Project Proposal

Leading Question

a. Our goal for this project is to analyze the correlation between why certain products are purchased after one another. In other words, we want to see if someone will buy a certain product given that they purchased another product in the past. We also want to see if this same trend exists among a large group of people wherein many individuals all buy one specific product and all buy another specific product later on.

Dataset Acquisition

a. We are using the Amazon product co-purchasing network dataset to analyze purchasing habits of consumers. This dataset shows if a product ‘i’ is frequently co-purchased with a product ‘j’ and the graph contains a directed edge from ‘i’ to ‘j’. The data was collected on March 2nd, 2003. http://snap.stanford.edu/data/amazon0302.html

Data Format

a. The data is sorted by ID #’s, each ID has title, group, salesrank, similar products, and categories it would be classified as. The source of the data is pulled from amazon. The input format of the dataset is a .txt file. The data is roughly 1 gigabyte with roughly 550,000 IDs. We plan on using all the data in order to create accurate predictions.

Data Correction

a. We will parse this input data by making sure to split each line and the whitespace in each line. Each line should only consist of two numbers, which should be checked by splitting whitespace and attempting to save them as numbers. If there are missing purchases for an ID, that line will be skipped and it will be assumed that the person did not buy anything there. There should not be any outliers for numbers, but if there is a non-integer anywhere, that line should also be skipped in order to not cause errors in the graph.

Data Storage

a. The data will be stored in a directed graph inside the code. A directed graph is one where nodes exist, and then the edges themselves have directions. For example, a node A can be connected by an edge to a Node B, but Node A points to node B, and Node B does not point to node A. Say the number of vertices in the graph is represented by a variable V, and the number of edges in a graph is represented by a variable E. The runtime and storage for the dataset would then be O(V + E) for BFS and DFS traversals on the directed graph. Auxiliary data structures will be required for the BFS and DFS traversals. The DFS traversal will require a stack data structure to successfully complete the traversal, and the BFS traversal will require a queue data structure to successfully traverse the graph.

Algorithm

a. After parsing through the Amazon co-purchasing dataset, we plan to to create a graph to hold user IDs and the items that specific user bought. Then, using a DFS traversal along with an iterator, we can traverse our new graph and visit the nodes that represent each individual user. Once we have done that, we can utilize Djikstra's algorithm to find a correlation between different users who share common purchases. We can do this by establishing a correlation coefficient of sorts to see how close different users' purchases are to each other given that they purchased a common product in the past.

Timeline

a. November 7-11 - set up all skeleton functions including all the algorithms (DFS, Djikstras)
b. November 14 - 18 - complete DFS algorithm and complete test cases,
c. November 21 - 25 - complete Djikstras along with test cases
d. November 28 - December 2nd - Ensure graph outputs properly along with all deliverables
e. December 5 - 8 Ensure all files and documents are properly uploaded to the git, Complete the final presentation and project report, ensure all steps of the final project are completed, We will be making sure that we are fulfilling all requirements as we go but we will also conduct one final check at the end