DATA 620 PROJECT 1 PROPOSAL

Group Members:

- Dhanya Nair
- Dirk Hartog
- Gillian McGovern

Title: Amazon Product Co-Purchasing Network Analysis

Research Question:

What is the purchasing pattern of Amazon customers?

Motivation:

This research question allows us to investigate how Amazon recommends their products to customers and gives us experience comparing centrality measures across categories of Amazon products.

Project Goals:

- Identify Amazon's most popular products
- Identify which products are more likely to lead to additional purchases, and therefore, should be marketed by Amazon
- Compare centrality measures across Amazon product categories
    - Explore which categories have frequently co-purchased products

Data Sources:

Amazon product co-purchasing network, March 02 2003 from Stanford Large Network Dataset Collection. Please see dataset link here.

Network was collected by crawling Amazon website. It is based on *Customers Who Bought This Item Also Bought* feature of the Amazon website. If a product i is frequently co-purchased with product j, the graph contains a directed edge from i to j.

The data was collected in March 02 2003.

Dataset statistics	
Nodes	262111
Edges	1234877
Nodes in largest WCC	262111 (1.000)
Edges in largest WCC	1234877 (1.000)
Nodes in largest SCC	241761 (0.922)

Edges in largest SCC	1131217 (0.916)
Average clustering coefficient	0.4198
Number of triangles	717719
Fraction of closed triangles	0.09339
Diameter (longest shortest path)	32
90-percentile effective diameter	11

Source (citation):

J. Leskovec, L. Adamic and B. Adamic. The Dynamics of Viral Marketing. ACM Transactions on the Web (ACM TWEB), 1(1), 2007.

EDA:

- The amazon dataset has 262,111 nodes which could represent products or entities like users or pages.
- It has 899,792 Edges which are possibly co-purchases, links, or recommendations.
- A Minimum Degree of 1 indicates there are nodes connected to only one other node — maybe niche or isolated items.
- A Maximum Degree of 420 suggests that it's a hub node and possibly a very popular item frequently bought with others.
- Node '14949' is the most connected product, with 420 edges. It likely represents a very popular product that co-occurs with many others in customer purchases.
- Only 20 of the 77 nodes form the largest connected component (LCC).
- The diameter which is the Longest shortest path between any two nodes in the LCC is 7 hops. For a 20-node network, this is relatively sparse, suggesting the need for longer paths between important nodes
- Average Clustering Coefficient: 0.0766 suggests a hub-and-spoke or tree-like structure rather than a tightly-knit community.

Planned Workflow:

- Read in the data (text file stored on GitHub)
    - Create subset of data
- Read in metadata
- Basic analysis
- Visualize the network
- For each of the nodes in the dataset, calculate:
    - Degree centrality
    - Eigenvector centrality

- Compare centrality measures across Amazon product categories
- Present findings (Jupyter Notebook report + GitHub)

Team Work:

Each team member will handle part of the workflow: data set up, graph visualizations, network metrics & comparisons, findings & conclusion