ADC project report - Group 9

Nuno Correia
Student number 58638, Portugal

Miguel Miranda
Student number 58246, Portugal

Abstract

Product recommendations on Amazon help users find products that are similar to the ones they're viewing/buying. The data set we chose represents nodes as products and edges as a relation that indicates that 2 products get frequently co-purchased together. Our project tries to explore how these products that get frequently co-purchased organize themselves in the biggest product categories that are present in the data. Utilizing various network measures studied in class we attempt to look for patterns in these product communities. The result we found show that products tend to be frequently co-purchased with other specific products and not with most products in their own category.

CCS Concepts

• **Networks** → **Network structure**.

Keywords

Network, Real network, Community, Clustering coefficient, Average path length, Density, Node distribution, Diameter, Assortativity

1 Literature Review

There hasn't been that much work done on this particular data set, as such we've chosen to review the parts of the paper present in SNAP that pertain to the analysis of the com-Amazon data set [2]. In this paper an analysis of community finding algorithms was done as a way to compare how these perform in various networks, and most importantly on the com-Amazon data set. Through an algorithm developed by the paper's authors, they were able to find communities with very high values of modularity [3]. Its worth noting as well that the top 5 thousand communities used in our project were taken from the result of this paper. As for the analysis of the network itself some noteworthy points found by the authors were:

- The network has 334863 nodes and 925872 edges;
- An average clustering coefficient of 0.3967;
- A diameter of 44;

These results were taken directly from the paper [3] and from the SNAP web page [2].

2 Subset of the network

2.1 Procedure

To create a subset of the full chosen network we did the following:

- Load the full Amazon product co-purchasing network with the *read_edgelist* function, from the *com − amazon.ungraph.txt* file, with the nodes representing individual products and the edges purchasing relations between each other;
- Load the top 5000 communities dataset, from the *com − amazon.top*5000*.cmty.txt* file, where each row corresponds to a community. From there split each one to the integer node id's and populate the dictionary of communities by size and the list of community sizes. Lastly, sort those structures;
- Next, we create a set of the top 2000 communities of those sorted top 5000 ones, using that to create a subgraph (*PARTITION_2K*) out of the complete network, which is then saved in a GraphML format;

2.2 Justification

We used this criterion, of reducing the network into a subset of the top 2000 communities, as a way to balance the computational cost of the exploration of the network with the ability to still perform in-depth analysis. As the 2000 top communities are chosen, we can focus on the communities with the biggest impact on the network structure. Comparing the distribution of community sizes of the top 5000 and 2000 communities, we can see they are very similar, ensuring the subset is still representative.

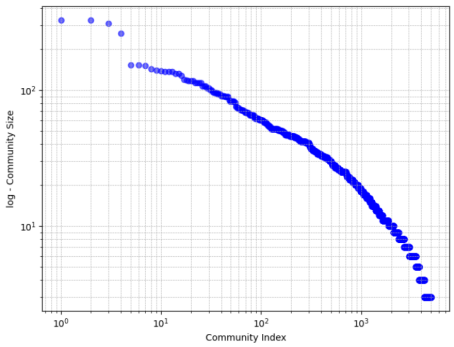

Figure 1: Community size of the top 5K communities

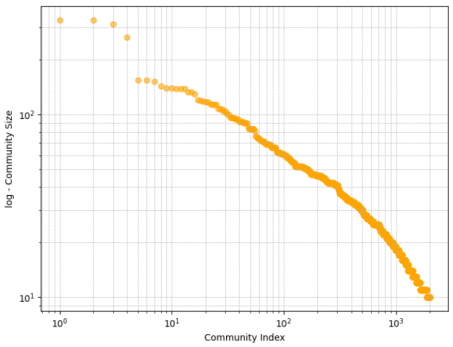

Figure 2: Community size of the top 2K communities

2.3 Graphical Representation

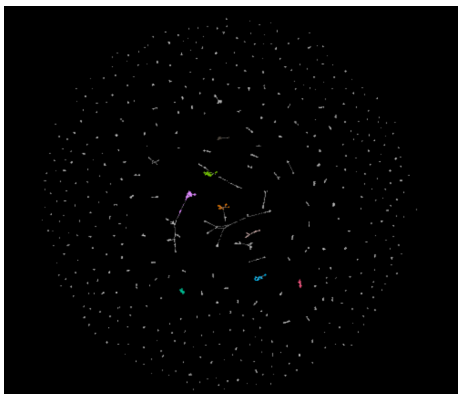

Figure 3: Graphical Representation of the 2k Partition

3 Analysis of the partition

Our entire analysis is based on the hypothesis that communities found on a graph are subgraphs of the network in which they belong. Therefore, we will analyze the communities as separate networks and average out their measures to obtain statistics that represent them. This is also done due to the fact that our partition is very sparse and disconnected. This means that some measures can't be calculated, so to have access to more information we chose the analyze the network this way.

3.1 Analysis of real network properties

To begin our analysis, we want to check whether our partition's various communities behave like typical real networks.

- Degree distribution
- Average node degree per community
- Average clustering coefficient (CC)
- Average Path Length (APL)
- Density and sparsity

We will include interpretations of these measures by themselves and from the point of view of the ground truth.

3.1.1 Degree distribution. The degree distribution of the ground-truth communities is as follows. It increases in frequency all the way until it reaches its maximum value of 3128 and the most frequent degree is 5. After it reaches its maximum value, it decays very fast until it reaches at around degree 20, where it stabilizes and stays constant. This distribution looks about standard for a real-world network.

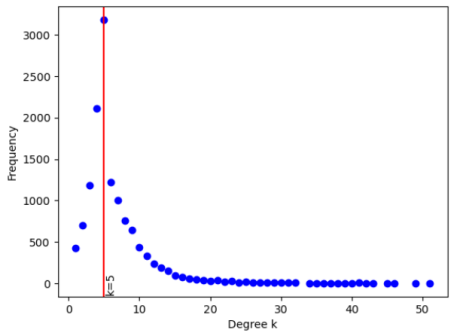

Figure 4: Degree distribution of the communities

3.1.2 Average clustering coefficient. Next, we discuss the average clustering coefficient per community. There is a clear pattern - before the mean, of the average clustering coefficient per community grows somewhat irregularly, until it reaches at around the mean, indicated by the red line, which is $< C >= 0.6366$. After the mean, it dips heavily at around $< C >= 0.6$, increases and then decays quite fast between $< C >= 0.7$ and $< C >= 0.9$. One thing of note is that there are very few communities that are below 0.3. This makes sense, since our partition uses the top two thousand communities in terms of size, that is, we have very big communities, which means that they have more nodes to create connections. Increasing the probability that a node and its neighbor's form.

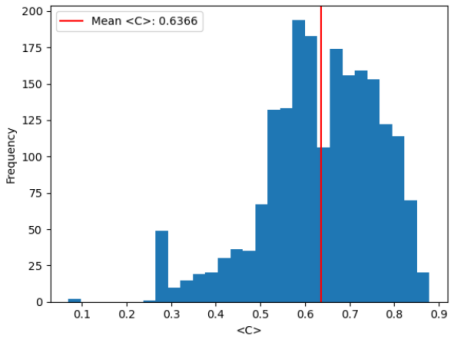

Figure 5: <C> in the top 2k communities

3.1.3 Average path length. The average path length (APL) of the entire network is about 2.0375, this value is indicated by the red line in the graph. Next, we analyse how the APL of each of the communities present in the top 2 thousand communities is represented. The APL values of a large portion of the communities are below the average APL per community, as show in the graph. This result is consistent with the small world characteristic of real-world networks, since the entire network has short paths, and the communities themselves also tend to have small paths.

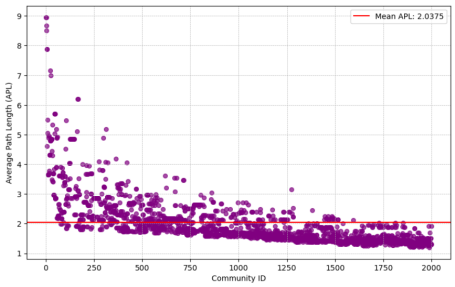

Figure 6: APL in the top 2k communities

3.1.4 Density and sparsity. The density of our entire partition network is density = 0.0005, which is sparse, and we can interpret it as most nodes do not connect to that many others. If we look at the density of all the communities in our partition, we get the following results. All the communities are sparse seeing as none of them have a value higher than 1. The average community density is <density> = 0.3726, while its quite high compared to the network average, it is still very much below 1, so we can conclude that the communities themselves are quite sparse too.

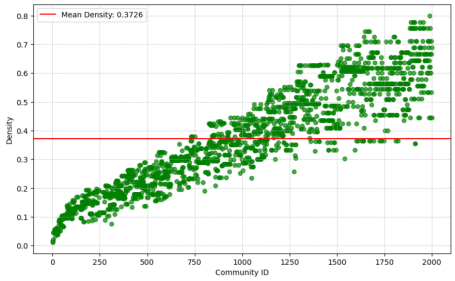

Figure 7: Density per community

3.1.5 Results of real network properties . We know from class that real networks have the following properties: very short paths (low APL), multiple triangles (high clustering coefficient) and they tend to be sparse. Our communities, when examined in isolation, confer all these properties, on average at least. These communities have an average $APL = 2.0735$, this means our typical path only has a distance of 2, rounded down. The average clustering coefficient of the communities is $< C >= 0.6366$, that is on average the communities have many triangles. As for the density, all communities in our partition are sparse by definition. When it comes to the distribution nodes, we only calculated it for the entire partition and not individually. However, the distribution of all nodes in the partition network is quite heterogeneous, growing rapidly until it reaches degree $k = 5$ and then it decays rapidly until it stabilizes and stays there. In conclusion, these communities found on the partition network seem to confer the properties of real-world networks, bar distribution of their nodes per community, that we don't know seeing as we didn't check that.

3.2 Analysis of other statistics

3.2.1 Diameter. The average diameter of the communities is $d = 3.9735$. Therefore, we can conclude that the longest paths are not that longer compared to the average APL of the communities, which is 2.0375. That is, the longest shortest path has one or two more additional nodes in its path. As for the distribution of the diameters, it begins high and then decays extremely fast for values bigger than the average.

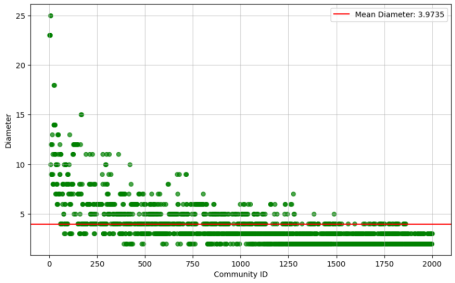

Figure 8: Diameter per community

3.2.2 Closeness. We can see that the average community closseness is 0.5581, with the peak in frequency around that value, suggesting that the communities have a balanced structure, with common central nodes. We can see some communities with low average node closseness, which can indicate communities with a large percentage of peripheral nodes.

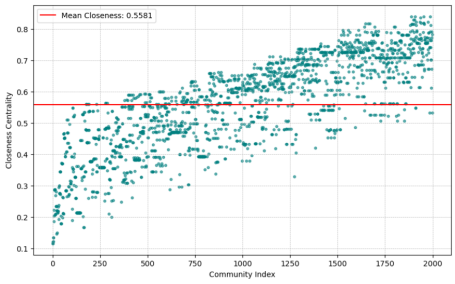

Figure 9: Average node closeness per community in the Top 2k Communities

3.2.3 Betweenness. The mean betweenness of the communities is very low: 0.0488. This is normal, given the skewed distribution of most of the nodes having minimal betweenness centrality and a minority having high values, like in real-world networks. Those nodes, with high betweenness, have a high influence over the spread of information throughout each community.

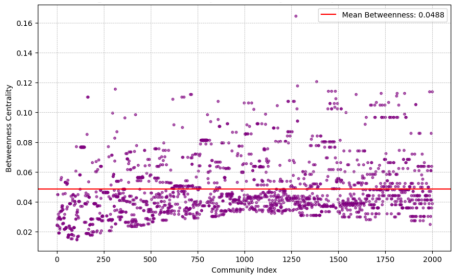

Figure 10: Average node betweenness per community in the Top 2k Communities

3.2.4 Assortativity. The assortativity for the whole network is −0.1018, while the mean assortativity for the communities is −0.2329. From this data we can see that both the networks and the communities are slightly disassortative, i.e. the highest degree nodes tend to connect to low degree ones. This applies in this network, as high-degree nodes such as popular products serve as hubs for less-connected nodes. Despite the communities' nature, there still exists a few communities with assortativity above 0.2, representing tight clusters of similar products.

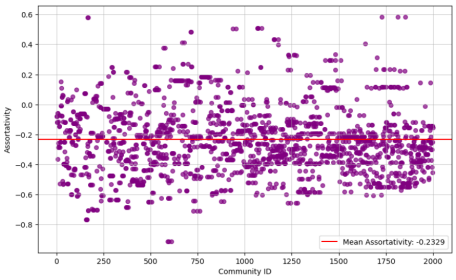

Figure 11: Assortativity per community

3.2.5 Weak ties. Weak ties can help us understand how different communities connect with each other, to better understand this we calculated the number of weak ties and how many different communities they connect. There are only 448 weak tie nodes, out of these 448 only 9 of them connect two other communities to their own. This shows us that the network is very disconnected, most components are isolated and don't connect with each other. Using this, we can say that there is a very low chance that a product from one category will be frequently co-purchased with a product from another category. Putting this into numbers, we have that there is a probability $p = 448/13118 \approx 0.0342$. That is, the probability that a product from one category is co-purchased frequently with a product from another category is around 0.0342.

Non Community Neighbors	Number of Nodes
1	439
2	9

Table 1: Table of Non Community Neighbors and Number of Nodes

4 Statistical analysis of calculated measures

4.1 Correlation between measures

Investigating correlations between node/communities' metrics, we created a Dataframe with the previously calculated values and statistics for:

- Nodes
- Edges
- Average Path Length
- Diameter
- Assortativity
- Average Clustering Coefficient
- Average Closeness
- Betweenness
- Density

Then we analyzed the correlations with values over 0.8, obtaining the following one's:

4.1.1 Correlation between APL and Diameter. Correlation between APL and Diameter at $R = 0.9748$ is the highest correlation. This makes sense, as both metrics are used to measure distance between nodes in a graph, so, generally, communities with higher overall distances will also have larger APL's.

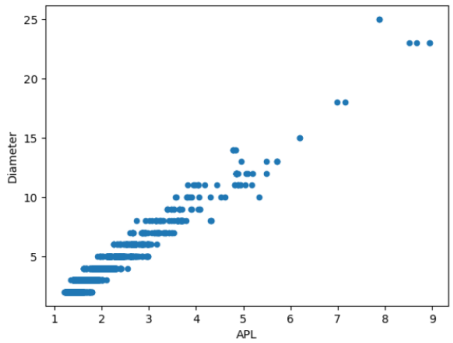

Figure 12: Correlation between APL and Diameter

4.1.2 Correlation between nodes and edges. Correlation between nodes and edges at $R = 0.9543$, as expected, given that generally communities with more nodes have more edges.

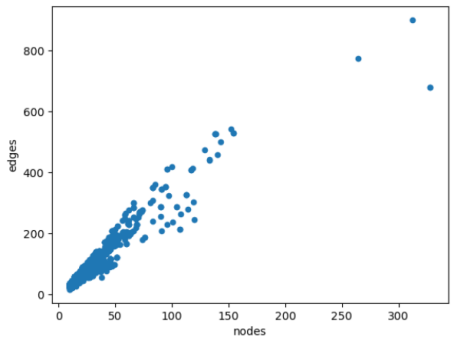

Figure 13: Correlation between nodes and edges

4.1.3 Correlation between Average Closeness and Density. Correlation between Average Closeness and Density at $R = 0.9260$. Communities with higher density have better connected nodes, which reduces the distances to other nodes, increasing the closeness centrality

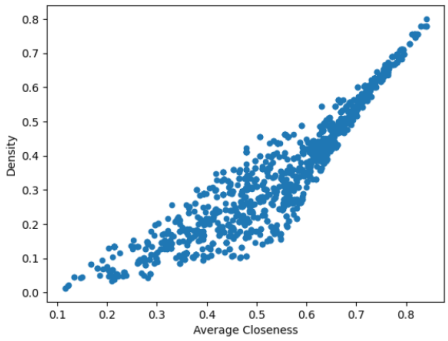

Figure 14: Correlation between Average Closeness and Density

4.1.4 Correlation between APL and Average Closeness. Correlation between APL and Average Closeness at $R = -0.8836$. Here we have a negative correlation, given that the longer a path length is, the further away the nodes are from each other generally, lowering the closeness centrality.

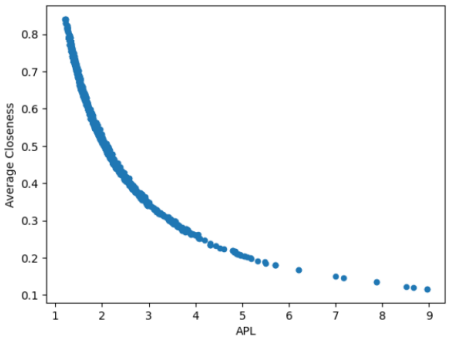

Figure 15: Correlation between APL and Average Closeness

4.1.5 Correlation between Diameter and Average Closeness. Correlation between Diameter and Average Closeness at $R = -0.8560$. Another negative correlation, since communities with higher longest shortest path indicate that the nodes are generally far away from each other.

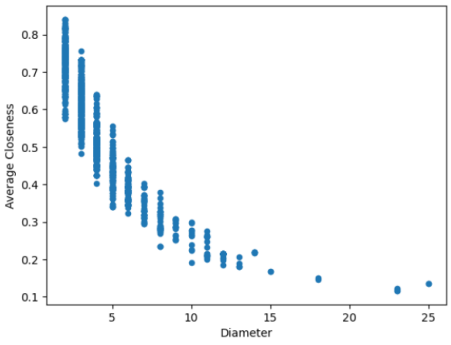

Figure 16: Correlation between Diameter and Average Closeness

4.1.6 Correlation between nodes and APL. Correlation between nodes and APL at $R = 0.8088$, which can be explained by the challenge to keep up an efficient connectivity as the size of the network grows, increasing the average path length.

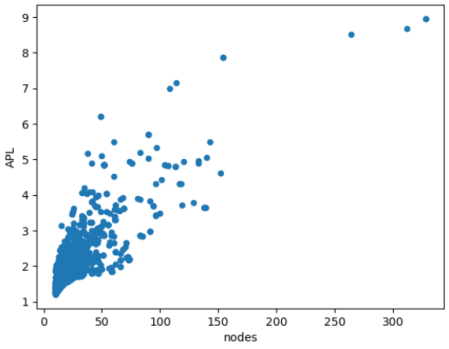

Figure 17: Correlation between nodes and APL

5 Conclusion

Applying the measures that we calculated to the ground-truth scenario we concluded that:

- The average product is frequently co-purchased with 5 other products, this is because $< k >= 5$;
- A product has a probability of 0.6366 of being co-purchased with two other products that are also co-purchased, this is because $< C >= 0.6366$;
- Most products in a category are not co-purchased together, they're only co-purchased with a select number of products. This is because community density is on average $< d >= 0.3726$;
- Most co-purchased products have a degree of separation of 2, since we have $APL = 2$, the "distance" between products tends to be 2;
- On average the path between the least related products is 4 steps;
- In most cases the centrality of each product in its category is balanced, not being to sparse, neither too much connected;

- Popular products generally connect to less frequently co-purchased products;
- The probability that a product from one category is co-purchased frequently with a product from another category is ≈ 0.0342.

All of the code written for this report is available along with all of the images in full size on the git repository created for this project [1].

References

[1] Github repository with the code, https://github.com/miguelmiranda22/ADC-group-9

[2] Webpage of the data set used, https://snap.stanford.edu/data/com-Amazon.html

[3] J. Yang and J. Leskovec, "Defining and Evaluating Network Communities based on Ground-truth," in *2012 IEEE 12th International Conference on Data Mining*, 2012, pp. 745–754. *doi: 10.1109/ICDM.2012.139.*