Leading Question

Analyzing amazon data to find the most popular product and how one product is similar to other.

Dataset Acquisition :

Data Format

We have acquired the amazon data (amazon0302) from the 'Stanford Large Dataset Collection,' http://snap.stanford.edu/data/. Each product is a node and they are liked by the frequency with which they both are bought together. There are 400,727 products in this dataset. We also plan to use the entire set of data. The graph had directed and unweighted nodes.

Data Correction

The data has been collected from a credible Stanford university website in 2012. If there are any missing entries in the data we will delete them or in the worst case assign the entries as a NULL value.

Data Storage

We will be using the matrix as our primary data structure. Our matrix storage time complexity will be O(n2)

Algorithm

Our program focuses on the implemenation of the Pagerank and Strongly Connected Components Algorithm. Pagerank algorithm will allow us to find the most popular product by counting the number and quality of links to a product. This algorithm has a time complexity of $O(E \cdot n)$ where E is the number of edges and n is the number of iterations. We use Kosaraju's BFS based strongly connected component algorithm on the directed graph to find how likely a set of products are to be bought together. To traverse through the data we will be using BFS. This algorithm to find the distance between the two products. The time complexity of BFS and strongly connected components algorithm is O(m+n). Where m stands for the number of vertices and n stands for the number of edges.

Timeline

week 1: Since we have now finalized on our idea, we will now be storing the data in a matrix (our primary data structure). we also beign working on the BFS traversal

week 2: By the second week we would have successully implemented a breadth first traversal (BFS) and work on the strongly connected components algorithm. We will also write test cases for debugging our code.

week 3: In this week we will be coding the Pagerank algorithm. Along with making the appropriae test cases to see how effiecent our code is.

week4: in the final week we will work on the test cases and just polising up the code. We will also work on the written and video report of the project.