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.

The table below will show the graph specifications:

Network	***N***	***L***	***<k>***	***ln N***
Circles of Twitter	81,306	1,342,310	33.01	11.30

Preliminary inspection of the data table reveals the **presence of a giant component** encompassing all nodes as **<k> > 1**. Furthermore, the absence of any disconnected components within the current graph (**as <k> > ln N**) implies that the **graph is fully connected**, with **the giant component itself constituting the entire graph**.

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*.

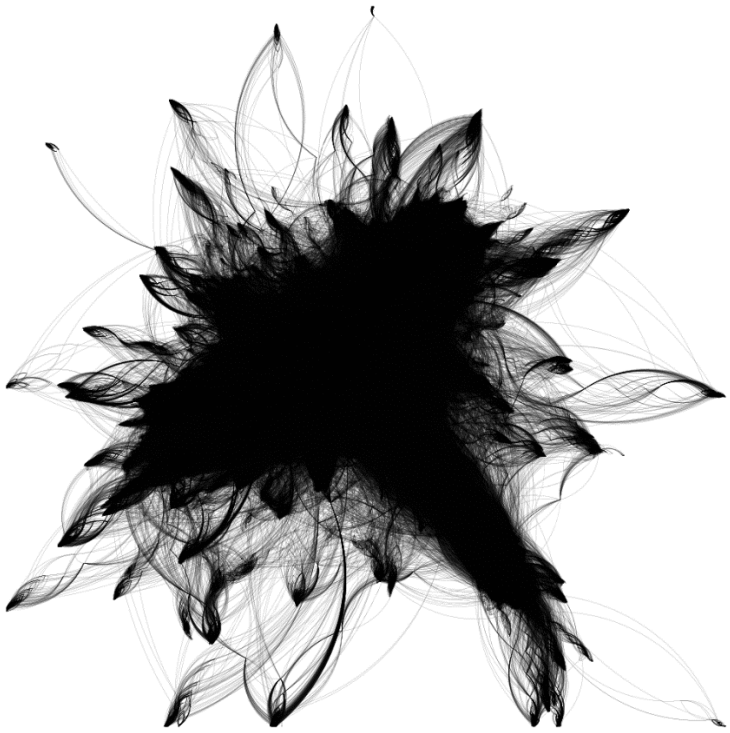


Figure 1

Giant Component

As it is found in the previous sections, it is known that the giant component is the graph itself, and by calculating the giant component using *NetworkX*, the seen results are proved by looking at the giant component specifications: $\boldsymbol{N = 81,306}$ and $\boldsymbol{L = 1,342,310}$.

Due to the substantial size of the graph and the coincidence of the giant component encompassing all nodes, **subsampling the graph becomes necessary**. This subsampling process aims to generate a smaller representative network that facilitates computational efficiency and reduces resource consumption during analysis.

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.

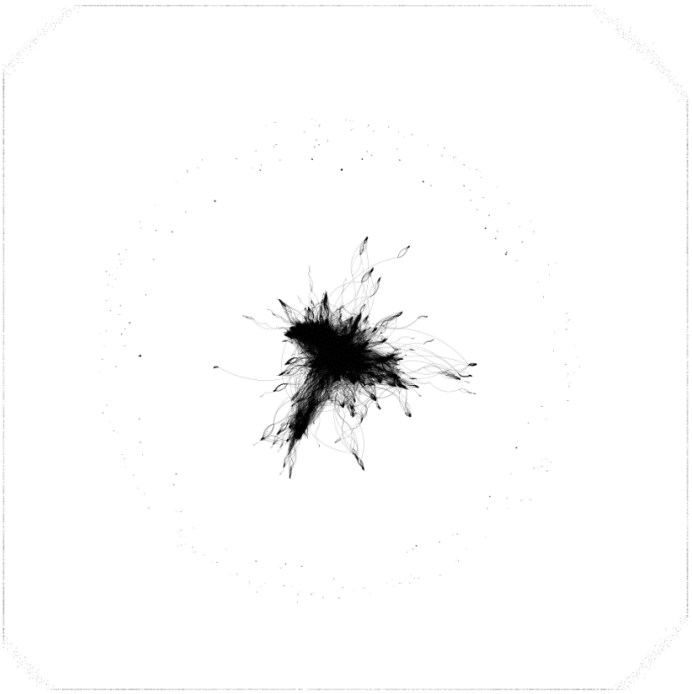


Figure 2

Network	***N***	***L***	***<k>***	***ln N***
Circles of Twitter - Subsampled	20,000	82,233	8.02	9.90

Consistent with our initial observations, examination of the subsampled network reveals the **persistence of a giant component**. However, unlike the original graph, this component is

not fully connected. The presence of additional disconnected components underscores the need to recompute the size and characteristics of the giant component within the subsampled network.

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 of 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*.

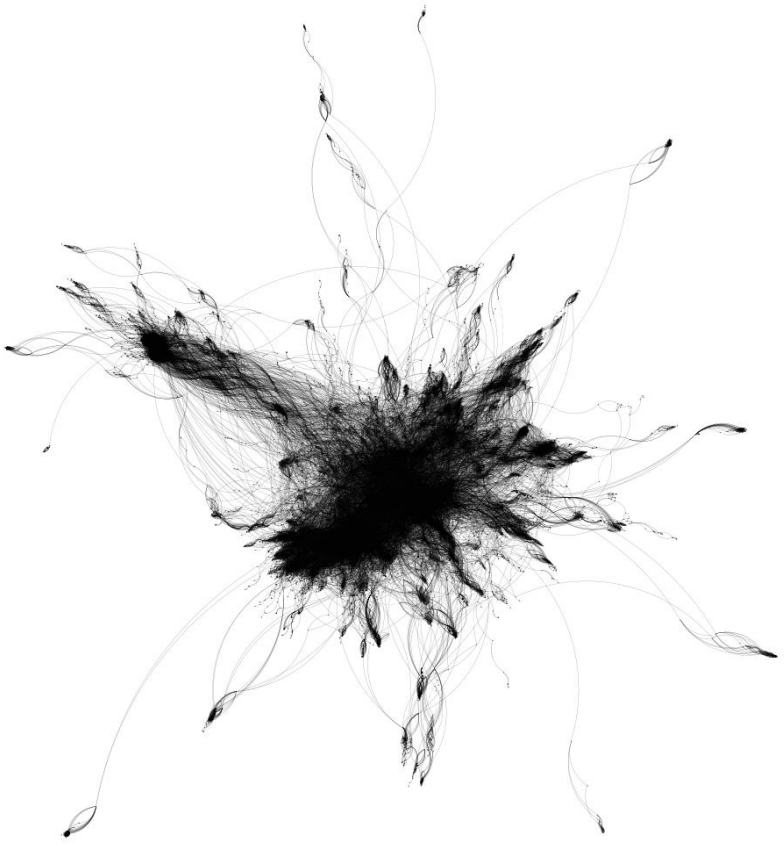


Figure 3

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.

Network	***N***	***L***	***<k>***	***ln N***
New Graph - Subsampled Giant Component	16,263	81,262	9.9934	9.6966

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**.

The diameter 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.

The **average closeness centrality** of the network was computed to be **0.2555**. This value suggests:

- **The network is moderately interconnected:** Nodes are not extremely close to each other, but they are not very far apart either.
- **Information propagation:** Information can propagate relatively quickly through the network. On average, nodes are not very far apart in terms of their connections.
- **Average Shortest Path Length:** This value can be calculated using the formula:
  - $\frac{1}{AC} \rightarrow AC: Average\ Closeness = \frac{1}{0.2555} = \mathbf{3.91\ steps}$
  - This means that on average, it takes about 3.91 steps to reach any node from a given node.

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.

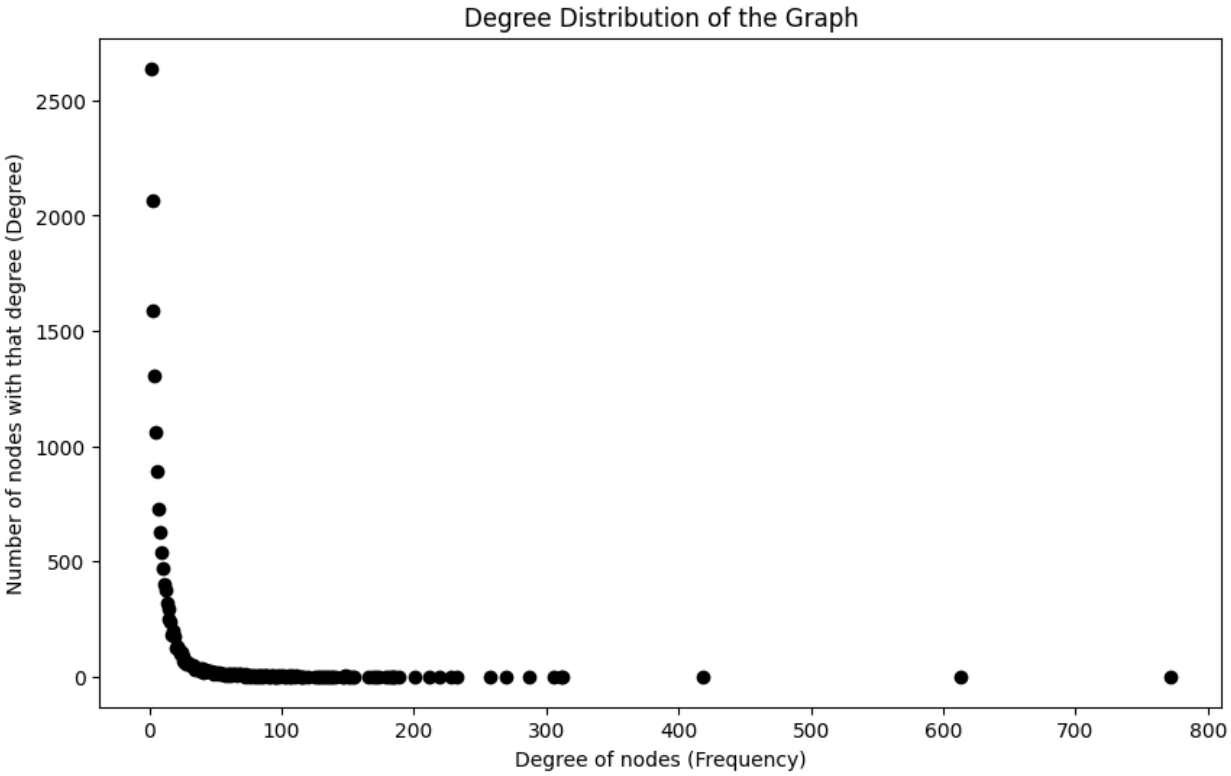
Global Clustering Coefficient (Transitivity)

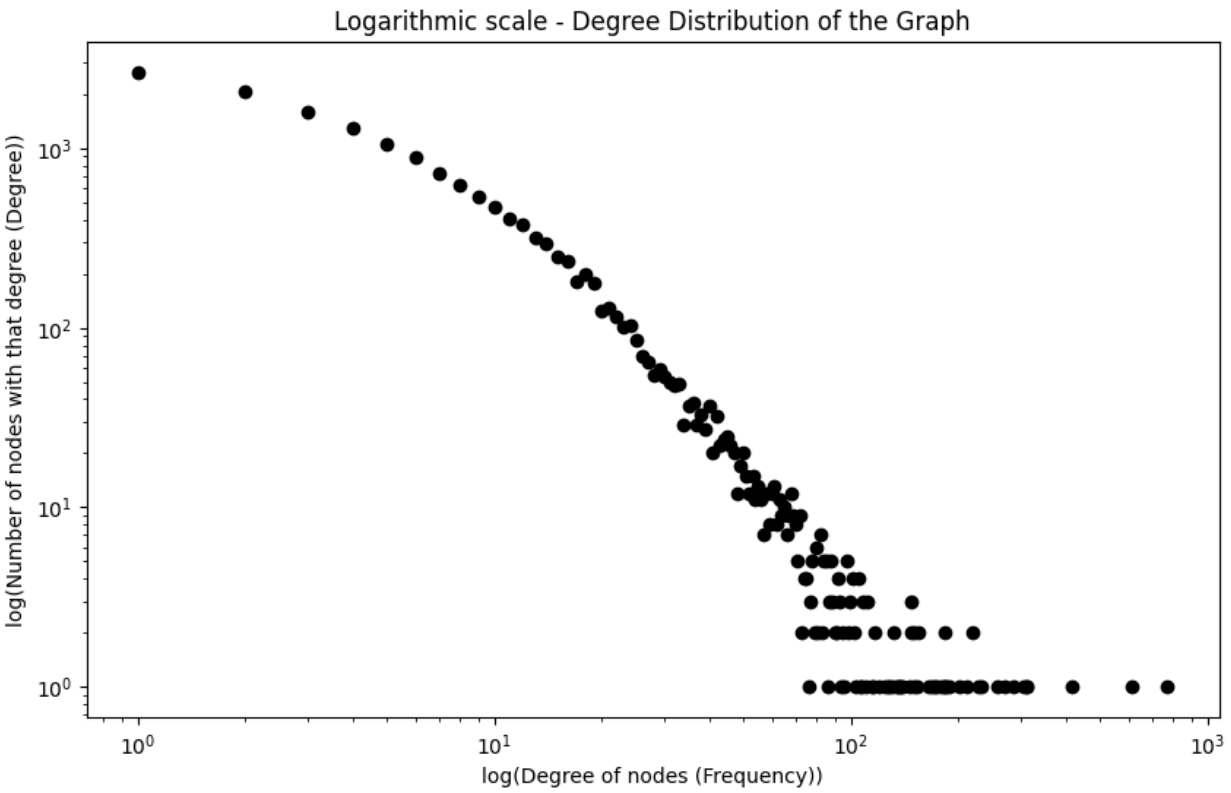
A **transitivity value of 0.1636** in the graph indicates a moderate level of clustering. This translates to roughly **16.36%** of connected triples forming triangles within the network (sets of three nodes where each node is connected to the other two).

- **Limited Close-Knit Communities:** The low transitivity value implies that users within Twitter Circles don't tend to form tightly knit groups where everyone follows each other. This could be because Circles are designed for more specific sharing rather than general social interaction.
- **Presence of Some Clustering:** However, the value is not zero, indicating some level of clustering does exist. This means there might be sub-groups of users with denser connections, like colleagues following each other within a Circle or friends following each other's close circles.

Degree Distribution

The 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.


Degree Distribution of the Graph
Number of nodes with that degree (Degree)
Degree of nodes (Frequency)
0
500
1000
1500
2000
2500
0
100
200
300
400
500
600
700
800


Logarithmic scale - Degree Distribution of the Graph
log(Number of nodes with that degree (Degree))
log(Degree of nodes (Frequency))
10^0
10^1
10^2
10^3

Conclusion

Overall, the Twitter Circles network exhibits properties consistent with social networks, including **small-world characteristics**, **moderate clustering**, and a **power-law degree distribution**, providing valuable insights into its structure and functionality.

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.