Araba Aidoo
December 9, 2024
Kontothanassis
DS210

Introduction to Facebook Social Network Analysis Project

My project centers on exploring this dataset to investigate the "six degrees of separation" theory within the Facebook social network. This theory posits that any two individuals are, on average, approximately six steps away from each other in a social connection chain. By analyzing the average shortest path length between nodes in the Facebook graph, I aim to measure how closely people are really connected to each other. The Facebook Social Network dataset, by Stanford's SNAP (Stanford Network Analysis Project), presents a compelling opportunity to delve into these dynamics. This dataset includes a social graph that maps user connections, making it a real-world example of social interactions and relationships.

The relevance of this study extends beyond academic interest, it has practical implications in various fields such as marketing, recommendation systems, and digital communication strategies. Insights from this analysis could help understand how information spreads within networks and how influencers can impact their communities. Moreover, my current internship is exploring the implementation of a following system between users and its potential effects on donations. This project will enrich that research by providing a deeper understanding of the interconnectedness inherent in social platforms.

Understanding social networks' structure and the typical distances between users can validate the small-world properties hypothesis, which suggests that most nodes (people) are connected by only a few intermediary steps. This exploration is not just about validating a mathematical concept but about uncovering the subtle nuances of human connections in the digital age.

Implementation

In my project, I delved deep into the network’s structure by parsing a dataset using a text file named facebook_combined.txt. This file starts off by listing the number of nodes, followed by subsequent lines that represent the connections between users, with each line denoting an edge in

the graph. I chose to represent this social network as an adjacency list because of its efficiency in managing sparse structures. This method allows quick access to any node's direct connections.

To handle large datasets effectively and ensure the performance isn't hindered by the size of the data, I adopted several optimization strategies. For instance, I implemented random sampling where the program randomly selects nodes to estimate the average number of direct connections each user has. This approach provides a broad view of the network's connectivity without the need to analyze every single node.

Additionally, I utilized Breadth-First Search (BFS) algorithms extensively. This choice was driven by BFS's capability to traverse wide sections of the graph efficiently, which is perfect for calculating the network's diameter. This measure gives us a clear picture of the network's reachability, so that I could assess how far apart individuals are within this social space.

Outputs and Analysis

The project generates several insightful metrics to understand the network's structure. It displays the number of connections for a random sample of 500 nodes, illustrating the range and distribution of user connectivity within the network. This includes an average degree of connectivity, which provides insights into the overall integration of users within the network. The clustering coefficient is measured to determine how likely it is that two acquaintances of a user are connected, suggesting the presence of tightly-knit communities. Additionally, the graph diameter, which represents the longest shortest path in the network, offers an idea of the network's breadth, and the identification of the most connected user highlights potential influencers within the network.

In my project, I've been able to uncover some fascinating patterns about how users connect and form communities. For instance, by examining a sample of 500 users, I noticed a wide range in the number of connections each user has. Some users are super connectors or hubs with a vast network of friends, while others have just a few connections. On average, each user has about 22.804 connections, which points to a well-connected network where information can travel quickly and effectively.

The clustering coefficient for the network is around 0.2280, indicating a moderate probability that any two friends of a user are also friends with each other. This typical feature of social networks suggests that users tend to form tight-knit groups, which helps in understanding how cohesive the network is and how social groups are structured within it.

The graph's diameter is 17 steps which means that the network, while expansive, still allows for relatively swift connections between individuals, aligning with the "six degrees of separation" theory commonly associated with social networks.

The most connected user in the network has 1,043 connections suggesting that this person is an influencer of some sort. These influencers are crucial not only for the robustness of the network but also show how vulnerable the network could be if such key nodes were targeted or removed.

The combination of these findings supports the small-world phenomenon typical of many social networks where despite large sizes, the networks maintain short path lengths and high clustering. This structure not only enhances the resilience of the network against random user drop-offs but also exposes potential vulnerabilities to targeted disruptions.

How to Run

You can clone this repository using **git clonehttps://github.com/araba13/210_final_project/.git** then navigate to the directory using **cd final_project** and finally I would suggest using **cargo run – release** to compiles the project and executes it.

Source

https://snap.stanford.edu/data/ego-Facebook.html