Data Analytics (2024-25)
Università della Svizzera italiana

College network

Course Assignment N.13

Paolo Deidda (paolo.deidda@usi.ch)
Andrea Luca Perugini (andrea.perugini@usi.ch)
https://github.com/USI-Projects-Collection/DA-College_network.git

April 17, 2025

Contents

Introduction

This document outlines the structure of the analysis on a dataset of private messages exchanged on a UC Irvine social network.

To run the code and reproduce the figures or outputs, please refer to the `README.md` file for setup and execution instructions.

1 Data Analysis

1.1 Setup and Data Loading

We utilized a dataset provided by the University of California, Irvine, comprising private messages exchanged among users within an online social network. The dataset includes 1899 unique nodes (users) and 59835 edges (messages). The data was loaded into a pandas DataFrame and preprocessed by converting timestamps from Unix time to readable datetime formats, facilitating temporal analysis.

1.2 Data Exploration

1.2.1 Basic Statistics

The dataset consists of 1899 unique users exchanging messages, resulting in 59,835 interactions. We computed essential statistics:

Total number of users (nodes): 1899

Total number of messages (edges): 59,835

Time span of data collection was determined by inspecting the minimum and maximum timestamps, indicating active periods of communication.

1.2.2 Message Frequency Over Time

To explore message distribution over time, messages were grouped and counted by day. Figure 1 clearly demonstrates peaks and troughs, possibly corresponding to academic activities or social events influencing user interactions. This analysis provides insights into temporal dynamics affecting messaging behavior within the network.

Figure 1: Daily Message Frequency

1.3 Network Construction

1.3.1 Degree Analysis

We analyzed node degrees to determine connectivity patterns within the network. Nodes with high degrees represent highly interactive users. The degree distribution shown in Figure 2 exhibits a typical scale-free network characteristic, with many users having few interactions and a small subset having numerous interactions, reflecting real-world social networks.

Figure 2: Degree Distribution

1.3.2 Identifying Isolated and One-Way Nodes

Further examination revealed isolated nodes (users who neither sent nor received messages) and one-way communication patterns (users who either only sent or only received messages). Identifying these nodes helps to understand user engagement and potential network fragmentation.

1.4 User Analysis

1.4.1 Top Senders and Receivers

We determined users with the highest message activity, distinguishing top senders and receivers. Understanding these roles highlights influential users within the network, possibly indicating user centrality or authority.

1.4.2 PageRank Centrality

We calculated PageRank scores to identify influential nodes based on network interactions. High PageRank scores correspond to users frequently contacted by others, emphasizing their significance within the social structure.

1.4.3 Edge Reciprocity

Edge reciprocity was analyzed to assess mutual interactions. High reciprocity indicates robust two-way communication, suggesting deeper social connections among users.

1.5 Temporal Analysis

1.5.1 Messages by Weekday

Analyzing message frequency by weekdays (Figure 3), we observed a clear pattern showing increased messaging during weekdays, which could be aligned with student schedules and reduced activity during weekends.

Figure 3: Messages by Weekday

1.5.2 Top User Activity Analysis

The hourly activity of top users was analyzed (Figure 4), revealing peak interaction times, which could indicate user availability or preferred communication periods. Such insights can guide future studies on user behavior or network usage patterns.

Figure 4: Hourly Activity of Top Users

2 Netwrok Analysis

2.1 Network Structure: Connectivity and Clustering

The network, while representing a connected social environment, exhibits structural nuances typical of large social graphs. Its overall **graph density** is low (approx. 0.0056), indicating a sparse network where only a fraction of potential connections exist.

Despite this sparsity, the network is largely interconnected when directionality is ignored, comprising only **4 weakly connected components (WCCs)**. The presence of a dominant giant component suggests that information can, in principle, diffuse widely. However, considering message directionality reveals a distinct core and periphery structure with **601 strongly connected components (SCCs)**. A large core SCC (containing 1294 nodes, 68% of users) facilitates abundant reciprocal communication, while numerous smaller SCCs highlight a significant periphery. This structure, conceptually illustrated in Figure 5, is crucial for understanding information propagation dynamics. All the other components contain either one or two users. On the other hand, considering the WCCs, there are only three pairs of two users that form separate components, while the rest of the users are part of the largest WCC. This indicates that almost all users are connected in at least one direction.

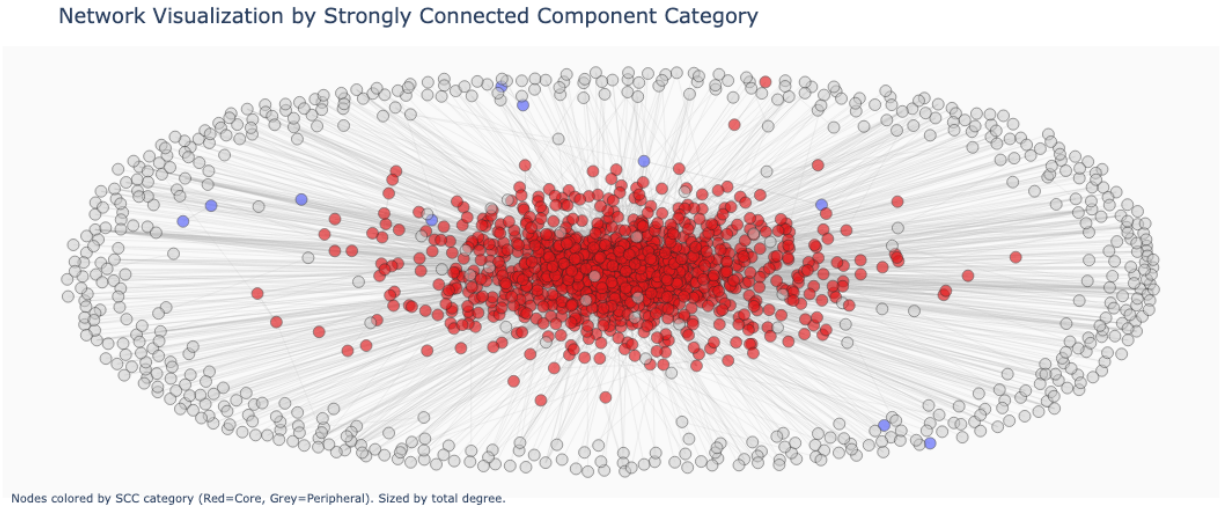

Figure 5: Network visualization highlighting Strongly Connected Components (SCCs). Nodes could be coloured by SCC category (e.g., Core vs. Periphery) and size reflecting degree.

At the local level, the network shows signs of social clustering. The **unweighted average clustering coefficient** (≈ 0.0872) suggests a moderate tendency for users to form triads—higher than random chance. Yet, the **weighted clustering coefficient** (considering message frequency) drops significantly (≈ 0.0018). This disparity implies that while structural triangles exist, they often lack intense, frequent communication loops. Strong, closed triads based on high message volume appear less common than the network's structure might suggest.

2.2 Identifying Influential Users: Centrality Analysis

Centrality metrics help identify users holding key positions. We examined several, focusing here on **Betweenness Centrality**, which identifies network 'brokers'. An approximated weighted calculation highlighted **User 323** (≈ 0.118), **User 1624** (≈ 0.086), and **User 103** (≈ 0.086) as the top intermediaries. These individuals frequently lie on the shortest communication paths, suggesting they bridge different social circles and potentially influence information flow across the network core.

Comparing various centrality measures (including degree, PageRank, and betweenness) via Spearman rank correlations revealed strong positive associations overall. Popularity (in-degree), activity (out-degree), and PageRank influence are highly correlated ($\rho > 0.84$ often), especially between weighted and unweighted versions. This indicates that different facets of importance often coincide. Betweenness centrality, while still strongly correlated ($\rho \approx 0.72 - 0.77$), showed slightly weaker links to degree/PageRank, suggesting its role in capturing brokerage is somewhat distinct. These strong interrelations paint a picture where different forms of network influence are significantly intertwined.

2.3 Network Patterns: Assortativity and Reciprocity

Examining network-wide connection patterns, we found the network to be **disassortative**, with a **degree assortativity coefficient** of approximately **-0.1375**. This negative value indicates a tendency for high-degree users (hubs) to connect with low-degree users, suggesting a "hub-and-spoke" topology rather than a core of interconnected hubs.

Contrasting this global pattern, interaction directness is high. The **graph reciprocity** is approximately **0.6364**, indicating a strong tendency for connections to be mutual; if A messages B, B is highly likely to message A back. This structural tendency towards mutual links is mirrored in individual behavior: a strong positive **Pearson correlation** ($\rho \approx 0.8311$) exists between user in-degree and out-degree (visualized in Figure 6). Users who receive messages from many people also tend to send them to many. Thus, while the network structure facilitates hub-spoke connections, the interactions themselves show strong mutuality, and active users engage vigorously in both sending and receiving.

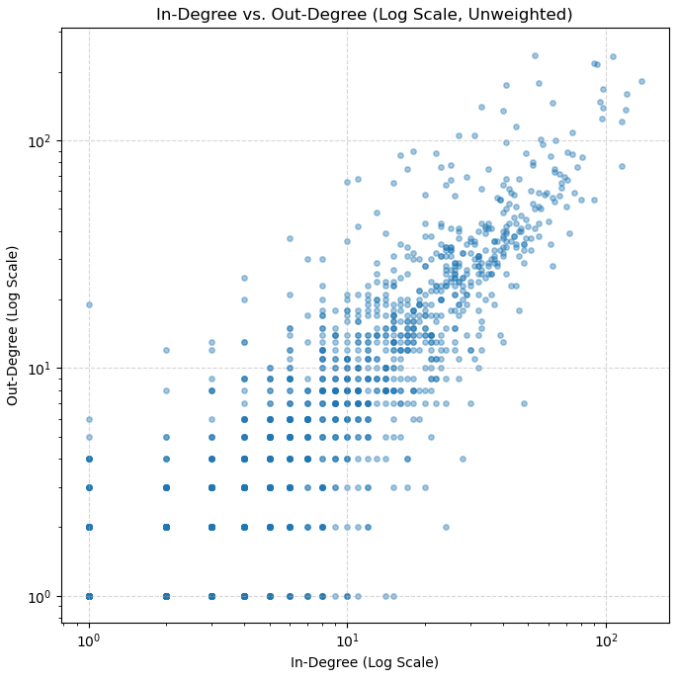

Figure 6: Scatter plot of user in-degree vs. out-degree. The strong positive correlation ($\rho \approx 0.83$) indicates symmetry in individual user activity.

2.4 Uncovering Social Groups: Community Structure

Beyond individual roles, networks often organize into meso-scale communities. Applying the **Louvain algorithm** directly to the directed graph (using message counts as weights) revealed such underlying organization, partitioning the network into **16 distinct communities**. The partition achieved a **modularity** score of approximately **0.3373**, indicating a meaningful structure where internal connections are significantly denser than expected by chance. This suggests genuine social clustering, visualized in Figure 7.

These communities exhibit significant **heterogeneity in size**. A few large communities (the top five containing **281, 214, 179, 173, and 162 nodes**) dominate the social landscape, while numerous smaller communities, including pairs, populate the periphery. This suggests a structure with core groups alongside more specialized or isolated clusters.

Analysis of internal characteristics (without detailing per-community stats) reveals further variation. Larger communities tend to be internally sparser (lower density) but contain active members (indicated by higher average internal degrees). Conversely, some smaller communities exhibit higher density, suggesting tighter-knit groups. Importantly, highly central users (in terms of global PageRank or Betweenness) do not appear concentrated within single communities but are distributed across several, potentially serving as bridges between these groups.

3 Final Considerations

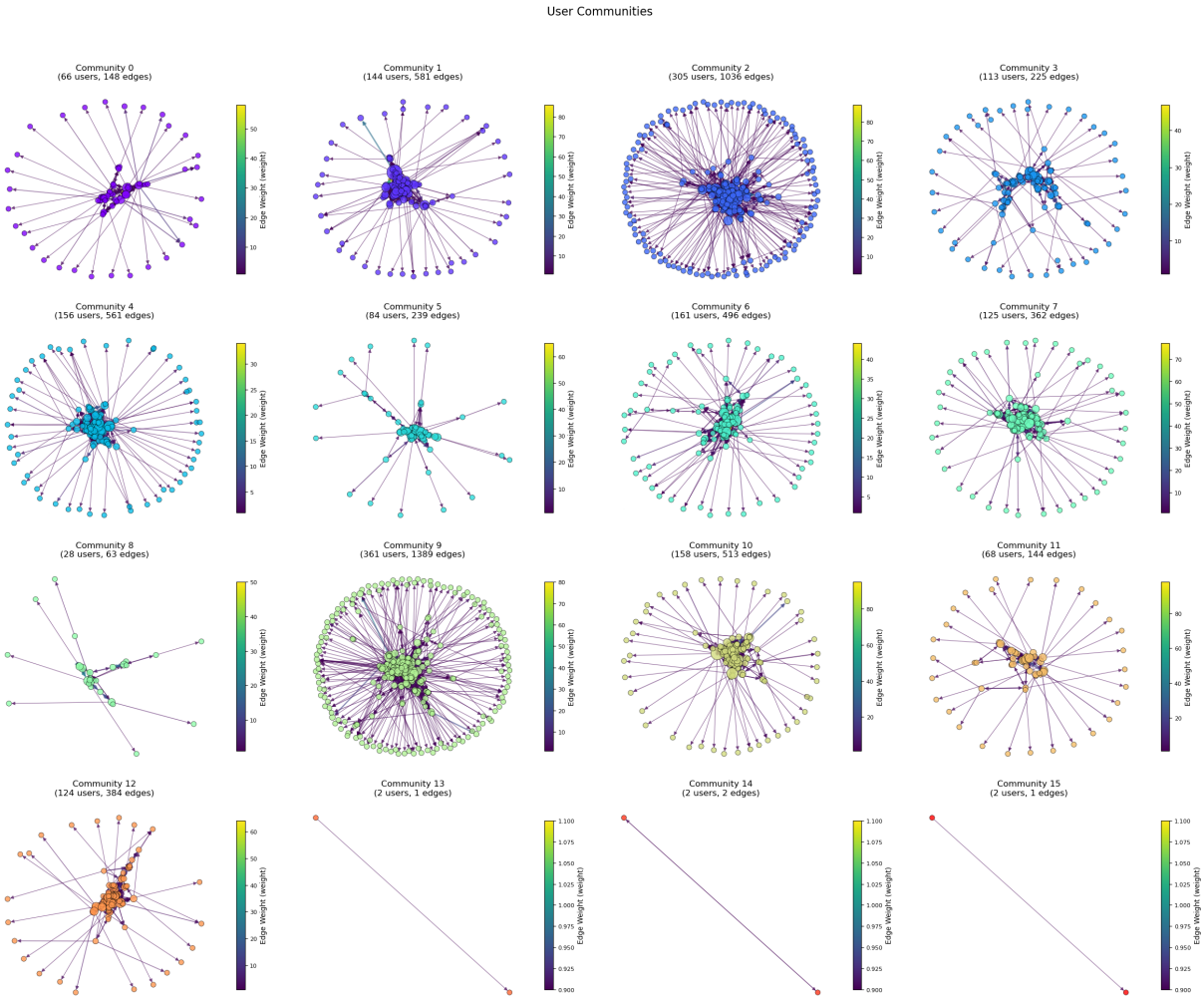

Figure 7: Network visualization coloured by detected community membership (16 communities found via Louvain).