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

To initiate the analysis, it was essential to first examine the structure of the dataset. The dataset contains private messages exchanged on an online social network at UC Irvine. Each entry represents a message characterized by the following attributes:

- `SRC`: Identifier of the sender
- `TGT`: Identifier of the receiver
- `UNIXTS`: Timestamp of the message in Unix time format

Given the non-human-readable format of UNIX timestamps, these were converted into datetime objects. To ensure temporal accuracy in the context of UC Irvine, all timestamps were further adjusted from UTC to the Pacific Time Zone (America/Los_Angeles), accounting for daylight saving time.

1.2 Data Exploration

As an initial step, the following questions were posed: *”How large is the dataset? How many users are involved and over what time span does the activity occur?”*

- Total number of messages: 59,835
- Total number of unique users: 1,899
- Temporal coverage: From April 15, 2004 to October 26, 2004

These findings confirmed that the dataset encompasses over six months of communication among nearly two thousand participants.

1.3 Message Volume Over Time

To identify communication trends, the number of messages exchanged per day was aggregated. As shown in Figure 1, notable fluctuations were observed, with certain peaks that may coincide with academic milestones or periods of heightened social activity.

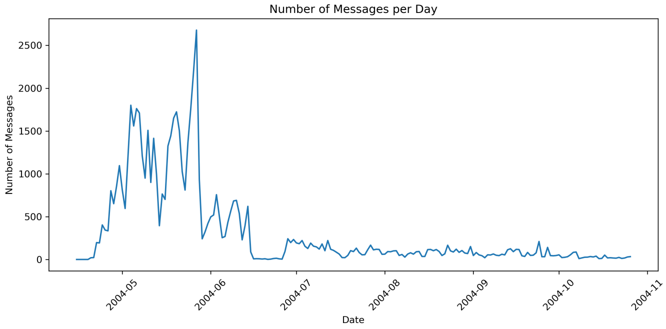

Figure 1: Daily message volume over time

1.4 Temporal Analysis

During the preliminary analysis of hourly messaging patterns, an anomaly was noted: a substantial number of messages appeared to be sent between 5:00 and 8:00 a.m. UTC—an unusual timeframe for student activity.

This observation suggested a potential misalignment between recorded and actual local times. Considering the dataset originates from UC Irvine, the hypothesis emerged that timestamps required conversion to Pacific Time. Upon applying this correction, the messaging peak appropriately shifted to between 10:00 a.m. and 1:00 p.m.

To develop a comprehensive understanding of temporal behavior, message frequencies were further analyzed across hours of the day, weekdays, and months. The resulting trends are presented collectively in Figure 2.

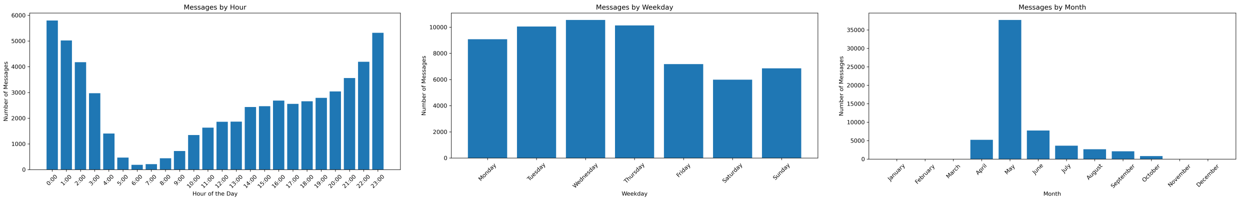

Figure 2: Temporal breakdown of message activity by hour, weekday, and month (adjusted to local time)

The analysis reveals heightened activity during mid-morning to early afternoon, reduced interaction over weekends, and seasonal variations, with communication peaking in spring and early autumn—likely aligned with the academic calendar.

1.5 Top Users vs Least Users Analysis

Following the identification of User 234 as one of the most active participants, a question arose:

"Do highly central users demonstrate distinct behavioral patterns compared to less active ones?"

To investigate this, PageRank scores were computed to identify the most and least central users. Subsequently, their hourly messaging distributions were analyzed.

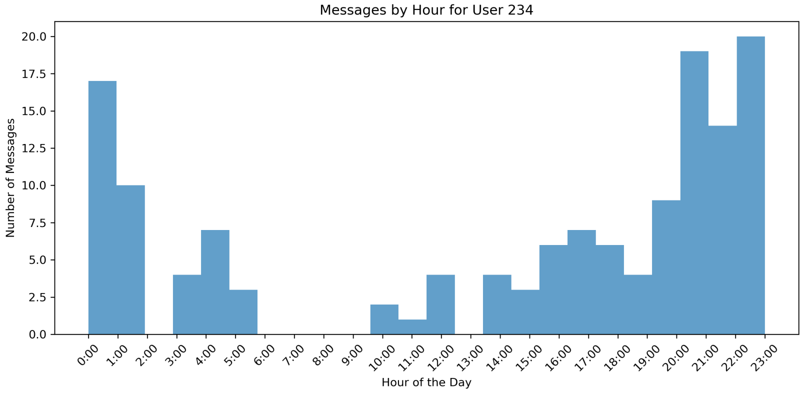

Figure 3: Hourly message distribution for User 234

As illustrated in Figure 3, User 234 maintained a consistent level of activity throughout the day, with noticeable intensity in the evening—potentially indicating a central role within the network.

To broaden the perspective, the average hourly activity of the top 10 users (by PageRank) was compared against that of the bottom 1500 users.

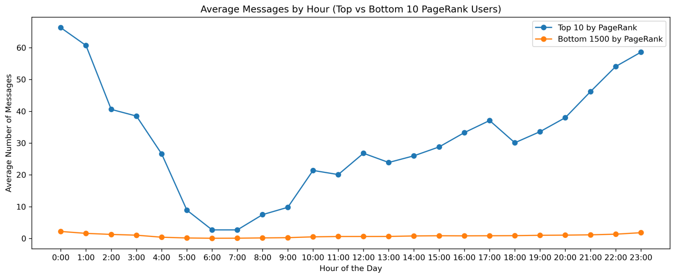

Figure 4: Comparison of average hourly activity between top and bottom PageRank users

As shown in Figure 4, top-ranked users exhibited a more consistent and structured communication pattern, particularly active during typical daytime hours. In contrast, lower-ranked users displayed sporadic and irregular messaging behavior, further reinforcing the relevance of centrality in understanding user roles within the network.

2 Netwrok Analysis

2.1 Network Structure Analysis

The network architecture exhibits key social network characteristics. The calculated **graph density** is `0.0056`, confirming a sparse structure typical for such networks where most potential connections are absent. Despite sparsity, analysis of **weakly connected components (WCCs)** reveals high overall connectivity, with a single giant component encompassing 1893 nodes (over 99.6% of users), facilitating potential information diffusion.

Considering message directionality (Figure 5), the **strongly connected components (SCCs)** analysis identified 601 components, dominated by a large core of 1294 users (≈ 68%) capable of reciprocal communication. The numerous smaller components highlight a significant periphery, indicating a distinct core-periphery structure critical to information flow dynamics.

Local structure analysis shows moderate **unweighted clustering** (`0.087`), suggesting inherent social grouping beyond random chance. However, the **weighted clustering coefficient** (using message counts) is markedly lower (`0.0018`). This significant difference implies that while structural triads are present, the intensity of communication within these local groups is often unevenly distributed.

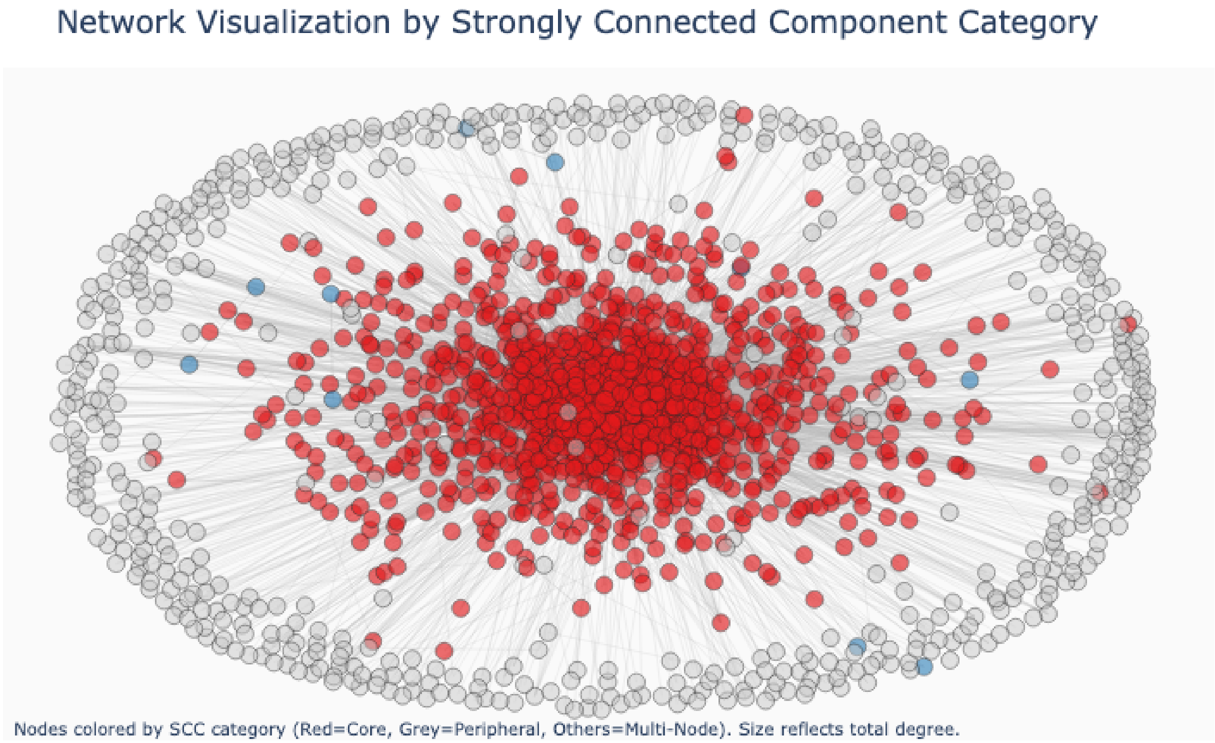
Network Visualization by Strongly Connected Component Category
Nodes colored by SCC category (Red=Core, Grey=Peripheral, Others=Multi-Node). Size reflects total degree.

Figure 5: Network visualization highlighting Strongly Connected Components (SCCs). Nodes coloured by SCC category (Red: Core; Grey: Peripheral; Others: Multi-Node SCCs). Size reflects degree.

2.2 Expanded Centrality Analysis

Beyond degree and PageRank (Section 1.4), **Eigenvector Centrality** and **Betweenness Centrality** provide further insights into node influence and structural roles.

Eigenvector centrality highlights influence via connections to other influential nodes. Top users like `User 1624` (score ≈ 0.47) and `User 398` (≈ 0.30) demonstrate high influence despite not having top degrees, indicating strategic positioning. Others, like `User 105` (≈ 0.27), combine high Eigenvector scores with high out-degree and prominence in other rankings.

Betweenness centrality identifies crucial network brokers. Top users `User 323` (≈ 0.12), `User 1624` (≈ 0.08), and `User 105` (≈ 0.08) score highly, reinforcing their central and bridging roles suggested by other measures. The appearance of other high-degree/PageRank users (like `User 32`, `User 103`) in the top 10 for Betweenness indicates that popular users can also be important intermediaries.

Analysis of **Spearman rank correlations** reveals strong positive associations between all centrality measures (most $\rho > 0.7$). In-degree shows particularly high correlation with PageRank ($\rho \approx 0.93$) and Eigenvector ($\rho \approx 0.90$). Betweenness centrality displays slightly weaker, yet still strong, correlations ($\rho \approx 0.72 - 0.81$), suggesting it captures the most distinct structural aspect (brokerage). Overall, these strong correlations indicate that different facets of importance (popularity, influence via connections, brokerage) are significantly intertwined in this network.

2.3 Network Patterns: Degree Correlation and Activity Symmetry

Examining network mixing patterns, the **degree assortativity coefficient** was `-0.1375`. This negative value indicates **disassortativity**: high-degree users tend to connect with low-degree users, suggesting a hub-and-spoke structure rather than interconnected cores of highly active users.

In contrast to this network-level pattern, individual user activity shows symmetry. A strong positive **Pearson correlation ($\rho \approx 0.83$)** exists between user in-degree and out-degree, meaning users receiving many messages typically also send many. Thus, while the network structure promotes connections between different activity levels, highly engaged individuals tend to be active communicators in both directions.

2.4 Community Structure

To uncover meso-scale organization, we applied the **Louvain algorithm** for community detection. The analysis partitioned the network into **16 distinct communities** with a **modularity score** of approximately `0.3607`. This modularity value indicates a meaningful, albeit imperfect, community structure significantly better than random partitioning.

The detected communities exhibit considerable heterogeneity. Sizes range from large groups (e.g., Community 9: 361 nodes, Community 2: 305 nodes) down to three 2-node communities (likely isolated pairs). Internal structure also varies; the largest communities tend to be internally sparse, while some smaller communities show higher internal density or higher average internal centrality metrics. This community structure is visualized in Figure 6, illustrating the arrangement and relative sizes of the main groups.

The presence of this discernible community structure provides valuable insights into the social segmentation within the network.

3 Final Considerations

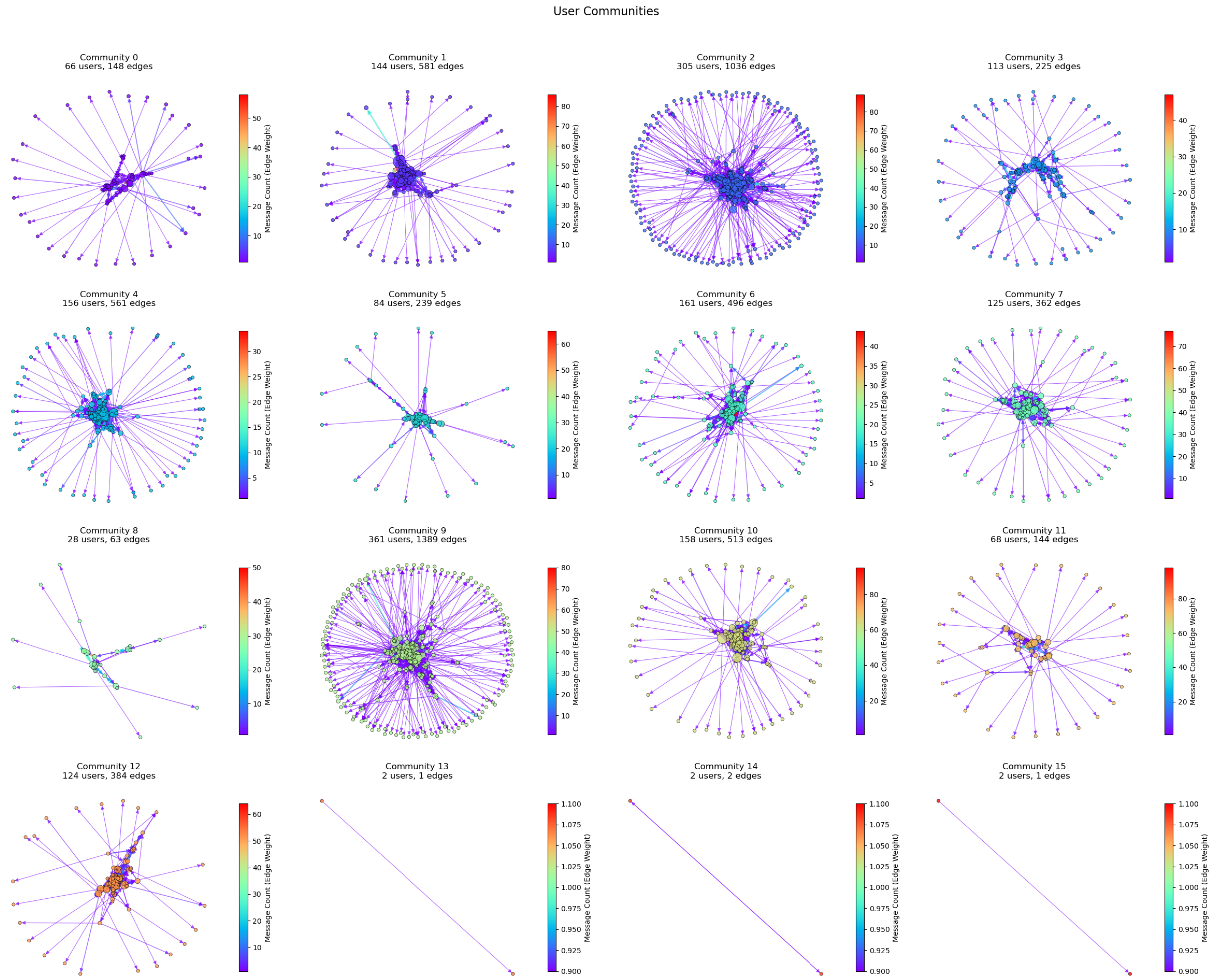

Figure 6: Network visualization with nodes coloured by detected community membership (Louvain algorithm).