Assignment No. 1

Compute different measures on a large graph

Network:

- Twitter Social Circles

Group Members:

- Parsa Moslem (5755015)
- Amir Mohammad Azimi (5795736)

Course:

- Network Analysis

The course is part of the below degree:

- Computer Science – Software & Security Engineering

Academic Year: 2023 - 2024

UNIVERSITÀ DEGLI STUDI
DI GENOVA

Dibris

Table of Contents

Introduction

In the field of network science, complex systems are often modeled by graphs, where nodes represent entities and edges represent the relationships between them.

This report delves into the intricate network of **Twitter users**, utilizing a unique dataset comprised of **81,306 nodes** and **1,342,310 edges**. This data originates from **Twitter's** "**circles**" feature, also known as "lists," which allows users to categorize other users into specific groups.

Graph Visualization

Following the import of the Twitter network data into ***Gephi***, a visual exploration of the graph was conducted to gain preliminary insights into the network structure. The resulting visualization is presented in *Figure 1*.

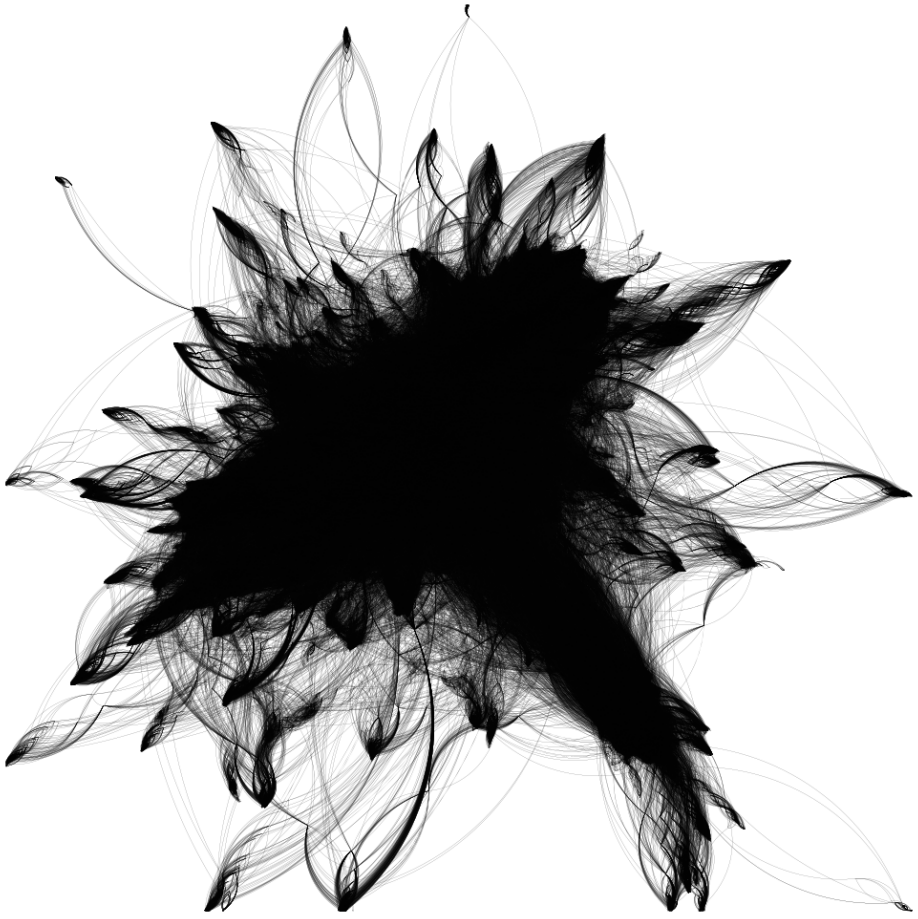


Figure 1

Giant Component

Our analysis of the network's connectivity revealed the presence of a single, all-encompassing connected component. This implies that every node within the network is reachable from every other node through a series of connected edges. In essence, **the giant component in this case coincides with the entire network itself.** This finding was corroborated using the ***NetworkX*** library, which identified the giant component as encompassing all nodes and edges present in the original dataset. The achieved giant component has **81,306 nodes** and **1,342,310 edges**.

Data Subsampling

Due to the computational demands associated with analyzing large network datasets, a subsampling approach was employed. This involved randomly selecting a representative subset of **20,000 nodes** from the original dataset containing 81,306 nodes. This resulted in a smaller network with **82,233 edges**, which is a subset of the original 1,768,149 edges. This subsampling strategy aimed to achieve a balance between **computational efficiency** and **the preservation of key network properties** for the analysis. The subsampled network was subsequently visualized using Gephi software. The resulting visualization is presented in *Figure 2*. This visualization can provide valuable insights into the structural properties of the Twitter network at a smaller scale.

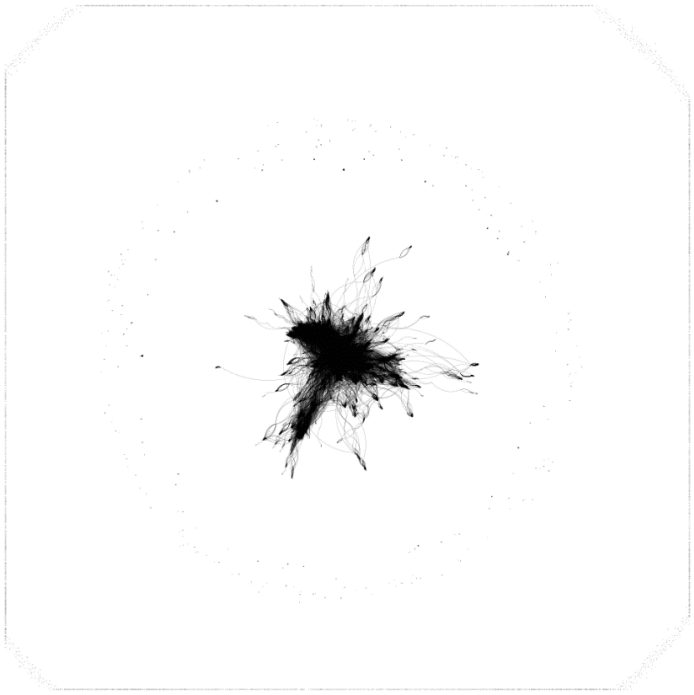


Figure 2

Subsampled Data - Giant Component

Visual inspection of *Figure 2* reveals that the subsampled network exhibits a non-trivial structure. Unlike the original network, it does not possess a single, all-encompassing connected component. In other words, the giant component in this case represents a subset of the entire network, encompassing a smaller number of nodes and edges compared to the original dataset. The analysis oNAf the subsampled network identified a giant component containing **16,263 nodes** and **81,262 edges**. This indicates that within the subsample, this connected component encompasses the largest number of nodes and edges.

To gain further insights into the structure of the giant component, a visualization was generated using Gephi software. This visualization is presented in *Figure 3*.

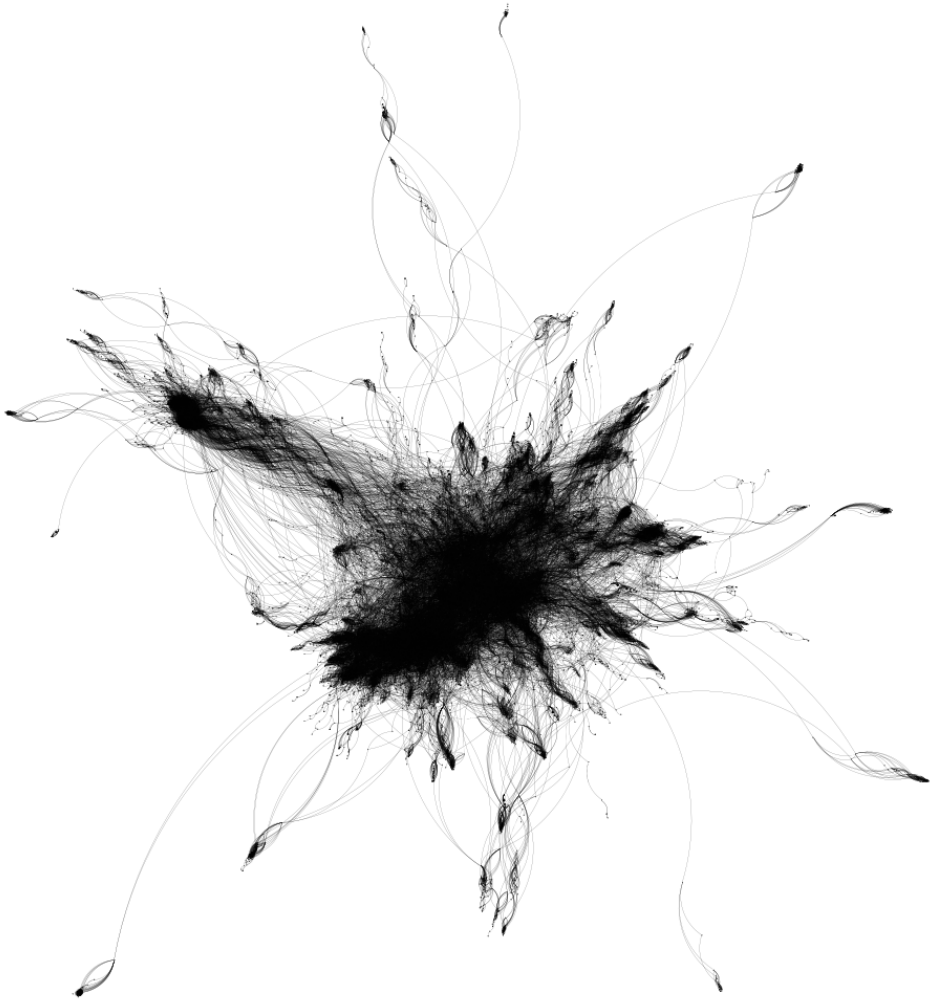


Figure 3

Graph Characterization

Following the construction or acquisition of a network dataset, a crucial step in network analysis involves graph characterization. This process aims to describe the key structural and statistical properties of the graph. By quantifying these features, we can gain a deeper understanding of the network's organization and functionality.

Average Degree

The average degree which is calculated for this network is equal to **9.9934** which means that, on average, each user in this network is directly connected to approximately 10 other users.

This relatively low average degree in a large network suggests that while there are some users with many connections, most users have only a few connections. This is a common characteristic of social networks, where a small number of users (e.g., influencers) have many followers, while most users have a smaller number of connections.

Approximate Diameter

Due to the computational demands associated with calculating the exact diameter for a network of this size (16,263 nodes and 81,262 edges), an approximate approach was employed. This technique involved **randomly selecting a subset of 100 nodes from the giant component**. Subsequently, the diameter was calculated for this subsample. While not the exact diameter of the entire network, this approach provides a close estimate in a significantly reduced timeframe. The application of the approximate diameter technique on the randomly chosen subset of 100 nodes resulted in a diameter of 4.

A diameter of 4 indicates that the maximum geodesic distance (shortest path length) between any two nodes within the giant component is 4. In simpler terms, no matter where you start in the network, you can reach any other node by traversing a maximum of 4 connections (edges) along the shortest path.

In conclusion, the findings from the network characterization and analysis, particularly the approximate diameter of 4 within the giant component, provide suggestive evidence that the network exhibits characteristics consistent with the **small-world phenomenon**. This implies that the network possesses a high degree of clustering (local connections) alongside a small number of long-range connections that enable relatively short path lengths between most nodes.

Centrality Measures

Understanding the distribution of centrality measures in the graph can reveal valuable information about how information flows, who the key players are, and how the network is structured.

Average Closeness

Closeness centrality of a node measures its average distance to all other nodes in the network, but average closeness centrality measures how closely knit the entire network is. In the context of social networks like Twitter, a high average closeness centrality could indicate that information can spread quickly through the network, as users are, on average, closely connected.

With an **average closeness** centrality of **0.2555**, there is a relatively well-connected network where information or influence can potentially propagate quickly. It means that 1 / 0.2555 = **3.91 steps** are needed to arrive a target node from any node that describes the graph is moderately interconnected. This could imply that information or trends can spread relatively quickly through the network, as users are, on average, not very far apart from each other in terms of their connections.

Average Betweenness

Betweenness centrality quantifies a **node's intermediacy** within a network. It reflects **the number of times that a particular node lies on the shortest paths between all other pairs of nodes**. **High betweenness centrality** for a node indicates that it acts as a **critical bridge** for communication flow within the network. Conversely, **low betweenness centrality** suggests that the node plays a **less significant role** in mediating information exchange.

In the context of network centralization, **average betweenness centrality** serves as a measure of the network's **overall reliance on intermediary nodes**. A network with a **high average betweenness centrality** is likely to be more **centralized**, implying a dependence on a select few nodes for efficient information transfer. Conversely, a network with a **low average betweenness centrality** suggests a more **decentralized** structure, where information can flow readily without the need for prominent intermediary nodes.

The observed **high average betweenness centrality (9.2025)** in our network suggests a **dependence on critical intermediary nodes** for efficient information flow. This indicates a **centralized network structure**, like the dynamic observed in Twitter Circles. In Twitter Circles, the circle owner acts as a gatekeeper, controlling information dissemination to the viewer members. Here, the high average betweenness centrality implies a network where a select few nodes play a similar role, bridging communication between other nodes.

Clustering Coefficient

An **average clustering coefficient (ACC) of 0.4457** within the network indicates a propensity for nodes to cluster together. This translates to approximately 44.57% of a node's neighbors also being neighbors with each other. This finding suggests the presence of densely connected communities within the network, characterized by significant overlap between the immediate neighborhoods of individual nodes.

In the context of Twitter Circles an **ACC of 0.45** might imply that users within distinct circles exhibit a moderate degree of overlap in their social connections. **This could be attributed to the inherent social dynamics where users tend to create circles with**

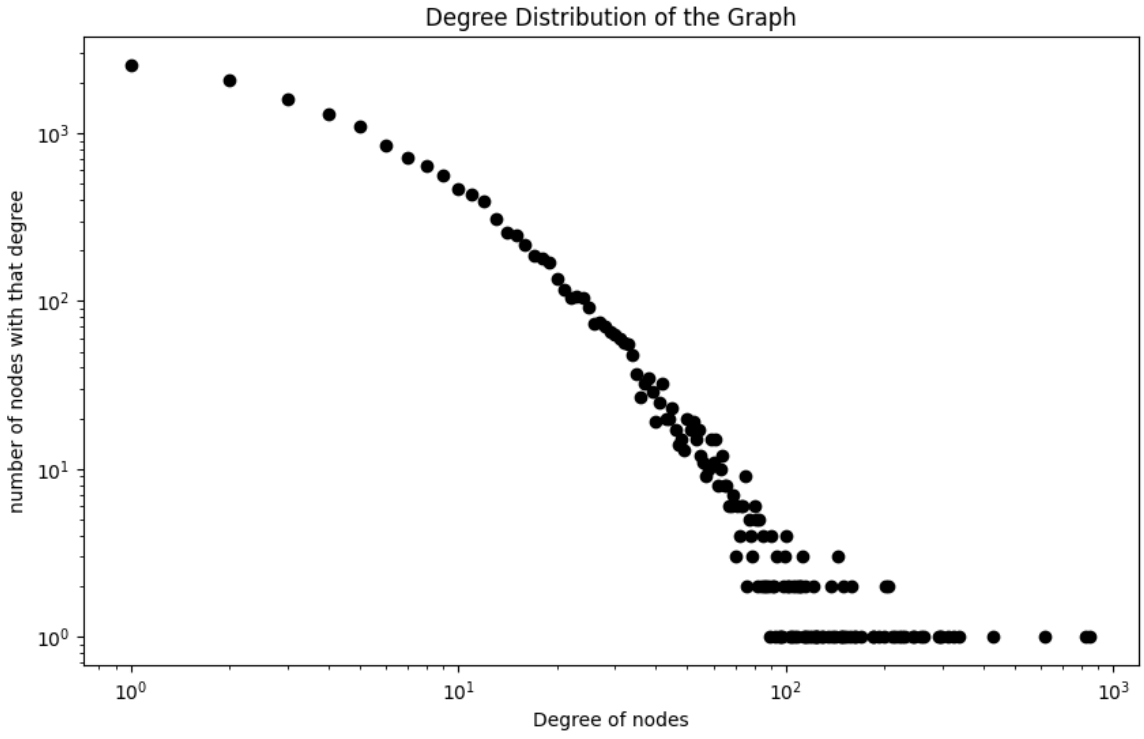
their existing friends, fostering a sense of familiarity and shared interests within the circle.

Transitivity

A **transitivity value of 0.1636** in the graph indicates a moderate level of clustering. This translates to roughly **16.36%** of connected triples forming triangles within the network. This implies that there is a **moderate probability** for two nodes connected to a common node to also be connected to each other, suggesting the presence of some level of clustering within the network. However, the relatively low transitivity value also indicates that **the network exhibits a heterogeneous structure**. This means **there is a significant presence of nodes** that **connect to others outside their immediate neighborhood**, deviating from the purely clustered pattern.

Degree Distribution

Characterization of the network's degree distribution offers valuable insights into its underlying structure. In this context, the Twitter Circle network is hypothesized to exhibit a **power-law distribution**. This implies that a significant portion of nodes will possess a relatively low degree (number of connections), while a smaller number of nodes will have a disproportionately high degree, functioning as hubs within the network. The figure below visually supports this hypothesis, demonstrating the presence of a tail in the distribution, indicative of a power-law relationship.



Conclusion

Our analysis of the Twitter Circles network reveals a fascinating interplay between global reach and local communities.

The network exhibits a single, large, connected component, implying that most users could reach each other through a series of connections.

This, coupled with the small-world effect indicated by the **approximate diameter of 4**, suggests an efficient network structure. Short paths connect most nodes despite the network's size, highlighting a balance between local clustering and long-range connections.

However, the high **average betweenness centrality (9.2025)** points towards a network reliant on critical intermediary nodes for information flow. This centralized structure is reminiscent of how circle owners in Twitter Circles controlled information dissemination.

The network also fosters strong community formation. The high **average clustering coefficient (0.4457)** and moderate **transitivity (0.1636)** suggest the presence of densely connected communities. Users likely formed circles with people they already knew, creating familiarity and shared interests within these groups.

Finally, the **power-law degree distribution** concept aligns with this network. A few highly connected hubs, likely the circle owners, would exist alongside many users with fewer connections, reflecting the viewer role within the circles.

In conclusion, the Twitter Circles network functioned as a small-world network with a centralized structure and strong community formation. Information flow likely relied on a select few influential users (hubs), while users also formed tight-knit communities based on shared connections.

Sources

The foundation for this analysis is a dataset inspired by the research presented in "Learning to Discover Social Circles in Ego Networks" by J. McAuley and J. Leskovec (NIPS, 2012) [1]. This work highlights the importance of identifying user-defined social circles within online social networks. Our dataset follows a similar structure, leveraging Twitter's "circles" functionality to represent these communities.

1. J. McAuley and J. Leskovec. Learning to Discover Social Circles in Ego Networks. NIPS, 2012.

Also, the data file is downloaded using this link.