Project Proposal
ID2211 - Data Mining

Andrei Iliescu, Leandro Duarte, Miguel Arroyo Marquez, Mingyang Chen

Project Title: Temporal Evolution of Political Communities on Reddit: A Graph-Based Analysis
Problem Statement: We aim to understand how political communities on Reddit form, change, and dissolve over time (2008-2019). We want to see if big political events like elections cause these communities to change their structure. We will track how groups of political subreddits cluster together based on shared users, and measure if these clusters become more separate (polarized) over time. This analysis will help us understand how online political discussion spaces react to real-world events.
Data: We will use the Reddit Politosphere dataset [1], a large-scale resource of online political discourse covering more than 600 political subreddits over a 12-year period (2008-2019). This dataset contains two primary types of information for each year:

1. **Network data files** (networks_YYYY.csv): These files contain weighted and unweighted network representations where nodes are subreddits and edges represent user overlap between them. Edge weights correspond to the number of users who posted at least 10 comments in both subreddits. The unweighted networks are derived using statistical network backboning techniques to filter out noise while preserving significant connections.
2. **Comment data files** (comments_YYYY-MM.bz2): These contain all comments posted in the political subreddits along with metadata such as creation time, author (pseudonymized), and other attributes. The dataset includes over 288 million comments across all years.

The dataset also provides valuable metadata for both subreddits and users. Subreddit metadata includes information about banned status, political affiliation (democratic/republican), and regional focus. User metadata, while preserving anonymity through pseudonymization, contains information about user types (e.g., bots, automoderators) and certain linguistic patterns in usernames.
Work Plan: Our work will progress through these main steps:

1. **Data preparation**: Load the yearly network files and preprocess them to ensure they are ready for analysis.
2. **Community detection**: Apply algorithms to find clusters of related subreddits for each year.
3. **Temporal analysis**: Track how identified communities change from year to year.
4. **Event correlation**: Create a timeline of major political events and check if they match with community changes.
5. **Analysis and visualization**: Calculate metrics about community structure and create visualizations showing how communities evolved.
6. **Report preparation**: Document our methods, results, and conclusions.

Methodology: We will implement multiple community detection algorithms from the course to analyze the networks:

1. **Label Propagation Algorithm**: A simple algorithm where nodes adopt the most common label among their neighbors. It's fast and scales well to large networks.
2. **Spectral Clustering**: We will construct the graph Laplacian matrix, find its eigenvalues and eigenvectors, and use the eigengap heuristic to determine the optimal number of communities.
3. **Louvain Method**: This algorithm tries to maximize modularity by iteratively moving nodes between communities.

For tracking communities over time, we will calculate the Jaccard similarity index between communities in consecutive years (e.g., 2008 and 2009). This tells us how much overlap exists between communities across years. We will say a community "continues" if its overlap with a community in the next year is above a certain threshold.

Evaluation: We will evaluate our analysis in several ways:

1. **Community quality**: We will use modularity (Q) scores to measure how well-defined our detected communities are. Higher modularity means better separation between communities.
2. **Temporal correlation**: We will check if significant changes in community structure happen around major political events like elections. We will measure the time distance between events and community shifts.
3. **Validation with metadata**: We will compare our detected communities with the known political affiliations of subreddits (from metadata) to see if our algorithms correctly group similar subreddits.
4. **Algorithm comparison**: We will compare results from different community detection methods to ensure our findings are not just artifacts of one algorithm.

Expected Outcomes: By the end of the project, we expect to deliver:

1. An analysis of how political communities on Reddit evolved over the 12-year period.
2. Evidence of whether political events cause measurable changes in online community structure.
3. Measurements of political polarization trends over time (whether communities became more or less separate).
4. Visualizations showing the evolution of communities and their responses to events.
5. Comparison of different community detection algorithms in their ability to track political communities.
6. Code that implements our methodology so others can reproduce our results.

References

[1] V. Hofmann, H. Schütze, and J. B. Pierrehumbert, “The reddit politosphere: A large-scale text and network resource of online political discourse,” *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 16, no. 1, pp. 1259–1267, May 2022. doi: 10.1609/icwsm.v16i1.19377. [Online]. Available: https://ojs.aaai.org/index.php/ICWSM/article/view/19377