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 Dataset Overview

To begin the analysis, I first wanted to understand the basic structure of the dataset. The dataset contains private messages exchanged on an online social network at UC Irvine. Each row represents a message with the following fields:

- `SRC`: ID of the sender
- `TGT`: ID of the receiver
- `UNIXTS`: timestamp of the message in Unix time

Since UNIX timestamps are not human-readable, I converted them into datetime objects. To make the time information meaningful in the local context of UC Irvine, I also converted the times from UTC to Pacific Time (America/Los_Angeles), accounting for daylight saving time.

1.2 First Impressions

I asked myself: *"How big is this dataset? How many users are involved and over what time span?"*

- Number of messages: 59,835
- Unique users: 1,899
- Time range: From April 15, 2004 to October 26, 2004

This confirmed that the data spans over six months and involves a sizable number of participants.

1.3 Message Volume Over Time

Curious about activity trends, I aggregated the number of messages sent per day. Figure 1 shows clear fluctuations, including spikes in activity that might correspond to academic events or social dynamics within the university.

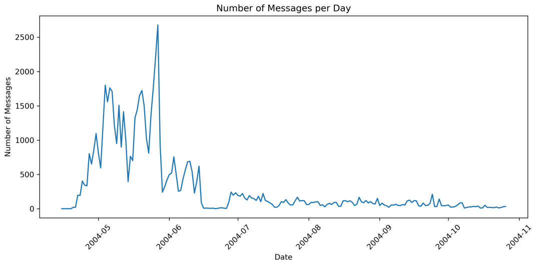

Figure 1: Daily message volume over time

1.4 Temporal Distribution by Hour, Weekday, and Month

Messages by Hour

Initially, I plotted the number of messages sent per hour of the day, using UTC timestamps. This produced a surprising result: a peak in activity between 5:00 and 8:00 a.m. This seemed odd for student behavior.

Hypothesis: The timestamps were in UTC. If users are based in California, messages should be shifted 7 hours backward.

After converting timestamps to Pacific Time, the new plot (Figure 2) showed a peak between 10:00 a.m. and 1:00 p.m., which is far more plausible.

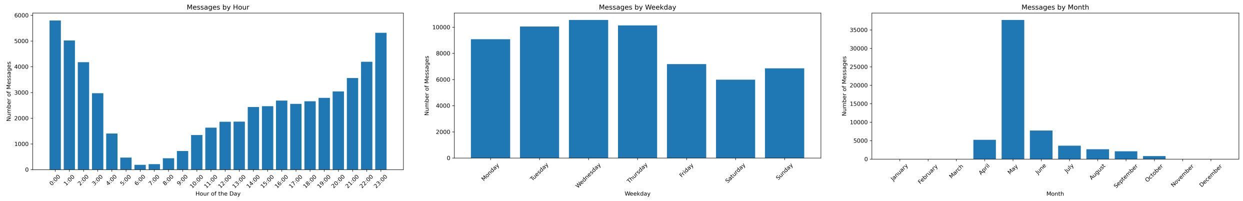

Figure 2: Messages by hour of day (after UTC to PST conversion)

Messages by Weekday

Next, I asked: *"Do students behave differently on weekdays compared to weekends?"*

Figure 2 reveals a drop in activity during weekends, suggesting that most messaging activity is related to academic or weekday social interactions.

Messages by Month

Figure 2 shows that messaging activity increased through the spring and summer months, peaking around mid-year and then tapering off.

1.5 Network Construction

I then asked: *"How are these users connected? Can we build a graph from the data?"*

To answer this, I constructed a directed graph using the sender and receiver columns. Repeated communications between users were aggregated as weighted edges.

- Nodes: 1,899
- Edges: 20,296
- Graph density: 0.0056

The graph is sparse, but this is typical in social networks where users only communicate with a small subset of others.

1.6 User-Level Analysis: Active vs Inactive Users

I wanted to find out: *"Are there certain users who dominate the conversation? What are their behavioral patterns?"*

To address this, I computed PageRank scores to identify the most central users. I focused on User 234 as one of the most active.

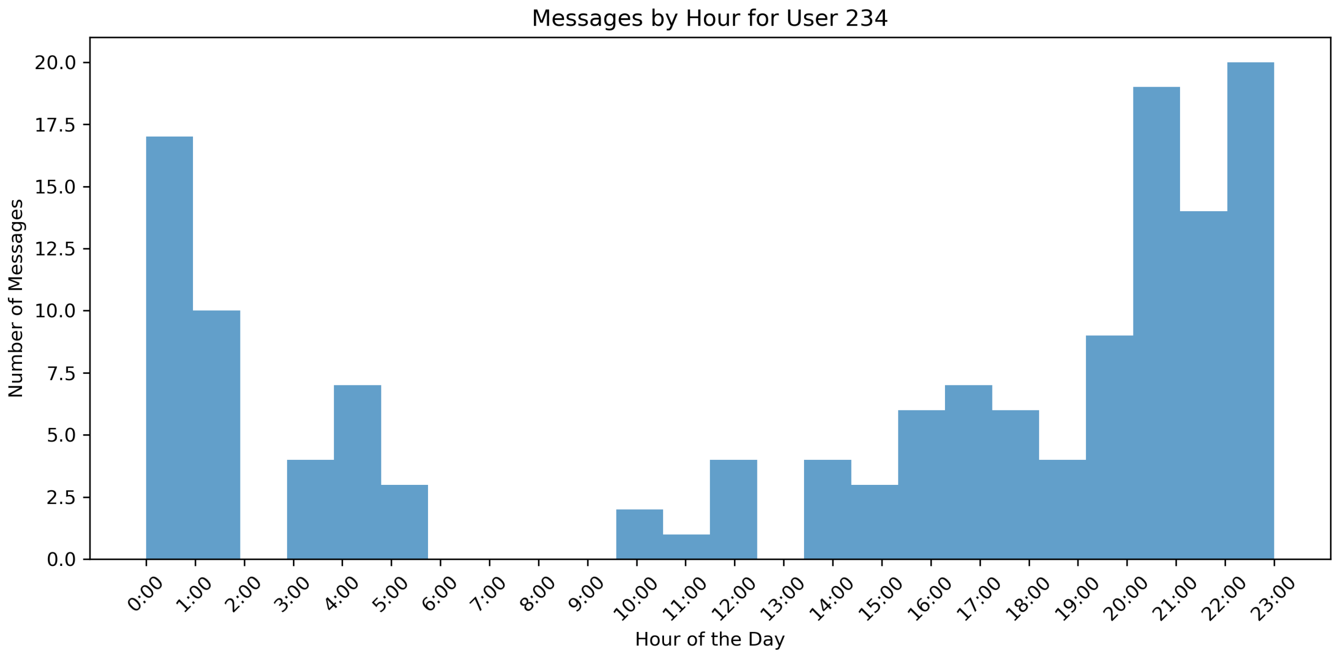

Figure 3: Messages by hour for User 234

As seen in Figure 3, this user was active at all hours, with a peak in the late evening.

1.7 Comparing Central vs Peripheral Users

Finally, I asked: *"Do central users have different messaging patterns compared to peripheral users?"*

I computed the average number of messages per hour for:

- The top 10 users by PageRank
- The bottom 1500 users by PageRank

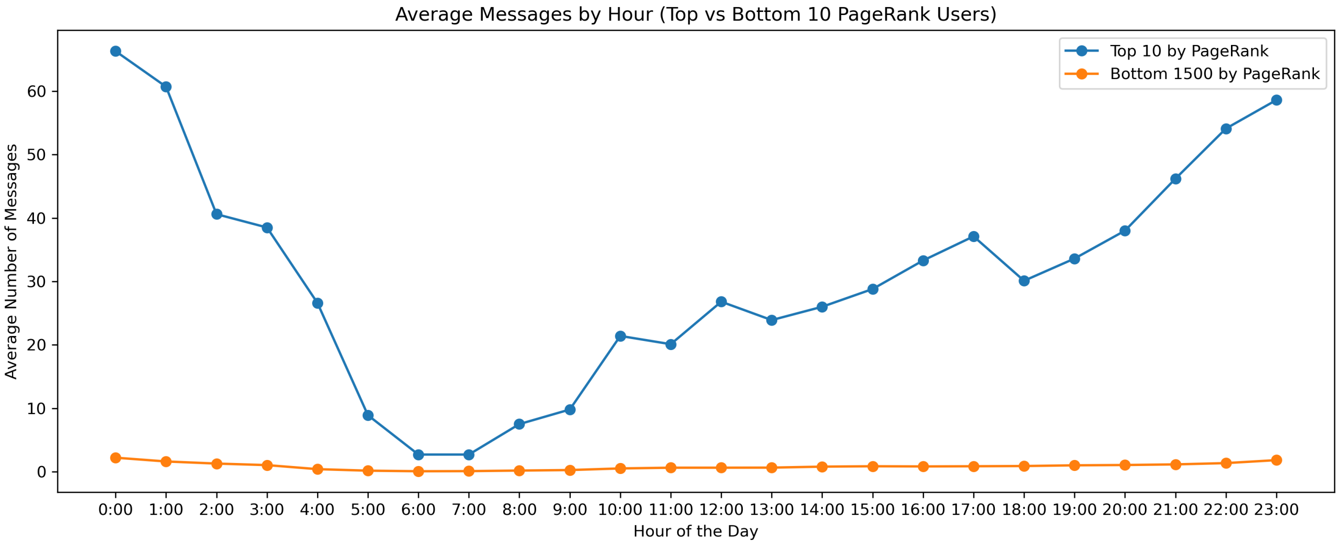

Figure 4: Average hourly activity: Top vs. Bottom PageRank users

The results (Figure 4) show that top users are more consistent and active throughout the day, especially during working hours, while peripheral users show sparse and inconsistent patterns.

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.

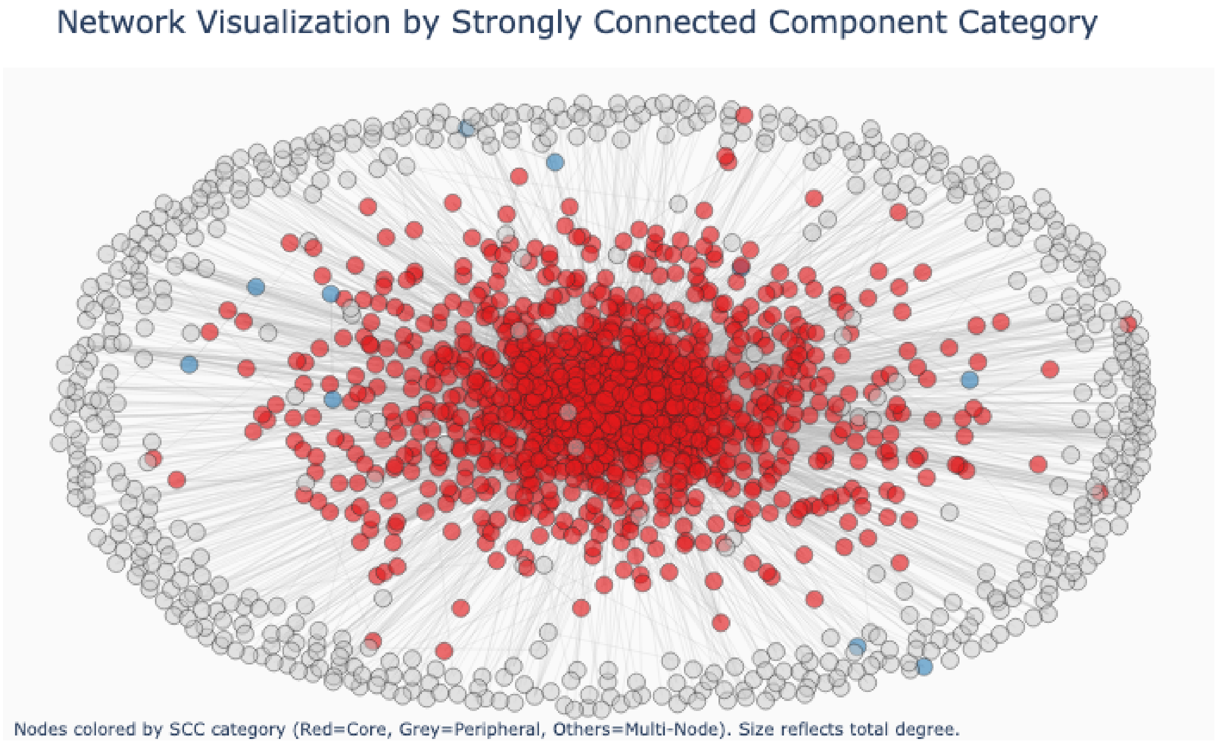
Network Visualization by Strongly Connected Component Category
Nodes colored by SCC category (Red=Core, Grey=Peripheral, Others=Multi-Node). Size reflects total degree.

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

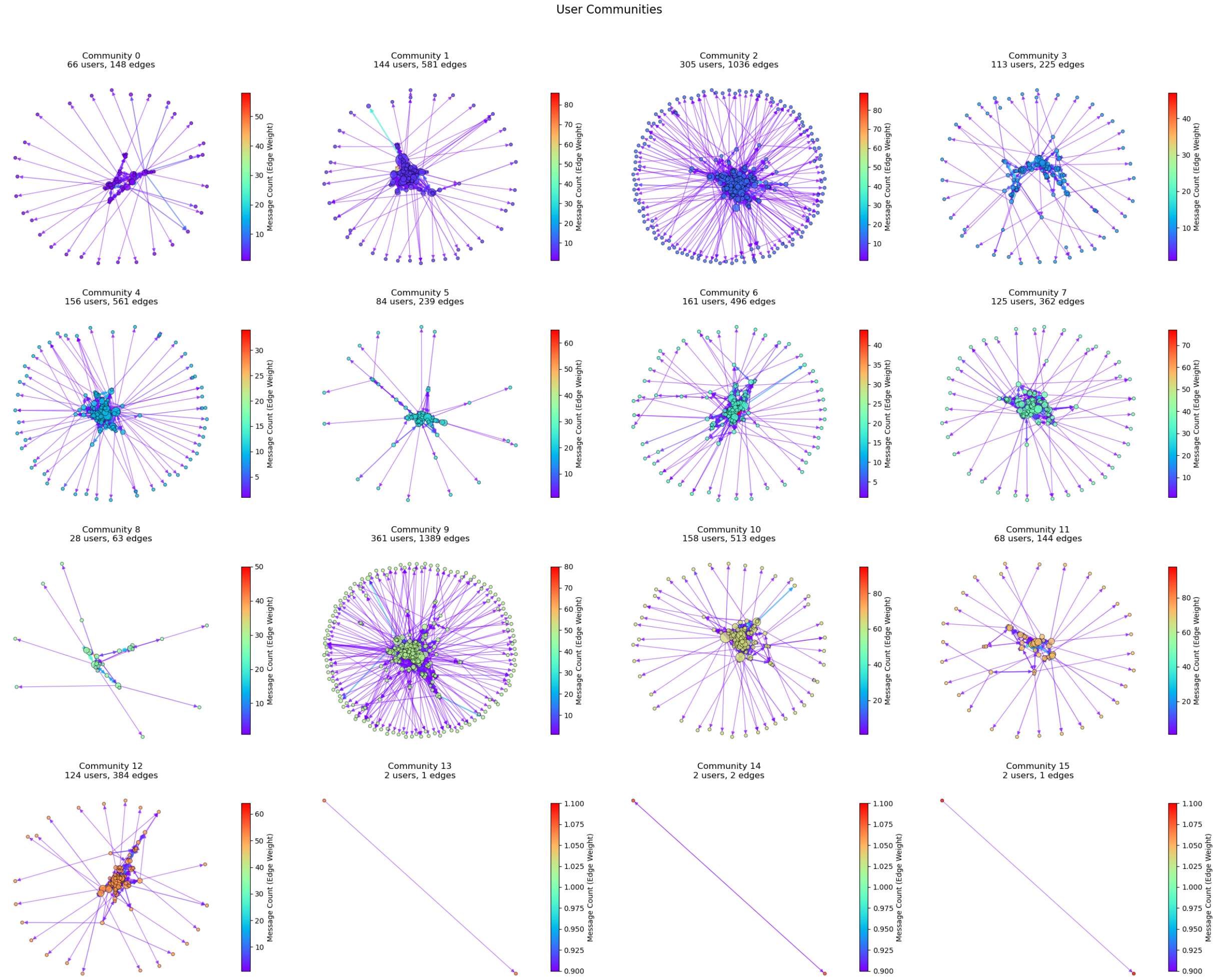

Figure 6: Network visualization with nodes coloured by detected community membership (Louvain algorithm).