Project Proposal: In-depth study of Amazon product flow

Darren Cheng, Steven Zhu, Ayman Momin

Nodes and links:

In this nuanced network, different types of products sold grouped by their geographical regions are considered to be the nodes. There will also be one node for each region to act as a distribution center and one node for each region to act as the sum of total sales. The links connect the products sold in each region to the distribution center, creating distribution channels.

Dataset:

For the dataset, in-depth research was done on Kaggle, and a dataset regarding Amazon sales during 2020[1] was found. This data set has attributes of focus, including product names, corresponding categories, selling prices, development costs, quantity, and brand. Further three more attributes of region of sale, center of distribution, and customers alongside product names form the components needed for the nodes and links of the network. These three attributes will be randomly generated and created as a part of the dataset.

Once the dataset is ready, Python libraries such as pandas and networkx will be used for data cleaning, visualization, and analysis. This step is crucial in performing the trend analysis section of the general marketing analysis, which would include product sales by time (month, quarter, year). In addition, customer behaviors will be studied to showcase the popularity of each product within different customer demographics and groups. Problems can be caused if the data generated does not follow a uniform distribution in the previous parts, which can cause bias.

The final network will be built in Gephi, presenting a great overview of the goods flow diagram. It would show clearly where goods are being distributed and sold, in what quantity and monetary amount. This network will be backed with further micro-analysis of customer purchase behaviors using Python, in the section above. It is vital to create a reliable dataset for Gephi to correctly draw out the network as intended, which is a process that can be done wrong.

Expected size of network:

The expected size of the network would be medium-scale. A rough estimate would be around 1,000 nodes with 3,000-5,000 edges. Not all data is available, so the region, distribution centers, and users need to be generated. The plan is to create a subset of nodes, the different product types sold, under the large singular sales sum node for each region.

Questions:

The most important question being explored in this project would be the sales volume of different products by region, and its correlation to the distribution of these goods. The second important question to explore is the marketing segmentation and trend analysis done with Python, as support for the network. The first question offers an in-depth overview of how the supply chain can be optimized to better manage resources, whilst the second optimizes marketing procedures, making sure more goods can reach as many customers as possible.

Citations for the dataset:

Amazon Product Dataset 2020. (2020, September 25). Kaggle.
https://www.kaggle.com/datasets/promptcloud/amazon-product-dataset-2020