Network Analysis of User Participation and Upvote Distribution in r/datascience

Faizan Waheed
fawaheed@iu.edu
March 16, 2025

Abstract

This study examines user engagement and recognition dynamics in the r/datascience subreddit by constructing two distinct networks: (1) the Participation Network, which captures user activity through posts and comments across five key flairs (Discussion, Projects, ML, AI, and Coding), and (2) the Upvote Network, which maps the distribution of upvotes received by users. The analysis reveals a clear distinction between engagement and recognition—while users actively participate across multiple flairs, upvotes are largely concentrated within a single dominant flair. Findings indicate that Projects and AI serve as discussion hubs in the Participation Network, fostering cross-topic interactions, whereas ML and Coding emerge as the most influential flairs in the Upvote Network, reflecting topic-driven recognition. Community detection results confirm that upvote-based communities are more distinct than participation-based ones, highlighting that engagement does not always equate to influence. Moreover, the higher assortativity and modularity in the upvote network reinforce the idea that recognition is flair-specific and driven by specialization rather than broad engagement. These insights suggest that while participation in online communities is interdisciplinary, recognition follows a specialization-driven reputation model, where technical expertise in specific domains yields greater visibility and influence

1. Introduction and Background

Reddit communities serve as vibrant spaces where users actively engage in discussions by posting and commenting. Understanding the dynamics of user participation and recognition within these communities offers valuable insights into how influence and interaction evolve. This study focuses on two primary aspects of engagement in the r/datascience subreddit:

- **Participation Network:** Analyzes how users contribute by posting and commenting across five key flairs (*Discussion, Projects, ML, AI, Coding*).
- **Upvote Network:** Examines how users gain recognition through upvotes received on their posts and comments.

By leveraging **graph-based network analysis**, this study explores whether active contributors also receive the most upvotes and how engagement translates into recognition. The structural differences between these networks provide insight into user behavior, identifying whether participation and influence align or diverge.

2. Data Collection

2.1 Data Source and Extraction Process

The dataset for this study was collected exclusively from the **r/datascience subreddit** using the **PRAW (Python Reddit API Wrapper)**. Two separate datasets were compiled:

1. **Comment Participation Dataset** – Captures interactions where users make posts or comment on posts categorized under different flairs.

2. **Upvote Distribution Dataset** – Records upvotes received by users for their posts and comments within specific flairs.

Data collection was restricted to **January 2025** to maintain consistency in analysis. The **Reddit API** was used to extract relevant discussions, focusing on posts and comments within the top five flairs. To ensure a representative sample, a minimum of **50 users per flair** was considered.

2.2 Network Construction

Each dataset was transformed into a graph structure suitable for network analysis. The networks consist of **two types of nodes** (*users and flairs*) and **edges** representing user interactions with flair categories.

Participation Network:

- **Nodes:** Users and the five flairs (*Discussion, Projects, ML, AI, Coding*).
- **Edges:** A user is connected to a flair if they posted or commented under that flair.
- **Node Attributes:** Number of **posts and comments** made by each user.

Upvote Network:

- **Nodes:** Users and the five flairs.
- **Edges:** A user is connected to a flair if they received upvotes from posts/comments under that flair.
- **Node Attributes:** Number of **upvotes received on posts and comments**.

2.3 Data Processing and Cleaning for Network Creation

Before visualization in **Gephi**, basic data cleaning and formatting steps were performed:

1. **Duplicate Removal:** Any redundant interactions (multiple comments/upvotes in the same flair) were aggregated.

2. **Standardizing Flair Labels:** Ensuring consistency in flair categorization for network mapping.

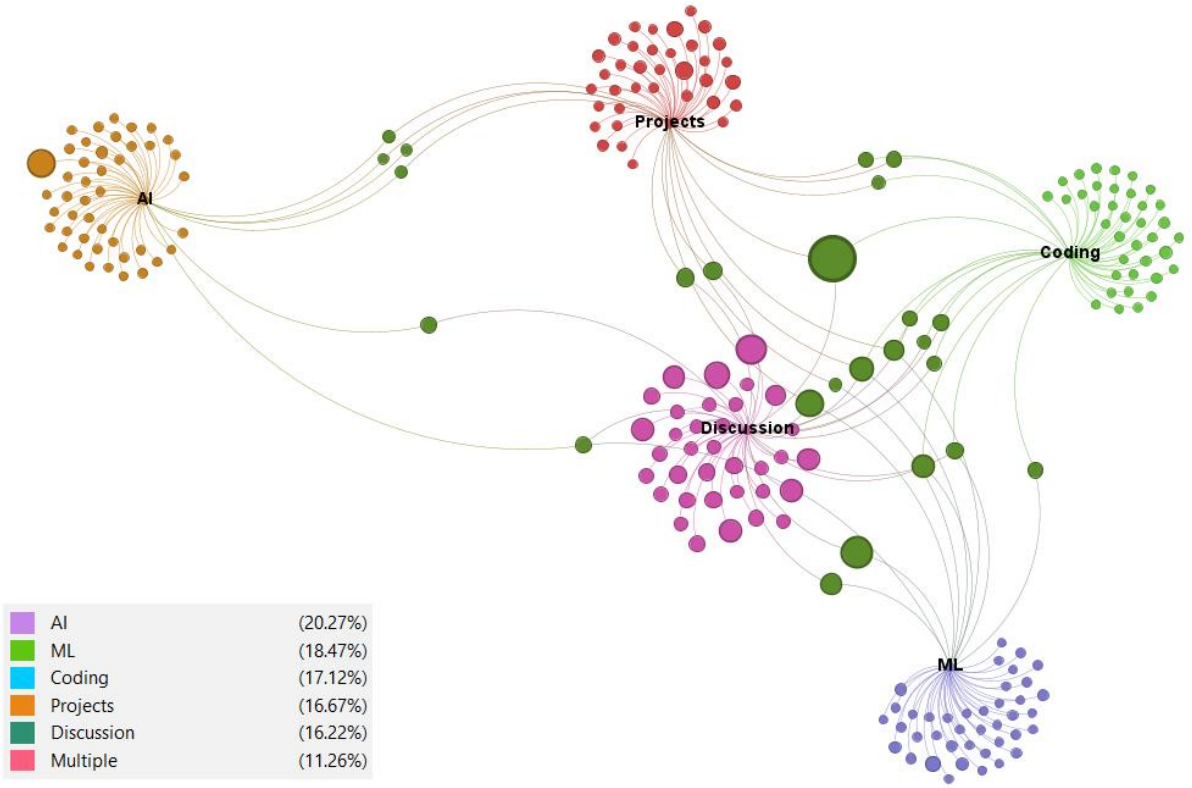
The cleaned datasets were then **exported as CSV files** containing **node attributes (user participation and upvotes) and edge relationships**. These structured datasets were used for **network visualization and analysis in Gephi**, allowing for deeper exploration of community structures and centrality measures.

3. Analysis

3.1 Individual Network Analysis

Comment Network (Participation Network)

The comment network consists of 222 nodes, including 217 users and 5 flair categories, with 250 edges representing user interactions through comments. The network has a very low density of 0.0102, indicating that only about 1% of all possible connections exist, confirming that user participation is relatively sparse and topic-focused. The average degree of 2.252 suggests that most users engage with approximately two different flairs through their commenting activity.



Community detection using **Louvain Modularity** resulted in a modularity score of **0.668**, showing that the commenting behavior strongly aligns with distinct flair-based topic groups. The **Label Propagation Algorithm (LPA)** also identified five well-defined communities, reinforcing that user engagement is primarily contained within the respective flairs. Additionally, the **Normalized Mutual Information (NMI) score of 0.845** indicates a strong alignment between detected communities and subreddit-defined flairs.

In terms of **node centrality**, the **Projects and AI** flairs have the highest **betweenness centrality**, suggesting they serve as key discussion hubs where users from multiple topics interact. The high **assortativity coefficient (0.720)** further confirms that users tend to engage within their assigned flair categories, reflecting a structured participation model. However, the

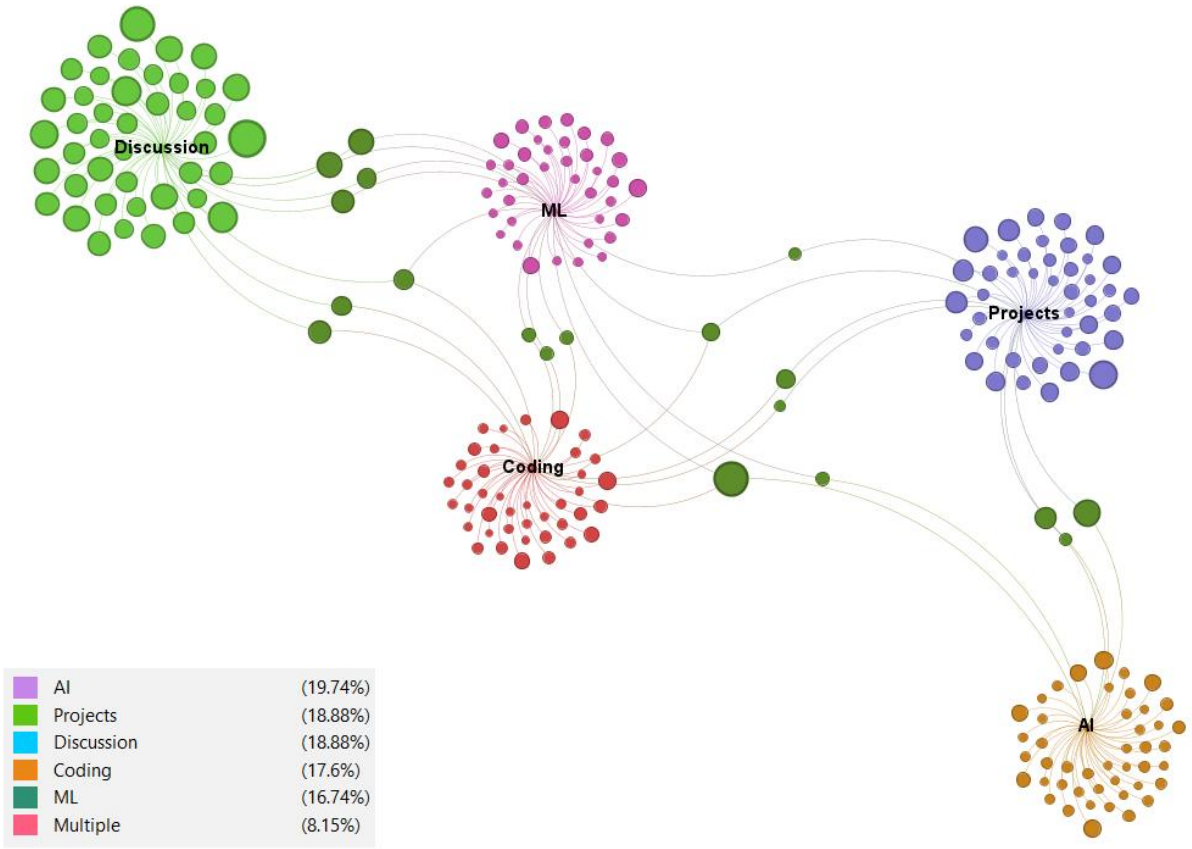
presence of **overlapping communities** suggests that some users are active across multiple flairs, bridging different discussions. The **average conductance of 0.132** indicates that while communities are distinguishable, there is some degree of interaction across flairs, making the discussions more interconnected.

Key Insights from the Comment Network

1. **Users Engage Across Multiple Topics** – The average degree of 2.252 suggests that users frequently participate in at least two flairs rather than limiting themselves to a single discussion category.
2. **Projects and AI Facilitate Cross-Topic Engagement** – Their high betweenness centrality implies that discussions in these areas often overlap with multiple domains, potentially due to interdisciplinary nature (e.g., AI applications in data science projects).
3. **Strong Flair-Based Segmentation with Some Overlap** – While the modularity and NMI scores confirm well-defined flair communities, the presence of bridging users highlights a level of interdisciplinary interaction.

<u>Upvote Network (Reception Network)</u>

The upvote network contains **233 nodes**, slightly more than the comment network due to some users receiving upvotes without actively commenting. The network maintains **250 edges**, mirroring the structure of the participation network but with **a slightly lower density (0.0092)**,



indicating a more selective engagement pattern where upvotes are concentrated among fewer

users. The **average degree of 2.146** suggests that users typically receive upvotes from fewer flairs compared to their commenting behavior.

Community detection using **Louvain Modularity** produced a **higher modularity score (0.712)** than the comment network, suggesting that upvotes tend to be more localized within specific flairs, leading to stronger community separation. The **NMI score of 0.880** further confirms that upvote-based communities align even more closely with the flair system than comment-based interactions.

A notable difference in **centrality measures** is that the **ML and Coding** flairs exhibit the highest **betweenness centrality**, implying that these technical discussion categories play a crucial role in knowledge dissemination and recognition within the subreddit. The **assortativity score of 0.801** suggests an even stronger within-flair interaction compared to the comment network, meaning upvotes are more likely to remain within the same flair. Additionally, the **average conductance of 0.088** indicates that upvote communities are even more distinct, reinforcing the idea that users tend to receive upvotes from a single dominant flair rather than multiple topics.

Key Insights from the Upvote Network

1. **More Distinct Flair-Based Communities** – The modularity of 0.712 confirms that upvote communities are even more separated compared to the comment network, meaning users receive recognition from a more concentrated audience.
2. **Technical Topics Drive Influence** – ML and Coding flairs dominate in betweenness centrality, showing that these fields attract the highest engagement in terms of upvotes.
3. **Upvotes Are More Flair-Concentrated** – The NMI score of 0.880 suggests that most users receive upvotes from a single flair rather than across multiple discussions, unlike comments where users spread their participation.

3.2 Inter-Network Comparison: Participation vs. Reception

A direct comparison between participation and upvote networks reveals **structural and behavioral differences** between engagement and recognition. The upvote network demonstrates a **higher modularity (0.712 vs. 0.668)**, meaning that users are more likely to receive upvotes from a **single flair** rather than commenting across multiple topics. Additionally, the **NMI score is higher in the upvote network (0.880 vs. 0.845)**, reinforcing the idea that **recognition is more specialized than participation**.

In terms of **key nodes**, the comment network highlights **Projects and AI** as critical **discussion hubs**, while the upvote network shows that **ML and Coding** receive the highest influence, suggesting that **technical content attracts more recognition**. The **assortativity score is also higher in the upvote network (0.801 vs. 0.720)**, indicating that **upvotes tend to stay within their respective flair communities more strictly than comments**.

The **average degree is slightly lower in the upvote network (2.146 vs. 2.252)**, meaning that users tend to receive **upvotes from fewer flairs than they comment on**, further supporting the observation that **participation is more distributed while recognition is more concentrated**. The **average conductance score of 0.088 in the upvote network vs. 0.132 in the comment network** suggests that **upvote-based communities are more strongly defined** and have **less interaction between different flair categories**.

Comparison of Participation and Upvote Networks

Metric	Participation Network	Upvote Network	Insight
Modularity	0.668	0.712	Upvotes form more distinct communities than comments, meaning users tend to receive upvotes from fewer flairs.
NMI (Flair Alignment)	0.845	0.880	Upvotes are more strictly confined to the flair where they were generated, while comments are slightly more dispersed.
Assortativity	0.720	0.801	Users commenting tend to engage in multiple flairs, while upvotes stay more within specific communities.
Average Degree	2.252	2.146	Users comment in slightly more flairs than they receive upvotes from.
Average Conductance	0.132	0.088	Upvote communities are more isolated, while comment-based interactions show more inter-flair engagement.
Key Central Nodes	AI, Projects	ML, Coding	AI & Projects facilitate discussions, while ML & Coding dominate recognition through upvotes.

4. Key Findings

1. **Participation is broad, while recognition is specialized** – Users engage in multiple flairs, but upvotes are concentrated within a single flair, indicating a stronger segmentation in recognition patterns.
2. **AI & Projects drive engagement; ML & Coding drive recognition** – AI and Projects act as discussion bridges, fostering interdisciplinary conversations, while ML and Coding attract the most upvotes, suggesting the community values technical contributions.
3. **Upvotes are more flair-specific than comments** – The higher modularity (0.712 vs. 0.668) and assortativity (0.801 vs. 0.720) in the upvote network show that recognition remains confined within specific flair communities more than participation.
4. **Users comment in multiple topics but receive upvotes from fewer** – The lower average degree in the upvote network (2.146 vs. 2.252) suggests that while users discuss various topics, recognition is tied to niche expertise.
5. **Flair-based segmentation is stronger in upvotes** – With an NMI of 0.880 in the upvote network vs. 0.845 in the participation network, upvotes align more closely with flair categories, reinforcing content specialization as a key factor in influence.

5. Recommendations

Content Strategy

- **Encourage More Cross-Flair Discussions** – Since AI and Projects act as discussion bridges, integrating topics such as "How AI is applied in ML projects" could increase interdisciplinary engagement.
- **Capitalize on ML and Coding for Upvote-Driven Content** – Given their dominance in upvote centrality, more technical guides, tools, and frameworks should be encouraged.

Community Management

- **Monitor ML as a Critical Bridge** – As ML is a key hub in the upvote network, ensuring balanced discussions and diverse content can prevent content monopolization.
- **Leverage Projects for Cross-Community Engagement** – Since Projects act as a bridge, cross-posting content that merges multiple flair topics can enhance interaction.

6. Conclusion

This study provides an in-depth network analysis of participation vs. recognition in the r/datascience subreddit. While users frequently comment across multiple flairs, upvotes tend to be concentrated within specific topics, indicating that recognition is more specialized than participation. The Participation Network reflects broad engagement, whereas the Upvote Network highlights a more polarized structure where users receive recognition in fewer flairs. These findings confirm that technical discussions (ML, Coding) dominate recognition, while Projects and AI facilitate broader engagement.

Future Research Directions

- **Analyzing Temporal Trends** – Investigate how participation and upvote trends evolve over time.
- **Predictive Modeling for Upvote Success** – Identify key factors that determine which posts receive the most upvotes.
- **Examining Cross-Posting Effects** – Evaluate whether posting across multiple flairs increases engagement and recognition.

These insights can be valuable for **moderators, content creators, and researchers** to optimize engagement strategies and foster a balanced discussion ecosystem within online communities.