Understanding Subreddit Networks: Analysis and Community Detection

RV Nayan, Mohd Adil, Shubham Daule, Aayush Kataria

May 20, 2025

1 Introduction

Reddit serves as a massive online forum organized into topic-specific communities called subreddits. Our project analyzes the network structure between subreddits by examining user participation patterns. The objectives are to create a list of subreddits and their users, then project the bipartite network into a subreddit interaction graph, Apply network analysis metrics such as centrality and clustering and detect subreddit community clusters and identify bridging subreddits.

2 Network Preperation

2.1 Data Source and Credits

All data used in this study is freely accessible via torrent. The dataset comprises Reddit data up to April 2023, collected by users **u/raiderbdev** and **u/Watchful1**, and is hosted openly on `academictorrents.com`. Their contributions have been invaluable in enabling and supporting the development of this project.

2.2 Construction

We processed 50 million potential subreddit pairs (for 10000 subreddits). We listed all the users who posted in each of these subreddits and tried to find common users. The intensity of the edge weight is later explained in later sections. We have listed only SFW subreddits that are listed in Reddits top communities page.

2.3 Optimisation while Calculating Edge Weights

While calculating similarity matrices, we had to compute the number of common users between pairs and also the sum of posts and users. Given that we had around 50 million pairs, these calculations had to be optimised for performance.

To achieve this, we applied several techniques:

1. **Hashing Usernames:** Instead of comparing long strings for equality (which is an $\mathcal{O}(L)$ operation, where L is the length of the string), we precomputed hashes of usernames. Comparing two integers is a single-instruction operation and thus significantly faster.

2. **Sorting and Two-Pointer Technique:** Once the usernames (represented as hashes) were sorted, we could efficiently compute the number of common users using the two-pointer technique. Here's a pseudocode explanation:

```
i = 0, j = 0
while i < A.size() and j < B.size():
    if A[i] == B[j]:
        common += 1
        i += 1
        j += 1
    else if A[i] < B[j]:
        i += 1
    else:
        j += 1
```

   This technique allows linear-time comparison of two sorted lists of user hashes.

3. **Parallelisation:** We further improved computation time by parallelising the similarity matrix construction using multi-threading support provided by C++ library `OpenMP`.

4. **Horizontal Scaling:** To handle massive data volume, we horizontally scaled the computation by dividing the user pairs across multiple machines. Each machine processed a subset of the pairs independently and in parallel, and the final results were aggregated.

To optimize computation, Applied thresholds based on minimum post counts, Filtered out known bots and Implemented the solution in C++ for efficiency.

Other languages would have required approximately 1200+ hours of computation time, making C++ the preferred choice for our analysis.

2.4 User Thresholding

As mentioned earlier, applying a threshold for minimum posts is essential for both computational efficiency and obtaining meaningful results. However, determining the optimal threshold value requires careful consideration. Through statistical analysis of all users who posted in subreddits, we established a data-driven threshold that balances inclusivity with data quality.

The percentile distribution of user activity reveals:

- 50th percentile: 2 posts
- 75th percentile: 5 posts
- 90th percentile: 16 posts

We considered only users with more than 3 posts ($>$ 50th-75th percentile) and excluded known bots to reduce noise. This threshold effectively removes low-activity users while retaining Reddit's core user base.

2.5 Similarity Metrics

We computed three similarity measures for subreddit pairs, where A and B represent the sets of users who have posted more **than three times** in each respective subreddit over the past five years. The term $p_{u,A}$ denotes the post count of user u in subreddit A.

The Jaccard similarity metric measures the proportion of shared users relative to all unique users in both subreddits:

$$\frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

where $|A|$ and $|B|$ represent the total number of active users in each subreddit.

The minimum-based similarity normalizes the shared user count by the smaller subreddit's user base:

$$\frac{|A \cap B|}{\min(|A|, |B|)}$$

The cosine similarity incorporates post frequency, calculating the cosine of the angle between the subreddits' activity vectors:

$$\frac{\sum p_{u,A} \cdot p_{u,B}}{\sqrt{(\sum p_{u,A}^2) \cdot (\sum p_{u,B}^2)}}$$

This measure accounts for both user overlap and relative posting activity patterns.

3 Analysis

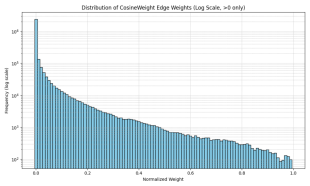
3.1 Frequency and Distribution

Looking at the distribution of Edge Weight Frequency show in Fig.2, its clear that Cosine 2a and Minimum Weight 2c metrics retain many high-degree nodes, suggesting they effectively capture network hubs, while Jaccard similarity 3c emphasizes the significance of individual nodes. This analysis includes all users who have posted more than 3 times in the past 5 years.

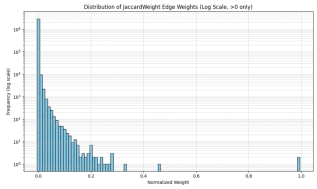
3.2 Betweenness Centrality

The subreddit nodes are ranked based on their Betweeness values for the three edge weight types Table 1

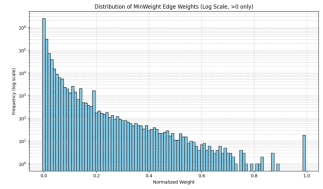
The variation suggests different metrics capture distinct connectivity patterns like Cosine emphasizes activity-weighted bridging, Jaccard highlights pure user overlap while Minimum-based detects asymmetric relationships.



(a) Cosine Weight Distribution



(b) Jaccard Weight Distribution



(c) Minimum Weight Distribution

Figure 1: frequency vs edge weights plot for similarity metrics

Cosine Node	Value	Jaccard Node	Value	Min Node	Value
r/pokemonshuffle	0.1228	r/furryartschool	0.0465	r/technology	0.0127
r/environment	0.1060	r/Quadrinhos	0.0371	r/furryartschool	0.0117
r/technology	0.0857	r/ImaginaryHellscape	0.0359	r/ImaginaryHellscape	0.0116
r/abandonedporn	0.0822	r/memescirclejerk	0.0356	r/Quadrinhos	0.0114
r/machining	0.0816	r/environment	0.0352	r/antiwork	0.0112
r/cars	0.0791	r/LiberalGooseGroup	0.0348	r/environment	0.0107
r/interestingasf**k	0.0781	r/interestingasf**k	0.0343	r/interestingasf**k	0.0106
r/india	0.0779	r/machining	0.0340	r/LiberalGooseGroup	0.0106
r/memes	0.0768	r/antiwork	0.0333	r/memescirclejerk	0.0103
r/natureismetal	0.0759	r/pokemonshuffle	0.0331	r/facepalm	0.0101

Table 1: Top 10 Subreddits by Betweenness Centrality Across Edge Weighting Schemes

3.3 Algorithm Performance

After evaluating both Leiden and Louvain clustering algorithms on our subreddit network, Louvain demonstrated superior performance across all three similarity metrics. We set edge weight thresholds to determine meaningful connections between nodes.

3.4 Cluster Detection

The Louvain algorithm identified distinct community structures for each similarity metric in Table 2

Our clustering 2 analysis revealed several important patterns in the subreddit network structure. Across all similarity metrics, mainstream communities like `r/memes` and `r/NoStupidQuestions` consistently emerged as dominant hubs, reflecting their universal appeal on the platform. The metrics each captured distinct aspects of community organization - Cosine similarity uniquely identified regional Indian communities such as `r/mumbai` and `r/CarsIndia`, while Minimum Weight clustering detected UK-focused subreddits including `r/CasualUK` and `r/unitedkingdom`. Jaccard similarity proved particularly effective at grouping specialized communities, clustering technical subreddits

Table 2: Top Subreddits by Clusters

Metric	Cluster	Top Subreddits (Degree)
Cosine	#1 (4897)	r/memes (5531), r/NoStupidQuestions (5481), r/aww (5412), r/Showerthoughts (5382), r/gaming (5256), r/teenagers (5079), r/mildlyinfuriating (4937), r/Music (4742), r/tipofmytongue (4691), r/StarWars (4531)
	#2 (967)	r/pokemongo (4496), r/buildapc (4441), r/lego (4118), r/iphone (4087), r/ADHD (3977), r/NintendoSwitch (3918), r/CasualUK (3891), r/harrypotter (3863), r/MechanicalKeyboards (3779), r/RocketLeague (3738)
	#3 (18)	r/mumbai (2722), r/CarsIndia (2135), r/uttarpradesh (1578), r/nagpur (1520), r/StartUpIndia (1448), /r/punjabi (491), r/cricketworldcup (396), r/Indiangamers (347), r/indiaplace (238), r/Haryana (184)
Jaccard	#1 (1464)	r/memes (5531), r/NoStupidQuestions (5481), r/aww (5412), r/Showerthoughts (5382), r/gaming (5256), r/teenagers (5079), r/mildlyinfuriating (4937), r/Music (4742), r/tipofmytongue (4691), r/StarWars (4531)
	#2 (1190)	r/privacy (3460), r/DataHoarder (3419), r/environment (2933), r/Economics (2918), r/Python (2791), r/androiddev (2779), r/linux (2773), r/finance (2735), r/economy (2599), r/linuxquestions (2573)
	#3 (868)	r/DisneyWorld (2417), r/TheCloneWars (2310), r/Barbie (2277), r/90scartoons (2223), r/moviecritic (2179), r/thevenomsite (1970), r/Arrowverse (1919), r/muppets (1863), r/Dolls (1829), r/hbo (1789)
Min Weight	#1 (1817)	r/memes (5531), r/NoStupidQuestions (5481), r/Showerthoughts (5382), r/gaming (5256), r/teenagers (5079), r/mildlyinfuriating (4937), r/tipofmytongue (4691), r/StarWars (4531), r/pokemongo (4496), r/MadeMeSmile (4388)
	#2 (1082)	r/dating_advice (3511), r/formula1 (3508), r/2007scape (3502), r/doordash_drivers (3412), r/Superstonk (3393), r/loseit (3370), r/lakers (3336), r/DotA2 (3249), r/MakeupAddiction (3182), r/MMA (3122)
	#3 (959)	r/antiwork (4463), r/interestingasf**k (4320), r/news (4319), r/nba (3966), r/CasualUK (3891), r/europe (3861), r/Philippines (3861), r/PoliticalHumor (3827), r/Christianity (3759), r/unitedkingdom (3750)

like r/privacy and r/Python together while also organizing entertainment niches such as r/DisneyWorld and r/Barbie.

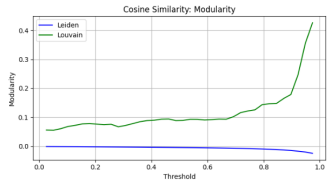
We observed significant variation in cluster sizes, with Cosine producing one very large cluster (4,897 nodes) alongside much smaller ones, while Jaccard created more balanced groupings. Notably, gaming-related subreddits including r/gaming, r/pokemongo, and r/StarWars appeared prominently across all metrics, underscoring gaming's central role in Reddit's ecosystem. These results collectively demonstrate how different similarity metrics reveal complementary aspects of community structure, with Cosine capturing geographic affinities, Jaccard reflecting topical specialization, and Minimum Weight highlighting activity patterns.

3.5 Powerlaw

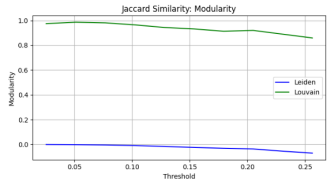
The empirical data Fig. 3 exhibits excellent agreement with the scale-free network generated by the power-law distribution, indicating a robust heavy-tailed degree distribution characteristic of real-world complex networks. The data was processed in carefully selected batches to ensure proper fitting conditions, while edge weight thresholds were applied to eliminate noise from spurious connections.

The power-law exponent $\alpha = 1.76$ holds particular significance as it governs the network's connectivity pattern. This value suggests a moderately steep decay in the probability of high-degree nodes, falling within the typical range ($2 < \alpha < 3$) observed in biological and technological networks. The minimum weight threshold of 2.64 represents the cutoff where edge weights demonstrate sufficient statistical significance to be included in the scale-free regime.

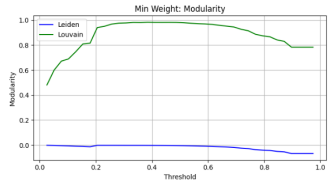
The Jaccard index value of 3.53 quantifies the structural similarity between the empirical network and the theoretical model, where higher values indicate stronger agreement in local connectivity patterns. These parameters collectively validate that the observed network maintains the fundamental properties of scale-free systems: the presence of hub nodes, resilience to random failures, and efficient information propagation through short path lengths.



(a) Cosine Weight Modularity



(b) Jaccard Weight Modularity



(c) Minimum Weight Modularity

Figure 2: Comparative performance of Louvain clustering across three similarity metrics: (a) Cosine similarity captures activity-weighted relationships, (b) Jaccard similarity measures pure user overlap, and (c) Minimum weight identifies asymmetric connections.

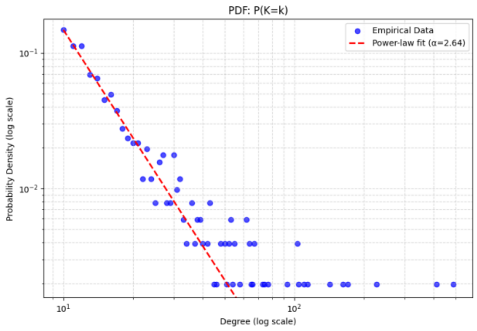
4 Network Visualization

The resultent network Fig.4 resulted in over 2.3 Million edges and over 5000 nodes. Due technical limitations and processing power, we had to remove few edges for better visualization.

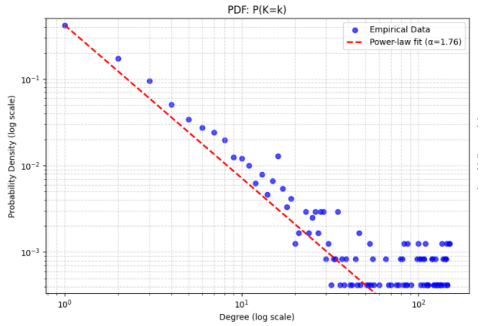
5 Conclusion

This study explored the Reddit network using three edge weighting strategies—**cosine similarity**, **Jaccard similarity**, and **minimum overlap**—to analyze centrality and community cohesion across subreddits. Betweenness centrality highlighted subreddits like `r/pokemonshuffle`, `r/environment`, and `r/technology` as key *bridges* that facilitate cross-cluster information flow, while edge-weight variations revealed differing interpretations of influence and community structure.

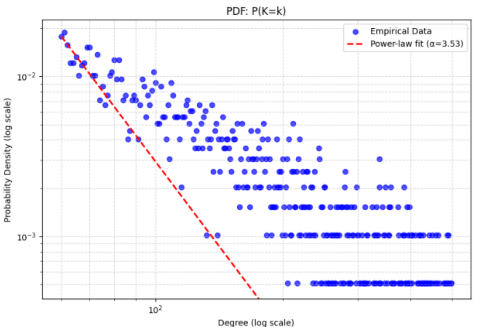
An analysis of clustering coefficients, ranging from 0.204 to 0.996, provided additional insight into local connectivity. Highly specialized communities such as `r/pokemonshuffle` and `r/machining` exhibit extremely high clustering (above 0.95), indicating dense local interactions and likely echo chamber behavior. In contrast, general-purpose or support-oriented subreddits like `r/buildapc`, `r/legaladvice`, and `r/NoStupidQuestions` show much lower clustering, suggesting broader reach with more diffuse engagement patterns.



(a) PowerLaw based on For the MinWeight



(b) PowerLaw based on For the Cosine Weight



(c) PowerLaw based on For the Jaccard Weight

Figure 3: Power Law Distributions

Figure 4: Visualized using Gephi

Interestingly, edge weighting schemes affect structural interpretations: cosine similarity tends to emphasize embedded central nodes, while Jaccard and Min-weighted graphs uncover decentralization and exclusive structural roles. The slight negative or near-zero correlation between betweenness and clustering across all methods indicates that structural bridges are often not part of tightly-knit communities.

These findings have practical implications. High-betweenness, low-clustering nodes are crucial for information diffusion and platform robustness, while high-clustering nodes reflect insular communities that may require targeted moderation or outreach.

Resources

- Drive link to the Analysis
- GitHub Repository preprocessing
- GitHub Repository for Data Fetch Automation
- Dataset Source