purity here means that the nodes or products within each cluster are similar to each other based on content also called similar items or part of frequent item-set transactions. We are looking at using these pure or almost ground truth cluster to get insights about the nodes based on their network structure alone. If we had metadata about the nodes, we could do a stronger analysis and provide concrete insights that could be used for recommendation and sales transaction updates. Since we don't know metadata about the node, we consider that the clusters formed are based on transaction similarity. Considering transaction similarity makes perfect sense because the edges are formed between nodes only if they have minimum of 3 transactions happening with both the products. Next step for providing recommendation is to rank nodes that are highly near or happens more time with the node in question. For this we use PageRank scoring technique. PageRank is a robust performance measure to use for ranking nodes with respect to each other for popularity based occurrences. Once we have identified the best cluster, we use PageRank as metric to recommend a list of products that are highly important or occurring with the node that is selected.

Algorithm to come up with this list is simple. If we want to provide 5 item recommendation with respect with the node in question on the premise that person who bought this item also bought these, we order the nodes in the community in descending order of their page ranks. Then we give the top most 5 items with high page rank with respect to the node. If the node selected has the high page rank, we recommend the next 5 nodes having highest page ranks among all nodes since the selected node has the highest PageRank.



By using this approach, we can recommend items based on transaction similarity with just the network structure between them and not knowing more about the product in question. Through this we understand network structure helps us get more information and insights even without knowing about the nodes in question or about how the network was formed. Here we knew that network was formed based on transactional similarity. Even if it weren't given, we would have interpreted it in 2 ways:

1. Transaction similarity
2. Node/item similarity

We could use this approach to not just recommend items but also to provide targeted promotions and pricing strategies. For example, if 2 nodes are transactional similar (beer and diaper), we could raise the price of beer and reduce the price of diaper. Hence, we could increase the willingness to pay of the customer by increasing his surplus on spending (through the discount). His increased surplus makes him willing to pay higher for the other item that frequently is purchased with the first item.

Another strategy where this could be used is to arrange the floor space or web page so that

similar transaction items are placed next to each other so that user sees it and hence will get an urge to buy it even if he had forgot about buying it.

Bibliography

[1] J. Yang and J. Leskovec. [Defining and Evaluating Network Communities based on Ground-truth](#). ICDM, 2012.

[2] Milo, Shen-Orr, Itzkovitz, Kashtan, Chklovskii, Alon. [Network Motifs: Simple Building Blocks of Complex Networks](#). Nature, Vol 298, 25 October 2002

Data Source: <https://snap.stanford.edu/data/com-Amazon.html>