CS 277
Project Report

The project aims to discover if there is an association between the Sales Rank of a product and its betweenness centrality measure when the product is stored in a graph network. The Graph dataset contains products as nodes and Product A linked towards (Outwards edge) node B when B is bought after product A was purchased. Here Sales rank is a metric at Amazon.com that explains the relationship among products within 1 category based on their sales performance[1]. The quantification of the above association that the project aims to discover solves an important issue in fields such as digital marketing, customer analytics where companies around the world spend millions of dollars to improve revenue. Though identifying popular and important products, they would be able to optimize their marketing budget improve the visibility of products that increases customer spending and thus improve profitability.

Through the project, I was able to establish that the betweenness centrality of the product is indicative of the product's sales rank or its popularity. From the results, The scaled average Betweenness Centrality of Popular products (low-ranked) is 1.82e-05 whereas the average scaled betweenness centrality of products that have a high rank and are unpopular is 9.51e-08. The ratio of Betweenness Centrality for unpopular to popular products comes out to be 191.5. The results were in line with the initial expectation that I had before initiating the project which was that products that may have high betweenness may have considerable influence within a network by virtue of their control over information passing between others.

	Dataset 1 - (Popular)	**Dataset 2 - (Medium)**	**Dataset 3 - (Not Popular)**
Sales Rank Range	59 < Rank <25833:	20830 < Rank < 33855	48460 < Rank < 65624
Betweneness Centrality	1.82E-05	2.00E-06	9.51E-08

Figure 1: Results Table

While there were different approaches to simulate the algorithms on the dataset. I selected the approach where I segregated the overall data into three subsets. These groups were derived from the complete dataset using sales rank. The first subset contains the most popular products (500) that have sales rank between 59 and 25833, 2nd group has been named as Dataset 2- a medium that contains products that have popularity in the intermediate level and the last group Dataset 3- Unpopular Products that have high sales rank between 48460 and 65625. The rationale behind choosing these ranks was that these ranks provide the topmost, middle, and bottom products

when the dataset is sorted on Rank. It is also to be noted that nodes or products that do not have an edge or co-purchased products have not been added to the graph.

To benchmark and verify that the implemented algorithm was running correctly, I referenced an example[2] of a directed graph provided on the neo4j website. neo4j is a Graph data platform used in the industry. The results obtained on a small network are in concurrence with the standard library's output.

Conclusion & Future Scope:

Through the project, I was able to discover the association between Prododucts's popularity and the betweenness centrality. As the popularity of the products improves the betweenness centrality measure also improves. Further scope of studies would be on quantifying this relationship between the above two quantities.

References:

1. https://sellics.com/blog-amazon-sales-rank/
2. https://neo4j.com/docs/graph-data-science/current/algorithms/betweenness-centrality/