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.

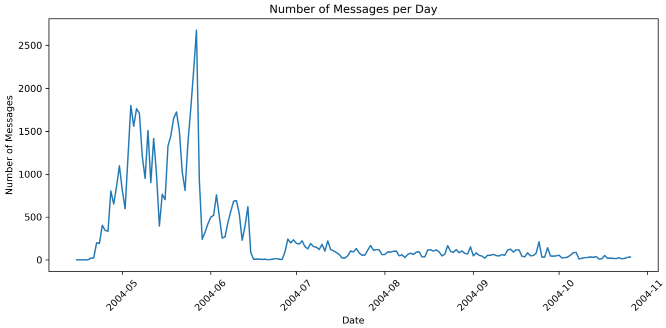

Figure 1: Daily message volume over time

It is worth noting that the observed period of high activity, especially between April and June 2004, corresponds to the spring academic term at UC Irvine. This aligns with the plausible hypothesis that the social platform was actively used during regular academic semesters, when students are most engaged.

1.4 Temporal Analysis

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

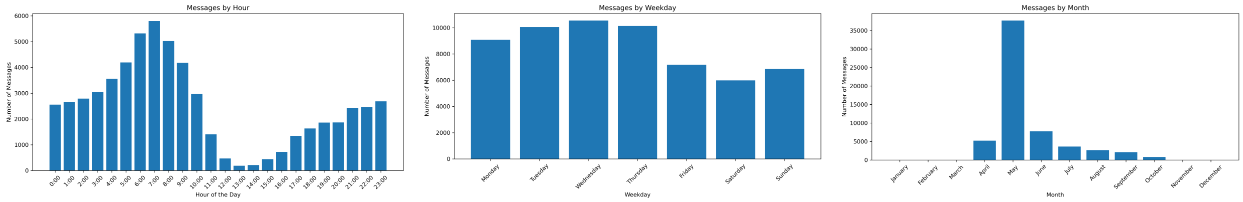

Figure 2: Temporal breakdown of message activity by hour, weekday, and month (adjusted to local time)

From this initial overview, it became evident that students were particularly active between 5:00 and 8:00 a.m. UTC. This observation appeared counterintuitive, as such early hours are not typically associated with peak student communication. It prompted further investigation to verify whether this pattern was the result of a timezone misalignment.

1.5 UTC Time Conversion

Upon reviewing the data origin, it was confirmed that the dataset is based at UC Irvine, which operates under the Pacific Time Zone (UTC–7). The recorded timestamps, however, were in UTC, thereby introducing a consistent 7-hour offset.

By applying an appropriate timezone conversion to Pacific Time, the messaging peak shifted to a more plausible interval between 10:00 a.m. and 1:00 p.m., aligning with expected behavioral patterns.

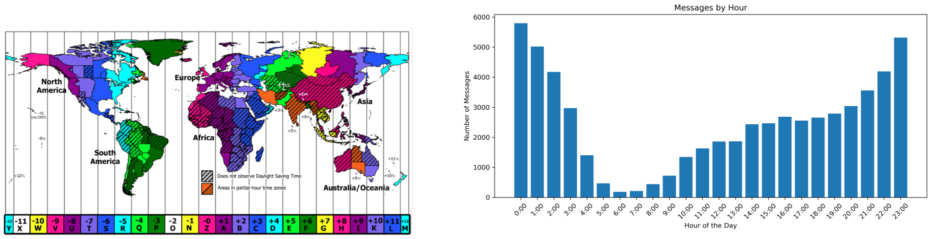

Figure 3: **Left**: UTC time zone map. **Right**: Hourly message distribution after conversion to local time

This correction confirmed that there was no error in the data. The original hourly anomaly was a consequence of overlooking the UTC format. Once adjusted, the temporal patterns aligned well with typical student communication habits, validating both the dataset and its preprocessing.

1.6 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.

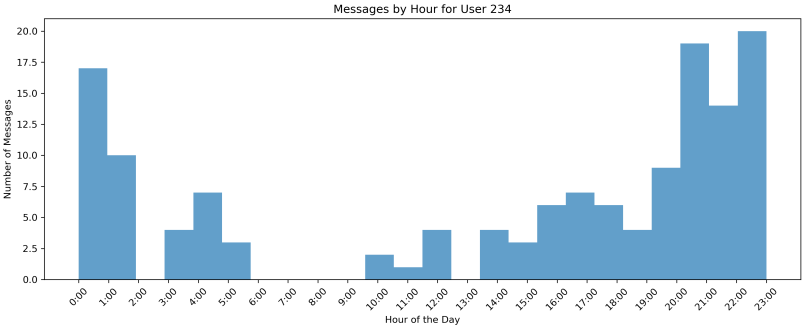

Figure 4: Hourly message distribution for User 234

As illustrated in Figure 4, 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.

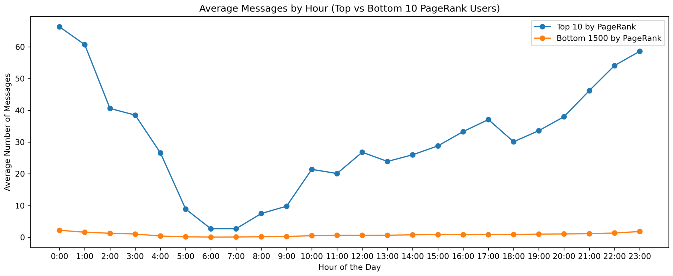

Figure 5: Comparison of average hourly activity between top and bottom PageRank users

As shown in Figure 5, 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.