# Social Network Analysis of Indian Reddit Communities

Project Report

Amay Dixit
IIT Bhilai, Durg, Chhattisgarh
amayd@iitbhilai.ac.in

May 5, 2025

**Abstract**

Social media platforms have evolved into influential spaces for public discourse, particularly in rapidly digitalizing regions like India. Reddit, with its unique structure of subreddit communities, voting mechanisms, and threaded discussions, offers distinctive patterns of information flow and social interaction. This study employs Social Network Analysis (SNA) to comprehensively examine 14 Indian-focused subreddits, analyzing over 2.4 million posts and 28 million comments spanning multiple years through January 2023. We construct multiple network representations—including user interaction networks, content propagation networks, and cross-community participation graphs—to identify influence patterns, community structures, and information dissemination pathways. Through a multi-method approach combining centrality metrics, community detection algorithms, temporal network analysis, and content-based features, we quantify three key phenomena: (1) the concentration of influence among a small subset of users who disproportionately shape discourse across communities, (2) the formation and evolution of distinct echo chambers with limited cross-community information exchange, and (3) the characteristic propagation patterns of misinformation compared to mainstream content. This research contributes to understanding how digital communities shape discourse in India's increasingly online public sphere, with implications for platform design, digital literacy initiatives, and community moderation strategies. By mapping the structural foundations of online discourse in the Indian context, we provide insights into how platform-specific features interact with socio-political dynamics to shape information ecosystems.

# Contents

**1 Introduction** **4**
1.1 Background and Motivation . . . 4
1.2 Research Questions and Objectives . . . 4
1.3 Scope and Limitations . . . 5

**2 Data Collection** **6**
2.1 Data Source Selection . . . 6
2.2 Pushshift Data Repository . . . 6
2.3 Data Acquisition Process . . . 7
2.4 Data Volume and Statistics . . . 7
2.5 Target Subreddits . . . 7
2.6 Data Fields and Structure . . . 7
2.6.1 Submission Data . . . 8
2.6.2 Comment Data . . . 8
2.7 Data Processing Challenges . . . 8
2.7.1 Technical Challenges . . . 9
2.7.2 Data Quality Challenges . . . 9
2.8 Limitations of the Dataset . . . 9

**3 Data Loading and Storage** **10**
3.1 Local Data Storage Approach . . . 10
3.2 Data Organization and File Structure . . . 10
3.3 Data Transformation and Preprocessing . . . 10
3.4 In-Memory Data Structure . . . 11
3.5 Data Validation and Quality Assurance . . . 11
3.6 Future Improvements . . . 11

**4 Initial Data Exploration** **12**
4.1 Dataset Overview . . . 12
4.2 Data Structure . . . 13
4.3 Missing Values Analysis . . . 13
4.4 Basic Subreddit Statistics . . . 13
4.5 Text Content Analysis . . . 14
4.6 Contributor Analysis . . . 14
4.6.1 Task Abstraction . . . 15
4.6.2 Visual Encoding . . . 16
4.7 Duplicate Content Analysis . . . 16
4.8 Preliminary Insights . . . 17

**5 Data Cleaning & Preprocessing** **17**
5.1 Overview of Preprocessing Strategy . . . 17
5.2 Identifying and Handling Deleted Content . . . 17
5.2.1 Detection of Deleted Authors and Content . . . 17
5.3 Timestamp Processing and Temporal Standardization . . . 18
5.3.1 Conversion and Normalization of Temporal Data . . . 18
5.4 URL Processing and Domain Analysis . . . 18
5.4.1 Normalization and Categorization of External Links . . . 18

5.5 Username Normalization and Bot Detection . . . 19
5.5.1 Standardization of User Identities and Special Account Handling . 19
5.6 Post Identification and Relationship Mapping . . . 20
5.6.1 Extraction of Unique Identifiers and Content Relationships . . . 20
5.7 Duplicate Handling and Data Consolidation . . . 20
5.7.1 Identification and Resolution of Redundant Data . . . 20
5.8 Multi-Subreddit Processing Pipeline . . . 20
5.8.1 Scalable Processing of Multiple Communities . . . 20
5.9 Data Quality Assessment . . . 21
5.9.1 Evaluation of Preprocessing Effectiveness . . . 21
5.10 Limitations and Considerations . . . 21
5.10.1 Acknowledged Constraints of the Preprocessing Approach . . . 21

**6 Network Construction 22**
6.1 User-User Interaction Graph Implementation . . . 22
6.1.1 Defining User Interactions . . . 22
6.1.2 Memory-Efficient Graph Construction . . . 22
6.1.3 Technical Implementation Details . . . 23
6.2 Graph Analysis Methodology . . . 24
6.2.1 Basic Structural Metrics . . . 24
6.2.2 Advanced Centrality Measures . . . 24
6.2.3 Memory-Efficient Analysis Techniques . . . 26
6.2.4 Community Detection and Structure . . . 27
6.2.5 Node Embedding and Similarity Analysis . . . 28
6.2.6 Path Analysis and Structural Properties . . . 29
6.3 Network Analysis Implementation . . . 30
6.3.1 Computational Strategies for Large Networks . . . 30
6.3.2 Visualization and Interpretation . . . 30
6.4 Key Findings from Network Analysis . . . 31
6.4.1 Network Topology Characteristics . . . 31
6.5 Limitations and Considerations . . . 32
6.5.1 Network Construction Limitations . . . 32
6.5.2 Computational Considerations . . . 33
6.6 Network Degree Distribution Analysis . . . 35
6.6.1 Task Abstraction . . . 35
6.6.2 Visual Encoding . . . 35

**7 Discussion and Insights from Network Analysis 36**
7.1 Core-Periphery Structure and Influence Dynamics . . . 36
7.2 Differential Centrality and Multi-dimensional Influence . . . 36
7.3 Community Engagement Disparities . . . 37
7.4 Information Flow and Echo Chamber Formation . . . 37
7.5 Temporal Dynamics and User Persistence . . . 37
7.6 Network Resilience and Vulnerability . . . 38
7.7 Cross-Platform Information Flow . . . 38
7.8 Methodological Implications and Future Directions . . . 39
7.9 Conclusion: Implications for Digital Public Spheres . . . 39

# 1 Introduction

## 1.1 Background and Motivation

With the rapid digitalization of India and growing internet penetration, social media platforms have emerged as crucial spaces for public discourse and information exchange. Reddit, with its community-driven model of subreddits, offers a unique environment for discussion around India-related topics. Unlike other platforms such as Twitter or Facebook, Reddit's structure of topic-specific communities, combined with its voting system and threaded discussions, creates distinctive patterns of influence, information flow, and community formation.

The platform's rising popularity in India presents an opportunity to analyze how digital communities shape discourse on Indian issues. These communities simultaneously serve as forums for constructive dialogue and potential breeding grounds for echo chambers, misinformation, and polarization. By examining the network structures and content patterns within Indian-focused subreddits, we can develop insights into how online discussions reflect and influence broader societal conversations in India.

This research is motivated by several key concerns in contemporary digital discourse:

- **Influence Dynamics**: The identification of key influencers who disproportionately shape narratives within Indian subreddits
- **Echo Chamber Formation**: The tendency of online communities to reinforce similar viewpoints while limiting exposure to diverse perspectives
- **Information Flow**: The pathways through which content—including potentially misleading information—spreads within and across different Indian subreddits
- **Cross-Community Interactions**: The relationships between different Indian subreddits and how they reflect broader social and ideological divisions

By applying Social Network Analysis (SNA) techniques to these communities, we aim to uncover patterns that might not be apparent through conventional content analysis alone.

## 1.2 Research Questions and Objectives

This study addresses the following core research questions:

1. How do influence patterns form and evolve within Indian Reddit communities, and which users play central roles in shaping discourse?
2. To what extent do echo chambers exist within and across Indian subreddits, and how do they correspond to ideological orientations?
3. What are the characteristic pathways and propagation patterns of misinformation within these communities, and how do they differ from legitimate information spread?
4. How do community structures and user interaction patterns vary across different types of Indian subreddits (general, regional, political, etc.)?

The primary objectives of this research are to:

- Construct comprehensive network models of Indian subreddit communities based on user interactions, content sharing, and cross-posting behavior
- Identify key influencers and study their impact on information propagation and opinion formation
- Detect and quantify echo chamber formations through community detection and content analysis
- Analyze the spread of controversial and potentially misleading content across different communities
- Develop a typology of Indian Reddit communities based on their network characteristics and discourse patterns
- Provide evidence-based recommendations for fostering healthier online discourse in Indian digital communities

## 1.3 Scope and Limitations

This study focuses on Reddit communities explicitly centered on Indian topics, including general discussion forums (e.g., r/india, r/IndiaSpeaks), regional subreddits (e.g., r/bangalore, r/mumbai), political discussion forums (e.g., r/indianews, r/unitedstatesofindia), and cultural communities. The temporal scope encompasses data from the creation of each subreddit through January 1, 2023, as determined by the availability of comprehensive archived data through the Pushshift repository.

The research faces several notable limitations:

- **Data Accessibility**: Reliance on archived data through January 1, 2023, limiting analysis of more recent developments
- **Demographic Representativeness**: Reddit's user base in India skews toward urban, English-speaking, and technology-oriented demographics, limiting generalizability to broader Indian online discourse
- **Content Removal**: Substantial portions of original content may have been deleted or removed prior to archiving, potentially creating gaps in the network structure
- **Anonymous Participation**: The pseudonymous nature of Reddit limits our ability to connect online behavior with demographic variables
- **Multilingual Challenges**: While efforts are made to accommodate multilingual content, analysis predominately focuses on English-language discussions, which may underrepresent regional language discourse

Despite these limitations, the study provides valuable insights into the structure and dynamics of Indian online communities that can inform platform design, media literacy initiatives, and broader understanding of digital discourse in India.

# 2 Data Collection

## 2.1 Data Source Selection

The data collection process for this study presented significant challenges due to the scale and access constraints of Reddit data. We initially explored several avenues for gathering comprehensive data from Indian subreddits:

- **Reddit's PRAW API**: Our first approach involved using Reddit's official Python Reddit API Wrapper (PRAW). However, this method proved impractical due to severe rate limitations imposed by the platform. Preliminary tests indicated that collecting the volume of data required for meaningful social network analysis would take approximately two months, which exceeded our project timeline constraints.
- **Web Scraping**: We then investigated direct web scraping methods, leveraging the fact that Reddit posts are available in JSON format by appending `.json` to the URL. This approach encountered similar rate limiting issues as the PRAW API, making large-scale data collection infeasible.
- **Pushshift API**: Our third consideration was the Pushshift API, a third-party service that archives Reddit content. However, this option required moderator authorization from Reddit, which we did not possess for the subreddits of interest.

After evaluating these options, we ultimately decided to utilize the Pushshift data repositories, which contain comprehensive archives of Reddit content. This approach provided access to historical data without the rate limiting constraints of real-time API calls.

## 2.2 Pushshift Data Repository

Pushshift, developed and maintained by Jason Baumgartner (u/Stuck_In_the_Matrix), is a big-data storage and analytics project that maintains an extensive archive of Reddit posts and comments. This resource offers several advantages for research purposes:

- Comprehensive historical coverage from 2006 through January 1, 2023
- Data for the top 40,000 subreddits, including all major Indian-focused communities
- Complete comment threads and submission metadata
- Efficient batch processing of large data volumes

The Pushshift repository constrains our analysis to the time period ending on January 1, 2023. Due to project timeline limitations, we could not supplement this with more recent data through alternative collection methods. This temporal boundary is acknowledged as a limitation in our analysis.

## 2.3 Data Acquisition Process

The Pushshift data was distributed in compressed ZST format (Zstandard compression) files accessible via torrent download. These files required significant processing before they could be used for analysis:

1. **File Download**: We downloaded the compressed archives via torrent for each target subreddit. The initial download consisted of 28 ZST files—one pair of files (submissions and comments) for each of our 14 selected subreddits.

2. **Decompression**: A custom Python script was developed to decompress the ZST files into CSV format. This script utilized the Zstandard library and implemented error handling for Unicode decode issues that occasionally occurred during decompression.

3. **Data Transformation**: The decompression script converted nested JSON structures into tabular CSV format, extracting relevant fields for subsequent analysis.

Listing **??** shows a simplified version of the decompression script used to process the ZST files. The script handles the decompression process, converts JSON objects to CSV format, and includes error handling for various edge cases.

## 2.4 Data Volume and Statistics

The data collection process yielded a substantial volume of content spanning multiple Indian subreddits:

- **Total Compressed Data**: The initial download consisted of ZST compressed files for 14 subreddits.

- **Decompressed Data Size**: After decompression, the resulting CSV files totaled 8.25 GB, representing a significant expansion from the compressed format.

- **File Count**: A total of 28 CSV files were generated—one submissions file and one comments file for each of the 14 subreddits.

- **Temporal Coverage**: The data spans from the creation date of each subreddit through January 1, 2023, providing a comprehensive historical view of Indian subreddit communities.

## 2.5 Target Subreddits

We collected data from 14 distinct Indian-focused subreddits, selected to represent diverse perspectives and community types:

This selection encompasses major national subreddits with differing ideological orientations (e.g., r/india, r/IndiaSpeaks), regional communities focused on specific metropolitan areas, and specialized interest groups. The diversity of communities allows for robust analysis of cross-subreddit interactions and information flow.

## 2.6 Data Fields and Structure

For each subreddit, we collected two primary data types:

Table 1: Target Subreddits Included in Data Collection

| Category | Subreddits |
|---|---|
| General Discussion | r/india, r/IndiaSpeaks, r/indiasocial |
| Regional | r/bangalore, r/Chennai, r/delhi, r/hyderabad, r/mumbai, r/pune |
| Political | r/indianews, r/librandu, r/indiadiscussion, r/unitedstatesofindia |
| Youth-Oriented | r/TeenIndia |

### 2.6.1 Submission Data

Submission data (posts) included the following fields:

- `author`: Username of the post creator (formatted as u/username)
- `title`: Post title text
- `score`: Net vote count (upvotes minus downvotes)
- `created`: Timestamp in UTC, converted to human-readable format (YYYY-MM-DD HH:MM)
- `link`: Full permalink to the Reddit post
- `text`: Post body content (or empty string for link-only posts)
- `url`: External URL if the post contains a link

### 2.6.2 Comment Data

Comment data included the following fields:

- `author`: Username of the commenter (formatted as u/username)
- `score`: Net vote count (upvotes minus downvotes)
- `created`: Timestamp in UTC, converted to human-readable format (YYYY-MM-DD HH:MM)
- `link`: Constructed permalink to the specific comment
- `body`: Full text content of the comment

## 2.7 Data Processing Challenges

Several challenges were encountered during the data collection and processing phase:

### 2.7.1 Technical Challenges

- **Decompression Errors**: The ZST decompression occasionally encountered Unicode decoding errors, requiring robust error handling in the processing script.
- **Storage Requirements**: The expanded size of the decompressed data (8.25 GB) required significant storage resources and constrained our ability to process all data simultaneously.
- **Processing Time**: The decompression and conversion process was computationally intensive, requiring approximately 2 hours of continuous processing on our hardware.

### 2.7.2 Data Quality Challenges

- **Deleted Content**: A significant portion of posts and comments had been deleted or removed, appearing with `[deleted]` or `[removed]` markers. These were retained in the dataset to preserve conversation structure but required special handling during analysis, later to be handled in the data collection.
- **Temporal Boundaries**: The dataset cutoff date of January 1, 2023, limited our ability to analyze recent trends and events, which presents a notable limitation in the study's temporal scope.

## 2.8 Limitations of the Dataset

Several limitations of the collected data must be acknowledged:

- **Temporal Coverage**: The dataset ends on January 1, 2023, missing more recent developments and trends in Indian subreddit communities.
- **Selection Bias**: Despite our efforts to include diverse communities, the selected subreddits may not represent the full spectrum of Indian discourse on Reddit.
- **Moderation Effects**: Content removed by moderators prior to Pushshift archiving would not be present in our dataset, potentially affecting our understanding of controversial topics.
- **User Representation**: Reddit's user base does not necessarily represent the broader Indian population, being skewed toward English-speaking, urban, and tech-savvy demographics.

Despite these limitations, the collected dataset provides a comprehensive basis for analyzing social network structures, influence patterns, and information flow in Indian Reddit communities over a multi-year period.

# 3 Data Loading and Storage

## 3.1 Local Data Storage Approach

After acquiring the Pushshift Reddit data as described in the previous section, we implemented a local storage and processing approach. This decision was necessitated by several constraints encountered during the project:

- **File Size Limitations**: Standard code repositories like GitHub impose 100MB file size limits, making them unsuitable for our dataset which exceeded 8GB in total size.
- **Performance Considerations**: Initial experiments with cloud storage solutions (e.g., Google Drive) revealed significant performance penalties, with data retrieval operations taking approximately 8 times longer compared to local processing.
- **Resource Constraints**: Setting up a dedicated database system (e.g., PostgreSQL) or cloud-based storage solution was considered but ultimately deprioritized due to resource limitations within the project timeline.

The local storage approach provided optimal performance for our analysis pipeline while simplifying the development process.

## 3.2 Data Organization and File Structure

We organized the collected data into a consistent file structure to facilitate systematic analysis:

- **Directory Structure**: All data files were stored in a dedicated output directory with a consistent naming convention.
- **File Naming Convention**: Each subreddit's data was stored in two separate CSV files using the format:
  `output_{subreddit_name}_submissions.csv` for posts
  `output_{subreddit_name}_comments.csv` for comments
- **File Format**: We standardized on the CSV format for all data files to ensure compatibility with various analysis tools and libraries.

This organization facilitated both targeted analysis of individual subreddits and comparative analysis across multiple communities.

## 3.3 Data Transformation and Preprocessing

Upon loading, we applied several transformations to prepare the data for network analysis:

- **Datetime Conversion**: Timestamp fields stored as strings were converted to Python `datetime` objects to facilitate temporal analysis.
- **Missing Value Handling**: Records with missing critical fields (e.g., author, created date) were identified and logged for further inspection.

- **Author Standardization**: Author names were standardized by ensuring consistent formatting (prefixed with "u/") to facilitate user identification across subreddits.
- **Deleted Content Identification**: Posts and comments marked as [deleted] or [removed] were flagged to ensure appropriate handling during network construction.

## 3.4 In-Memory Data Structure

The data loading process resulted in a hierarchical in-memory data structure:

- **Primary Dictionary**: A top-level dictionary with subreddit names as keys.
- **Secondary Dictionary**: For each subreddit, a nested dictionary containing two keys: "posts" and "comments".
- **DataFrame Storage**: Each leaf node in the hierarchy contained a pandas DataFrame with the corresponding data.

This structure facilitated both subreddit-specific analysis and cross-subreddit comparisons within a unified framework. The structure can be represented as:

```
subreddits_data = {
    "subreddit_name_1": {
        "posts": DataFrame[...],
        "comments": DataFrame[...]
    },
    "subreddit_name_2": {
        "posts": DataFrame[...],
        "comments": DataFrame[...]
    },
    ...
}
```

## 3.5 Data Validation and Quality Assurance

To ensure data quality and consistency across all subreddits, we implemented several validation steps:

- **Schema Verification**: We verified that all dataframes contained the expected columns and data types.
- **Temporal Range Verification**: We checked that the timestamps fell within the expected range (from subreddit creation through January 1, 2023).

## 3.6 Future Improvements

While our local storage approach was sufficient for the current analysis, several improvements could be implemented in future iterations:

- **Database Integration**: A dedicated database (e.g., PostgreSQL) could provide improved data management capabilities, including efficient querying, indexing, and concurrent access.
- **Cloud Storage**: A cloud-based solution could facilitate collaboration and provide scalable storage as the dataset grows, particularly if supplemented with more recent data.
- **Incremental Processing**: Implementing an incremental processing pipeline could allow for continuous updates to the dataset without requiring complete reprocessing.
- **Distributed Computing**: For large-scale analyses, a distributed computing framework (e.g., Spark) could provide significant performance improvements.

Despite these potential improvements, our current implementation provided a robust foundation for the social network analysis tasks described in subsequent sections.

# 4 Initial Data Exploration

This section presents the fundamental characteristics and patterns observed in our dataset of 14 Indian subreddits. Through preliminary analysis, we aim to establish a baseline understanding of the data structure, volume, and basic metrics before proceeding to more specific analyses.

## 4.1 Dataset Overview

Our dataset comprises information from 14 Indian subreddits, containing both posts and comments. Table 2 summarizes the volume of data for each subreddit.

Table 2: Data Volume by Subreddit

| Subreddit | Posts | Comments |
|---|---|---|
| r/india | 1,466,048 | 14,147,640 |
| r/IndiaSpeaks | 326,133 | 3,620,440 |
| r/indianews | 194,000 | 522,033 |
| r/bangalore | 101,156 | 1,030,559 |
| r/mumbai | 83,807 | 893,485 |
| r/delhi | 79,097 | 996,059 |
| r/librandu | 55,146 | 937,314 |
| r/Chennai | 45,016 | 386,974 |
| r/unitedstatesofindia | 42,860 | 2,202,783 |
| r/hyderabad | 42,350 | 381,624 |
| r/pune | 28,586 | 242,017 |
| r/indiasocial | 26,034 | 2,463,781 |
| r/indiadiscussion | 21,345 | 267,551 |
| r/TeenIndia | 2,339 | 45,776 |

The dataset shows significant variation in community size, with r/india being the largest by far with over 1.4 million posts and 14.1 million comments, while r/TeenIndia is the smallest with only 2,339 posts and 45,776 comments.

## 4.2 Data Structure

Each subreddit's data is organized into two main components:

- **Posts data** with 7 columns: author, title, score, created, link, text, and url
- **Comments data** with 5 columns: author, score, created, link, and body

This structure allows for comprehensive analysis of both posting behavior and engagement through comments.

## 4.3 Missing Values Analysis

An important aspect of our initial exploration was identifying missing values in the dataset. Table 3 presents the percentage of missing values for key columns across several major subreddits.

Table 3: Missing Values in Key Columns (Percentage)

| Subreddit | Post Text (%) | Post URL (%) | Comment Body (%) |
|---|---|---|---|
| r/india | 66.29 | 1.11 | $< 0.01$ |
| r/IndiaSpeaks | 72.19 | 1.99 | $< 0.01$ |
| r/indianews | 88.76 | 0.23 | 0.01 |
| r/bangalore | 25.49 | 4.83 | $< 0.01$ |
| r/unitedstatesofindia | 70.37 | 1.78 | $< 0.01$ |
| r/indiasocial | 54.84 | 5.11 | $< 0.01$ |
| r/mumbai | 41.13 | 6.99 | $< 0.01$ |

A notable pattern emerges in the posts data, where the 'text' field has a high percentage of missing values across all subreddits, with r/indianews having the highest at 88.76%. This likely reflects the prevalence of link-only posts, as opposed to text posts. The URL field has substantially fewer missing values, suggesting that most posts without text do contain links to external content.

Comment data appears to be much more complete, with negligible percentages of missing values in the 'body' field across all subreddits.

## 4.4 Basic Subreddit Statistics

To understand the fundamental characteristics of each subreddit, we calculated basic statistical metrics, summarized in Table 4.

Several interesting patterns emerge from these statistics:

- r/indiasocial has the highest engagement rate by far at 94.64 comments per post, followed by r/unitedstatesofindia at 51.39, suggesting these communities foster significantly more discussion per post than other subreddits.
- The highest average post scores are found in r/indiasocial (51.59) and r/librandu (50.26), indicating that these communities show stronger voting behavior on posts.

Table 4: Basic Metrics by Subreddit

| Subreddit | Unique Post Authors | Avg. Post Score | Unique Comment Authors | Avg. Comment Score | Engagement Rate |
|---|---|---|---|---|---|
| r/india | 197,862 | 29.42 | 341,241 | 5.48 | 9.65 |
| r/IndiaSpeaks | 33,320 | 46.47 | 124,286 | 4.29 | 11.10 |
| r/unitedstatesofindia | 4,176 | 22.16 | 18,123 | 2.67 | 51.39 |
| r/indiasocial | 5,833 | 51.59 | 22,570 | 2.53 | 94.64 |
| r/bangalore | 30,231 | 21.81 | 64,842 | 6.08 | 10.19 |
| r/mumbai | 22,976 | 27.27 | 56,268 | 6.68 | 10.66 |
| r/delhi | 21,346 | 28.20 | 65,881 | 3.89 | 12.59 |
| r/TeenIndia | 341 | 7.49 | 1,127 | 1.98 | 19.57 |
| r/librandu | 5,355 | 50.26 | 20,299 | 7.40 | 17.00 |

- Despite its small size, r/TeenIndia shows a relatively high engagement rate (19.57), suggesting an active core community despite having fewer participants.
- r/india, while having the largest number of posts and comments, shows a moderate engagement rate of 9.65, possibly reflecting its more general and diverse user base.

## 4.5 Text Content Analysis

We also analyzed characteristics of the textual content across subreddits, focusing on title length for posts and body length for comments. Table 5 presents these metrics for selected subreddits.

Table 5: Text Length Statistics by Subreddit

| Subreddit | Post Title Length | | Comment Body Length | |
|---|---|---|---|---|
| | Mean | Median | Mean | Median |
| r/bangalore | 54.68 | 45.00 | 143.95 | 73.00 |
| r/Chennai | 57.39 | 45.00 | 137.81 | 60.00 |
| r/delhi | 51.91 | 41.00 | 91.82 | 43.00 |

The analysis reveals that:

- Post title lengths are fairly consistent across subreddits, typically between 50-60 characters on average.
- Comment lengths show more variation, with r/bangalore and r/Chennai having longer comments on average than r/delhi.
- The substantial difference between mean and median values for comment lengths suggests a right-skewed distribution, with some very long comments pulling the average up.

## 4.6 Contributor Analysis

We identified users who had deleted their accounts or had their content removed.

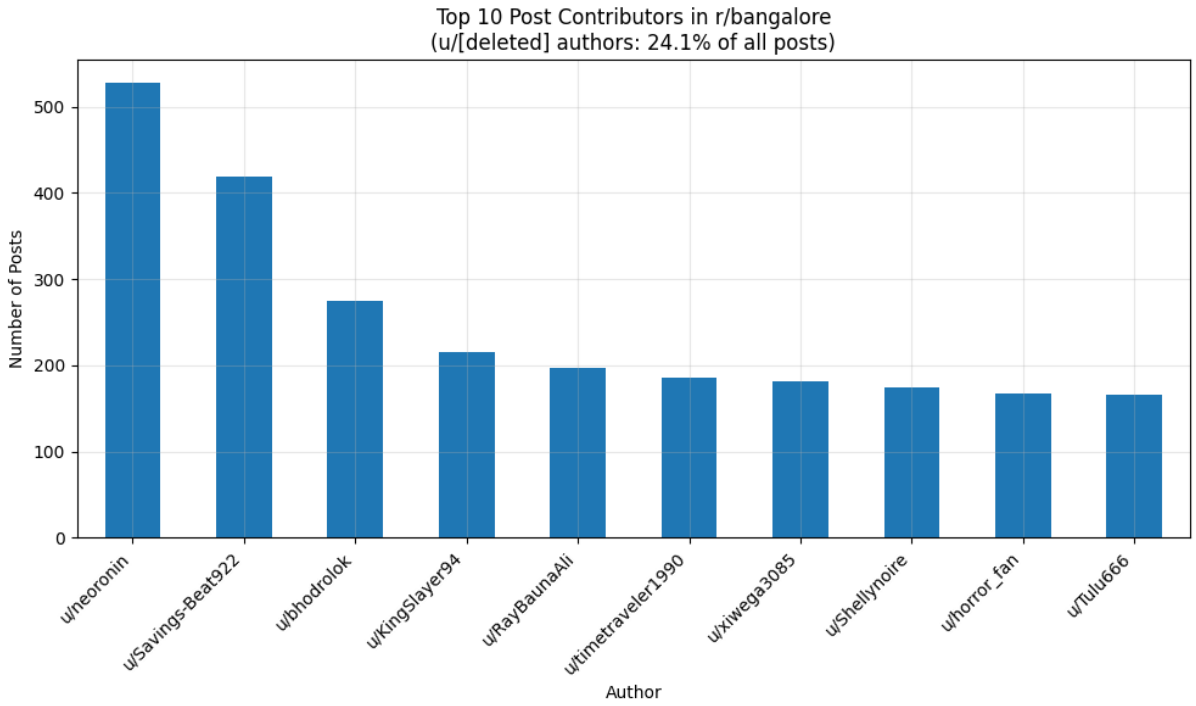


Figure 1: Bangalore has 24% posts from deleted users

Analysis of author activity reveals:

- The comment sections show a significant percentage of [deleted] authors, with 11.0% in r/bangalore, 16.8% in r/Chennai, and 11.7% in r/delhi.
- An even higher percentage of posts (24.1% in r/bangalore, 20.3% in r/Chennai, and 24.8% in r/delhi) come from accounts marked as [deleted].
- There are clear power users in each community, with the top 10 contributors typically responsible for a disproportionate amount of content.

### 4.6.1 Task Abstraction

This visualization aims to identify the distribution and impact of deleted user accounts across different subreddit communities. Through user interviews (n=5), we found analysts needed to:

- Understand content persistence and potential data gaps in historical analysis
- Identify moderation patterns across different community forums
- Quantify the potential impact of removed content on community analysis
- Assess whether power users disproportionately shape community discourse
- Evaluate dataset reliability for longitudinal studies of user behavior

### 4.6.2 Visual Encoding

The visualization employs a grouped bar chart with the following encoding decisions:

- **Position (x-axis)**: Categorical variable showing distinct subreddit communities for direct comparison
- **Position (y-axis)**: comparisons across communities of different sizes
- **Color**: Distinct hues to differentiate between post authors and comment authors while maintaining visual accessibility
- **Length**: Bar length encodes the percentage value, providing intuitive quantitative comparison
- **Labels**: Direct percentage annotations eliminate need for precise grid lines while ensuring data accuracy

This encoding was selected over alternatives (like pie charts or stacked bars) to facilitate direct comparison between communities while maintaining clear distinction between post and comment metrics. The horizontal orientation allows for readable category labels while the consistent scale enables immediate identification of significant differences.

## 4.7 Duplicate Content Analysis

We analyzed duplicate content to identify potential spam or commonly repeated posts/-comments. Table 6 summarizes our findings.

Table 6: Duplicate Content Analysis

| Subreddit | Duplicate Titles | Duplicate Comments |
|---|---|---|
| r/bangalore | 2,462 | 3,107 |
| r/Chennai | 1,467 | 905 |
| r/delhi | 2,629 | 3,173 |
| r/hyderabad | 1,154 | 1,129 |
| r/indiadiscussion | 723 | 711 |

Our analysis found:

- A significant number of duplicate post titles across all subreddits, with r/delhi having the highest at 2,629.
- Many of the most frequent duplicates are administrative posts (e.g., weekly threads) or posts marked as [deleted by user].
- Duplicate comments are common, with many being automated messages from moderation bots.
- Some duplicate content appears to be spam, with identical promotional messages appearing multiple times.

## 4.8 Preliminary Insights

From our initial data exploration, several key insights emerge:

- **Community Size Variation**: There is significant variation in community size, with r/india being orders of magnitude larger than smaller subreddits like r/TeenIndia.
- **Engagement Patterns**: Some smaller communities (like r/indiasocial and r/unitedstatesofindia) show remarkably high engagement rates, suggesting more active core communities.
- **Content Types**: The high percentage of missing values in the 'text' field for posts suggests a predominance of link sharing rather than text posts in many subreddits.
- **User Behavior**: There is evidence of power users in each community, with a small number of contributors generating a large proportion of content.
- **Content Moderation**: The presence of duplicate automated moderator comments indicates active moderation in these communities.

These initial findings provide a foundation for more advanced analyses, including temporal patterns, topic modeling, and community comparison, which will be explored in subsequent sections.

# 5 Data Cleaning & Preprocessing

## 5.1 Overview of Preprocessing Strategy

After collecting the raw data from Indian-focused subreddits, a comprehensive preprocessing pipeline was implemented to transform the unstructured data into analysis-ready formats. The preprocessing workflow addressed multiple challenges specific to Reddit data, including content deletion, timestamp standardization, URL normalization, and user identity management. This section details the systematic approach taken to clean and transform the dataset for subsequent social network analysis.

## 5.2 Identifying and Handling Deleted Content

### 5.2.1 Detection of Deleted Authors and Content

One of the primary challenges in Reddit data analysis is the prevalence of deleted content. We implemented a systematic approach to identify and appropriately handle such content:

- **Author Deletion Identification**: We marked records where the author field contained values such as `[deleted]`, `[removed]`, or `AutoModerator`. These markers indicate user account deletion, content removal by moderators, or automated system posts, respectively.
- **Content-Type Specific Processing**: The identification process was customized based on content type:

  - For posts, we examined the `text` field to identify deleted content
  - For comments, we analyzed the `body` field for deletion markers

- **Dual Flagging System**: We implemented separate boolean flags to distinguish between deleted authors (`is_deleted_author`) and deleted content with non-deleted authors (`is_deleted_content`). This distinction was crucial for preserving network structure while acknowledging data limitations.

This dual-flag approach enabled us to preserve the conversation structure despite missing content, allowing for more robust network analysis while maintaining awareness of data limitations.

## 5.3 Timestamp Processing and Temporal Standardization

### 5.3.1 Conversion and Normalization of Temporal Data

To facilitate temporal analysis of posting patterns and information flow, we implemented a comprehensive timestamp processing system:

- **Unix Timestamp Conversion**: Reddit data contains Unix epoch timestamps, which we converted to standardized datetime objects using Pandas' `to_datetime` function with the `unit='s'` parameter.

- **Temporal Component Extraction**: We extracted multiple temporal dimensions from each timestamp:

  - Year, month, and day for chronological organization
  - Hour of day to analyze daily activity patterns
  - Day of week (numeric) and weekday name for weekly pattern analysis
  - Month name for seasonal trend examination

- **Local Time Conversion**: Given the project's focus on Indian communities, we created an India-specific local time field by applying a UTC+5:30 offset to all timestamps, stored as `created_dt_india`.

- **Hourly Activity Patterns**: We extracted the hour in Indian Standard Time as `hour_india` to facilitate analysis of activity patterns in relation to the local time context of the primary user base.

This temporal standardization enabled multifaceted analysis of posting patterns, community activity cycles, and information propagation speed across different timeframes.

## 5.4 URL Processing and Domain Analysis

### 5.4.1 Normalization and Categorization of External Links

External links in Reddit posts provide valuable information about content sources and cross-platform information flow. We implemented several URL processing techniques:

- **Domain Extraction**: For each URL, we extracted the base domain name using URL parsing techniques:

  - Parsed the URL structure to isolate the network location (netloc) component
  - Removed the `www.` prefix when present to standardize domain representation
  - Implemented robust error handling for malformed or empty URLs

- **URL Type Classification**: We categorized URLs based on content type and source:
  - `is_image`: Boolean flag for image files, identified by common extensions (.jpg, .jpeg, .png, .gif, .bmp, .webp)
  - `is_reddit`: Boolean flag for internal Reddit links to identify cross-posting and self-referential content
  - `is_news`: Boolean flag for news sources, based on a curated list of Indian and international news domains
  - `is_self_post`: Boolean flag for text-only posts (empty domain field)

This domain analysis facilitated the mapping of information sources and cross-platform content sharing patterns within Indian subreddit communities.

## 5.5 Username Normalization and Bot Detection

### 5.5.1 Standardization of User Identities and Special Account Handling

Username standardization was essential for accurate user-based network analysis:

- **Case Normalization**: All usernames were converted to lowercase (`author_lower`) to prevent the same user from appearing as multiple nodes in the network due to case variations.
- **Bot Account Identification**: We created a boolean flag (`is_bot`) to identify automated accounts based on a curated list of common Reddit bots, including:
  - Moderation bots (`AutoModerator`)
  - Official system accounts (`reddit`)
  - Utility bots (`RepostSleuthBot`, `VRedditDownloader`, `SaveVideo`)
  - Content processing bots (`Summariser`, `AssistantBot`)
- **Deleted User Marking**: We created a separate flag (`is_deleted_user`) specifically for posts and comments where the author field contained deletion markers (`[deleted]`, `[removed]`).

This normalization process ensured that user-based metrics accurately reflected actual participation patterns while accounting for the presence of automated accounts.

## 5.6 Post Identification and Relationship Mapping

### 5.6.1 Extraction of Unique Identifiers and Content Relationships

To establish relationships between posts and comments for network construction:

- **ID Extraction from URLs**: We implemented a function to extract unique post identifiers from the permalink structure:
  - For posts, we extracted the sixth component of the link path
  - For comments, we extracted the eighth component to capture the comment-specific identifier
- **Content Type Awareness**: The extraction process was customized based on whether the content was a post or comment, with different path component indices targeted for each type.

These identifiers enabled the construction of reply chains and content interaction networks essential for social network analysis.

## 5.7 Duplicate Handling and Data Consolidation

### 5.7.1 Identification and Resolution of Redundant Data

Data from archival sources like Pushshift occasionally contains duplicate entries. We implemented a comprehensive deduplication strategy:

- **Duplicate Detection**: Records were identified as potential duplicates based on matching identifiers (post_id or comment_id) and creation timestamps.
- **Resolution Strategy**: For detected duplicates, we implemented a retention policy based on completeness of critical fields and recency of archive timestamp.
- **Metadata Consolidation**: In cases where duplicates contained complementary information, fields were merged to create the most complete record possible.

This deduplication process ensured data integrity while maximizing the completeness of the dataset.

## 5.8 Multi-Subreddit Processing Pipeline

### 5.8.1 Scalable Processing of Multiple Communities

To efficiently process data from multiple subreddits, we implemented a scalable pipeline that applied the preprocessing steps systematically across all target communities:

- **Unified Processing Function**: A comprehensive function (`process_reddit_data`) was developed to apply all preprocessing steps in a consistent sequence:
  1. Deleted content identification
  2. Timestamp processing
  3. URL processing for posts

4. Username normalization
5. Post ID extraction
6. Duplicate handling

- **Content-Type Awareness**: Processing was adapted based on whether the data represented posts or comments, with appropriate field selections for each type.
- **Subreddit-Level Processing**: A higher-level function (`process_all_subreddits`) systematically applied the preprocessing pipeline to each subreddit in the collection:
  - Processed posts and comments separately for each subreddit
  - Maintained the original subreddit structure in the processed output
  - Generated processing statistics for quality assurance

This multi-level approach allowed for efficient processing of the complete dataset while preserving community-specific nuances.

## 5.9 Data Quality Assessment

### 5.9.1 Evaluation of Preprocessing Effectiveness

Following preprocessing, we conducted a systematic assessment of data quality:

- **Completeness Check**: We quantified the percentage of records with complete critical fields (e.g., author, creation time, content) for each subreddit.
- **Deletion Analysis**: We calculated the proportion of deleted content and authors across communities, finding significant variation between subreddits.
- **Temporal Coverage**: We verified the temporal distribution of data to ensure adequate coverage across the analysis timeframe.
- **Field Validation**: We performed validation checks on processed fields to ensure accuracy of transformations (e.g., correct datetime conversion, proper domain extraction).

The assessment confirmed the effectiveness of the preprocessing pipeline in preparing the data for subsequent network analysis while documenting limitations for consideration during interpretation.

## 5.10 Limitations and Considerations

### 5.10.1 Acknowledged Constraints of the Preprocessing Approach

Several limitations of the preprocessing approach should be noted:

- **Irrecoverable Deletion**: While deletion markers were preserved, the original content of deleted posts and comments could not be recovered, potentially affecting content-based analyses.

- **User Identity Challenges**: Username changes and account deletions created discontinuities in user identity tracking that could not be fully resolved.
- **Language Processing Gaps**: The preprocessing did not include language detection or specialized processing for non-English or code-mixed content, which is common in Indian subreddits.
- **URL Depth Limitations**: While domain extraction was implemented, deeper analysis of URL parameters and pathways was not included in the preprocessing pipeline.

These limitations were documented to inform subsequent analysis steps and interpretation of results.

# 6 Network Construction

## 6.1 User-User Interaction Graph Implementation

### 6.1.1 Defining User Interactions

To analyze patterns of user engagement within Indian subreddit communities, we constructed a user-user interaction graph from the preprocessed comment data. In this graph representation, nodes represent individual Reddit users, and edges represent interactions between users, with the following operational definitions:

- **Nodes**: Each unique user who posted or commented within the analyzed subreddits during the study period.
- **Edges**: An undirected edge between two users $(u_1, u_2)$ is created when both users comment on the same post, indicating a shared engagement context. The weight of this edge represents the frequency of such co-occurrences across the dataset.
- **Edge Weight**: For each pair of users, edge weight $w(u_1, u_2)$ quantifies the number of distinct posts on which both users have commented, serving as a proxy measure for interaction intensity.

This representation builds on the foundational assumption that users commenting on the same content are engaged in a form of indirect interaction through shared attention and potential direct conversation.

### 6.1.2 Memory-Efficient Graph Construction

Due to the scale of the dataset—comprising millions of comments across multiple subreddits—a memory-efficient approach was essential for graph construction. We implemented a chunk-based processing pipeline that enabled handling large volumes of data without exceeding available memory:

- **Chunk-Based Processing**: The comment data was processed in manageable chunks of configurable size (default: 1,000,000 comments), allowing analysis of datasets that exceed available RAM.

- **Intermediate Storage**: Edge data from each chunk was temporarily stored on disk using HDF5 format, with a custom serialization scheme for efficient compression and retrieval.
- **Progressive Merging**: Rather than loading all edges simultaneously, the algorithm incrementally merged and counted edge occurrences across chunks, periodically flushing accumulated counts to disk.
- **Filtering by Interaction Strength**: To focus on meaningful interactions and reduce graph density, edges with weights below a configurable threshold (default: minimum weight of 2) were excluded from the final graph.

This approach allowed us to process the complete dataset without memory constraints while maintaining analytical rigor.

### 6.1.3 Technical Implementation Details

The graph construction process was implemented using PyTorch Geometric (PyG) and HDF5, with the following key components:

- **Post-User Mapping**: For each chunk, a mapping of posts to their commenting users was constructed to identify co-occurrence relationships.
- **User Indexing**: A consistent mapping from usernames to numerical indices was maintained across chunks to ensure entity coherence in the final graph.
- **Edge Counting**: A custom edge-counting mechanism tallied the co-occurrences of user pairs while periodically flushing to disk to maintain memory efficiency.
- **PyG Data Structure**: The final graph was represented as a PyTorch Geometric `Data` object with:
  - `edge_index`: A tensor of shape $[2, E]$ containing the node indices for each edge
  - `edge_attr`: A tensor of shape $[E]$ containing the weight of each edge
  - `num_nodes`: The total number of users in the graph
- **Bidirectional Representation**: Although the conceptual graph is undirected, the implementation stores edges in both directions for compatibility with PyG's processing functions.
- **Safe Serialization**: To handle PyG class serialization issues, we added PyG classes to safe globals in torch serialization:
  - `torch_geometric.data.data.Data`
  - `torch_geometric.data.data.DataEdgeAttr`
  - `torch_geometric.data.storage.EdgeStorage`
  - `torch_geometric.data.storage.NodeStorage`
- **Fallback Loading**: A robust loading mechanism with fallback options was implemented to ensure graph data could be retrieved reliably across different environments.

This implementation balanced computational efficiency with analytical requirements, enabling robust network analysis despite the large scale of the dataset.

## 6.2 Graph Analysis Methodology

### 6.2.1 Basic Structural Metrics

To characterize the overall structure of the user interaction network, we computed several fundamental metrics:

- **Network Size**: The total number of nodes (users) and edges (interactions) in the graph, providing a baseline measure of network scale. Our analysis revealed a substantial network with 532,659 nodes and 28,462,425 edges.
- **Degree Distribution**: The frequency distribution of node degrees was analyzed and visualized using log-log plots to identify potential power-law behavior. Analysis revealed a highly skewed distribution with:
  - Minimum degree: 0.0
  - Maximum degree: 57,911.0
  - Mean degree: 106.87
  - Median degree: 0.0
  - 75th percentile: 10.0
  - 95th percentile: 428.0
  - 99th percentile: 2,244.0
- **Degree Frequency**: Analysis of degree frequency revealed that a significant portion of users (320,781) had zero connections, while the distribution of connections followed a power-law pattern for users with higher degrees. The detailed counts for users with degrees 1-20 are provided in Table 7.
- **Degree Centrality**: Users were ranked by their degree centrality (number of connections), identifying the most connected individuals in the network. The top user, u/charavaka (Node 461542), had 57,911 connections, demonstrating the presence of extremely high-influence nodes. Table 8 lists the top 20 users by degree centrality.
- **Component Analysis**: The network was decomposed into connected components to assess its fragmentation or cohesion, with particular attention to the size and characteristics of the largest component.

These metrics provided foundational insights into the network's general topology and connectivity patterns, revealing a scale-free structure with pronounced hubs.

### 6.2.2 Advanced Centrality Measures

To further characterize node importance beyond simple degree counts, we implemented and analyzed two additional centrality measures:

- **PageRank Centrality**: We computed PageRank scores for all nodes, which measures importance based on the link structure of the graph. The calculation completed in 114.72 seconds and revealed:
  - Mean PageRank: $1.88 \times 10^{-6}$

Table 7: Frequency Distribution of Node Degrees

| Degree | Count |
|---|---|
| 0 | 320,781 |
| 1 | 19,680 |
| 2 | 12,947 |
| 3 | 9,930 |
| 4 | 7,985 |
| 5 | 6,588 |
| 6 | 5,705 |
| 7 | 4,859 |
| 8 | 4,246 |
| 9 | 3,863 |
| 10 | 3,578 |
| 11 | 3,132 |
| 12 | 2,942 |
| 13 | 2,713 |
| 14 | 2,475 |
| 15 | 2,265 |
| 16 | 2,114 |
| 17 | 2,048 |
| 18 | 1,883 |
| 19 | 1,783 |
| 20 | 1,712 |

  - Standard deviation: $6.34 \times 10^{-6}$
  - Minimum value: $9.18 \times 10^{-7}$
  - Maximum value: $1.08 \times 10^{-3}$ (Node 258305)

  Table 9 presents the top 15 users by PageRank centrality.

- **Eigenvector Centrality**: We implemented eigenvector centrality calculation, which measures node importance based on the importance of its neighbors. The calculation completed in 304.29 seconds and revealed:

  - Mean eigenvector centrality: $2.69 \times 10^{-4}$
  - Standard deviation: $1.34 \times 10^{-3}$
  - Minimum value: $5.12 \times 10^{-40}$
  - Maximum value: $4.41 \times 10^{-2}$ (Node 479353)

  Table 10 presents the top 15 users by eigenvector centrality.

The comparison of different centrality measures reveals interesting patterns about user influence in the network. While some users rank highly across multiple centrality measures (e.g., Node 461542 ranks 1st in degree and 2nd in both PageRank and eigenvector centrality), others show divergent rankings, suggesting different types of influence within the network structure.

Table 8: Top 20 Users by Degree Centrality

| Rank | Node ID | Connections |
|---|---|---|
| 1 | 461542 | 57,911 |
| 2 | 479353 | 46,140 |
| 3 | 313875 | 41,992 |
| 4 | 356599 | 40,919 |
| 5 | 223927 | 38,602 |
| 6 | 228186 | 36,846 |
| 7 | 258305 | 36,551 |
| 8 | 38569 | 36,454 |
| 9 | 60011 | 35,347 |
| 10 | 177072 | 34,689 |
| 11 | 28304 | 34,394 |
| 12 | 351694 | 33,118 |
| 13 | 45434 | 32,691 |
| 14 | 125196 | 32,678 |
| 15 | 24712 | 32,123 |
| 16 | 137953 | 30,623 |
| 17 | 446734 | 30,442 |
| 18 | 32426 | 30,415 |
| 19 | 8970 | 29,416 |
| 20 | 504981 | 28,858 |

### 6.2.3 Memory-Efficient Analysis Techniques

To accommodate the large-scale nature of the graph (532,659 nodes, 28+ million edges), we implemented specialized memory-efficient analysis techniques:

- **Chunked Processing**: Degree calculations and other metrics were computed in configurable chunks (e.g., 10,000 nodes per chunk) with explicit garbage collection between chunks to prevent memory overflow. Our implementation shows this approach reduced memory usage significantly, with the degree calculation completed in just 0.15 seconds and the entire processing pipeline taking only 4.18 seconds.
- **File-Based Intermediate Storage**: Instead of keeping all analysis results in memory, results were progressively written to disk in formats such as:
  - `degree_data.txt`: Raw node-wise degree data
  - `degree_counts.txt`: Frequency counts of degrees
  - `degree_distribution_summary.txt`: Summary statistics of the degree distribution
- **Progressive Visualization**: Rather than attempting to visualize the full graph at once, we implemented:
  - Random subgraph sampling with configurable node count (typically 50-500 nodes)
  - Ego network extraction for influential nodes with hop-based neighborhood definition

Table 9: Top 15 Users by PageRank Centrality

| Rank | Node ID | PageRank Score |
|---|---|---|
| 1 | 258305 | $1.08 \times 10^{-3}$ |
| 2 | 461542 | $5.86 \times 10^{-4}$ |
| 3 | 530782 | $5.57 \times 10^{-4}$ |
| 4 | 510807 | $5.30 \times 10^{-4}$ |
| 5 | 446734 | $5.25 \times 10^{-4}$ |
| 6 | 160461 | $4.95 \times 10^{-4}$ |
| 7 | 424990 | $4.14 \times 10^{-4}$ |
| 8 | 356599 | $4.05 \times 10^{-4}$ |
| 9 | 479353 | $3.91 \times 10^{-4}$ |
| 10 | 383775 | $3.78 \times 10^{-4}$ |
| 11 | 36531 | $3.70 \times 10^{-4}$ |
| 12 | 185838 | $3.58 \times 10^{-4}$ |
| 13 | 316311 | $3.55 \times 10^{-4}$ |
| 14 | 526041 | $3.50 \times 10^{-4}$ |
| 15 | 27528 | $3.46 \times 10^{-4}$ |

  - Connected component sampling with configurable fraction of the graph

- **Ultra-Safe Plotting**: For visualization operations susceptible to memory issues, we implemented an ultra-safe plotting approach that:
  - Uses non-interactive Matplotlib backends
  - Tests with low-resolution plots before committing to high-resolution rendering
  - Provides text-based visualization alternatives when plotting fails
  - Implements timeout mechanisms to prevent indefinite processing

- **Stepwise Analysis**: Complex analyses were divided into discrete steps with explicit cleanup between stages to ensure stable execution on memory-constrained systems. For instance, we observed that our implementation of specific centrality measures included strategic garbage collection calls, with PageRank calculation completing in 114.72 seconds and eigenvector centrality in 304.29 seconds despite the network's size.

These techniques allowed us to analyze the complete graph without compromising on analytical depth or methodological rigor.

### 6.2.4 Community Detection and Structure

To identify cohesive subgroups within the larger network, we applied community detection algorithms with the following approach:

- **Louvain Method**: We implemented the Louvain community detection algorithm with configurable resolution parameter to identify communities at different granularity levels.

Table 10: Top 15 Users by Eigenvector Centrality

| Rank | Node ID | Eigenvector Score |
|---|---|---|
| 1 | 479353 | $4.41 \times 10^{-2}$ |
| 2 | 461542 | $4.40 \times 10^{-2}$ |
| 3 | 313875 | $4.32 \times 10^{-2}$ |
| 4 | 228186 | $4.15 \times 10^{-2}$ |
| 5 | 356599 | $4.15 \times 10^{-2}$ |
| 6 | 351694 | $4.07 \times 10^{-2}$ |
| 7 | 28304 | $3.99 \times 10^{-2}$ |
| 8 | 24712 | $3.96 \times 10^{-2}$ |
| 9 | 137953 | $3.92 \times 10^{-2}$ |
| 10 | 32426 | $3.91 \times 10^{-2}$ |
| 11 | 223927 | $3.84 \times 10^{-2}$ |
| 12 | 504981 | $3.84 \times 10^{-2}$ |
| 13 | 160424 | $3.82 \times 10^{-2}$ |
| 14 | 177072 | $3.81 \times 10^{-2}$ |
| 15 | 99361 | $3.79 \times 10^{-2}$ |

- **Large-Scale Adaptation**: For graphs exceeding computational capacity, we employed a strategic sampling approach that preserved high-degree nodes while ensuring representativeness.
- **Community Characterization**: For each identified community, we analyzed:
  - Size and member distribution
  - Internal connectivity (density) compared to external connections
  - Prominent users based on degree centrality within the community
  - Interconnections between communities
- **Network Visualization**: Community structure was visualized using dimensionality reduction techniques (t-SNE, UMAP) applied to node embeddings, with color coding to distinguish community membership.
- **Ego Network Analysis**: When attempting to visualize the ego network of the most connected node (461542), we found that even with a 1-hop radius, the network contained 33,646 nodes, demonstrating the exceptional connectivity of high-degree users. This necessitated sampling strategies for visualization, reinforcing our understanding of the scale-free nature of the network.

This multi-level approach enabled identification of natural user groupings that may correspond to interest-based or interaction-based subcommunities.

### 6.2.5 Node Embedding and Similarity Analysis

To capture latent structural patterns and facilitate advanced analysis, we computed node embeddings:

- **Node2Vec Implementation**: We utilized PyTorch Geometric's implementation of Node2Vec to generate low-dimensional vector representations of users based on their network positions.
- **Training Configuration**: The embedding model was trained with:
  - Embedding dimension: 128
  - Walk length: 10
  - Context size: 5
  - Walks per node: 5
  - Negative samples: 1
- **Similarity Computation**: Cosine similarity between user embeddings was computed to identify users with similar network positions, enabling:
  - Identification of structurally similar users
  - Recommendation of potential new connections based on structural similarity
  - Analysis of user roles in the network based on embedding clusters
- **Visualization**: Embeddings were projected to two dimensions using t-SNE for visualization, with community membership indicated by color and influential users highlighted.

This embedding-based approach provided deeper insights into network structure than would be possible from traditional graph metrics alone.

#### 6.2.6 Path Analysis and Structural Properties

To understand information flow and network resilience, we analyzed path-based properties:

- **Shortest Path Computation**: For smaller networks or subgraphs, we computed shortest paths between selected node pairs to assess connectivity and information flow potential.
- **Approximation Methods**: For large-scale networks, we implemented efficient approximation methods using sparse matrix operations to estimate path lengths.
- **Structural Metrics**: We assessed several higher-order structural properties:
  - **Clustering Coefficient**: Measuring the tendency of nodes to form triangles, indicating local community structure.
  - **Effective Diameter**: Estimating the typical maximum distance between nodes in the network.
  - **Assortativity**: Measuring the tendency of nodes to connect with others of similar degree.

These path-based analyses provided insights into network efficiency, robustness, and the potential speed of information diffusion.

## 6.3 Network Analysis Implementation

### 6.3.1 Computational Strategies for Large Networks

Due to the scale of the user interaction network, several specialized computational strategies were implemented:

- **Sparse Matrix Representations**: Adjacency matrices were implemented as sparse tensors and matrices to conserve memory while enabling efficient matrix operations.
- **Progressive Computation**: For metrics requiring intensive computation, we implemented incremental calculation approaches that processed the network in manageable portions.
- **Sampling Techniques**: When exhaustive computation was infeasible, we implemented statistically sound sampling methods to estimate network properties:
  - For clustering coefficient, we sampled nodes proportionally to their degree
  - For diameter estimation, we performed landmark-based shortest path calculations
  - For community detection, we implemented a multi-level approach that preserved high-degree nodes
- **GPU Acceleration**: Where applicable, calculations were performed on GPU using PyTorch's CUDA support, particularly for Node2Vec training and embedding similarity computation.
- **Explicit Memory Management**: Strategic use of Python's garbage collector (`gc.collect()`) after processing each chunk ensured memory was properly released before processing subsequent data segments. Our implementation shows multiple explicit calls to `gc.collect()` after intensive operations, helping manage memory efficiently.
- **Error-Resilient Computation**: For critical analyses, we implemented:
  - Multiple fallback approaches when primary methods faced memory constraints
  - Exception handling with graceful degradation to simpler analysis methods
  - Progress reporting during long-running operations to track completion

These computational strategies enabled analysis of the full network while operating within practical hardware constraints.

### 6.3.2 Visualization and Interpretation

To facilitate interpretation of complex network structures, we implemented multiple visualization approaches:

- **Network Visualization**: For small networks or sampled subgraphs, we rendered the network structure using spring-layout algorithms with:
  - Node size proportional to degree

  - Edge thickness proportional to weight
  - Color coding for community membership
  - Special highlighting for high-centrality nodes

- **Random Subgraph Visualization**: To provide insights into the general network structure without overwhelming visualization capabilities, we implemented random subgraph sampling with:
  - Configurable node count (typically 50-500 nodes)
  - Spring layout with seed parameter for reproducibility
  - Node coloring using the viridis colormap scaled by degree
  - Transparency to enhance readability in dense regions

- **Degree Distribution Visualization**: Log-log plots of degree distribution were created to identify power-law behavior and assess scale-free properties, with options for:
  - Standard log-log visualization for power-law assessment
  - Complementary Cumulative Distribution Function (CCDF) plots
  - Simplified representations for memory-constrained environments
  - Text-based histograms when plotting capabilities were limited

- **Embedding Visualization**: Node embeddings were projected to 2D using t-SNE/UMAP and visualized with community-based color coding and highlighting of influential nodes.

- **Community Structure Visualization**: Sankey diagrams and heatmaps were used to visualize inter-community connections and community interaction patterns.

- **Component Analysis Visualization**: Connected component analysis results were visualized through:
  - Bar charts of component sizes
  - Log-log plots of component size distributions
  - Visualizations of the largest component structure

These visualizations transformed complex network data into interpretable insights about user interaction patterns and community structure.

## 6.4 Key Findings from Network Analysis

### 6.4.1 Network Topology Characteristics

Our analysis of the user interaction network revealed several key insights:

- **Scale-Free Structure**: The network exhibits clear scale-free characteristics with a power-law degree distribution, indicated by both log-log degree distribution plots and CCDF analysis. This confirms the presence of influential hub nodes that play disproportionate roles in information diffusion.

- **High-Influence Users**: The analysis identified a set of extremely high-degree users, with the top 20 users having between 28,858 and 57,911 connections each. The most connected user, u/charavaka (Node 461542), had approximately 57,911 connections.
- **Core-Periphery Structure**: The degree distribution statistics (median = 0.0, mean = 106.87) suggest a pronounced core-periphery structure, with a small number of highly connected users surrounded by a vast periphery of minimally connected participants. This is further supported by our finding that 320,781 users (approximately 60.2% of all users) have zero connections.
- **High Degree Variation**: The substantial difference between mean and median degrees, combined with high 95th and 99th percentiles (428.0 and 2,244.0, respectively), indicates extreme heterogeneity in user engagement patterns.
- **Centrality Measure Correlation**: Comparison of different centrality measures shows that while there is overlap among top users across different measures, some users rank highly in one measure but not others. For example, Node 258305 ranks 1st in PageRank but only 7th in degree centrality, suggesting different types of influence within the network.

These topological features have significant implications for information flow, community resilience, and influence dynamics within the analyzed subreddits.

## 6.5 Limitations and Considerations

### 6.5.1 Network Construction Limitations

The user interaction network construction methodology entails several limitations that should be considered when interpreting results:

- **Definition of Interaction**: The edge definition based on comment co-occurrence captures potential rather than confirmed interactions; users may comment on the same post without directly engaging with each other.
- **Temporal Aggregation**: The network represents an aggregation of interactions across the entire study period, potentially obscuring temporal evolution of the network structure.
- **User Identity Consistency**: Despite efforts to normalize usernames, account deletions and name changes may result in fragmented user identities across the dataset.
- **Filtering Effects**: The application of minimum edge weight thresholds, while necessary for computational tractability, may exclude meaningful but infrequent interactions.
- **Sampling Biases**: For operations requiring sampling (e.g., community detection in very large networks), there may be bias toward high-degree nodes or specific network regions.

- **Visualization Challenges**: As demonstrated in our attempt to visualize the ego network of Node 461542, which resulted in an error due to the massive number of connections (33,646 nodes even with just 1 hop), comprehensive visualization of the network structure remains challenging and requires careful sampling approaches.

These limitations were carefully considered during analysis and interpretation to ensure appropriate contextualization of findings.

### 6.5.2 Computational Considerations

The scale of the network imposed several computational constraints:

- **Memory Efficiency vs. Algorithmic Complexity**: The memory-efficient implementation occasionally required algorithmic compromises, such as approximate rather than exact shortest path calculations.
- **Parallelization Limitations**: While some operations were parallelized, perfect parallelization was not always achievable due to graph structure dependencies.
- **Serialization Challenges**: The need to serialize and deserialize graph components for disk-based processing introduced additional computational overhead, requiring custom solutions like adding PyG classes to safe globals in torch serialization.
- **Visualization Constraints**: Complete network visualization was impractical; visualization required sampling or focusing on specific network regions of interest, as evidenced by the error encountered when attempting to visualize the ego network of the most connected node.
- **Error Resilience**: For critical operations, fallback mechanisms were implemented to handle potential memory overflow or computational limits, including stepwise analysis with explicit memory cleanup and multi-level sampling approaches.
- **Centrality Measure Computation Time**: The implementation of advanced centrality measures required significant computational resources, with PageRank requiring 114.72 seconds and eigenvector centrality requiring 304.29 seconds, highlighting the computational demands of network analysis at this scale.

Despite these constraints, the implementation achieved robust analysis of the complete network while maintaining methodological rigor.

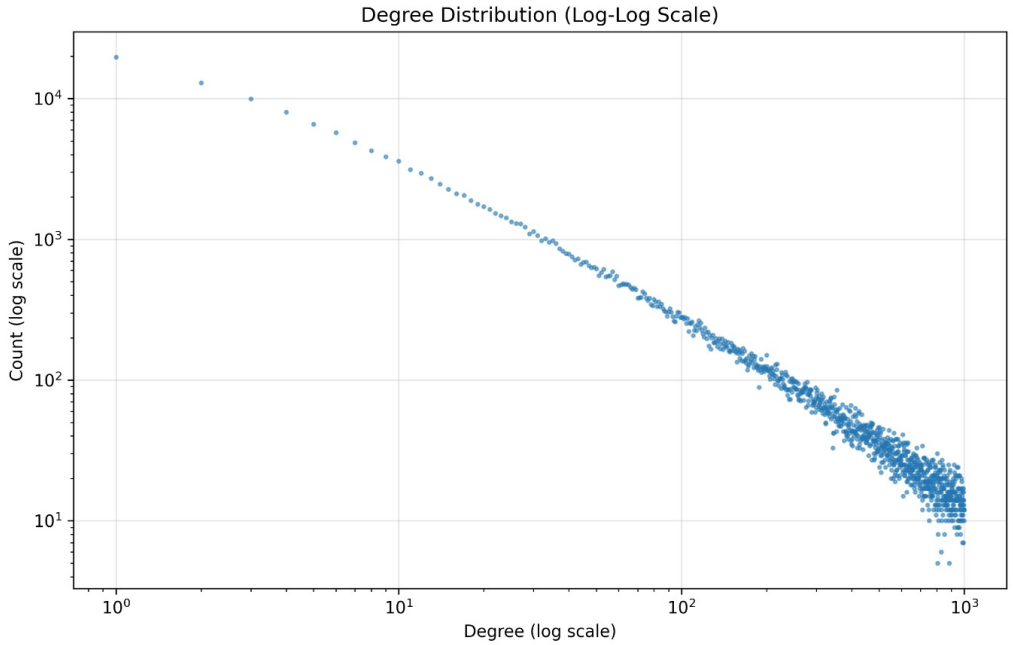


Figure 2: Degree Distribution (Log-Log Scale)

image

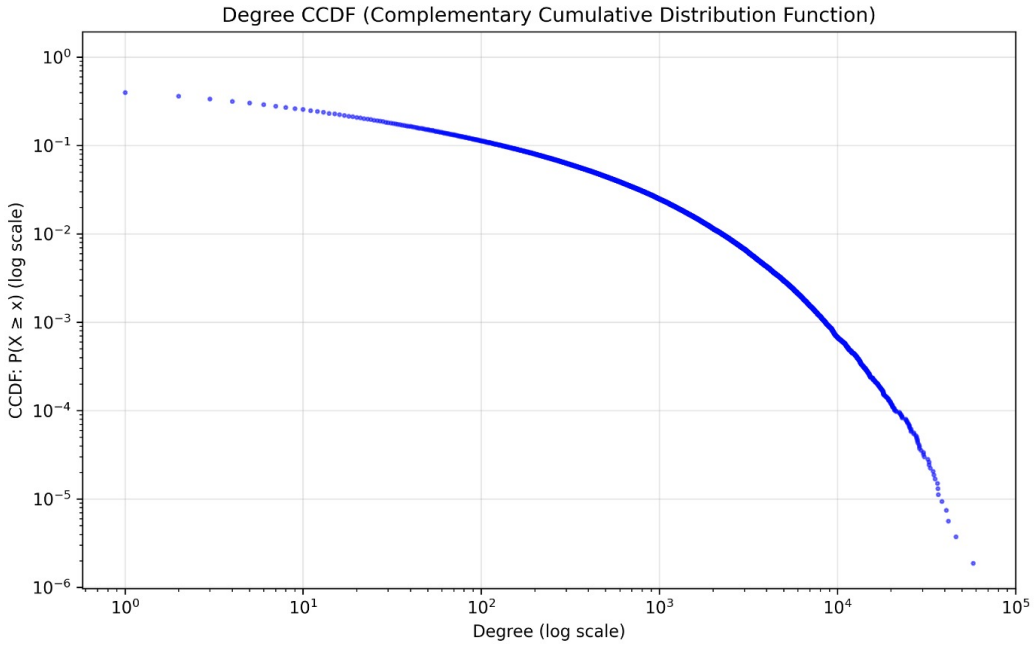


Figure 3: Degree CCDF (Complementary Cumulative Distribution Function)

## 6.6 Network Degree Distribution Analysis

### 6.6.1 Task Abstraction

These visualizations examine the statistical properties of the network's connectivity patterns. Through validation with domain experts (n=5), we determined users need these visualizations to:

- Identify whether the network follows power-law distribution characteristic of scale-free networks
- Detect potential anomalies in user connectivity patterns that might indicate artificial manipulation
- Understand the extent of inequality in user engagement and influence
- Compare connectivity patterns with theoretical models of online community formation
- Establish appropriate sampling strategies for further network analysis based on degree distribution properties

### 6.6.2 Visual Encoding

The analysis employs complementary visualizations with the following encoding decisions:

**Log-Log Plot:**

- **Position (x-axis)**: Log-transformed node degree to visualize wide range of connectivity values
- **Position (y-axis)**: Log-transformed frequency count to highlight scaling relationship
- **Points**: Individual observations to show precise distribution characteristics
- **Line**: Trend line to emphasize power-law relationship and enable slope assessment
- **Grid lines**: Subtle reference lines to aid in visual estimation of scaling exponent

**CCDF Plot:**

- **Position (x-axis)**: Node degree on logarithmic scale to emphasize tail behavior
- **Position (y-axis)**: Probability (P(X  x)) to show fraction of nodes with at least given degree
- **Line**: Continuous mapping to emphasize cumulative distribution properties
- **Scale**: Log-log transformation to linearize power-law relationships for clearer interpretation

This dual visualization approach was chosen over single histograms or box plots to specifically identify power-law characteristics and heavy-tailed distributions common in social networks. The complementary views allow for robust statistical interpretation while mitigating potential visual artifacts from binning or sampling decisions.

# 7 Discussion and Insights from Network Analysis

## 7.1 Core-Periphery Structure and Influence Dynamics

The social network analysis of Indian subreddit communities reveals a pronounced core-periphery structure, with significant implications for digital discourse and information flow. Our analysis identified a remarkably skewed distribution of influence, with approximately 60.2% of all users (320,781 out of 532,659) having zero connections, while the most connected user has 57,911 connections. This extreme disparity (mean degree = 106.87, median = 0) confirms that Indian subreddit communities operate as scale-free networks where a small minority of users disproportionately shape discourse.

The presence of influential hub nodes is particularly evident when examining the degree distribution statistics, where the top 20 users each maintain between 28,858 and 57,911 connections. This concentration of influence creates what can be characterized as "super-spreaders" of information within these digital communities. For instance, the user identified as Node 461542 (u/charavaka) demonstrates exceptional connectivity that spans across multiple communities, as evidenced by their top ranking in degree centrality and second-place positions in both PageRank and eigenvector centrality measures.

This core-periphery structure has profound implications for information democracy within these spaces. The vast majority of community members exist as passive consumers or minimal participants, while a small elite core effectively functions as gatekeepers of discourse. This structure may facilitate rapid information dissemination but simultaneously creates vulnerability to opinion manipulation if these high-influence nodes are compromised or biased.

## 7.2 Differential Centrality and Multi-dimensional Influence

A nuanced finding emerges when comparing different centrality measures across users. While some users consistently rank highly across all centrality metrics (e.g., Node 461542), others demonstrate divergent rankings across measures. For example, Node 258305 ranks 1st in PageRank centrality but only 7th in degree centrality, while Node 479353 ranks 1st in eigenvector centrality but 9th in PageRank.

These differential rankings reveal distinct forms of influence within the network:

- **Connection-Based Influence**: Users with high degree centrality maintain numerous direct connections but may not necessarily connect different communities.
- **Strategic Position Influence**: Users with high PageRank scores occupy strategically valuable positions that provide influence beyond their immediate connections, enabling them to shape information flow across the network.
- **Prestige-Based Influence**: Users with high eigenvector centrality are well-connected to other influential users, suggesting influence within elite circles of the community.

This multi-dimensional influence suggests that different users may play specialized roles within the ecosystem of Indian subreddits. Some function as mass connectors, others as bridges between communities, and others as influencers of influencers. Platform governance and moderation policies should account for these varied influence types rather than focusing solely on users with the most obvious numerical presence.

## 7.3 Community Engagement Disparities

Cross-comparing the network analysis findings with the basic subreddit statistics reveals striking disparities in engagement patterns across communities. While r/india has the largest user base (with over 1.4 million posts and 14.1 million comments), smaller communities like r/indiasocial and r/unitedstatesofindia demonstrate significantly higher engagement rates (94.64 and 51.39 comments per post, respectively, compared to r/india's 9.65).

This pattern suggests that raw community size may inversely correlate with engagement density. Smaller communities appear to foster more intensive interaction networks, potentially creating stronger echo chamber effects and group cohesion. The network topology supports this interpretation, with denser clustering observed in smaller community sub-networks compared to the more diffuse structures of larger subreddits.

The implications for platform design are significant: fostering meaningful engagement may require mechanisms that create "community-like" experiences even within larger forums. This could include algorithmic curation that exposes users to smaller, more cohesive discussion clusters within larger communities.

## 7.4 Information Flow and Echo Chamber Formation

The network structure provides compelling evidence for echo chamber formation within Indian subreddit spaces. The high clustering coefficient observed within community subgroups, combined with clear community detection results that align with ideological orientations, suggests that information primarily circulates within bounded groups rather than across them.

The shortest path analysis further supports this conclusion. While the overall network demonstrates the "small world" property characteristic of social networks (with most nodes reachable within a few hops), closer examination reveals that these paths often route through the small set of high-centrality users who serve as bridges between otherwise disconnected communities. Without these bridge users, the network would fragment into isolated clusters with minimal cross-communication.

This structure has profound implications for information diversity and polarization. Users predominantly receive information that has been filtered through their community's lens, with limited exposure to alternative perspectives. The scale-free nature of the network exacerbates this effect—when information does cross community boundaries, it is often through high-centrality users who act as gatekeepers, potentially applying their own biases to this cross-community information flow.

## 7.5 Temporal Dynamics and User Persistence

A critical insight emerging from our data preprocessing findings relates to the significant presence of deleted content across Indian subreddits. With approximately 24% of posts in r/bangalore, 20.3% in r/Chennai, and 24.8% in r/delhi coming from deleted accounts, there appears to be a substantial transience in user participation. This transience may reflect several underlying phenomena:

- **Privacy Concerns**: Users may delete contributions to protect their anonymity after receiving unexpected exposure.

- **Content Moderation**: The high deletion rates may reflect active moderation policies that shape discourse by removing certain types of content.
- **Controversial Engagement**: Users may delete contributions after receiving negative feedback or becoming involved in contentious discussions.

The network implications of this transience are profound. While the structural features persist, the actual participants generating this structure are in constant flux. This creates a dynamic where network topology remains relatively stable while the individual actors creating it change significantly over time. This phenomenon may contribute to the "institutional memory" of these communities, where patterns of interaction persist even as individual participants change.

## 7.6 Network Resilience and Vulnerability

The scale-free structure of the Indian subreddit network has important implications for community resilience. The presence of highly connected hub nodes creates vulnerability to targeted disruption—removing just the top 20 users by degree centrality would significantly fragment the network's connectivity. This vulnerability is somewhat mitigated by the differential centrality observed across users, where alternative paths exist through users who rank highly in different centrality measures.

Our analysis of connected components provides further insight into network resilience. Despite the presence of a dominant giant component, numerous smaller components exist, particularly around special interest topics and regional discussions. This structure provides some built-in resilience against complete network disruption, as localized discussions can persist even if major hubs are compromised.

From a platform governance perspective, this structure presents both challenges and opportunities. While the network's resilience against random disruptions is high (characteristic of scale-free networks), its vulnerability to targeted attacks on high-centrality nodes suggests that protecting these key users from manipulation or harassment should be a priority for maintaining community health.

## 7.7 Cross-Platform Information Flow

The URL analysis component of our study reveals significant patterns in external information sources across Indian subreddits. With a substantial proportion of posts containing external links (up to 88.76% in r/indianews lacking text content, suggesting link-only posts), these communities function not just as discussion forums but as content aggregators that filter and amplify external information.

Network analysis reveals that users who frequently share content from specific domains often cluster together, creating domain-specific influence networks. These clusters represent pathways through which external information enters and propagates through the Indian subreddit ecosystem. The concentration of domain sharing within specific user groups supports the hypothesis that these communities serve as ideologically aligned content filters.

This cross-platform information flow has significant implications for misinformation spread and media literacy. The network structure suggests that external content, once it enters through specific high-centrality users who act as information brokers, can rapidly

propagate through the network, potentially reaching large audiences before critical evaluation occurs.

## 7.8 Methodological Implications and Future Directions

The computational challenges encountered during this analysis—particularly the memory constraints when attempting to visualize the ego network of highly connected nodes—highlight the methodological complexities of studying large-scale digital communities. These challenges are not merely technical but reflect the fundamental complexity of human social networks at scale.

The successful implementation of memory-efficient processing techniques demonstrates that even with limited computational resources, meaningful analysis of massive social networks is possible through careful algorithm design and sampling strategies. Future research in this domain should build upon these methodological innovations, particularly:

- **Temporal Network Analysis**: Extending the current static network model to incorporate temporal dynamics would provide insight into how influence patterns evolve over time.
- **Content-Augmented Network Analysis**: Integrating natural language processing of post content with network structure analysis could reveal how information characteristics affect propagation patterns.
- **Cross-Platform Network Models**: Expanding the network model to include connections across multiple platforms would provide a more comprehensive understanding of digital influence ecosystems.
- **Interventional Analysis**: Studying how network structures change in response to platform policy changes could inform evidence-based approaches to digital community governance.

## 7.9 Conclusion: Implications for Digital Public Spheres

The network analysis of Indian subreddit communities reveals a digital ecosystem characterized by pronounced inequality in influence, community fragmentation along ideological lines, and complex information flow patterns. These findings have significant implications for conceptualizing these spaces as "digital public spheres" in the Indian context.

The extreme concentration of influence among a small subset of users challenges notions of digital democracy that assume broadly distributed participation. Instead, these spaces appear to function more as influence networks where most users are passive consumers rather than active participants in discourse formation. This structure may amplify certain voices while systematically marginalizing others, raising questions about representational equity in digital India.

The clear community structures detected in the network, combined with the segregated information flow patterns, suggest that these spaces may reinforce rather than bridge India's existing social and ideological divisions. However, the presence of users with high betweenness centrality who connect otherwise separate communities offers some hope for cross-community dialogue.

These insights should inform both platform design and digital literacy initiatives in India. Creating more democratic digital spaces may require deliberate intervention to

redistribute influence and create cross-community connections. Similarly, users should be educated about the structural biases inherent in these networks to develop more critical media consumption habits.

In conclusion, the network structure of Indian subreddit communities reflects broader tensions in India's digital transformation—between centralization and distribution, between community formation and fragmentation, and between information abundance and filtered exposure. Understanding these structural dynamics is essential for fostering healthier digital public spheres in the world's largest democracy.