The goal of the project is analyzing amazon sales dataset[1] and finding the relationship between the products.

## Dataset

The dataset is in CSV format and it contains each product’s information, such as price, name, category, and a list of buyers. Preprocessing was done by three steps using a command "cargo run --bin preprocessing".

1. user\_id column is comma seperated list of ids, so it is splitted into rows.
2. price column is prefixed with “₹” and seperated by comma in thousands, so those characters are removed and final value is converted to float.
3. As the dataset contains products of 8 different categories, price range is too wide. Therefore prices are normalized by dividing with the highest price of the products.

The result of preprocessing is saved as amazon\_cleaned.csv file and it is used for analysis.

## Analysis

Analysis is done in Rust by running a command "cargo run --bin amazon". Firstly, csv file is read and each row is converted to Product struct. Each row are converted to a directional graph using petgraph crate. During the process, each product is added to a HashMap with its id as key to avoid adding identical node to the graph. After that, each product’s buyers are added to a HashMap with their ids as key. This process is done to make it easier to find the relationship between the products.

The graph only contains vertices at the moment. Edges are added by checking each product’s buyers. If a buyer bought another product, an edge is added between those products. The weight of the edge is the number of buyers who bought both products. After building the entire graph, subgraphs are found using petgraph’s connected\_components function. If a subgraph contains less than 3 verticies, it is removed from the graph. The reason behind this is that some products do not share buyers with other products, so they are not related to any other product. In addition, clustering coefficient is calculated for each subgraph. If the coefficient is less than 0.3, I regarded it as a weak relationship and removed the subgraph from the graph. The reason behind this is that the coefficient is calculated by dividing the number of edges between the neighbors of a vertex by the maximum number of edges between the neighbors. If the coefficient is low, it means that the vertex is not well connected to its neighbors. Therefore, it is not related to other products.

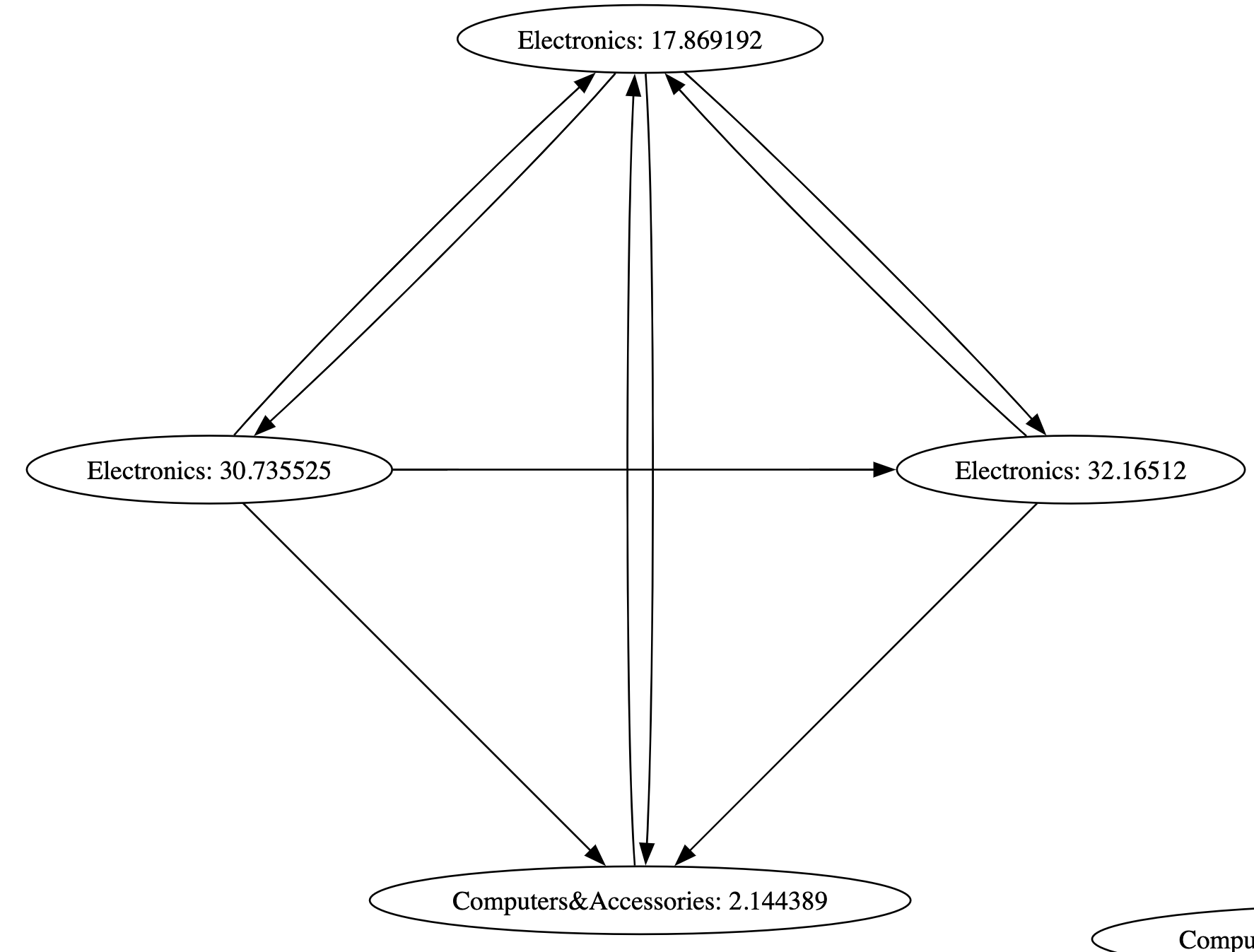
Finally, The graph is saved as amazon\_graph.dot file and it is visualized using Graphviz.

## Result

The result of the analysis is shown in the graph below. The graph is too big to show in this report, so it is saved as graph.svg file. The graph is also saved as graph.dot file, so it can be opened in Graphviz.

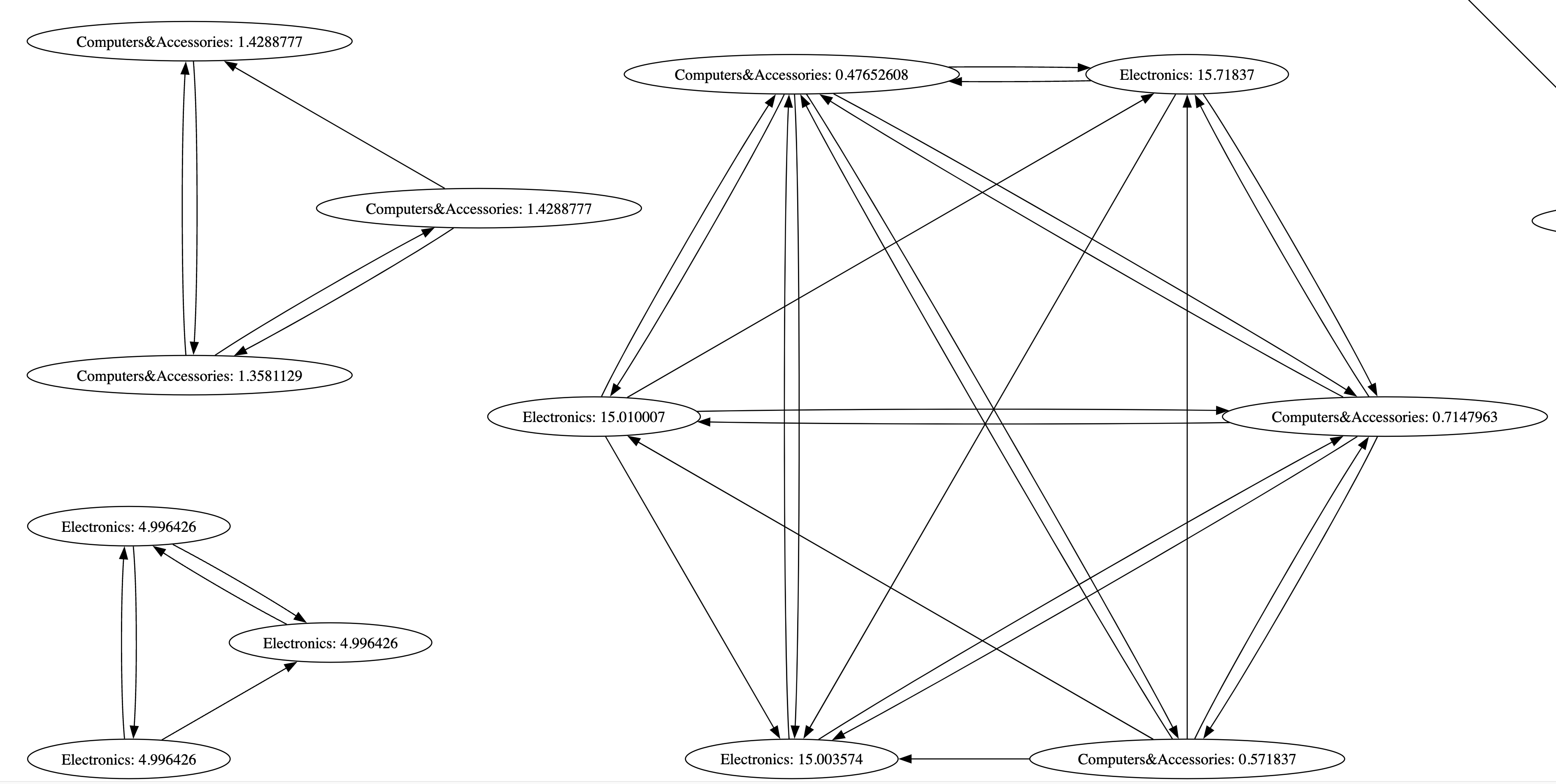
graph.svg

Key findings are as follows:

* Only 2 categories are highly related to each other. They are Electronics and Computers & Accessories.
* Only 2 users bought 3 products in Home&Kitchen category. Furthermore they did not buy any other products from other categories. Therefore they are not included in the graph.
* The largest subgraph contains 10 products.
* There is a tendency in subgraphs that prices of products are similar to each other. For example, the subgraph that contains the most expensive product has the second most expensive product.

Alt text

This tendency is also observed in other subgraphs.



Alt text

The result may be caused by the fact that the dataset is not large enough. As it is shown in the table below, only 3 categories contains more than 3,000 products, which means there is asymmetry in the dataset. Therefore, the result may be biased towards those categories.

category Electronics 4095 Computers&Accessories 3595 Home&Kitchen 3513 OfficeProducts 248 MusicalInstruments 16 HomeImprovement 16 Toys&Games 8 Car&Motorbike 8 Health&PersonalCare 4 Name: count, dtype: int64

[1] (Dataset source)