Final Report

Introduction

The dataset is from Facebook and consists of “circles” and “friends lists” collected through a Facebook application utilized by survey participants. It includes 4039 node features representing user profiles, 88234 circles indicating social connections, and ego networks reflecting individual users and their connections within the dataset. To ensure privacy protection, the IDs assigned to each user have been replaced with new identifiers. Moreover, the interpretation of the features has been hided. Therefore, while it is feasible to discern similarities between users' attributes, the exact nature of these attributes remains unsure, preserving the privacy of individuals involved in the dataset.

This project aims to observe individuals' connection to others on social media and provides insight into complicated human relationships, societal structures, and online behavior. If users are in the same friend circle, they might share the same interests, engage in similar activities, or have common social affiliations. This project explores the intricate dynamics of Facebook friendships, with a focus on understanding the phenomenon of "friend-of-friend" relationships. Utilizing a dataset comprising 'circles' or 'friends lists', I analyze the connections that define the social graph of this digital ecosystem.

Data Analysis

Proportion of friends’ friend

In the dataset, the proportion of friends' friends that are also your friends is 0.514302596286771. This number is relatively high, indicating a significant overlap in friendship connections within the social network. This tells us that Facebook users in this dataset tend to have mutual connections with their friends.

A higher proportion indicates a more closely connected network, which allows for efficient social interactions and information sharing. This means that the average distance between individuals in the network is likely to be shorter, as information and messages can spread more quickly from one user to another. The network can function more effectively with fewer intermediaries required to relay information. This will be further discussed in the “path length” section.

In networks with a high proportion of mutual connections, people tend to come across similar information and opinions. This can lead to the creation of echo chambers, where

individuals are exposed only to content that aligns with their existing beliefs or preferences.

Similarity of connection

Mutual connections reinforce group identities, as people tend to interact more often with similar-minded peers, forming cohesive communities. This reinforcement can contribute to the formation of opinion clusters and echo chambers within the network.

To better understand the clusters or communities within the network, I utilized Jaccard similarity, a commonly used metric in network analysis. Jaccard similarity calculates the similarity between sets of connections in the graph by dividing the intersection of the sets by their union. By applying this calculation to pairs of nodes in the network, I could identify nodes with high similarity scores, indicating that they are likely to share common neighbors or similar connections.

Most similar and dissimilar nodes

Output:
Most similar: (749, 775) (Similarity: 1)
Most dissimilar: (1590, 481) (Similarity: 0)

It's important to note that while specific pairs are presented, there are multiple pairs of users exhibiting similar or dissimilar connections, and the output is randomly chosen.

With a similarity score of 1, the pair of users (749, 775) share identical sets of connections, indicating a high level of commonality in their interests, affiliations, or activities. However, we don't know how those factors influence their similarity due to privacy protection. Nevertheless, such high similarity underscores the potential for shared engagement patterns or community affiliations between these users.

On the other hand, the pairs of users (1590, 481) exhibit a similarity score of 0, indicating a complete absence of overlapping attributes. This disparity suggests that these two individuals may belong to distinct communities or subgroups in the network. The low similarity between these nodes not only highlights their diverse characteristics but also implies divergent perspectives and interests.

Clustering coefficient

A node with a high clustering coefficient means that its neighbors are more likely to be directly connected. This could relate to a dense local connectivity within the

neighborhood of these nodes. When we consider the clustering coefficient, it is important to note that nodes with high centrality can have a big impact on the clustering patterns in their local neighborhoods. Specifically, nodes that have a high degree of centrality have many connections, which can increase the likelihood of densely connected clusters forming around them.

Based on the data, the average clustering coefficient is 0.6055. This relatively high number indicates that nodes in the network tend to form tight clusters, where neighbors of a node are likely to be connected. Additionally, a high clustering coefficient can suggest that a network can withstand random failures. In networks with high clustering coefficients, if a node fails, its neighbors are likely to remain connected to each other, preserving the overall integrity of the local cluster.

Node importance

Nodes with higher visit counts in the random walk sequences are considered more important or central in the network. This helps to determine the significance of individual nodes. Highly central nodes sometimes referred to as influential nodes, are essential for mediating connections and facilitating communication between different parts of the network. These nodes, characterized by their significant visit counts, represent individuals who have many connections or are part of multiple social circles.

One specific output:
Node Importance: Vertex 2469: Visits 1
Node Importance: Vertex 345: Visits 1
Node Importance: Vertex 1718: Visits 1
Node Importance: Vertex 282: Visits 1
Node Importance: Vertex 0: Visits 2
Node Importance: Vertex 1621: Visits 1
Node Importance: Vertex 253: Visits 1
Node Importance: Vertex 1287: Visits 1
Node Importance: Vertex 1590: Visits 1

Nodes that have been visited multiple times during a random walk are likely to be more central. In the random output, nodes with a visit count of 1 may still be important, but they are not as central as the node with a visit count of 2 (Vertex 0). Node 0, with a visit count of 2, appears to be more significant or central in the network compared to the other nodes listed, as it has been visited twice during the random walk.

Random walks and path length

The random walk explores the social network by moving from one user's profile to another based on friendship connections. Each "Current vertex" represents a Facebook user profile visited during the random walk and the path length indicates the distance traveled through friendship connections.

Output:
Current vertex: 2833
Current vertex: 3291
Current vertex: 3313
Current vertex: 2875
Current vertex: 2310
Current vertex: 2304
Current vertex: 2327
Current vertex: 1981
Current vertex: 2384
Current vertex: 2267
Path Length: 10

The random walk moves each step based on the friendship connections in the graph, and this specific output doesn't contain repeated vertex so the random walk continued until it visited 10 different vertices without revisiting any previously visited vertex. Additionally, the path length matches the number of steps corresponding to this point, with no revisited vertices during the random walk. This suggests that the network is connected enough for the walker to reach 10 distinct vertices without encountering any dead ends or isolated parts of the network.

Given that the diameter of the network, which represents the longest shortest path between any pair of nodes, is 8, a path length of 10 would be longer than the longest shortest path in the network. This suggests that the path length of 10 is relatively long compared to the shortest paths between nodes. Additionally, considering the 90-percentile effective diameter of 4.7, which indicates that 90% of node pairs can be reached within an average of 4.7 steps, a path length of 10 is substantially longer than the typical path length between most nodes in the network. Therefore, it implies that there are areas of the network that are less densely connected or more distant from the starting point, contributing to the longer path length.

Conclusion

This network exhibits a phenomenon similar to the small-world phenomenon, where most nodes are closely connected to their neighbors, resulting in a high clustering coefficient. However, there are also a few long-range connections that enable short

paths between two nodes. These far-reaching links contribute to a longer path length, but they can coexist with high local connectivity. These connections may bridge different clusters together, leading to a higher average path length than expected based solely on local clustering. The network may be divided into distinct communities or modules with strong internal connections, resulting in high clustering. However, there may be weak connections between these communities. In such cases, traversing between different communities can result in longer paths, which contributes to a longer average path length. Therefore, this allows for effective transmission of information and has different levels of impact on various communities.