Temporal Evolution of Political Communities on Reddit: A Graph-Based Analysis

Andrei Iliescu
KTH Royal Institute of Technology
iliescu@kth.se

Leandro Duarte
KTH Royal Institute of Technology
leandrod@kth.se

Miguel Arroyo Marquez
KTH Royal Institute of Technology
miguelam@kth.se

Mingyang Chen
KTH Royal Institute of Technology
minchen@kth.se

Abstract

We investigate real-world political shockwaves by leveraging the 'Reddit Politosphere' dataset, a collection of over 600 political subreddits, their interaction networks, and user comments from 2008 to 2019. Our analysis reveals that major U.S. events, particularly elections and pivotal national incidents, trigger distinct shifts in Reddit's topical discussions, community structures, and expressed opinions. Through network visualizations and by analyzing evolving community structures, we map these online responses. These findings demonstrate the value of graph-based methods for mapping large-scale political discourse online and provide new insights into the dynamics of polarization, radicalization, and user migration within digital political ecosystems.

1 Introduction

1.1 Motivation

Social media platforms have become key areas for public deliberation, where everyday users, political elites, journalists, and automated actors continuously co-produce a running commentary on current events. Though these discussions do not happen in a vacuum, there is a reciprocal relationship between online talk and real-life political discourse [2].

Understanding this dual role is especially urgent for political information ecosystems. Online political talk feeds into offline outcomes—voting, protest, policy support—and vice-versa. At the same time, the affordances of different sites create distinct conversational ecologies.

Reddit is an especially revealing case: it aggregates more than a decade of timestamped conversations in thousands of topical "subreddits," yet it also allows users to roam freely across communities under a single pseudonymous identity.

This combination of persistent identity and voluntary association means that we can watch political communities form, splinter, and re-assemble at a resolution that surveys or news coverage cannot approach.

With this ground truth, we aim to use data mining techniques to assess the evolution of political communities through topological analysis across multiple election cycles. Our goal is to identify structural shifts in community formation, detect patterns of radicalization or the emergence of echo chambers and trace user migration. By doing so, we hope to uncover how large-scale political discourse adapts to offline shocks and contributes to broader trends in polarization and democratic engagement.

1.2 Problem Definition

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.

2 Related Work

Researchers have extensively investigated political communities on Reddit to identify patterns of polarization and hostility. Efstratiou et al. (2023) examined the interactions within political echo chambers on Reddit, discovering that hostility frequently arises within politically similar groups rather than between opposing factions. Their findings challenge traditional perspectives on polarization by highlighting complexities at the user interaction level rather than broadly defined group interactions [3].

Additionally, Guimarães et al. (2019) systematically studied Reddit's political discussions, categorizing conversations into four distinct types: harmonies, discrepancies, disruptions, and disputes. By incorporating sentiment and topic variations, their research illustrates nuanced interaction patterns beyond simple agreement or disagreement, providing a deeper understanding of online political discourse [4].

In another relevant study, Soliman et al. (2019) analyzed political communities on Reddit by examining differences between left-leaning and right-leaning subreddits. Their findings highlight significant distinctions in content sharing, cross-community engagement, and patterns of attention distribution. Notably, right-leaning communities were found to utilize more derogatory language and exhibit stronger interconnections compared to left-leaning communities, emphasizing the distinct nature of online political behavior [8].

Furthermore, statistical analyses of Reddit's network data have provided critical insights into structural aspects of these online communities. Metrics such as clustering coefficients, modularity, and transitivity have been effectively utilized to quantify community cohesion, polarization, and overall network stability. Such structural metrics are essential in understanding the evolution of online political communities, especially in response to major political events [5].

Methodologically, previous studies commonly apply algorithms including label propagation, spectral clustering, and the Louvain method to identify communities and analyze their structures. Techniques such as PageRank have also been employed to determine influential nodes and clusters within these political networks. Temporal analyses, leveraging measures like Jaccard similarity, are instrumental in detecting shifts and realignments within online communities around critical political events [8].

Our study builds upon these foundational insights by employing advanced network analyses, including attribute assortativity and small-world indices. This approach aims to offer novel perspectives on the temporal dynamics of Reddit's political communities. Through these methodologies, we seek to measure not only polarization but also community volatility and realignment, enhancing the broader understanding of how online political communities adapt to both external political events and internal discourse shifts.

3 Background

To effectively analyze the temporal evolution and structural dynamics of political communities on Reddit, we utilize several mathematical and algorithmic frameworks essential to network analysis and clustering. Below we describe these concepts and algorithms formally:

3.1 Giant Connected Component (GCC)

In graph theory, the giant connected component is the largest connected subgraph within a graph. Formally, a graph $G = (V, E)$ comprises nodes V and edges E, and the giant connected component is a subgraph $G' = (V', E')$ with $V' \subseteq V$ and $E' \subseteq E$, where every vertex in V' can reach any other vertex in V' via some path, and $|V'|$ is maximized.

3.2 Community Detection Algorithms

Label Propagation The Label Propagation Algorithm (LPA) is a semi-supervised machine learning algorithm used for community detection. Each node initially carries a unique label, which iteratively updates by adopting the most frequent label among its neighbors until labels stabilize. The updating rule can be expressed as:

$$x_i = \arg\max_{l \in L} \sum_{j \in N(i)} \delta(x_j, l)$$

where x_i is the label of node i, $N(i)$ is the neighborhood of node i, L is the set of labels, and δ is the Kronecker delta function.

Louvain Method The Louvain method optimizes the modularity Q, defined as:

$$Q = \frac{1}{2m} \sum_{ij} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j)$$

where A_{ij} represents the edge weight between nodes i and j, k_i and k_j represent node degrees, m is the sum of all edge weights, and c_i, c_j are community assignments. This method maximizes modularity by aggregating nodes iteratively into larger communities until no further modularity improvement is possible.

3.3 Jaccard Overlap

The Jaccard similarity index measures similarity between two sets A and B:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

This metric is useful for quantifying overlap and stability of communities across time periods.

3.4 Graph Embeddings (node2vec)

Node2vec creates continuous, vector-space embeddings for nodes by exploring neighborhoods through biased random walks. The embedding $f : V \to \mathbb{R}^d$ maximizes the probability of preserving network neighborhoods, formulated as:

$$\max_f \sum_{u \in V} \log P(N_S(u) | f(u))$$

where $N_S(u)$ is the network neighborhood of node u.

Together, these mathematical and algorithmic foundations enable us to robustly characterize the evolving nature of political discourse and community dynamics on Reddit.

4 Dataset and Initial Findings

4.1 Dataset used

We use the Reddit Politosphere dataset [6], which covers more than 600 political subreddits over a 12-year period (2008-2019). The dataset contains:

- Network data files
- Comments per user, per subreddit, per month for each year
- User metadata files

4.1.1 Network files

The network files are 3: a weighted graph, an unweighted graph and a networks metadata file.

Weighted Graph:
The weighted graph was constructed based on user comments. Pairs of subreddits were connected if they shared users who had posted at least 10 comments in each of the two subreddits.

Unweighted Graph:
The unweighted graph is based off filtering the weighted graph. This was done using the noise-corrected backbone method [1], which accounts for expected overlap based on subreddit size and retains only statistically significant edges.

The optimal significance threshold was determined using the Kneedle algorithm[7]; this ensured a balance between retained edges and nodes.

Node 1	Node 2	Weighted	Unweighted
Anarchism	Economics	20	0
Economy	Ronpaul	2	0
...	...	...	...

Table 1: Head of the network metadata file (2008)

Various network metrics were then computed, including average node degree, average shortest path length, and network density. Community structure was analyzed using the Louvain method to compute modularity Q Q; values greater than 0.3 across all years indicate a high degree of fragmentation and potential polarization.

4.1.2 User Metadata files

The user metadata is a singular JSON file, which includes an anonymized identifier that relates to the user comments.

Author	automoderator	bot	gender	angry	anti	...	trump
"7W7I3"	0	0	0	0	0	...	0
"IZ5YE"	0	0	0	0	0	...	0

Table 2: Head of the user metadata file

As seen by the table there is a lot of binary features. These features were created by using a sentiment analysis/regex evaluation on the username. For example, if there was any clear military denominations such as "c[a]pt", "sgt" the feature "military" would be marked as positive "1". The goal is to still retain any pertinent information from the user name while keeping the user non-identifiable from the dataset.

4.2 Initial findings from the dataset

We started with a preliminary analysis of the already constructed graphs before running any sort of data mining algorithms.

First we focused on how the graph has been growing in size through out the years.

The instability of subreddit communities becomes particularly evident during politically sensitive periods, when existing communities may fragment into smaller, more ideologically focused subgroups, and entirely new subreddits may emerge in response to evolving public discourse.

So we filtered subreddits which have not survived at least 5 years. After the filtering we were left with 131 nodes/subreddits from original 600.

Number of subreddits present in at least 5 consecutive years: 131

year	nodes	edges	total_weight
2008	9	34	2081
2009	14	79	5191
2010	25	209	8140
2011	56	788	23066
2012	86	1769	44011
2013	110	2588	46098
2014	135	3386	43288
2015	170	4488	61122
2016	256	9992	252137
2017	300	13913	281783
2018	321	15716	270496
2019	416	25500	406020

Table 3: Yearly summary of network size and total edge weight

	Total
Unique subreddits (nodes)	600
Unique interactions (edges)	39863

Table 4: Overall network summary across all years

This lead to a retention of about 33% of the original edges and a retention of $\sim$22% in the number of nodes.

The goal was to remove those spontaneous subreddits for this initial evaluation. For further analysis, considering that the loss is around 73% in edges and nodes, we will try a tighter 3-year window further down the line.

Within this filter we then tried to list the most "popular" subreddits by using the summation of its weights with other nodes.

	Shape
Original dataset:	78462, 5
Filtered dataset:	25954, 5

Table 5: Subreddits present in at least 5 consecutive years

Subreddit	Total Connection Weight
politics	10500
worldnews	9800
news	8700
SandersForPresident	8500
Conservative	8300
Ask_Politics	8200
The_Donald	8000
politicalhumor	7900
LateStageCapitalism	7700
neoliberal	7500

Table 6: Top 10 subreddits by total connection weight

In table 6 we see the obvious subreddits — those which have been on the site the longest: `r/politics`, `r/worldnews`, ... taking the top spots.

What did catch our attention was the strength of `r/The_Donald`, a relatively young subreddit in comparison to the rest. This includes `r/SandersForPresident`, which was created three years before `r/The_Donald`. Clearly indicating the gravity of 2016 elections for this website and the online political landscape. Another thing that stood out was the one-sidedness, considering that `r/HillaryClinton` is not visible in this top 10 list.

We then plotted weights over time for each popular subreddit. You can see this in Figure 1

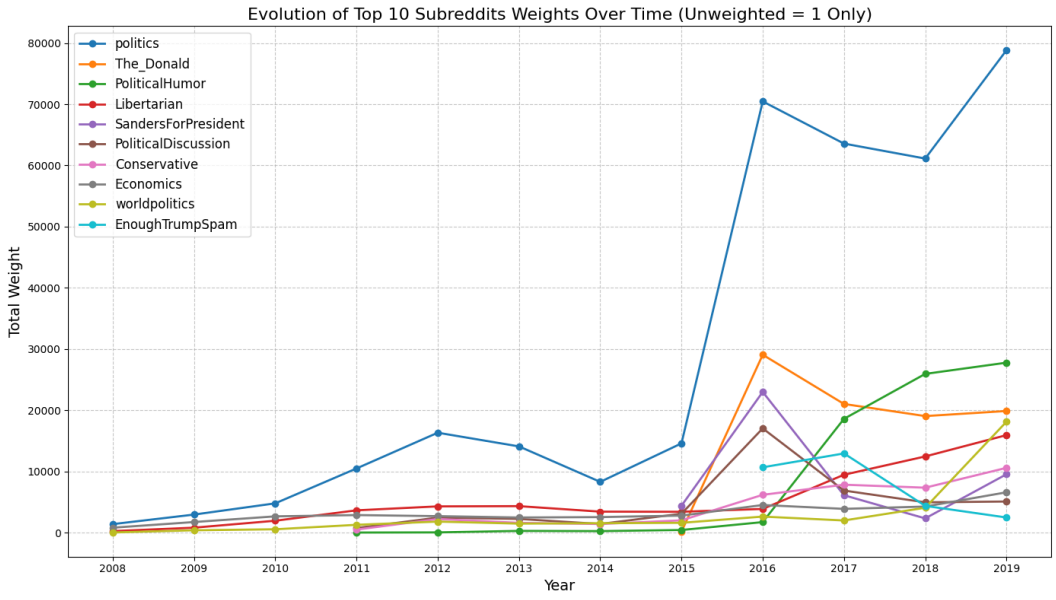


Figure 1: Evolution of weights over time

Here we can see a clear spike within the 2016 election cycle, which confirms the previously mentioned importance of this time for online political discussion. Although, there is an argument to be made about automated actors (bots) hitting the scene. Possibly explaining part of the spike. Furthermore, focusing on the graph we can see that `r/worldpolitics` doesn't really spike up, rather grows at a more sustained pace post american elections. Possibly a reaction to the elected candidate. Additionally, another subreddit emerges: `r/EnoughTrumpSpam`. Its very existence carries a certain irony, as it protests the abundance of Trump-related content — something arguably inevitable following his election as president. Eventhough, may be indicative of something larger that could possibly picked up with certain analysis.

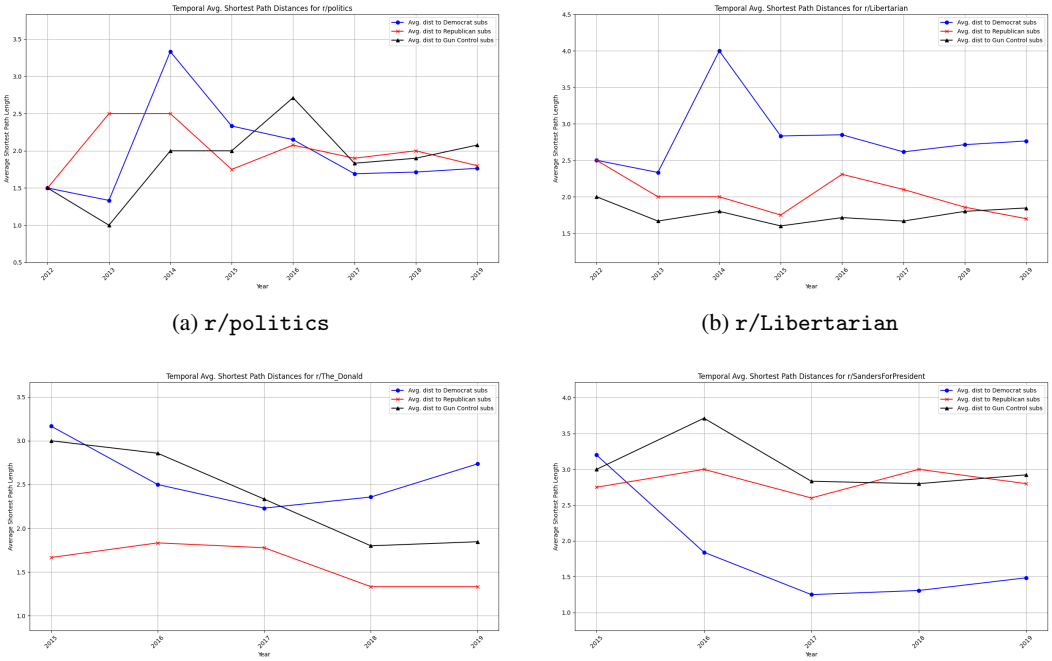
This preliminary evaluation and brainstorming will prepare us for the evaluation of the topologies discussed further in this paper.

5 Experiments

5.1 Temporal Proximity Analysis via Shortest Paths

To gauge the shifting allegiances and topical focus within the Reddit politosphere, we analyzed the temporal evolution of "distances" between key subreddits and distinct ideological/topical groups. We operationalized distance as the average shortest path length in yearly unweighted interaction graphs. The shortest path length, representing the minimum number of subreddit 'hops' connecting two communities, was computed using NetworkX. For unweighted graphs, as used here, NetworkX employs an efficient algorithm based on Breadth-First Search (BFS), typically a bidirectional BFS that explores from both the source and target nodes simultaneously. This path length serves as a proxy for their relational proximity.

We tracked the average shortest path from influential subreddits — `r/politics`, `r/Libertarian`, `r/The_Donald`, and `r/SandersForPresident` — to three curated target groups: Democrat-affiliated, Republican-affiliated, and Gun-Control-related subreddits. The analysis spans from 2012 (or 2015 for newer subreddits like `r/The_Donald` and `r/SandersForPresident`) to 2019.



(a) `r/politics` (b) `r/Libertarian`

(c) `r/The_Donald` (d) `r/SandersForPresident`

Figure 2: Average shortest path distances from main subreddits to Democrat-affiliated (blue), Republican-affiliated (red), and Gun-Control-related (black) subreddits. Lower values indicate closer proximity. Key events: 2012 (Obama re-elected, Sandy Hook shooting), 2015-2016 (Trump's rise, Pulse/San Bernardino shootings), 2017 (Las Vegas shooting), 2019 (El Paso/Dayton shootings, Trump impeachment).

The trajectories reveal distinct community alignments and reactions to major socio-political events:

- `r/politics`: Shows dynamic proximity. Notably, it neared Democrat-affiliated and Gun Control subreddits around 2012-2013, coinciding with Obama's re-election and heightened gun control discourse post-Sandy Hook. A shift towards Republican-affiliated subreddits is visible approaching the 2016 election.
- `r/Libertarian`: Consistently maintains closer ties to Republican-affiliated and Gun Control subreddits, reflecting ideological alignment.
- `r/The_Donald`: Unsurprisingly, demonstrates strong and persistent proximity to Republican-affiliated subreddits throughout its existence.
- `r/SandersForPresident`: Remains closely aligned with Democrat-affiliated subreddits, mirroring its political orientation during its active period.

This shortest path analysis provides a quantitative lens on how online political spheres reconfigure, tighten, or distance themselves in response to the real-world political landscape, offering insights into polarization and topical gravity. Additionally, we start seeing patterns in spikes and dips from this calculated average, showing heightened discussion of these topics. Notably, there is a drop within the path lengths for Gun Control around 2013, which coincides with the Sandy Hook shooting. We will use these findings

5.2 Analysis of Topology

Following our analysis, we found some interesting topological correlations between the different metadata properties. Keeping with the gun control example; in 2016 the subreddits labeled as related to gun control were topologically closer to the Republican subreddits. This is visible in Figure 3. This proximity may reflect the polarized nature of the gun control debate during the 2016 election cycle, when the issue was a major political flashpoint. While the Democratic Party generally

advocated for stricter gun regulations, many online discussions around gun rights—especially those opposing regulatory measures—were often co-opted or dominated by Republican-aligned rhetoric. As a result, even subreddits discussing gun control may have attracted participants with conservative viewpoints or became arenas for debate that leaned rightward. This clustering suggests that ideological boundaries on Reddit were not strictly issue-based, but instead reflected broader political alignments and echo chambers that influenced how topics were discussed and framed within different subreddit communities.

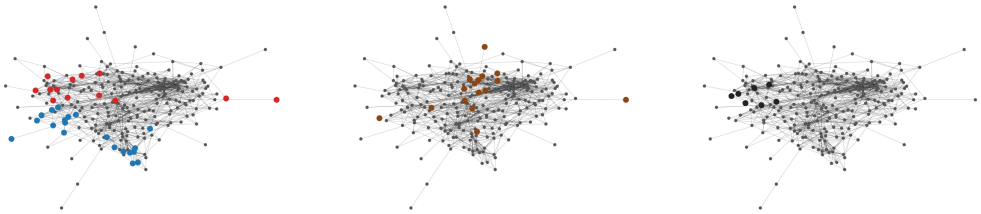


Figure 3: Network of subreddits in 2016: the left graph is democratic (blue) and republican (red), the middle graph is radical subreddits that got banned, and the right graph is subreddits related to gun control.

6 Conclusions and Next Steps

Our journey through the Reddit politosphere, as detailed in this report, has already yielded compelling insights into the platform's evolving political communities. Preliminary analyses clearly demonstrate temporal patterns in subreddit interactions that correlate strongly with major real-world political events, particularly U.S. election cycles and significant incidents like the Sandy Hook and Pulse shootings. We've navigated the initial phases of dataset discovery and processing, establishing an analytical foundation that allows us to define and track influential subreddits based on their network properties.

Our first experiment, focusing on temporal shortest path analysis, confirms our ability to detect the resonance of impactful offline events within online discussions; the data reflects topical focus around critical junctures such as 2013 and 2016. Complementing this, visualizations of yearly networks provided confirmation of subreddit clustering around shared topics, alongside polarization, with Democrat and Republican affiliated communities often forming distinct, relatively isolated clusters.

Currently, we are applying community detection experiments, including label propagation, the Louvain method, and spectral clustering. Initial runs suggest that the Louvain method offers the most promising path for clustering nodes within these complex yearly interaction networks. However, our work with label propagation indicates a need to refine the initial labeling strategy to better achieve our goal of categorizing all subreddits into meaningful topical or ideological groups such as Democrat, Republican, Gun Control, or Radical.

Looking ahead, our next steps will involve a more granular analysis. We plan to zero in on specific, highly impactful years like 2013 and 2016, selecting subreddit clusters, identified primarily via the Louvain method, that appear most responsive to the major events of those periods. With these targeted subreddit-year combinations, we will pivot to analyzing their textual content. By applying simple topic modeling techniques to the comment data from these specific subreddits during relevant months, we aim to confirm that their discussions indeed centered on the events we identified as critical. Following topical confirmation, a sentiment analysis model will be employed to gauge the collective reaction and emotional tone within these communities concerning those events. This focused approach will allow us to test concrete hypotheses, such as how mass shootings ignite polarized discussions on gun control, or how public opinion within these online enclaves shifts towards presidential candidates like Obama and Trump around their respective election periods.

This focused analysis will deliver concrete insights into how Reddit communities evolve and their sentiment is shaped.

References

[1] Michele Coscia and Frank MH Neffke. Network backboning with noisy data. In *2017 IEEE 33rd international conference on data engineering (ICDE)*, pages 425–436. IEEE, 2017.

[2] Michela Del Vicario, Sabrina Gaito, Walter Quattrociocchi, Matteo Zignani, and Fabiana Zollo. Public discourse and news consumption on online social media: A quantitative, cross-platform analysis of the italian referendum. *arXiv preprint arXiv:1702.06016*, 2017.

[3] Alexandros Efstratiou, Jeremy Blackburn, Tristan Caulfield, Gianluca Stringhini, Savvas Zannettou, and Emiliano De Cristofaro. Non-polar opposites: Analyzing the relationship between echo chambers and hostile intergroup interactions on reddit. *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, 17:197–208, 2023.

[4] Anna Guimarães, Oana Balalau, Erisa Terolli, and Gerhard Weikum. Analyzing the traits and anomalies of political discussions on reddit. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, pages 205–213, 2019.

[5] Amira Haji, Benjamin Rosenbaum, Jordan Hartmann, and David Aldous. A statistical analysis of network data from reddit, 2017. Available at: https://www.causeweb.org/usproc/sites/default/files/usresp/2017-1/statistical-analysis-network.pdf.

[6] Valentin Hofmann, Hinrich Schütze, and Janet Pierrehumbert. The Reddit Politosphere: A large-scale text and network resource of online political discourse. In *Proceedings of the International AAAI Conference on Web and Social Media 16*, 2022.

[7] Ville Satopaa, Jeannie Albrecht, David Irwin, and Barath Raghavan. Finding a" kneedle" in a haystack: Detecting knee points in system behavior. In *2011 31st international conference on distributed computing systems workshops*, pages 166–171. IEEE, 2011.

[8] Ahmed Soliman, Jan Hafer, and Florian Lemmerich. A characterization of political communities on reddit. In *Proceedings of the 30th ACM Conference on Hypertext and Social Media*, HT '19, page 259–263, 2019.