Social Media Analytics - Group Assignment

Krishan Gupta, Lakshya Agarwal, Om Sangwan, Nandani Yadav,
Yash Joshi, Yiyi Yang

28th March, 2024

Section 2

Overview

In Section 2 of our analysis, we focus on the practical application of social network analytics to identify influencers within a specific online community. This section builds on the theoretical foundations discussed in class and applies methodologies to real-world data from Reddit. Our goal is to leverage network analysis techniques to pinpoint key individuals in the subreddit "IndiaInvestments" who wield significant influence over discussions and user engagement.

Methodology

Data Collection and Preparation

We collected data on submissions and comments from the "IndiaInvestments" subreddit, ensuring to exclude any entries by deleted users to maintain the integrity of our analysis. The collected data were then transformed to highlight the relationships between authors, submissions, and comments, setting the stage for network construction.

Network Construction

Using the NetworkX library, we constructed a directed graph representing the interaction dynamics within the subreddit. Nodes in the graph correspond to users (authors of comments and submissions), while edges represent the direction of communication (e.g., a comment on a post). This graph served as the basis for our subsequent analysis.

Centrality Measures

To understand the influence dynamics within the network, we calculated three centrality measures for each node:

- **Degree Centrality**: Reflects the number of connections each node has, indicating general activity and visibility within the network. For a graph $G = (V, E)$ with $|V|$ vertices, the degree centrality for a vertex v is defined as the fraction of nodes it is connected to:

$$C_D(v) = \frac{deg(v)}{|V| - 1}$$

  where $deg(v)$ is the degree of vertex v, and $|V| - 1$ accounts for v itself not being included in its own degree count. This can be calculated in NetworkX using the `degree_centrality(G)` function.

- **Betweenness Centrality**: Measures the extent to which a node lies on the shortest path between other nodes, highlighting those who play a crucial role in information flow. It quantifies the number of times a node acts as a bridge along the shortest path between two other nodes and is given by:

$$C_B(v) = \sum_{s,t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

  where σ_{st} is the total number of shortest paths from node s to node t and $\sigma_{st}(v)$ is the number of those paths that pass through v. NetworkX provides the `betweenness_centrality(G)` function for this purpose.

- **Closeness Centrality**: Captures how close a node is to all other nodes in the network, reflecting the ability to quickly interact with the entire community. It is calculated as the inverse of the average shortest path length from a given node to all other reachable nodes:

$$C_C(v) = \frac{|V| - 1}{\sum_{u \in V \setminus \{v\}} d(v, u)}$$

  where $d(v, u)$ is the shortest-path distance between v and u, and $|V| - 1$ is the normalization factor. The `closeness_centrality(G)` function in NetworkX can be used to calculate this measure.

Identifying the Most Influential Nodes

The most influential nodes in a network might be identified through a composite score derived from various centrality measures. For instance, one might average the standardized scores of degree, betweenness, and closeness centralities for each node:

$$Score(v) = \alpha \cdot C_D(v) + \beta \cdot C_B(v) + \gamma \cdot C_C(v)$$

where α, β, and γ are weights that reflect the relative importance of each centrality measure in determining influence. The nodes with the highest composite scores are then considered the most influential. Adjusting the weights allows for emphasizing different aspects of influence, such as network position or connectivity.

Results and Visualization

Our analysis identified key influencers within the "IndiaInvestments" subreddit, with "AutoModerator," "crimelabs786," and "srinivesh" emerging as the top figures based on a combination of centrality measures and user engagement metrics. These users not only engage frequently with the community through submissions and comments but also hold strategic positions within the network that facilitate widespread information dissemination.

To visually represent the influence dynamics within the subreddit, we generated a plot showcasing the top 100 most influential users based on our composite score, which incorporates centrality measures and user activity data. This visualization provides a clear depiction of the network's structure, highlighting the central role played by the top influencers in connecting various subgroups within the community.

The plot below illustrates the interconnectedness of users within the "IndiaInvestments" subreddit, with node sizes reflecting the degree of influence. Larger nodes represent users with higher centrality scores and user engagement, indicating their pivotal role in the network.

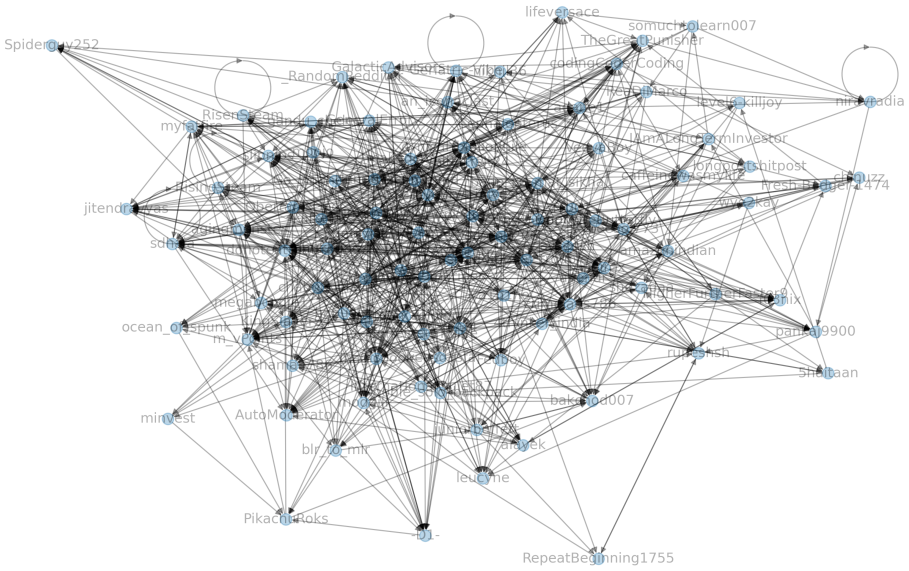

Figure 1: Top 100 Influential Users in the "IndiaInvestments" Subreddit

Implications for Business and Research

Understanding the roles these influencers play can be valuable for businesses looking to engage with communities or for researchers studying information dissemination in social networks. By targeting influencers for outreach or monitoring their interactions for insights, stakeholders can more effectively navigate and leverage the dynamics of online communities.

Conclusion

The application of social network analysis techniques to Reddit data demonstrates the power of centrality measures and PCA in uncovering the underlying structure of online communities and