NETWORK ANALYSIS PROJECT OVERVIEW

Project Title: Network Analysis of MOOC User Interactions

Introduction

This project aims to conduct a network analysis using a dataset from the Stanford Large Network Dataset Collection (https://snap.stanford.edu/data/). The specific dataset chosen is the MOOC User Action Dataset, which captures user interactions in a massive open online course (MOOC) platform. By analyzing this social network, we aim to uncover interaction patterns, engagement trends, and user behavior within the learning environment.

Dataset Description

The dataset used for this analysis is a merged version containing the following columns:

- ACTIONID: A unique identifier for each recorded action.
- USERID: The ID of the user performing the action.
- TARGETID: The ID of the user or entity receiving the action.
- TIMESTAMP: The time when the interaction occurred.
- FEATURE0, FEATURE1, FEATURE2, FEATURE3: Additional numerical attributes describing user behavior.
- LABEL: A classification label for the action.

Research Questions

The analysis is guided by the following six research questions.

For User Activity and Engagement:

1. *How does user activity evolve over time, and are there significant trends or seasonal variations? This helps us understand general engagement patterns and peak activity periods.*

2. *Are there specific users who contribute disproportionately to the network, and what roles do they play (hubs, influencers, or passive participants)? This will identify power users based on network centrality and activity level.*

For Network Structure and Interaction Patterns

3. *What is the overall structure of the user interaction network, and how densely connected is it? This explores network density, clustering, and overall connectivity.*

4. *How do interactions between users change over time, and can we detect evolving communities or shifts in engagement? This incorporates temporal network analysis to see how user relationships form and dissolve over time.*

For Temporal and Predictive Analysis

5. *Can we predict future engagement based on past user activity, and what factors most influence a user's likelihood to remain active? This will help build predictive models on retention and dropout behavior.*

6. *What external or internal factors (e.g., time of day, day of the week, type of action) most significantly impact user interaction levels? This helps optimize platform engagement strategies by identifying the best times to encourage participation.*

Methods and Analytical Approach

To answer these research questions, we apply the following network analysis techniques:

1. Data Preprocessing and Exploration

a. Load the dataset and inspect missing values, duplicates, and data consistency.
b. Convert timestamps to a suitable format for time-based analysis.
c. Filter relevant interactions based on defined criteria (e.g., user-to-user interactions only).

2. Network Construction

a. Build a directed or undirected network graph using NetworkX.
b. Define nodes (users) and edges (interactions) with weights representing interaction frequency.

3. Network Visualization

a. Generate basic network graphs to understand structural properties.
b. Improve visualization clarity by adjusting layout algorithms and filtering low-degree nodes.
c. Use edge weights to highlight stronger connections.

4. Network Metrics & Structural Insights

a. Compute degree centrality, betweenness centrality, and closeness centrality to identify influential users.

b. Apply community detection algorithms (e.g., Louvain method) to find user groups.

c. Analyze temporal patterns to observe changes in network dynamics over time.

5. Data-Driven Findings

a. Compare the most active users with the most central users.

b. Identify clusters of closely interacting users.

c. Evaluate how interaction frequency impacts network influence.

d. Explore the role of user attributes in shaping interaction behavior.

Key Findings and Insights

a. Major interaction hubs: Identification of highly connected users.

b. Community structures: Group formations based on interaction patterns.

c. Temporal evolution: Changes in user engagement over time.

d. Correlation between user attributes and interaction frequency.

e. Insights on user influence: Key players shaping network dynamics.

Data Visualizations & Network Graphs

a. Initial raw network visualization to examine overall connectivity.

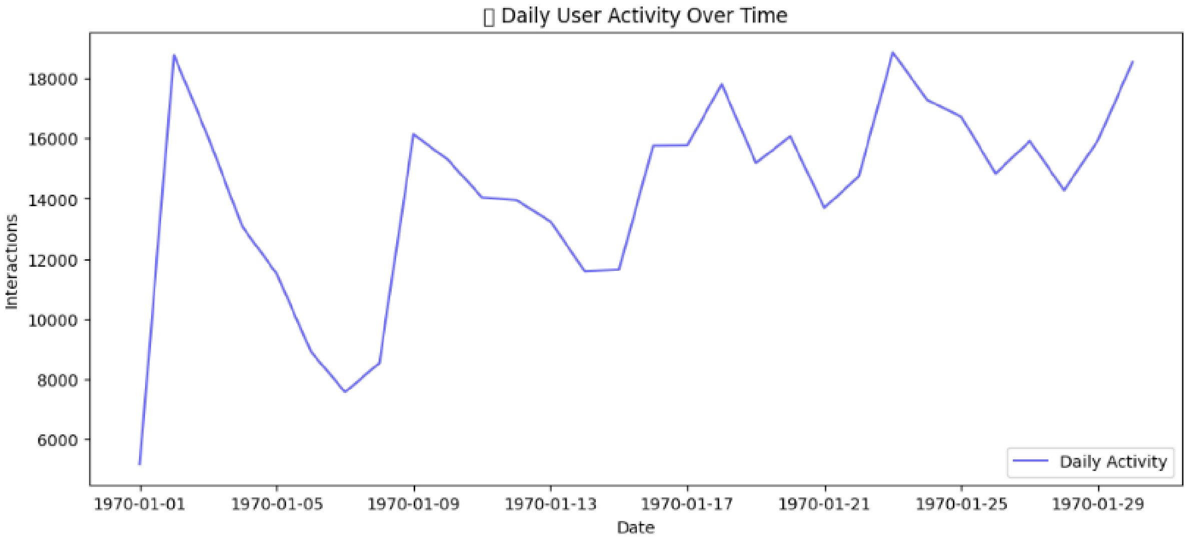
b. Filtered network graph focusing on key interactions.

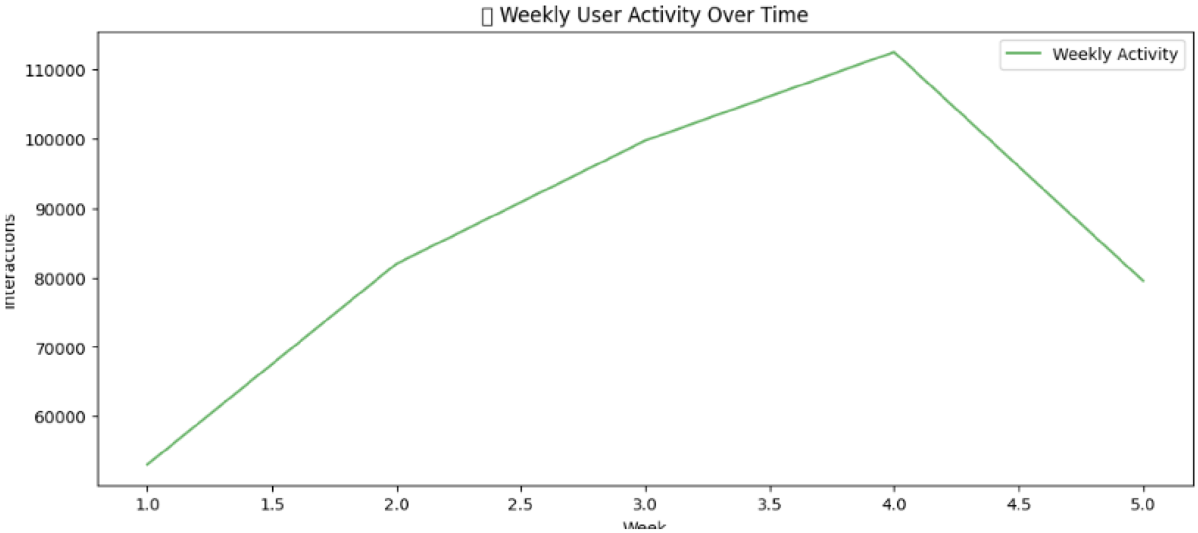
c. Weighted graph visualization with edge thickness representing interaction strength.

d. Community detection visualization to highlight different user groups.

e. Temporal interaction trends plotted to observe engagement patterns.

1. Daily Activity Trend





Interpretation:

These two-line charts depict user activity over time, measured by interactions at both daily and weekly intervals.

Top Chart: Daily User Activity Over Time

- The X-axis represents the date.

- The Y-axis represents the number of interactions (user engagement, clicks, messages, or activity events).

- The blue line represents fluctuations in daily activity, showing spikes and dips, suggesting varying user engagement levels on different days.

- The up-and-down pattern indicates periodic fluctuations, possibly influenced by weekends, holidays, or system events.

Bottom Chart: Weekly User Activity Over Time

- The X-axis represents weeks, likely numbered sequentially.

- The Y-axis represents the number of interactions over a week.

- The green line shows a steady increase in activity over the first four weeks, peaking at Week 4, followed by a decline in Week 5.

- This trend suggests that user engagement grew steadily but then dropped, possibly due to saturation, seasonal effects, or external influences.

Potential Issues & Next Steps:

1. Understanding Peaks and Drops:

   - Investigate what caused the activity spike in Week 4 and the drop in Week 5. Possible reasons include:

      - Marketing campaigns

      - Feature launches

      - External events affecting user behavior

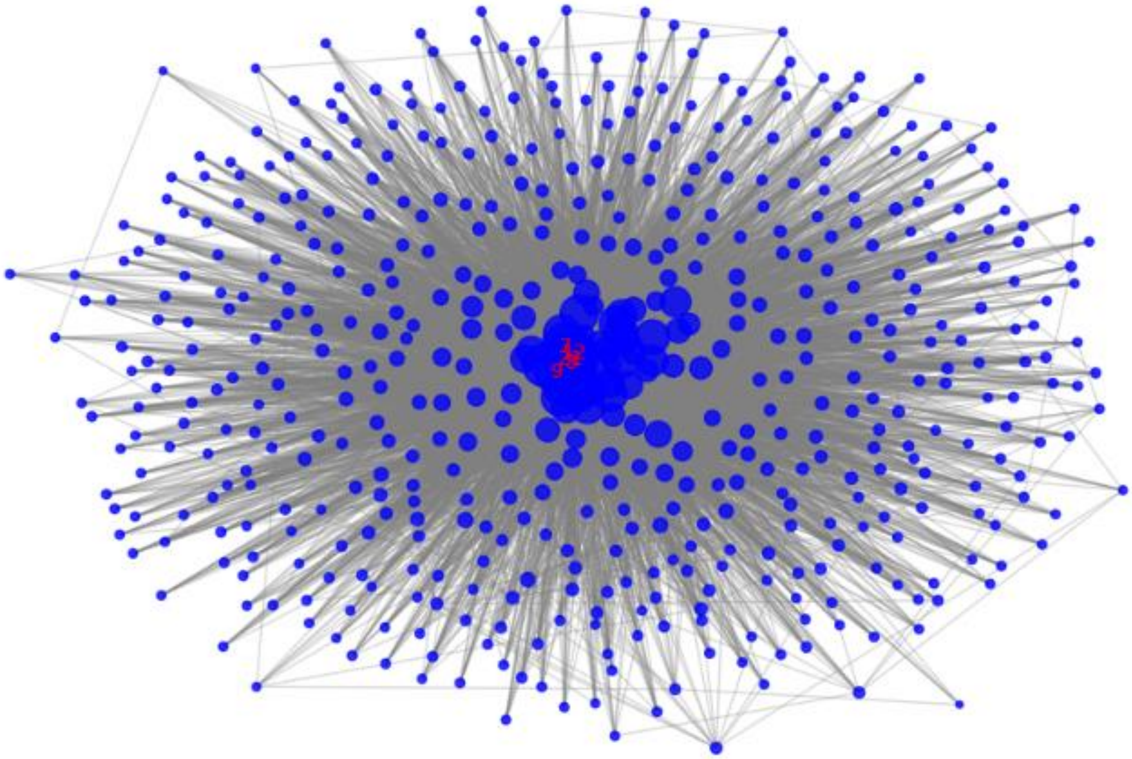
2. Further Analysis:

   - Compute moving averages to smooth fluctuations.

   - Perform anomaly detection to identify unusual peaks or drops.

   - Segment users by region, time zone, or engagement level to find behavioral patterns.

2. Network Graph Highlighting Central Nodes



Interpretation:

This graph represents a highly connected network, likely depicting user interactions, connections, or relationships.

Key Observations:

1. Centralized Core (Dense Cluster in the Middle)

- The center of the graph has many tightly connected nodes, forming a core network.
- These nodes (shown in red) are likely the most influential or highly connected entities.

2. Radial Structure (Hub and Spoke Pattern)

- The network has a hub-and-spoke appearance, with nodes on the periphery having fewer connections but still linked to the central hub.

- This suggests a structure where a core group of users drive most interactions, while others engage at a lower level.

3. High-Degree Nodes (Influential Users or Entities)

- The larger nodes in the center likely represent entities with high degree centrality (many connections).

- These could be super-users, key accounts, or influencers in the network.

4. Peripheral Nodes (Outliers or Less Active Participants)

- The outermost nodes have fewer direct connections, meaning they might be less active or infrequent participants.

Potential Next Steps for Analysis:

Identify the most central/influential users using:

- Degree Centrality: Who has the most connections?
- Betweenness Centrality: Who acts as a bridge between groups?
- Eigenvector Centrality: Who is connected to the most important nodes?

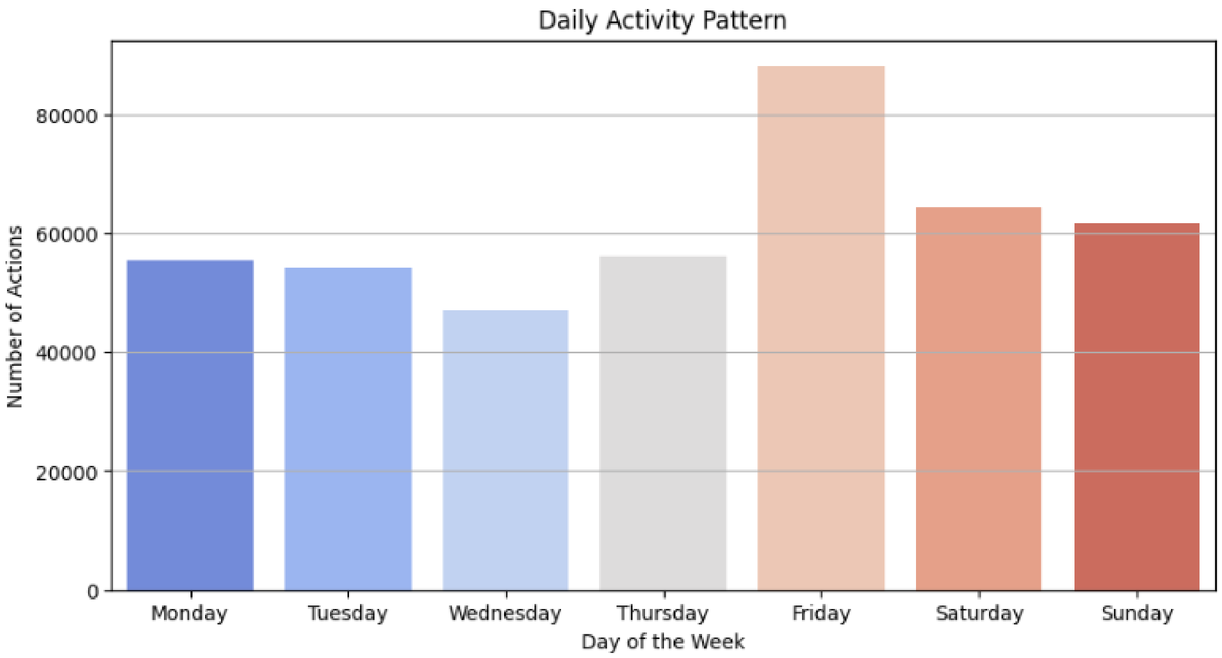
Detect Network Communities

- Use clustering algorithms (Louvain, Girvan-Newman) to identify subgroups.
- This will help understand how users form groups or interact within smaller communities.

Analyze Peripheral vs. Core Behavior

- Investigate what differentiates highly connected nodes from peripheral ones.

3. Daily Activity Pattern



Interpretation:

This bar chart visualizes user activity in the MOOC social network dataset across different days of the week.

Key Observations:

1. Friday has the highest activity. The number of actions peaks on Friday, surpassing 80,000 interactions.

2. There is a midweek dip on Wednesday. Wednesday shows the lowest activity, with fewer than 50,000 interactions.

3. Weekend engagement is high. Saturday and Sunday maintain relatively strong engagement, though lower than Friday.

4. Monday to Thursday maintained a notable level of consistency. Activity levels from Monday to Thursday are fairly stable, with slight variations.

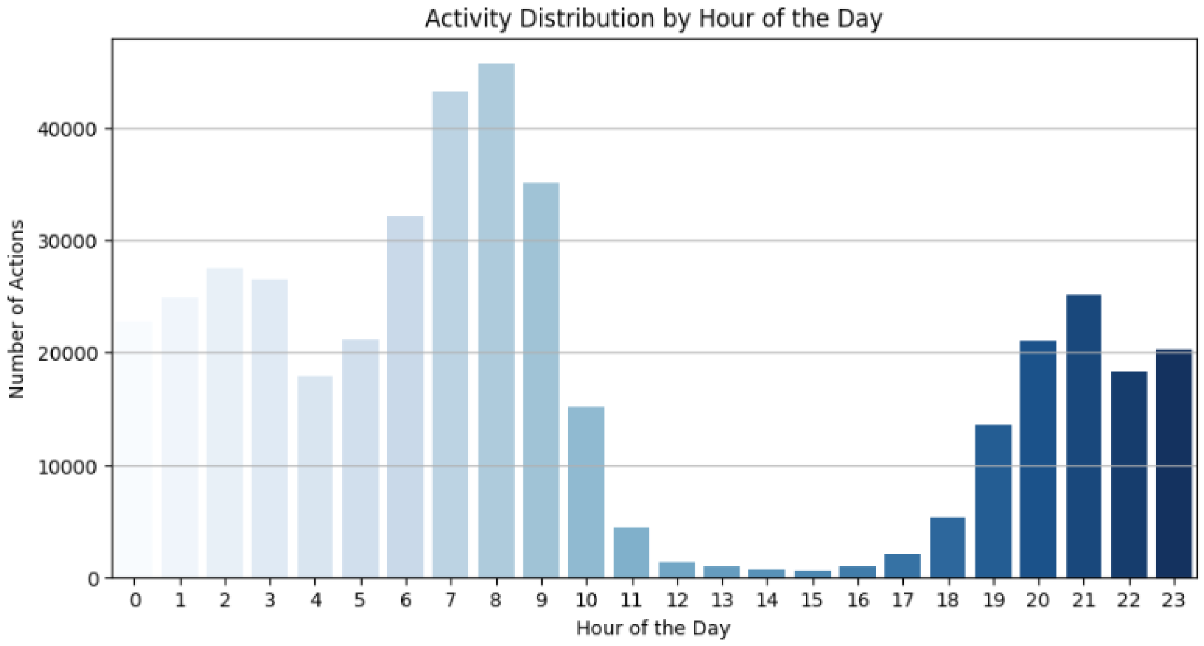
Possible Explanations:

- **Friday peak:** Could indicate weekly deadlines, assignments, or discussion activities peaking before the weekend.
- **Wednesday dip:** Might be a midweek slump where users take a break before ramping up towards the end of the week.
- **High weekend engagement:** Suggests that students or users continue interacting over the weekend, possibly catching up on coursework.

Next Steps for Analysis:

- **Compare activity types per day:** Find out whether certain activities (e.g., forum discussions, submissions) are more common on specific days.
- **Time-of-day analysis:** Find out if engagement varies within each day.
- **User segmentation:** Find out if different user groups (e.g., active vs. passive users) follow the same pattern.

4. Activity Distribution by Hour of the Day



Interpretation:

This bar chart displays the number of user actions in the MOOC social network dataset, distributed across different hours of the day.

Key Observations:

1. Morning peak (7 am – 9 am)

- The highest activity occurs between 7 am and 9 am, with a sharp peak around 8 am (approximately 45,000 actions).

- This suggests that users are highly active in the early morning, likely engaging in learning activities or discussions.

2. Afternoon lull (10 am – 6 pm)

- Activity drops significantly after 10 am, reaching its lowest point between 12 v and 6 pm.

- This may indicate that users are occupied with other tasks, such as work or school.

3. Evening rebound (7 pm – 11 pm)

- A secondary surge in activity begins around 7 pm, peaking between 8 pm and 10 pm.

- This pattern suggests users return to the platform in the evening, possibly for revision or interaction.

4. Late-night engagement (10 pm – 11 pm)

- A moderate level of activity continues late at night, indicating that some users prefer studying or engaging in discussions before bed.

Possible Explanations:

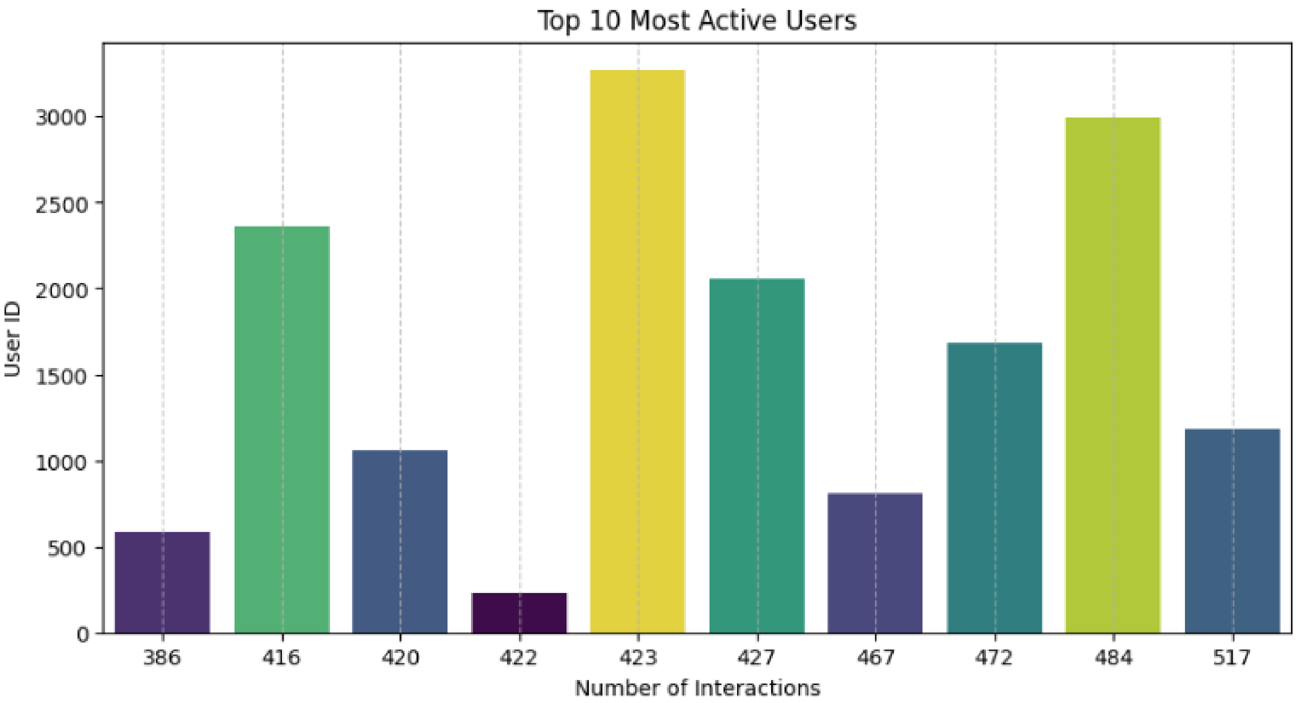
- For the Morning peak (7–9 am): Users may engage with coursework or discussions before starting their daily commitments.
- For the Afternoon drop (10 am–6 pm): Users might be busy with work, school, or other responsibilities.

- For the Evening rise (7–11 pm): A strong return to the platform, likely for review, assignments, or forum participation.
- For the Late-night activity (10–11 pm): Night owls or users in different time zones might contribute to sustained engagement.

Next Steps for Analysis:

- Segment by user type: Find out whether different user groups (e.g., students vs. professionals) follow the same pattern.
- Compare weekdays vs. weekends: Find out whether evening activities are stronger on weekdays due to work/school constraints.
- Analyze activity types: Find out whether users mainly submit assignments, discuss in forums, or watch content during these peak hours.

5. Top 10 Most Active Users



Interpretation:

This bar chart presents the top 10 most active users in the dataset based on the number of interactions they have made.

Key Observations:

1. User 423 is the most active

   - This user has over 3,000 interactions, making them the most engaged participant.

2. User 484 is also highly active

   - Similar to User 423, User 484 has a high number of interactions, standing out among the top 10.

3. Users 416 and 427 also have strong engagement

   - These users have over 2,000 interactions, making them among the most active.

4. Some users have significantly lower interactions

   - Users like 422 and 467 have much fewer interactions compared to the most active users.

   - This suggests a large variation in engagement levels among the top 10 users.

Possible Explanations for High Activity:

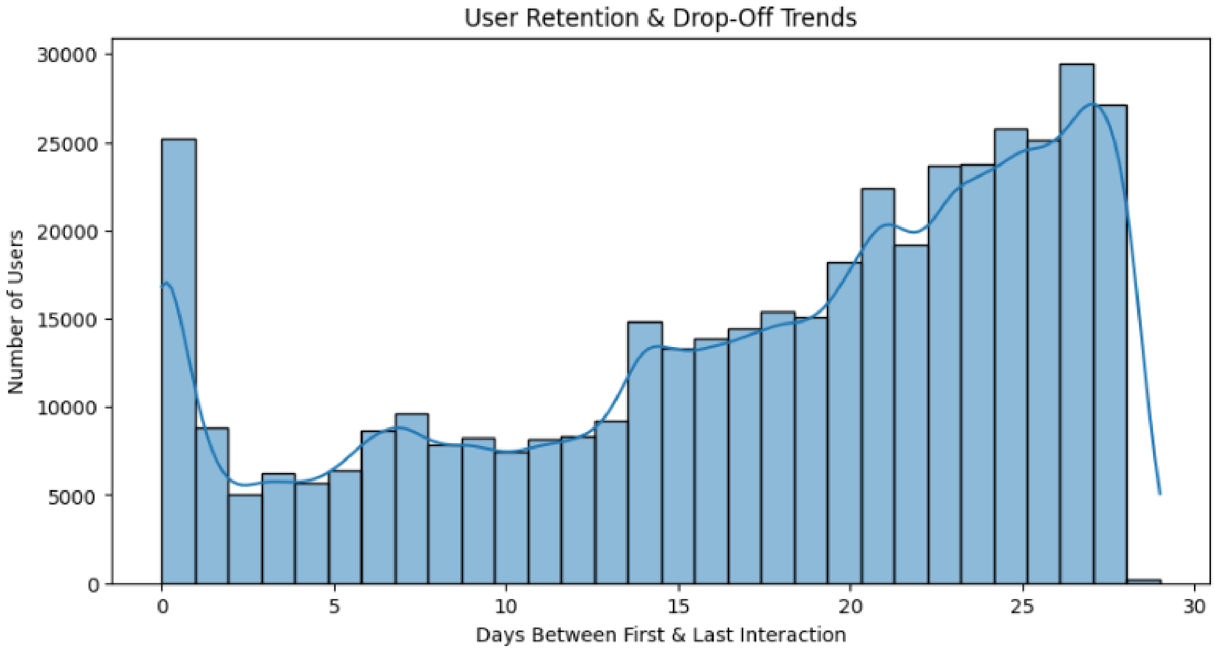
- These users might be frequent contributors, participating in discussions, assignments, or forums.
- They could be mentors, instructors, or power users who engage more than the average participant.

Next Steps for Analysis:

- Behavioral Analysis: Confirm what types of interactions these users are making (posts, comments, submissions).

- Engagement Impact: Are these users influencing overall participation in the network?
- Comparison with General Users: How do their activity levels compare to the average user?

6. User Retention & Drop-Off Trends



Interpretation:

This histogram visualizes the number of users based on the days between their first and last interaction, offering insights into user retention and drop-off trends over time.

Key Observations:

1. A sharp peak at Day 0 (approximately 25,000 users)

- This suggests that many users only interact once and never return, indicating a high one-time user drop-off.

- This could mean that users are not finding value initially or facing barriers to continued engagement.

2. A dip in retention after the first day

- After Day 0, the number of retained users drops sharply but then stabilizes.

3. Gradual increase in user retention over time

- Between Days 5–20, the number of retained users fluctuates but shows an upward trend.

- This suggests that users who stay engaged beyond the first few days are likely to continue participating.

4. Peak retention at Days 25-27 (approximately 30,000 users)

- This implies that a significant number of users stay engaged for almost a full month.

- This could be due to structured programs, challenges, or deadlines that keep them active.

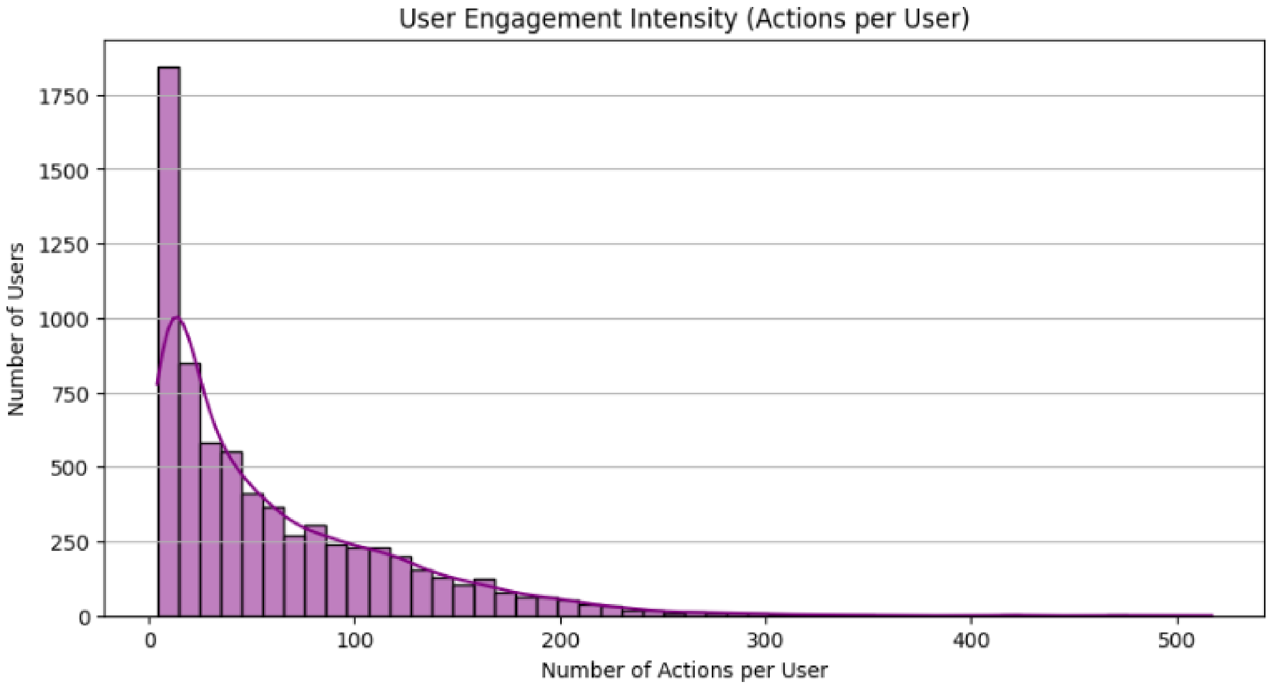
5. Sharp drop after Day 30

- This suggests that most users drop off after a month, possibly due to program completion or loss of interest.

Possible Explanations & Next Steps:

- Onboarding Optimization: Since many users drop off at Day 0, improving the first-time user experience could boost retention.
- Engagement Programs: The rise in retention suggests that users who stay beyond a few days are more likely to stay longer. Programs to encourage participation beyond Day 1 could help.
- Understanding Long-Term Drop-Off: The drop after Day 30 may indicate the end of a cycle (internship, challenge, or event). If so, strategies for continued engagement beyond the program could help retain users.

7. User Engagement Intensity (Actions per User)



Interpretation:

This histogram visualizes the distribution of user actions, indicating how active users are in terms of the number of actions they perform.

Key Observations:

1. Highly Skewed Distribution (Right-Skewed)

- The vast majority of users perform very few actions.

- A sharp peak near zero suggests that many users engage minimally, possibly performing only one or two actions before dropping off.

2. Gradual Decline in User Count as Actions Increase

- As the number of actions increases, fewer users remain engaged.

- There is a steady decline in participation, meaning that only a small fraction of users are highly engaged.

3. Long Tail of Highly Active Users

- Some users have performed hundreds of actions, but they are rare.
- This suggests the presence of a small, highly engaged user base that interacts frequently.

Possible Explanations & Next Steps:

Improve Early Engagement:

- The large number of low-action users suggests potential onboarding or engagement issues.
- Consider implementing strategies to encourage continued interaction (e.g., gamification, notifications, task reminders).

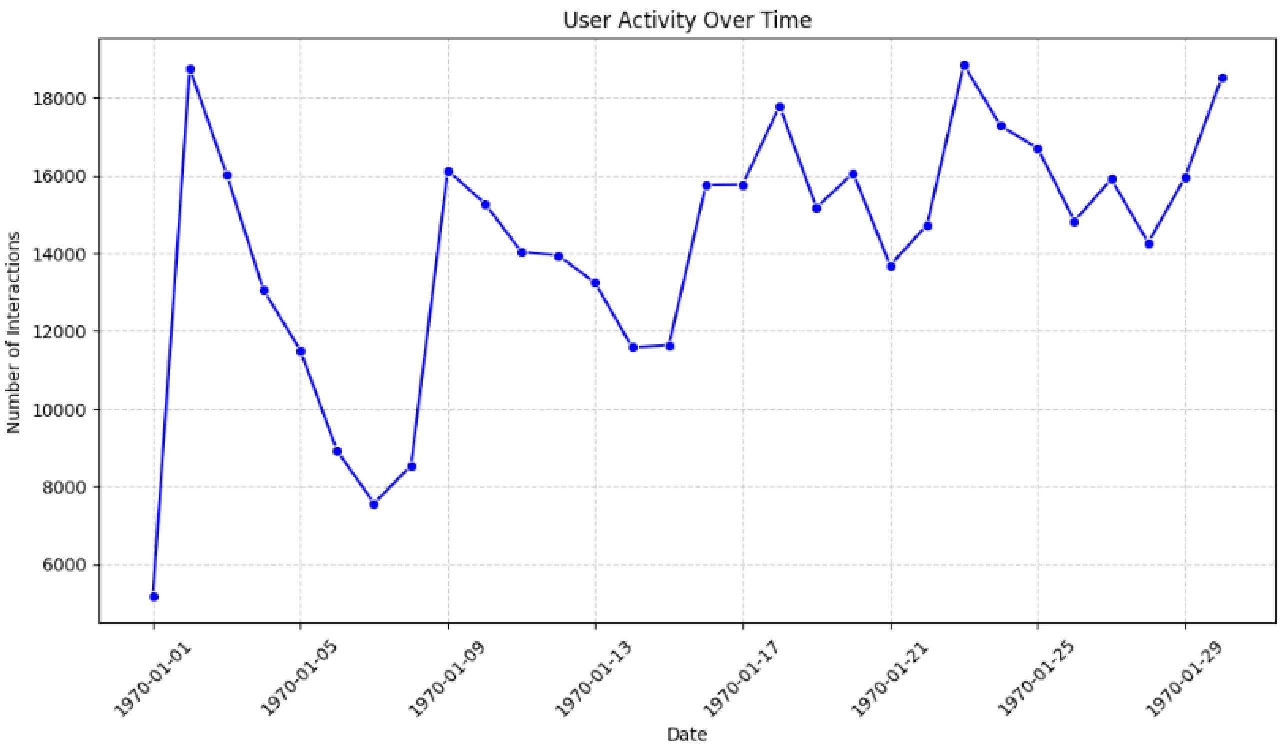
Identify & Reward Power Users:

- Since a small group of users drives most of the engagement, targeting them for exclusive benefits, leadership roles, or feedback loops can help sustain their activity.

Investigate Drop-Off Reasons:

- Understanding why many users interact only a few times before disengaging can help improve retention strategies.
- Surveys, feedback, or A/B testing could reveal potential friction points.

8. User Activity Over Time



Interpretation:

This line chart visualizes user interactions over time, showing trends in engagement levels.

Key Observations:

1. Fluctuating Activity Levels

   - The number of interactions varies significantly over time.

   - Peaks and drops suggest periods of high and low user engagement.

2. Sharp Initial Spike & Drop

   - There is a huge surge in interactions at the beginning, followed by a steep decline.

- This could indicate a launch event, promotional activity, or a new feature rollout that initially attracted users before interest dropped.

3. Regular Peaks & Dips

- Activity shows periodic increases and decreases, which might correspond to:
  - Scheduled events (e.g., weekly challenges, deadlines).
  - User behavior patterns (e.g., workweek vs. weekends).

4. Recent Increase in Engagement

- Towards the end of the time period, interactions rise again.
- This suggests renewed interest, possibly due to new features, promotions, or community engagement efforts.

Next Steps & Recommendations:

Identify Key Drivers

- Analyze what caused the major peaks (e.g., feature releases, promotions, external factors).
- Investigate why engagement dropped after the peak and how to maintain momentum.

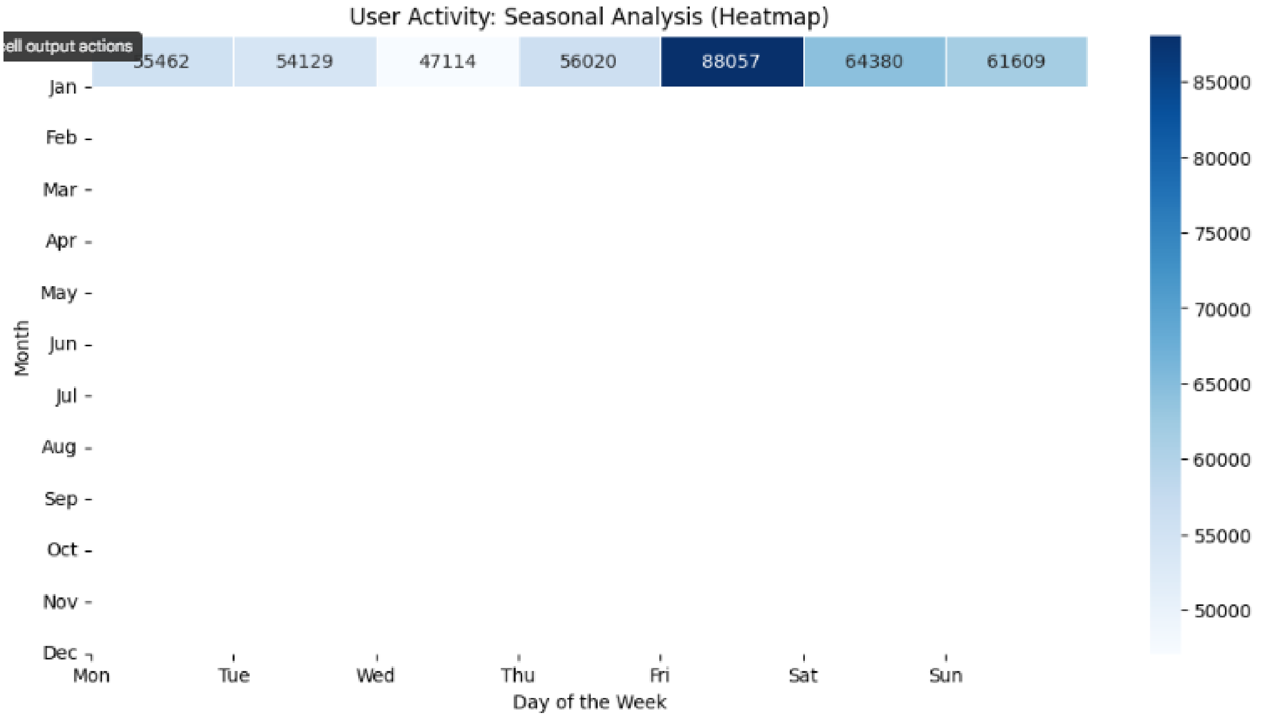
Understand Cyclic Trends

- If engagement fluctuates weekly, aligning content, updates, or engagement activities with user behavior patterns could help.

Improve Retention Strategies

- Since engagement has a fluctuating pattern, user retention campaigns (e.g., notifications, email reminders, incentives) could help sustain activity.

9. User Activity: Seasonal Analysis (Heatmap)



Interpretation:

This heatmap visualizes user activity across different days of the week and months, but it seems that only data for January is available. The color intensity indicates the level of user activity, with darker shades representing higher activity levels.

Key Observations:

1. Friday Sees the Highest Activity

- Friday (88,057 interactions) has the highest user activity, as seen from the darkest blue shade.

- This suggests that users are most engaged on Fridays—possibly due to end-of-week trends or specific platform-related events.

2. Wednesday Has the Lowest Engagement

- Wednesday (47,114 interactions) has the lowest activity among all recorded days.

- This might indicate a midweek slump in engagement.

3. Weekends See Moderate Activity

- Saturday (64,380 interactions) and Sunday (61,609 interactions) show a moderate level of engagement.

- Users might be more active on weekends than midweek but less engaged than on Fridays.

4. Missing Data for Other Months

- The chart seems to have only one row (January), so it's unclear if this pattern holds throughout the year.

- If the missing months are due to a data issue, it would be good to check whether the dataset needs updating.

Insights & Recommendations:

Leverage Peak Activity on Fridays:

- Schedule important content releases, feature launches, or engagement campaigns on Fridays to maximize impact.

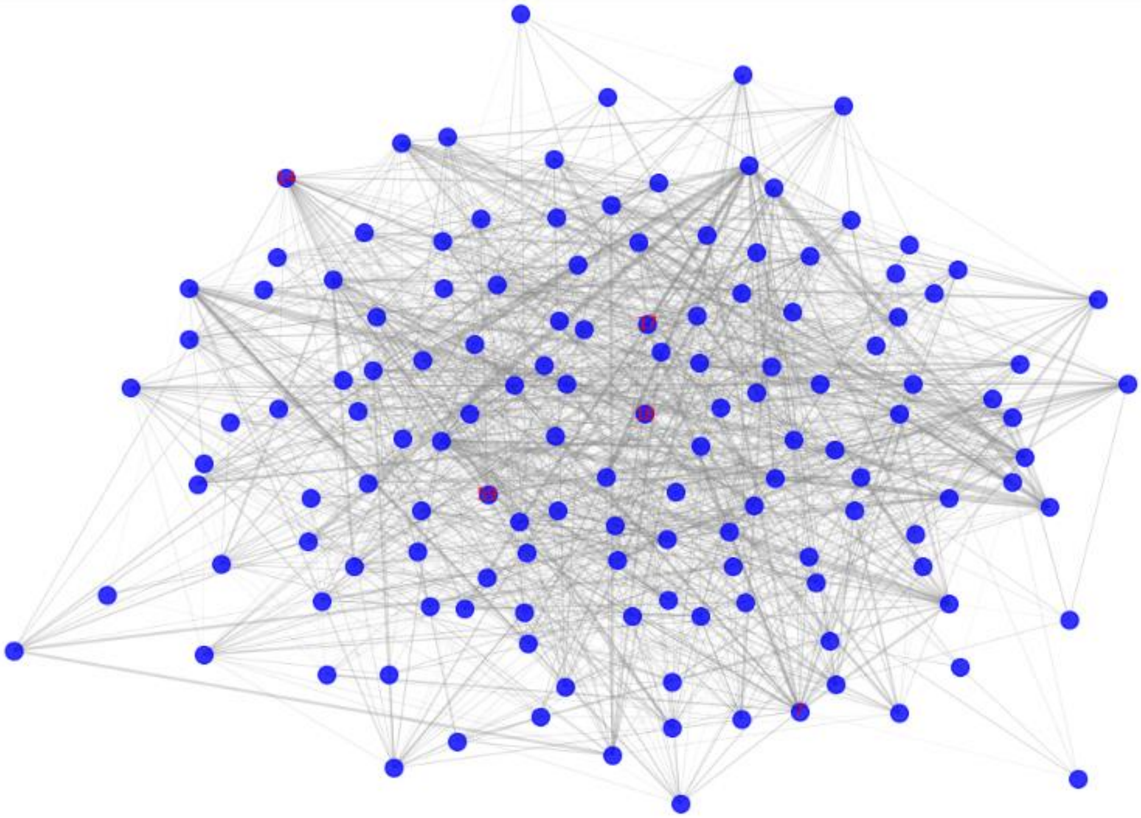
Boost Midweek Engagement:

- Introduce engagement strategies for Wednesdays, such as promotions, notifications, or interactive features, to counteract the dip.

Explore Monthly Trends (if available):

- If data for other months can be retrieved, it would be useful to check for seasonal variations in engagement.

10. Optimized Weighted User Interaction Network



Interpretation:

This visualization represents a network graph, where:

- Nodes (blue circles) represent users, actions, or entities in the network.

- Edges (gray lines) represent relationships, interactions, or connections between these entities.

- Highlighted nodes (red or purple ones) likely represent key players, such as high-degree (highly connected) nodes or influential users.

Key Observations:

1. Dense Connectivity

- The network appears highly connected, meaning most users have multiple interactions or relationships.

- The dense central region suggests a core group of users with many interconnections.

2. Potential Hubs (Red Nodes)

- A few nodes are highlighted in red/purple, indicating high centrality (important or influential nodes).

- These could be super-connectors (users who engage with many others) or key facilitators in the network.

3. Peripheral Nodes

- Some nodes are located towards the edges with fewer connections, possibly new or less active users.

Insights & Next Steps:

Identify Key Players:

- Nodes with the highest connections or centrality can represent power users, influencers, or major contributors.

- Running centrality analysis (degree, betweenness, closeness centrality) could quantify this influence.

Cluster Analysis:

- Check for sub-communities or clusters within the network.

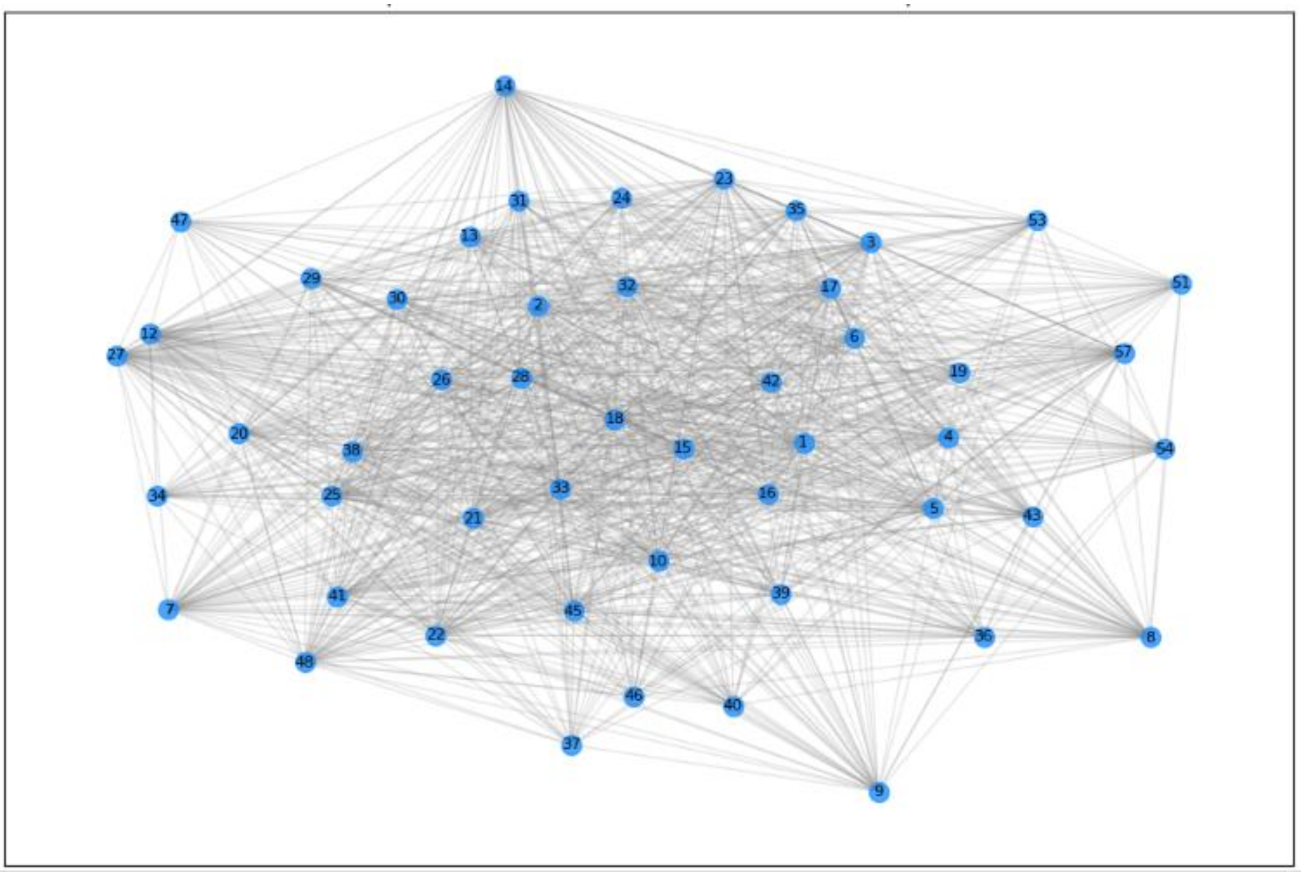
- This can reveal natural user groupings, such as mentor-mentee relationships or topic-based discussions.

Engagement Strategy:

- Users in the core drive interactions, so incentivizing them could boost overall engagement.

- For peripheral users, targeted engagement strategies (e.g., recommendations, outreach) could help integrate them better into the network.

11. Top 50 Most Connected Users - Less Dense Graph



Interpretation:

This is another network graph, where:

- Nodes (blue circles with numbers) represent entities (likely users, actions, or accounts).
- Edges (gray lines) indicate relationships or interactions between these entities.

Key Observations:

1. Highly Connected Network
   - The dense structure suggests strong interconnectivity between nodes.
   - Many nodes have multiple direct connections, forming a tight-knit network.

2. Central & Peripheral Nodes

- Certain nodes (e.g., 1, 2, 15, 26) appear in the core, meaning they are likely highly influential or central in the network.

- Nodes along the edges of the graph have fewer connections, meaning they could represent less active participants or isolated entities.

3. Possible Communities or Subgroups

- Though the graph is highly interconnected, there could be subgroups or clusters within it.

- Community detection algorithms (like Louvain method or Girvan-Newman clustering) could help identify these groups.

Potential Next Steps for Analysis:

Centrality Analysis

- Compute degree centrality to find the most connected users.

- Use betweenness centrality to identify key bridge nodes that connect different groups.

Community Detection

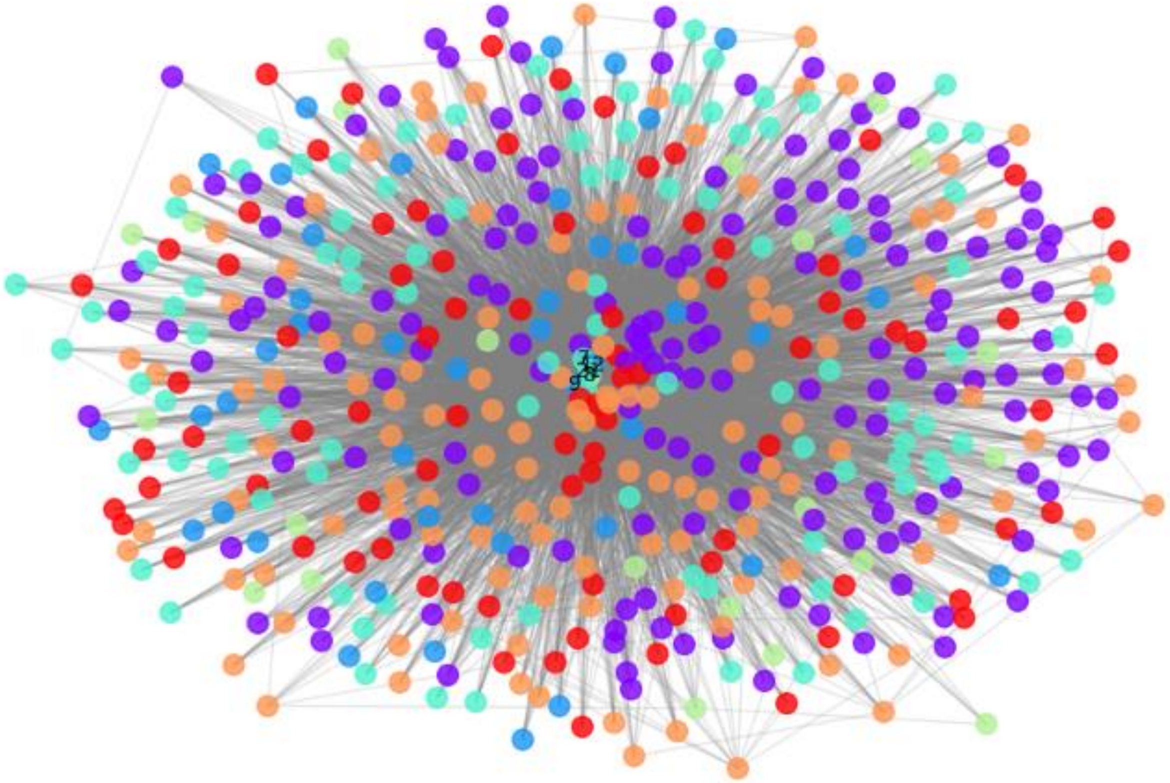
- Run clustering algorithms to segment the network and find distinct user groups.

Network Density & Influence Mapping

- Check how dense the network is—how many actual connections exist compared to possible ones.

- Identify influential users or nodes driving most interactions.

12. Community Detection in User Interaction Network



Interpretation:

This network graph is a community-detected social or interaction network, where nodes (circles) represent entities (users, accounts, or elements), and edges (lines) represent relationships or interactions.

Key Observations:

1. Community Clustering (Different Colors Indicate Groups)

- Nodes are colored differently, likely representing distinct communities or clusters.

- The presence of multiple colors suggests modular structures within the network, meaning sub-groups exist with stronger internal interactions.

- Common community detection algorithms include Louvain, Girvan-Newman, and Label Propagation.

2. Highly Connected Core (Dense Center vs. Sparse Edges)

- The central nodes are densely connected, indicating influential or high-degree nodes (users or entities with many interactions).

- These nodes likely act as bridges or hubs connecting different communities.

3. Peripheral Nodes (Less Engaged Participants)

- The outermost nodes are loosely connected, meaning they interact with only a few others or are less active participants.

- This could represent casual users, new participants, or isolated entities in the network.

4. Network Structure (Hub-and-Spoke with Dense Center)

- The hub-and-spoke pattern suggests a centralized network, where a core group dominates interactions while peripheral nodes have limited influence.

- If this were a social network, the core nodes could be power users, influencers, or key accounts.

Potential Next Steps for Analysis:

Identify Key Influencers:

- Compute degree centrality (who has the most connections?).
- Compute betweenness centrality (who acts as a bridge between communities?).

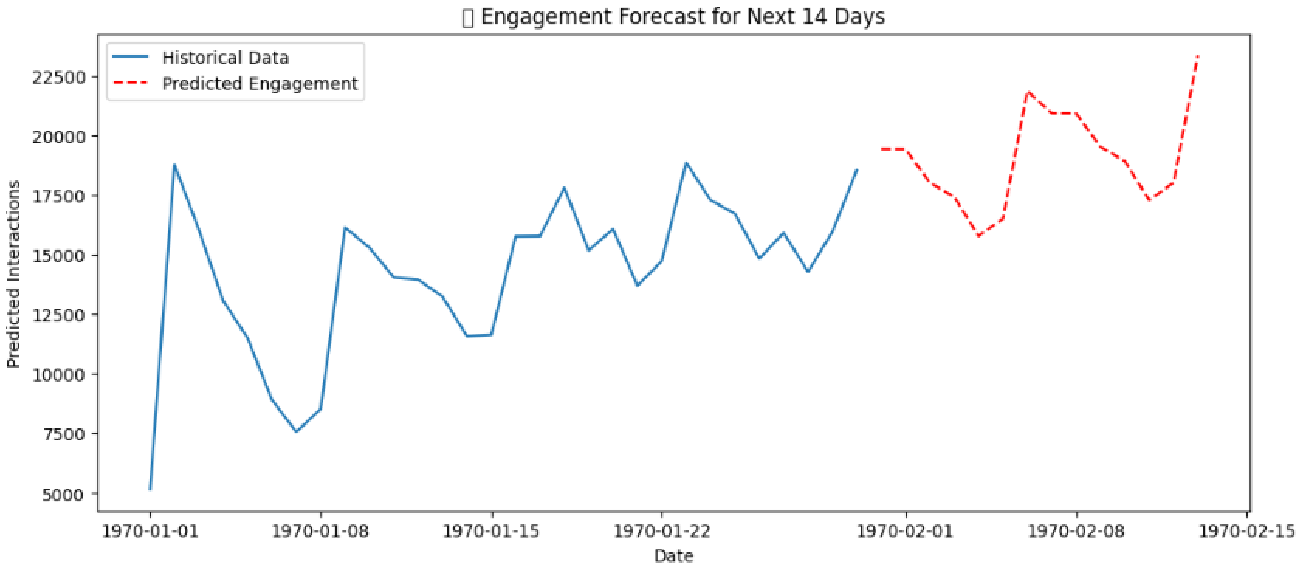
Community Detection Analysis:

- Extract sub-networks to analyze individual clusters.
- Understand how different groups interact or remain isolated.

Network Density & Structural Metrics:

- Compute clustering coefficient to measure how interconnected the nodes are.
- Analyze graph density to understand overall connectivity.

13. Engagement Forecast for Next 14 Days



Interpretation:

This line chart shows the engagement forecast for the next 14 days, comparing historical data with predicted engagement levels over time.

Key Elements:

1. X-axis (Date):

- Represents time in daily intervals.

- The dates appear incorrectly formatted (showing "1970"), likely due to a timestamp issue where dates were not properly processed.

2. Y-axis (Predicted Interactions):

- Represents the number of predicted interactions (e.g., user engagement, clicks, messages, or activities).

3. Two Line Trends:

- Blue Line (Historical Data):

  - Shows past engagement trends, fluctuating over time with noticeable peaks and dips.

- Red Dashed Line (Predicted Engagement):

- Represents the forecasted user engagement for the next 14 days, showing an increasing trend with fluctuations.

Insights & Observations:

Past Trends (Historical Data):

- Shows spikes and drops in engagement—possibly due to seasonal trends, promotions, or platform activity cycles.

- Similar behavior was observed in previous daily activity charts.

Future Trends (Predicted Engagement):

- The forecast suggests increased engagement in the next two weeks, with some ups and downs but an overall upward trajectory.

- Peaks may indicate expected high-traffic days.

- The model likely uses time series forecasting techniques (e.g., ARIMA, Prophet, or LSTMs) to predict engagement.

Potential Issues & Next Steps:

Fix the Date Formatting Issue:

- The X-axis still shows 1970, meaning timestamps might not have been correctly converted.

- Ensure proper datetime conversion in Python (e.g., `pd.to_datetime()`).

Validate Forecast Accuracy:

- Compare predicted values with actual engagement data (once available) to assess model accuracy.

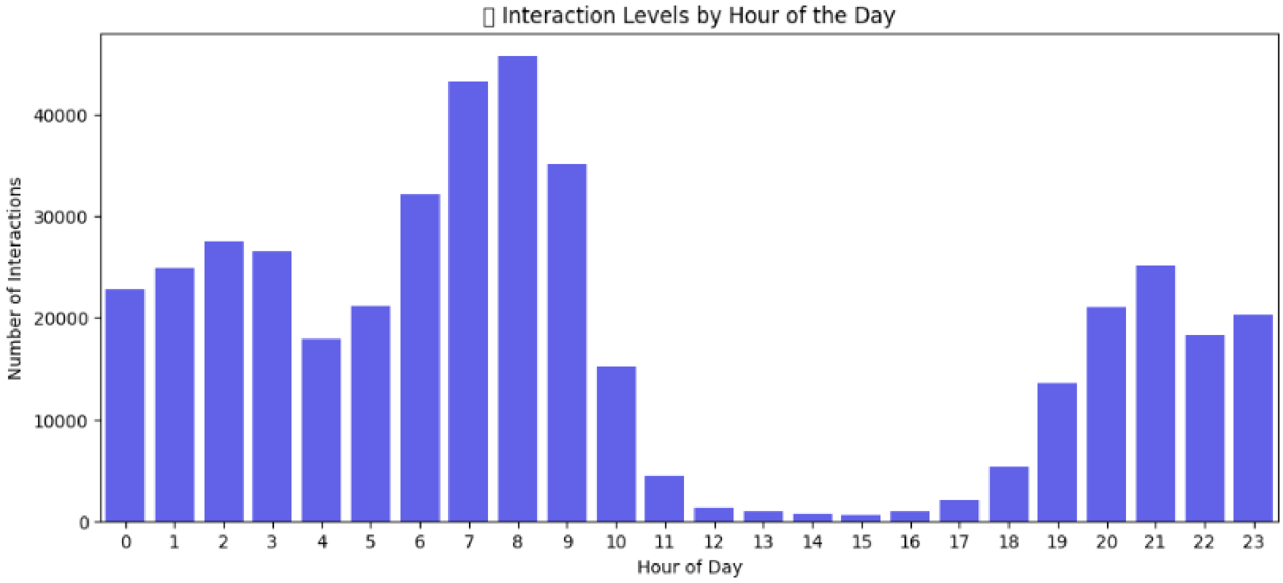
- Adjust forecasting parameters if needed.

Investigate the Factors Behind the Trend:

- Identify what drives engagement—weekends, promotions, or specific user behaviors.

- Use seasonal decomposition analysis to check if engagement follows a weekly or monthly pattern.

14. Interaction Levels by Hour of the Day



Interpretation:

This bar chart visualizes the number of interactions (e.g., user activity, clicks, messages, etc.) at different hours of the day.

Key Observations:

1. X-axis (Hour of the Day):

   - Represents time in hourly increments from 0 (midnight) to 23 (11 pm).

2. Y-axis (Number of Interactions):

   - Represents user engagement levels (e.g., interactions, logins, or messages).

3. Peak Activity Periods:

   - The highest engagement occurs around 7 am - 9 am, with the peak at 8 am (approximately 45,000 interactions).

   - Another smaller evening peak happens between 8 pm - 10 pm.

4. Lowest Activity Periods:

- Very low interaction between 12 pm - 5 pm (early afternoon), possibly indicating user inactivity or a break period.

- The lowest point is around 2-5 pm, with minimal engagement.

Insights & Possible Explanations:

Morning Peak (7-9 am):

- Users might be starting their day, checking messages, or engaging before work/school.

- Could be morning rush hour engagement (e.g., news consumption, social media, work-related tasks).

Afternoon Drop (12-5 pm):

- Users may be busy with work, school, or daily responsibilities, reducing engagement.

- Some platforms see lower activity in midday hours, especially for social or entertainment-based apps.

Evening Rise (8-10 pm):

- Users return to engagement after work/school, possibly for leisure activities, social media, or entertainment.

Potential Next Steps & Recommendations:

Target High Engagement Hours:

- If this data is for a business or platform, it should focus content releases, promotions, or major updates around the peak hours (7-9 am & 8-10 pm).

Understand the Afternoon Drop:

- Investigate why engagement is low in the afternoon—possible reasons: work distractions, lack of notifications, or competition from other platforms.

Compare Different Days of the Week:

- Check if the pattern is consistent across weekdays vs. weekends.

Conclusion

This project leverages network analysis techniques to extract valuable insights from the MOOC User Action Dataset. By systematically constructing and analyzing the network, we uncover key interaction patterns, central users, and evolving engagement trends within the platform. The findings provide actionable insights that can help improve online learning environments by enhancing engagement strategies, fostering collaboration, and identifying influential users.

Through a structured approach, we document the rationale behind each step, ensuring transparency and reproducibility of the analysis. This study demonstrates the power of network analysis in understanding complex user interactions in digital learning ecosystems.