Analysis of E-Commerce Product Graphs

Shalin Shah
AI Sciences
Target Corporation
Sunnyvale, CA 94086, USA
`shalin.shah@target.com`

November 4, 2019

Abstract

Consumer behavior in retail stores gives rise to product graphs based on co-purchasing or co-viewing behavior. These product graphs can be analyzed using the known methods of graph analysis. In this paper, we analyze the product graph at Target Corporation based on the Erdős-Renyi random graph model. In particular, we compute clustering coefficients of actual and random graphs, and we find that the clustering coefficients of actual graphs are much higher than random graphs. We conduct the analysis on the entire set of products and also on a per category basis and find interesting results. We also compute the degree distribution and we find that the degree distribution is a power law as expected from real world networks, contrasting with the ER random graph.

1 Introduction

Social networks exhibit strong clustering properties as explained by the principle of triadic closure. Two individuals with a mutual friend are highly likely to form a friendship on a social networking website. A friend of a friend is a friend, and this also creates structural balance.

The clustering coefficient (CC) measures the density of edges in a graph. A CC of 1 indicates that all nodes are connected to each other i.e. the graph is a clique. A CC of 0 indicates a graph of formation of stars in which there are no triangles, and hence no indication of clustering of the nodes.

Assortive mixing in a graph causes similar nodes to connect together. It is our conjecture that products in a particular category tend to mix more as compared to a graph produced using products of all categories. Thus, we expect the clustering coefficient of a graph constructed using products of a particular category to be higher. Also, as compared to the Erdős-Renyi random graph, we expect our product graphs to have higher clustering coefficients as well as a different degree distribution i.e. a power law.

We use one week of data from the logs of Target.com online product views and purchases. We do not distinguish between a view and a purchase.

The work in [1] does some analysis of graphs of retail of a book seller. They analyze both the product graphs as well as consumer graphs, both of which are induced by the bipartite consumer-product graph. We follow a similar approach to construct our graphs. Books such as [2] and [3] have a good coverage on general networks as well as the metrics that can be generated from them.

2 The Clustering Coefficient

The clustering coefficient of a graph is described by the following equation:

$$CC = (3 * T_r) / (T_t)$$

where T_r is the number of triangles in the graph and T_t is the number of triplets (open or closed).

The clustering coefficient is then 1.0 if the graph is a giant clique and 0.0 if the graph only has star formations i.e. there is no clustering of the nodes.

The clustering coefficient depends directly on the number of edges of the network.

3 The Erdős-Renyí Random Graph Model

We compare the results of each category with the [4] ER random graph model $G(n, p)$ where n is the number of vertices and p is the probability of an edge between any of the $2 * \binom{n}{2}$ pairs of edges (in an undirected graph created from directed edges).

So, the probability is then:

$$p = e / (n * (n - 1))$$

where e is the number of edges in the original graph and n is the vertices in the original graph. The graph is then constructed by adding randomly chosen edges from the $2 * \binom{n}{2}$ edges.

The degree of any node is then binomial which is approximated by the Poisson distribution when $n \rightarrow \infty$. This is why the ER random graph model is called the Poisson random graph model.

4 Results on the Product Graph at Target Corporation

4.1 Global Clustering Coefficients of Actual and Random Graphs

Figure 1 shows the clustering coefficients of graphs of some categories. In all but the baby category, the clustering coefficient is higher than the one in the corresponding random graph (with the same number of edges).

Figure 2 shows the probability of an edge in the graph of several categories. The category baby has the highest probability of almost 10% and grocery has the next highest probability. The probability represents the number of co-viewed or co-bought products divided by the total number of edges possible in the undirected graph ($n * (n - 1)$), where n is the number of vertices).

Figure 1: Clustering Coefficients Across Categories

4.2 Predictive Power of the Local Clustering Coefficient

The local clustering coefficient is calculated per node in the graph. It is defined as the number of triangles formed in the neighborhood of a node divided by the number of triplets.

$$CC_{local} = (3 * T_{local}) / (T_{local})$$

Figure 2: Probability of an Edge in the Graph Across Categories

where $T_{r_{local}}$ is the number of triangles in the immediate neighborhood of a node and $T_{t_{local}}$ is the number of triplets in the neighborhood (open or closed).

We evaluated the predictive value of the local clustering coefficient by using it for product recommendations for Target.com. We keep 20% of the consumers as a held out test set and evaluate the recommendations using the hit rate @ position 10. We compare the hit rates with the ones calculated in [5] which notes that the maximum hit rate is 31.1%. We found that using only the local clustering coefficient (one dimension as compared to 250 dimensions), we can achieve a hit rate of 0.7% @ position 10.

4.3 Degrees of Actual and Random Graphs

The average degree on the actual graph is about 40 in all categories within a one week period of views and purchases on Target.com. Figure 5 shows the top six products with the highest degrees on Target.com (for this figure, we ignore all edges that did not occur at least 10 times i.e. at least 10 guests co-viewed the items). This is significantly different from the random graph (Figure 6).

Also, the degree distribution is shown on figure 7. The degree distribution of the actual graph is clearly a power law. The degree distribution of the random graph is quite

Figure 3: Number of Products in each Category

different. (Note that in figure 4, I have truncated the long tail of both distributions at 1000. The distribution of the actual graph goes on to a much longer tail. But the entropy calculation in the discussion section uses the entire 500k nodes).

4.4 Average Path Lengths of Actual Graphs Across Categories

We randomly sample 1000 nodes from each category and compute all pairs shortest paths. The sampling is done to keep all categories comparable and to avoid using infeasible computational resources (similar to [6]). For each pair of nodes in the sample, we compute the shortest path between them and then compute the mean of the shortest paths. The results are shown in figure 7 (Note that all categories includes all items and not just those shown in the figure 7. There are a few other categories that we have not included in the individual category numbers).

The average shortest path lengths do follow the "its a small world" or "six degrees of separation" principles. There is not much variability in the path lengths across categories. Also, we found that the random graphs (with the same number of edges and nodes) have slightly lower average path lengths (similar to what [1] found), but the values are quite similar which is why we do not show them in figure 7.

Just something to note is that electronics and baby have slightly lower average path

Figure 4: Degree Distribution of Actual and Random Graphs

lengths even if they have a very different number of items in the catalog and a very different probability of an edge. This could be because of the high clustering coefficients for both categories.

4.5 Discussion

The probability of an edge (figure 2) and the clustering coefficient (CC) (figure 1) are the important properties of a graph. These two measures have a correlation of 0.85. Both of these measures represent the amount of cohesion in the graph between products. The probability of an edge measures the number of times two products are viewed or bought together, and the CC also measures the amount of cohesion in the products of the graph.

Figure 2 shows the probability of an edge in the graph. Baby, toys and kitchen have a good probability of an edge. Apparel, furniture and home have a higher probability of an edge as compared to all categories, but there is an opportunity to co-promote other items of the same category using recommender systems.

We calculate the clustering coefficient (CC) for both the actual graph and the ER random graph. We expect that the CC for the random graph to be significantly less than for the actual graph, which is the case for all categories except baby. For baby, the CC for

Figure 5: Top Six Products with the highest Product Graph Degrees on the Actual Graph

the actual graph is the highest, which shows that consumers are very active within the category. But the random graph has a higher CC. This could be because the probability of an edge for baby is significantly higher for the baby category. But this also means that there is an opportunity for increasing the cohesion between products for the baby category.

The difference between the CC for the actual graph and the random graph measures the amount of information in the actual graph. For electronics, this difference is the least, which shows that consumers who buy electronics do so cohesively with a much higher amount of online research activity within the category. Same is the case for apparel, furniture and home.

Grocery, kitchen and toys seem to have a good amount of clustering, but if we compare the CC of these categories with the CC of all categories, there is an opportunity to increase the cohesion (especially because the CC of the random graph for these categories is also quite high).

The degree distribution of the actual graph and the random graph is shown in figure 4. The distribution of degrees on the actual graph has a power law distribution, which might indicate that the graph is a scale-free network with significant amount of preferential attachment (the tendency of forming larger and larger clusters of products). The average degree is about the same for the random graph and the actual graph. The distribution of the random graph is centered on the average degree, dropping symmetrically on both sides, which is expected as the edges are chosen randomly. The top six

Figure 6: Top Six Products with the highest Product Graph Degrees on the Random Graph

products with the highest degree for the actual graph are shown on figure 5. Women's clothing appears to cluster the most with other products. Contrasting this with the random graph, this is not by chance, and there is a lot of information in the degrees of the actual graph. If we consider the risk of losing cohesion in the network as products are removed or changed, a scale-free network has a much lower risk as compared to the random graph. Decisions on which items to remove or change and what risk it poses should be made using the degrees of the products in the graph. Care should be taken when the hubs, like the ones shown in figure 5 are removed or changed.

The information entropy of the degree distribution is defined as:

$$I = - \sum_{i=1}^n p_i \log_2(p_i)$$

$$p_i = \binom{n-1}{k_i} p^{k_i} (1-p)^{n-k_i-1}$$

where n is the number of nodes in the graph, k_i is the degree of the i^{th} node and p is the probability of an edge in the network. Entropy is on a logarithmic scale, which means that differences in the entropy should be interpreted as exponentially different. The information entropy measures the amount of information in the network. The information entropy of the degree distribution of the actual graph is 1274 while that of the ER random graph is almost zero ($5.897721e-31$). The difference between the values is important and not the values themselves. In the random graph, the edges are chosen randomly, so there is almost no variability in the degrees which is the reason for the

Figure 7: Average Shortest Path Lengths across Categories

very low entropy.

5 Conclusion

In this paper, we analyzed the product graph at Target corporation, formed using co-viewing and co-purchasing behavior in the consumer-product bipartite graph. We find that the clustering coefficient is quite different across product categories and we discuss some of the findings. We find that the degree distribution and the entropy of the degree distribution are vastly different between the actual and random graphs. The path lengths show that in a random sample of 1000 nodes from the actual graph, the average is less than 5 for all categories implying the small world property that social networks exhibit.

Future work could be to do some analysis on the mixing of categories. Specifically, it might be useful to see how the products cluster when the graph is constructed using pairs of categories. For instance, it might be the case that furniture and home mix really strongly and the CC of the combined graph might be higher than the graphs constructed from the individual categories.

References

- [1] Zan Huang, Daniel D Zeng, and Hsinchun Chen. Analyzing consumer-product graphs: Empirical findings and applications in recommender systems. *Management science*, 53(7):1146–1164, 2007.
- [2] Mark Newman. *Networks*. Oxford university press, 2018.
- [3] David Easley, Jon Kleinberg, et al. *Networks, crowds, and markets*, volume 8. Cambridge university press Cambridge, 2010.
- [4] Paul Erdős and Alfréd Rényi. On random graphs’, math. *Debrecen*, 6:290–297, 1959.
- [5] Shalin Shah and Venkataramana Kini. Hebbian graph embeddings, 2019.
- [6] Adriana Iamnitchi, Matei Ripeanu, and Ian Foster. Small-world file-sharing communities. In *IEEE INFOCOM 2004*, volume 2, pages 952–963. IEEE, 2004.
- [7] Zan Huang. Link prediction based on graph topology: The predictive value of generalized clustering coefficient. *Available at SSRN 1634014*, 2010.
- [8] Yuchen Zhang, Amr Ahmed, Vanja Josifovski, and Alexander Smola. Taxonomy discovery for personalized recommendation. In *Proceedings of the 7th ACM international conference on Web search and data mining*, pages 243–252. ACM, 2014.
- [9] Winfried Just, Hannah Callender, and M Drew LaMar. Clustering coefficients. *Department of Mathematics, Ohio University, Athens. Dosegljivo*, 22(3):2017, 2015.