6/22/2024

UNIVERSIDAD POLITÉCNICA DE YUCATÁN

BIS Universities

Social Network Analysis
Social Circles

Miguel Sánchez Piña

Data Engineering 8°B

Social Network Analysis
Professor Didier Gamboa

Contents

Abstract.

This document has the intended objective of analyzing a dataset derived from Facebook, comprising 'circles' or 'friends lists', where each circle represents a group of users interconnected through mutual friendships. The dataset includes anonymized node features, circle affiliations, and ego networks, providing a comprehensive view of digital social interactions.

The dataset features two distinct sets of entities: users and circles. Users are exclusively connected to circles, reflecting a typical structure in social network datasets where nodes (users) are linked to different types of entities (circles).

During the analysis this document will be employing tools like NetworkX and matplotlib, visual representations of the network elucidate user-group affiliations and network structures. This graphical approach facilitates the identification of community formations and central users within the dataset. Applying Graph Theory principles, such as degree distribution and community detection algorithms, we will reveal patterns of connectivity and group cohesion within the bipartite network. These insights contribute to understanding user behavior and network dynamics on digital platforms.

Keywords: Social Network Analysis, Bipartite Networks, Facebook Dataset, Graph Theory, Community Detection.

Introduction.

The dataset named “Social Circles: Facebook” by Jure Leskovec consists of 'circles' or 'friends lists' from Facebook. These circles represent groups of friends connected through mutual friendships. The dataset includes node features, which are profiles of the users, circles indicating groups of friends, and ego networks, which are the networks of friends for each individual in the dataset. The data has been anonymized by replacing Facebook-internal IDs with new values, and the feature vectors have been anonymized as well, making it possible to determine whether two users share the same characteristics without revealing the actual characteristics.

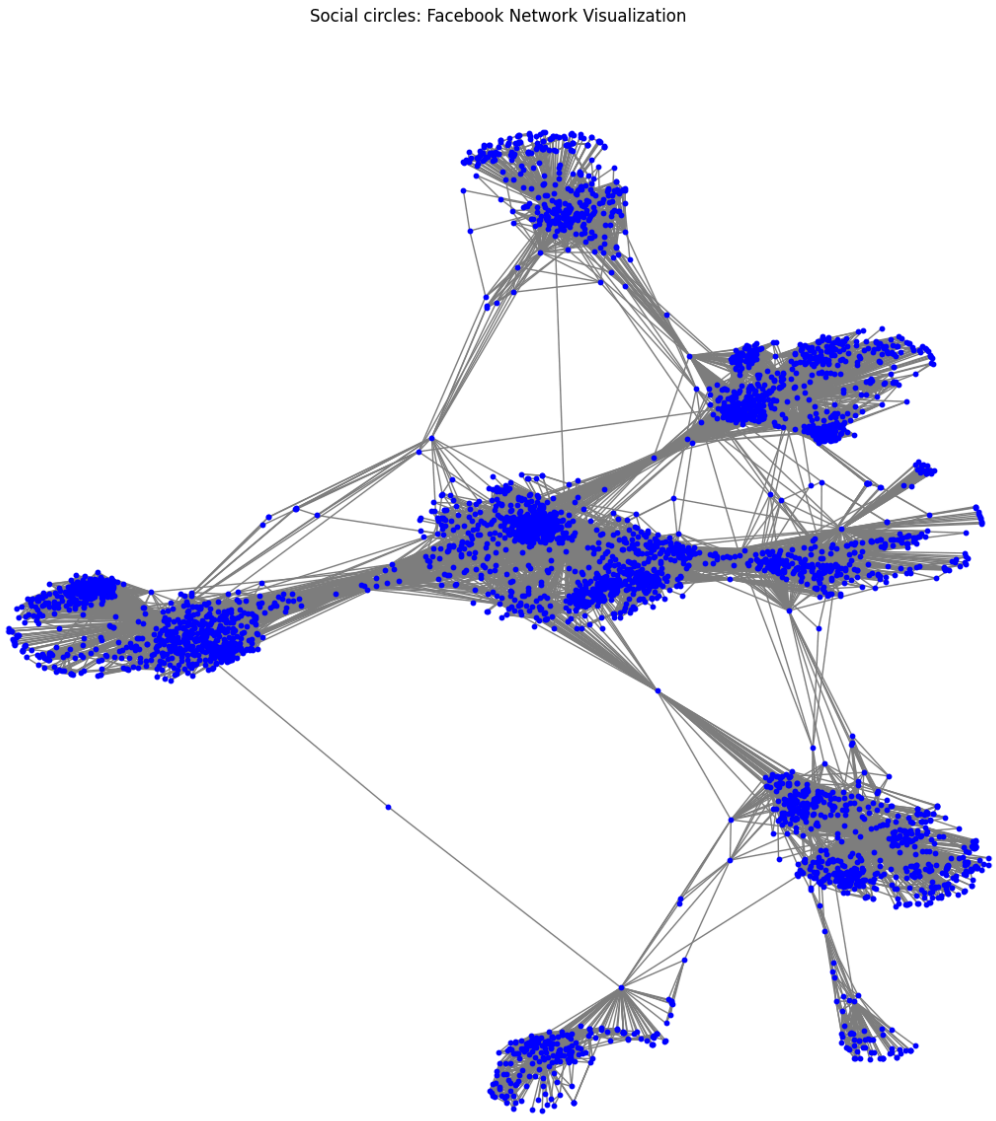
Analyzing social networks like Facebook provides valuable insights into the patterns of social interaction, information spread, and community formation. Such analyses can be applied in various fields including marketing, where understanding how information spreads can improve targeting strategies, or sociology, where insights into social behavior and community dynamics are sought. The dataset, while anonymized, still holds significant value for research in these areas.

Although specific papers using this exact dataset are not mentioned, similar Facebook datasets have been widely used in academic research. Studies utilizing such datasets typically focus on community detection, influence propagation, and behavior prediction, offering a

wide range of applications from enhancing recommendation systems to understanding social dynamics and public opinion formation.

This Facebook dataset can be classified as a bipartite network. In this context, the bipartite nature arises because the network consists of two distinct types of entities: users (nodes) and circles (groups of users). Users are connected to circles rather than directly to other users within the dataset. This classification is typical for networks where nodes are associated with different types of entities, such as users belonging to multiple groups or communities.

The reason for classifying the network as bipartite is that it follows the structure where nodes (users) can only connect to circles (friends lists), and there are no direct connections between users within the dataset itself. This structure is common in social networks where the relationships are defined by group memberships rather than direct user-to-user connections.



This graphical representation helps in understanding the network's structure and the relationships between users and their respective circles. For instance, a simple approach to visualize the network could involve plotting the nodes (users) and edges (connections to circles) to observe how users are grouped into different circles.

Based on Graph Theory principles, several important observations can be made. Analyzing the degree distribution, which is the distribution of the number of circles each user is part of, can reveal insights into user influence and centrality within the network. Users with high degrees are part of many circles and may be more influential.

Network Characteristics.

The dataset consists of 4039 nodes (users) and 88234 edges (friendship connections), indicating a large and densely interconnected network. The **average path length** is a crucial metric for understanding network efficiency , this path length measures the average number of steps along the shortest paths for all possible pairs of nodes. The **average path length** of 3.6925 that this network has indicates the average number of steps along the shortest paths for all possible pairs of nodes, which suggests that the shorter average path equals to a quicker communication and more efficient interaction between users within the network.

The **clustering coefficient** is another important metric, reflecting the tendency of nodes to cluster together. With an average clustering coefficient of 0.6055, this network exhibits a notable tendency for nodes to form tightly interconnected clusters or communities. This high clustering coefficient suggests that users in the network are more likely to be connected to each other through mutual friends, indicating the presence of cohesive subgroups or communities.

Distance metrics such as **diameter** and **90-percentile effective diameter** provide insights into the network's structural properties:

- The **diameter**, which is the longest shortest path between any pair of nodes, is 8 in this network. This means that the maximum number of steps required to connect any two users in terms of friendship connections is 8, indicating a relatively small diameter for a network of this size.
- The **90-percentile effective diameter** is 4.7, indicating that 90% of all pairs of nodes are reachable within an average of approximately 4.7 steps. This metric highlights the network's efficiency in terms of connectivity and the ability to quickly reach other nodes within the network.

These metrics show the structure and dynamics of the social network analysis. The high clustering coefficient suggests strong community structures, while the diameter and effective diameter metrics underscore efficient connectivity and short paths between users.

Centrality Measure.

Closeness centrality measures how quickly a node can interact with all other nodes in the network, with higher values indicating nodes that are more central and accessible. Node 107 emerges as the most central node in terms of closeness centrality, scoring approximately 0.460. This suggests that Node 107 is positioned to quickly disseminate information or influence throughout the network, due to its minimal average path length to other nodes. Following closely are nodes like 58, 428, and 563, each demonstrating strong closeness centrality values around 0.397 to 0.395, indicating how crucial roles are in maintaining efficient communication pathways within the network. Node 1684 has a closeness centrality of 0.394, underscoring its significance in facilitating rapid interactions across the network; this of course helps to point to certain profiles that act as bridges and can create edges between different nodes or profiles.

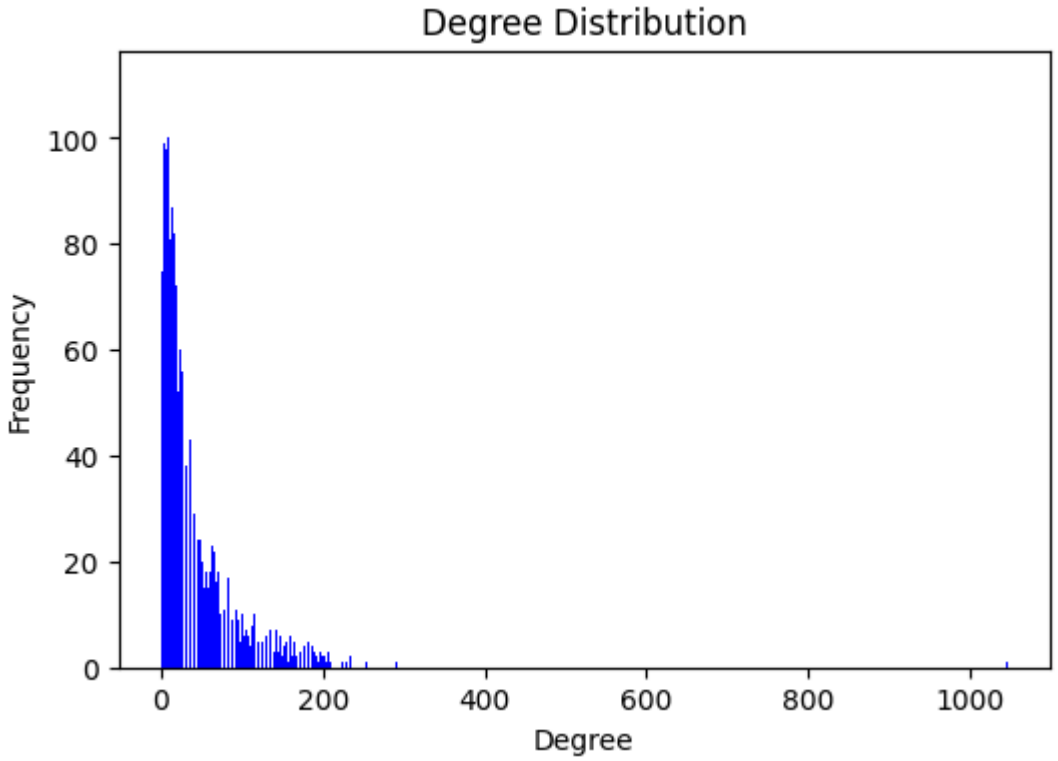
Degree centrality, which quantifies the number of direct connections (edges) each node has, highlights nodes that are highly connected within the network. Node 107 has the highest degree centrality at approximately 0.259, indicating it has the most connections among all nodes. This high degree centrality positions Node 107 as a key hub in the network, likely central to various social interactions and information dissemination activities. Node 1684 follows with a degree centrality of 0.196, indicating significant connectivity and influence within its network neighborhood. Nodes like 1912, 3437, and 0 also show notable degrees of centrality.

Eigenvector centrality evaluates nodes based on the quality of their connections, considering connections to other influential nodes. Node 1912 emerges as the most influential node by eigenvector centrality, scoring approximately 0.095. This shows the direct connections and how connected to other nodes that themselves have considerable influence within the network. Nodes 2266, 2206, 2233, and 2464 also demonstrate significant eigenvector centrality values, suggesting they play critical roles in maintaining network cohesion and influencing dynamics across different parts of the network.

These measures highlight nodes that are pivotal in maintaining network connectivity, driving information dissemination, and influencing community interactions.

Degree Distribution.

The following degree distribution plot reveals how the number of connections (degree) is distributed among nodes in the network. This means that the plot shows a pattern where there are numerous nodes with low degrees (few connections) and a smaller number of nodes with high degrees (many connections). This distribution often follows a power-law shape, characteristic of many real-world networks.

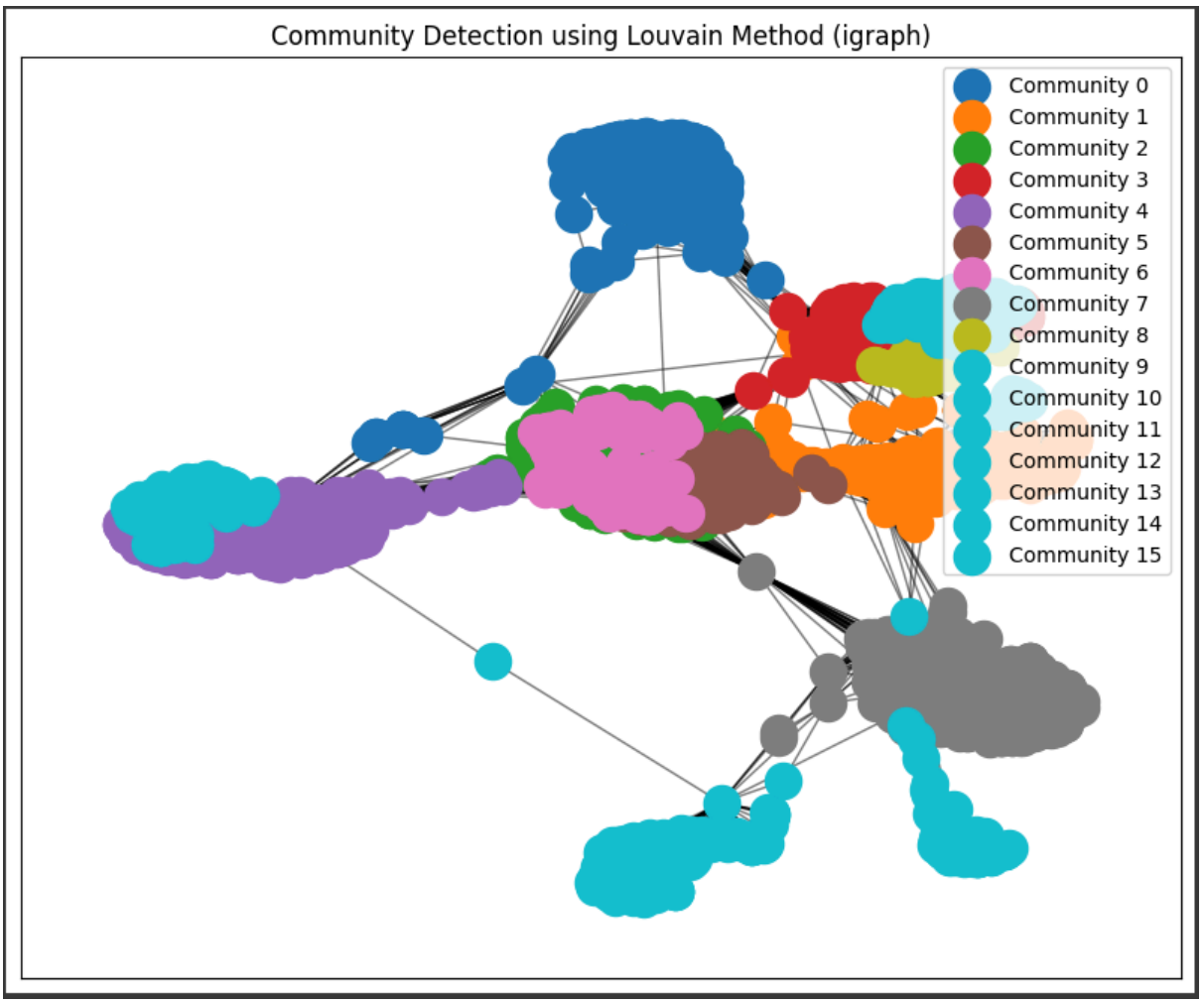


The plot shows a skewed distribution where most nodes have relatively few connections, while a few nodes (hubs) have significantly more connections than the average. This pattern proves to us that some nodes play crucial roles in maintaining network connectivity and facilitating communication as mentioned before in the document.

Some nodes with high degrees (hubs) are most important in the network. They serve as central points for information flow, communication, and influence. These hubs can efficiently spread information due to their extensive connectivity, making them influential actors within the network structure.

Community Detection.

Community detection will help to identify groups of nodes that are densely connected internally but sparsely connected to nodes in other groups. This analysis helps uncover underlying structures, functional units, or communities within the network.



The Louvain method was chosen for community detection in the Facebook network dataset due to its efficiency, ability to optimize modularity, hierarchical nature, broad applicability across different network types, and availability in widely-used software packages. These factors collectively make it a robust choice for exploring community structures and understanding the organizational principles of complex networks.

The Facebook network dataset was partitioned into distinct communities. Among the identified communities, Community 4 stands out as the largest, encompassing a significant portion of the network's nodes and edges. This observation suggests that Community 4 is densely interconnected internally while maintaining fewer connections with nodes outside the community.

Conclusions.

The analysis of the “Social Circles: Facebook” network dataset has provided valuable lessons into its structural characteristics and functional dynamics. Comprising 4039 nodes and 88234 edges, the network exhibits a densely connected nature typical of social networks. This dataset, derived from anonymized circles (friends lists) and ego networks, works to show us the foundational resource for understanding social relationships and communication patterns within online communities.

Key network metrics such as average path length, clustering coefficient, and centrality measures have revealed important aspects of its topology. With an average path length of approximately 3.69, the network facilitates efficient information propagation, crucial for viral content spread and communication efficiency among users. The high clustering coefficient of 0.61 indicates the presence of tightly-knit clusters or communities where nodes tend to form interconnected groups, fostering cohesive interactions and shared interests.

Centrality measures such as degree centrality, eigenvector centrality, and closeness centrality have identified influential nodes and hubs within the network. Nodes like Node 107 and Node 1684 exhibit high degree centrality, signifying their extensive connections within the network. Meanwhile, nodes such as Node 1912 and Node 2266, highlighted by eigenvector centrality, are influential due to their connections with other well-connected nodes. These metrics underscore the importance of certain nodes in facilitating information flow and maintaining network cohesion.

The degree distribution analysis has shown a power-law distribution, where most nodes have low degrees while a few nodes (hubs) have exceptionally high degrees. This distribution pattern is characteristic of many real-world networks and underscores the presence of influential hubs capable of exerting significant influence over the network's dynamics. Such hubs play critical roles in information dissemination and maintaining network resilience against disruptions.

Community detection using the Louvain method has made more clear the network's modular structure. By partitioning the network into distinct communities based on connectivity patterns, the analysis has identified cohesive groups of nodes with shared interests or functional roles. These communities not only provide insights into user behavior and interaction patterns but also offer opportunities for targeted interventions and community engagement strategies tailored to specific groups.

References.

SNAP: Network datasets: Social circles. (s. f.). https://snap.stanford.edu/data/ego-Facebook.html

Derr, A. (2023, 14 septiembre). *Social Network Analysis 101: Ultimate Guide - Visible Network Labs*. Visible Network Labs. https://visiblenetworklabs.com/guides/social-network-analysis-101/

Joshi, P. (2024, 23 febrero). *Community Detection: Getting Started within Graphs and Networks*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2020/04/community-detection-graphs-networks/

LatentView, T. (2024, 30 enero). *A Guide to Social Network Analysis and its Use Cases*. LatentView Analytics. https://www.latentview.com/blog/a-guide-to-social-network-analysis-and-its-use-cases/