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 Initial Exploration

Before diving into the analysis, we started by taking a look at the structure of the dataset. It contains private messages exchanged on an online social network at UC Irvine. Each row represents a message with three main pieces of information:

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

Since Unix time isn't very human-readable, we converted the timestamps into proper datetime objects. We also adjusted them to Pacific Time (America/Los_Angeles), which is the timezone used at UC Irvine. This made it easier to interpret the data accurately.

As a first step, we asked ourselves a few basic questions: *"How big is this dataset? How many users are involved? Over what time period were these messages exchanged?"*

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

These numbers show that the dataset covers just over six months of messaging among nearly 2,000 users.

1.2 Message Volume Over Time

To get an idea of how activity changed over time, we counted how many messages were sent each day. As shown in Figure 1, there are clear fluctuations, with some noticeable spikes. These might match up with academic deadlines or social events on campus.

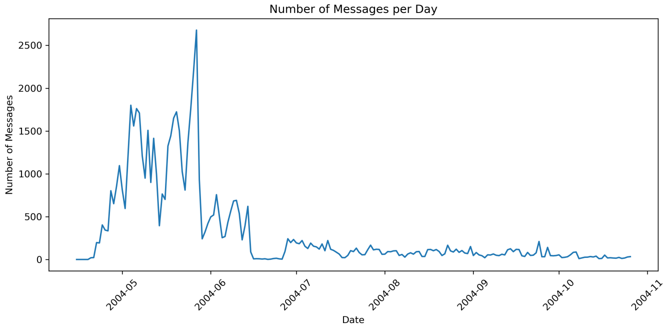

Figure 1: Daily message volume over time

One thing worth pointing out is that the busiest period falls between April and June 2004, which lines up with UC Irvine's spring term. This supports the idea that the platform was most used during the academic semester.

1.3 Temporal Analysis

Next, we wanted to see if students followed any specific patterns when sending messages. We looked at three time-based breakdowns: by hour of the day, by weekday, and by month. The results are shown in Figure 2.

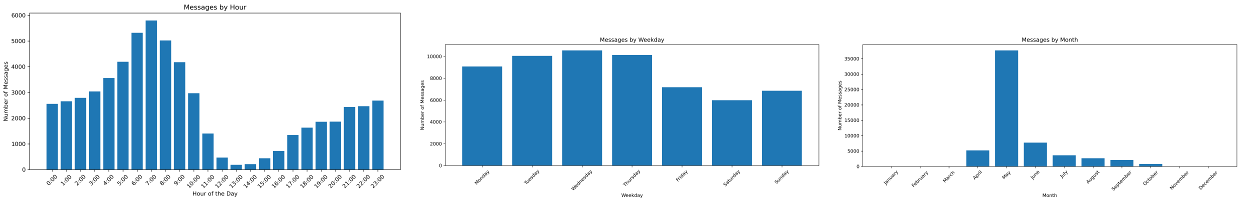

Figure 2: Temporal breakdown of message activity by hour, weekday, and month (adjusted to local time)

However, something strange popped up right away: a lot of messages were sent between 5:00 and 8:00 a.m. UTC. That didn't seem right—why would students be so active that early in the morning? That made us think there might be an issue with the timezone.

1.4 UTC Time Conversion

Since the dataset is from UC Irvine, which is in the Pacific Time Zone (UTC–7), and the timestamps were in UTC, we figured we needed to adjust the time accordingly. After converting to Pacific Time, the peak shifted to around 10 a.m. to 1 p.m., which made a lot more sense.

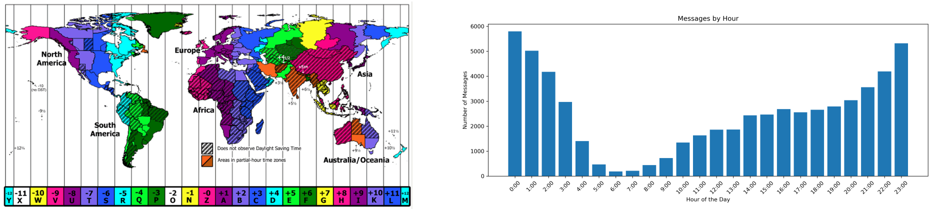

Figure 3: **Left**: UTC time zone map. **Right**: Hourly message distribution after conversion to local time

This small fix cleared up the confusion. The data was fine—it just needed the correct timezone. With this adjustment, the messaging behavior matched what you'd expect from a student population.

1.5 Top Users vs Least Users Analysis

While exploring the data, we noticed that User 234 was particularly active. This got us thinking:

"Do the most central users behave differently from users who are active but not very well connected in the network?"

To find out, we used the PageRank algorithm to rank all users and then selected:

- The top 10 users by PageRank
- The bottom 10 users by PageRank among those with non-zero out-degree (so, users who actually sent messages)

We then compared how these two groups behave throughout the day by averaging their message activity hour-by-hour.

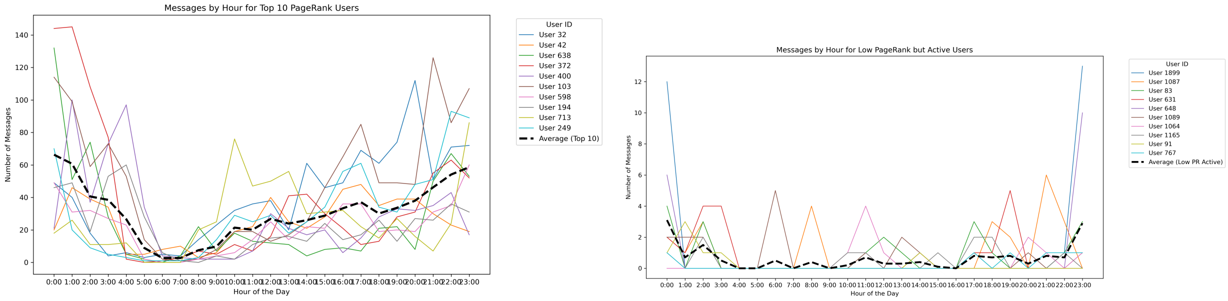

Figure 4: Comparison of average hourly activity between top 10 and bottom 10 active PageRank users

Looking at Figure 4, it seems like the top users have a more stable and consistent messaging pattern throughout the day. Their activity spans many hours with smaller fluctuations. On the other hand, the less central users show a more sporadic and concentrated behavior, often limited to specific times. This could mean that more central users are generally more engaged in ongoing conversations, while less connected users tend to message in bursts.

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 ($\approx$ 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.

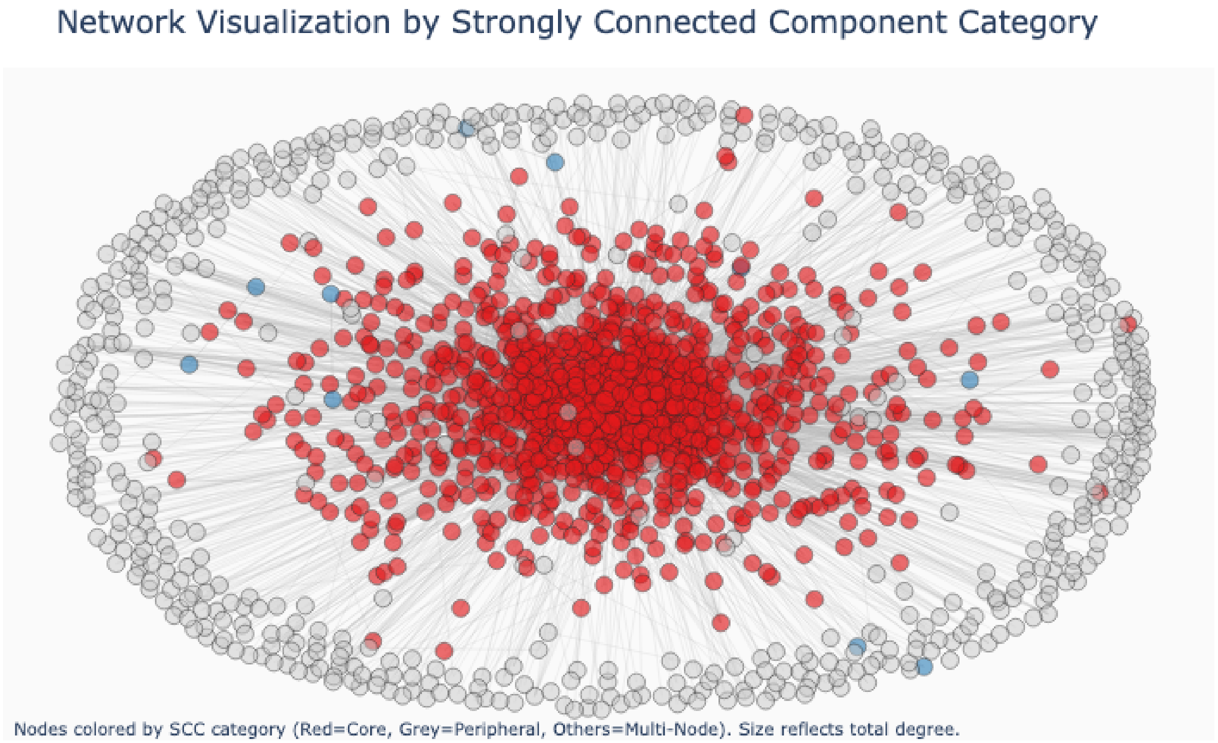
Network Visualization by Strongly Connected Component Category
Nodes colored by SCC category (Red=Core, Grey=Peripheral, Others=Multi-Node). Size reflects total degree.

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 $\approx$ 0.47) and `User 398` ($\approx$ 0.30) demonstrate high influence despite not having top degrees, indicating strategic positioning. Others, like `User 105` ($\approx$ 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` ($\approx$ 0.12), `User 1624` ($\approx$ 0.08), and `User 105` ($\approx$ 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

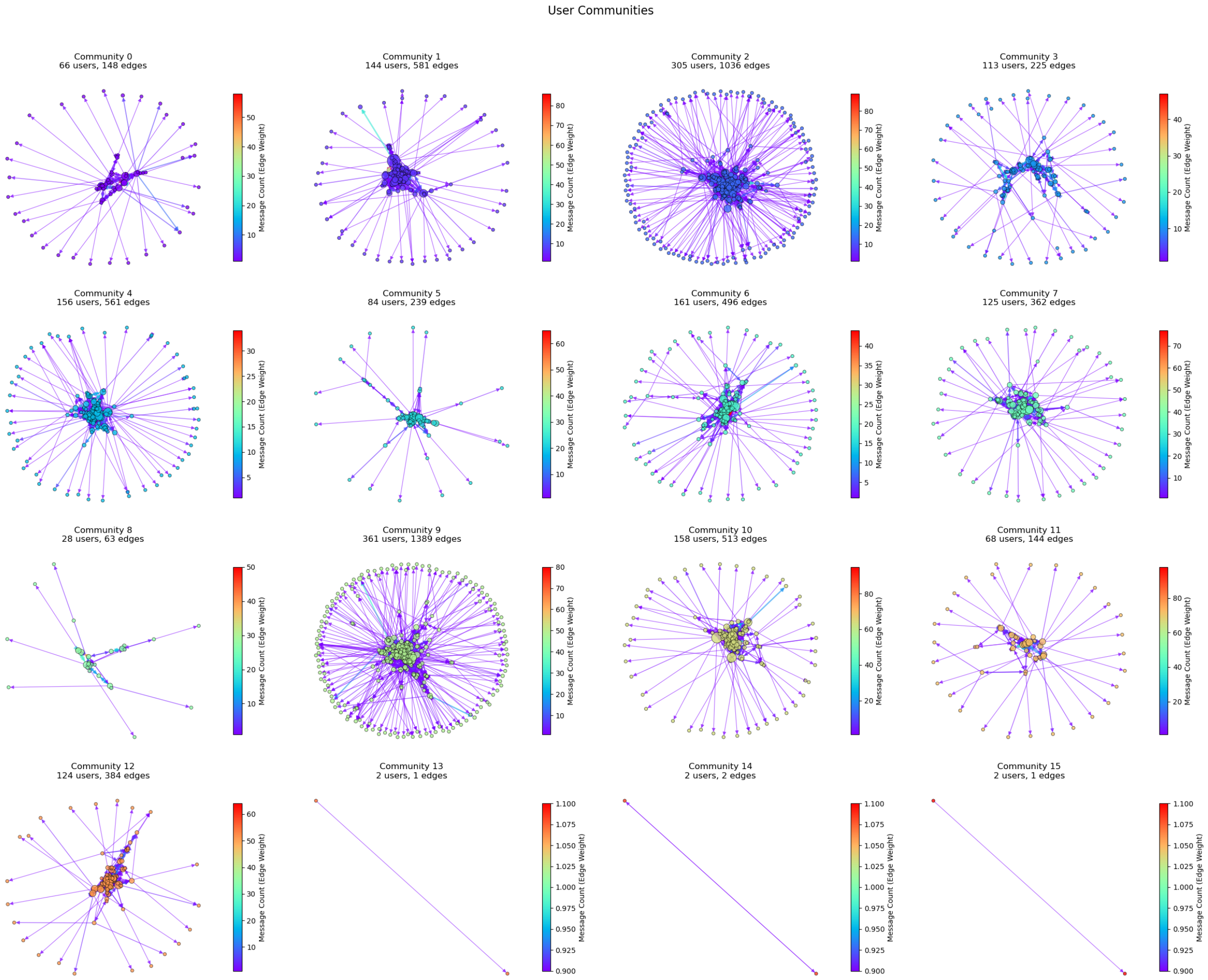

Figure 6: Network visualization with nodes coloured by detected community membership (Louvain algorithm).