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

Setup

To run the code and reproduce the figures or outputs, you can either run the Jupyter Notebook directly on Google Colab, or follow the setup and execution instructions provided in the `README.md` file included in the repository.

Introduction

This document presents an analysis of a dataset of private messages exchanged on a UC Irvine social network. The dataset includes 1,899 users and 59,835 messages, each represented as a directed interaction from a sender (SRC) to a recipient (TGT) with a Unix timestamp (UNIXTS).

The report is structured into two main parts. The first part focuses on data exploration and temporal activity analysis, where we study messaging behavior over time and among different types of users. The second part covers network-based analysis, where we investigate the structure of the social graph, user centrality, interaction patterns, and community structures

1 Data Analysis

1.1 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.

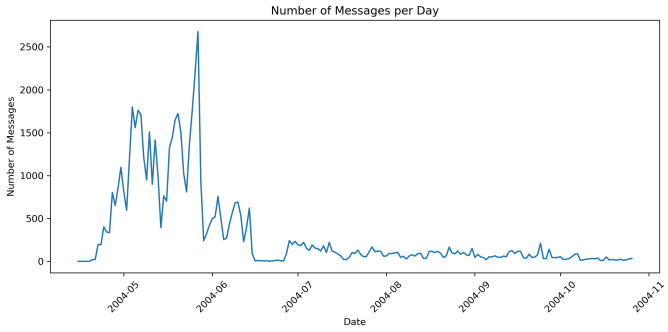

Figure 1: Daily message volume over time

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.

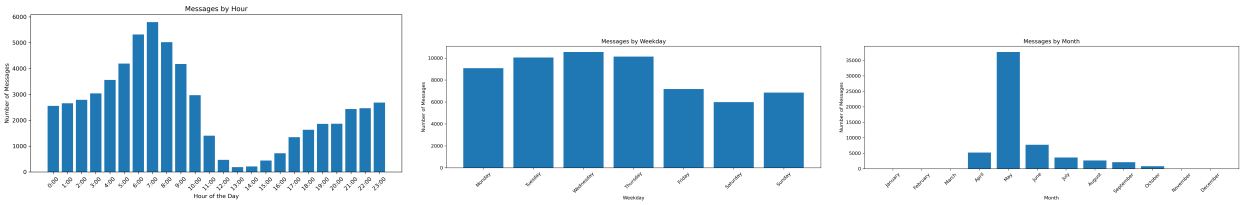

Figure 2: Temporal breakdown of message activity by hour, weekday, and month (adjusted to local time)

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.

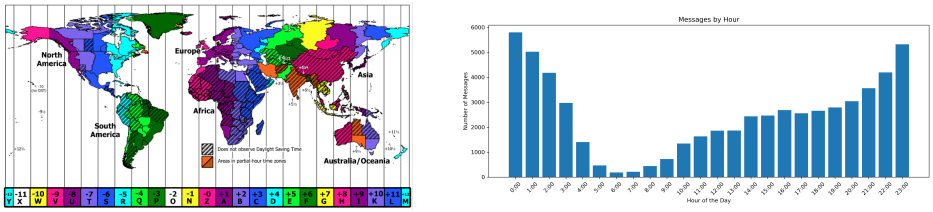

Figure 3: **Left**: UTC time zone map. **Right**: Hourly message distribution after conversion to local time

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.

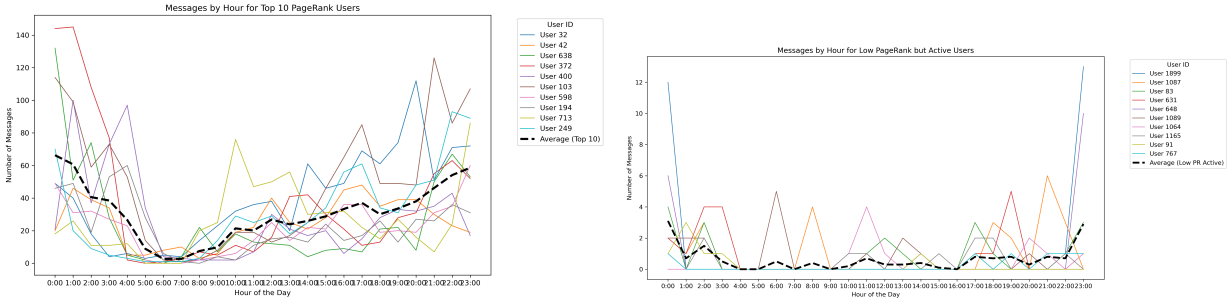

Figure 4: Comparison of average hourly activity between top 10 and bottom 10 active PageRank users

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: 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.

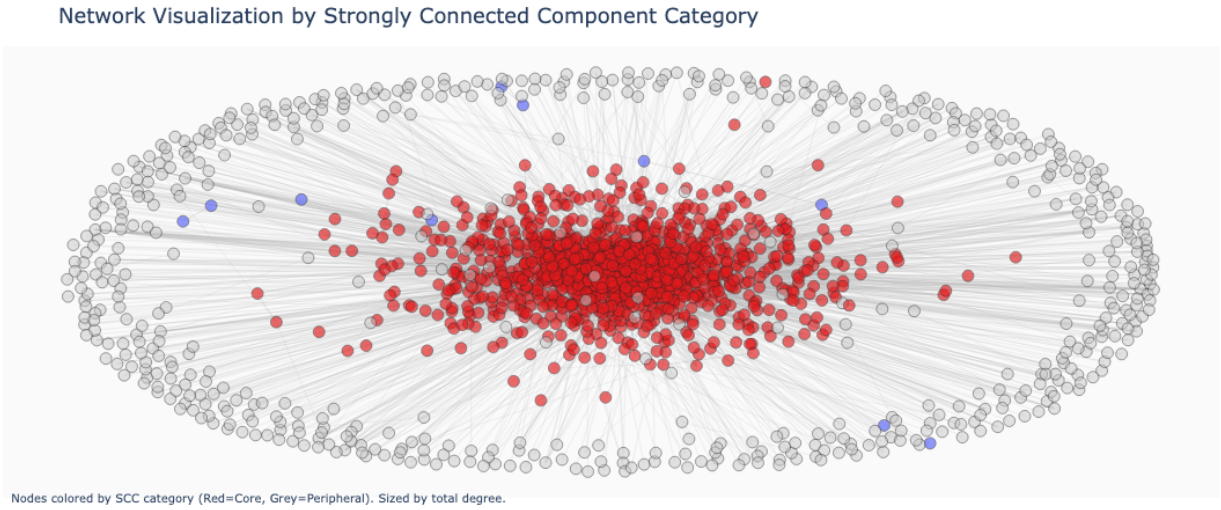

Figure 5: Network visualization highlighting Strongly Connected Components (SCCs). Nodes could be coloured by SCC category (e.g., Core vs. Periphery) and size reflecting degree.

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.

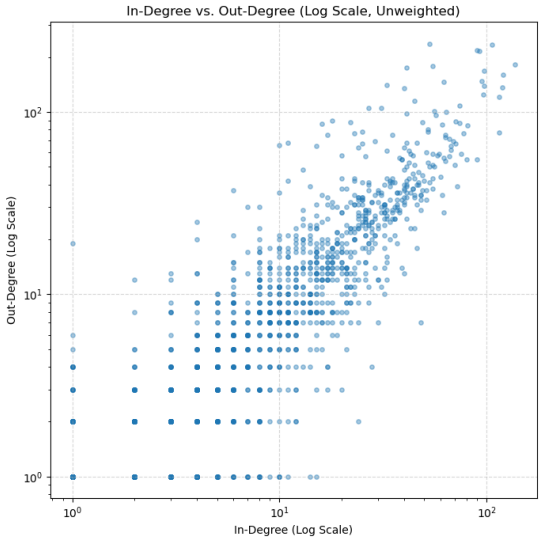

Figure 6: Scatter plot of user in-degree vs. out-degree.

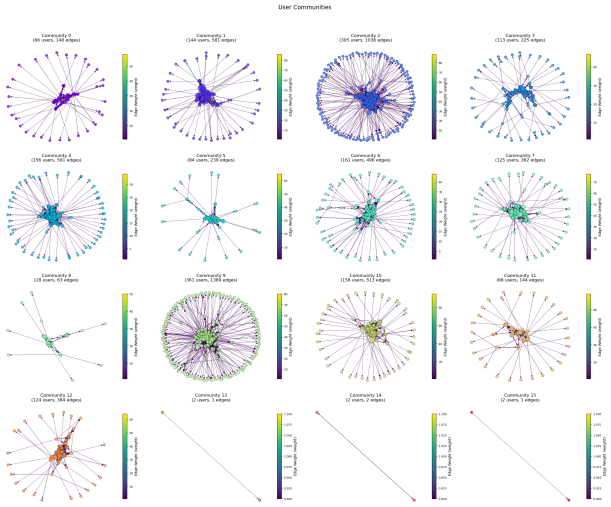

Figure 7: Network visualization coloured by community membership.

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.