Graph Motif Generation for Social Media Platforms Using Topological Data Analysis

Ryan Clement
Halicioğlu Data Science Institute
University of California San Diego
rrclement@ucsd.edu

1 Introduction

Social media platforms have become central to information consumption and public discourse. However, these platforms are increasingly criticized for fostering ideological echo chambers and radicalization pipelines via algorithm-driven content recommendations. These phenomena—commonly referred to as "rabbit holes" or "pipelines"—occur when users are funneled into content spheres that reinforce existing beliefs while isolating them from alternative viewpoints. This project seeks to investigate the structural patterns within social media interaction graphs, particularly Reddit, that contribute to these phenomena. Leveraging graph motif analysis and topological data analysis (TDA), the project aims to identify the subgraph patterns and higher-order topological features that underlie content hubs, echo chambers, and radicalization pipelines.

Key Research Questions:

- Are there specific graph motifs that serve as content hubs or bottlenecks?
- Can we identify subgraphs that signify epistemic echo chambers?
- What topological signatures characterize radicalization pathways?

2 Objectives

The project has three main objectives:

- Graph Motif Detection: Identify statistically significant subgraph patterns (motifs) in Reddit's interaction graph that correspond to echo chambers or radicalization pipelines.
- Topological Analysis: Apply TDA methods—such as persistent homology—to uncover higher-order structures invisible to classical graph analysis.
- Visualization: Create intuitive visualizations of motifs and topological features to better understand their network roles.

3 Dataset Description

The analysis is based on the Reddit Hyperlink Network dataset. Each node is a subreddit, and each directed edge represents a hyperlink from one subreddit to another (i.e., when a post in one subreddit links to a post in another).

- Node: Subreddit name
- Edge: Directed, from source to target subreddit
- Edge attributes:

  - Timestamp
  - Weight (number of hyperlinks/interactions)

This dataset enables the construction of a multi-layered, directed graph reflecting the flow of information and community interconnections on Reddit.

4 Methodology

4.1 Data Preprocessing

The Reddit Hyperlink Network is loaded and processed in large chunks to construct a weighted, directed graph using NetworkX:

```
import pandas as pd
import networkx as nx

# Read in chunks and build graph
for chunk in pd.read_csv('soc-redditHyperlinks-body.tsv', sep='\t',
    chunksize=100_000):
    for _, row in chunk.iterrows():
        ...
```

Listing 1: Sample Python code to load data

- Edge weights represent connection strength (frequency of hyperlinks).
- The final graph is serialized for efficient reuse.

Graph Summary

- **Number of nodes:** 35,776
- **Number of edges:** 137,821
- **Average in-degree:** 3.85
- **Average out-degree:** 3.85
- **Density:** 0.000108
- **Edge weight statistics:**
  - Minimum: 1
  - Maximum: 548
  - Mean: 2.08
  - Median: 1.0

Top 5 nodes by in-degree:

Subreddit	In-degree
askreddit	2161
iama	1646
pics	953
videos	879
todayilearned	816

Top 5 nodes by out-degree:

Subreddit	Out-degree
subredditdrama	1350
copypasta	636
drama	600
subredditoftheday	559
outoftheloop	507

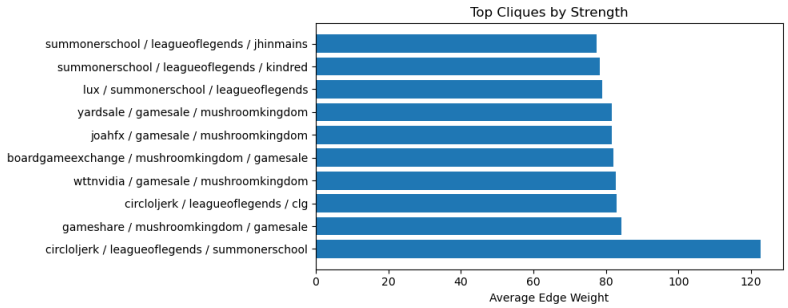

Figure 1: Enter Caption

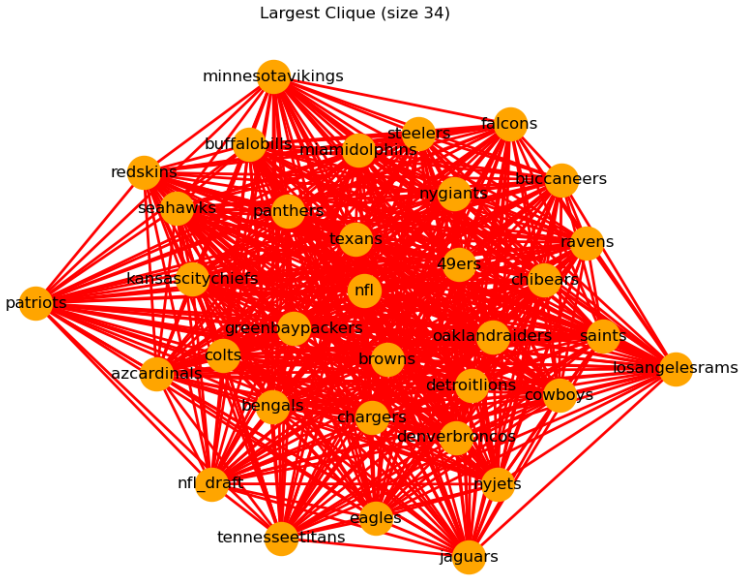

Figure 2: Enter Caption

4.2 Graph Motif Analysis

Graph motif analysis aims to find recurring small subgraphs (motifs) that occur more frequently than expected by chance. These motifs can represent functional "units" of network structure.

- Closed loops/cycles (potential echo chambers)
- Directed paths (potential radicalization pipelines)
- Cliques (tightly-knit groups)

Motif statistics and clique analysis were conducted using NetworkX functions, along with custom analyses for edge weights and in/out flows.

4.3 Persistent Homology

Persistent homology tracks the appearance and disappearance ("birth" and "death") of features as the network is filtered by edge weight or distance. Using ripser and gudhi, persistence diagrams were computed on sampled subgraphs.

- H_0 (Connected components): Measures fragmentation (community structure)
- H_1 (Cycles/loops): Measures feedback, echo chamber potential

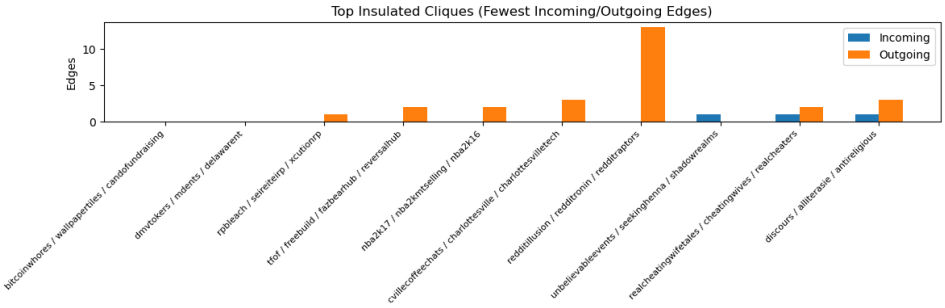

Figure 3: Enter Caption

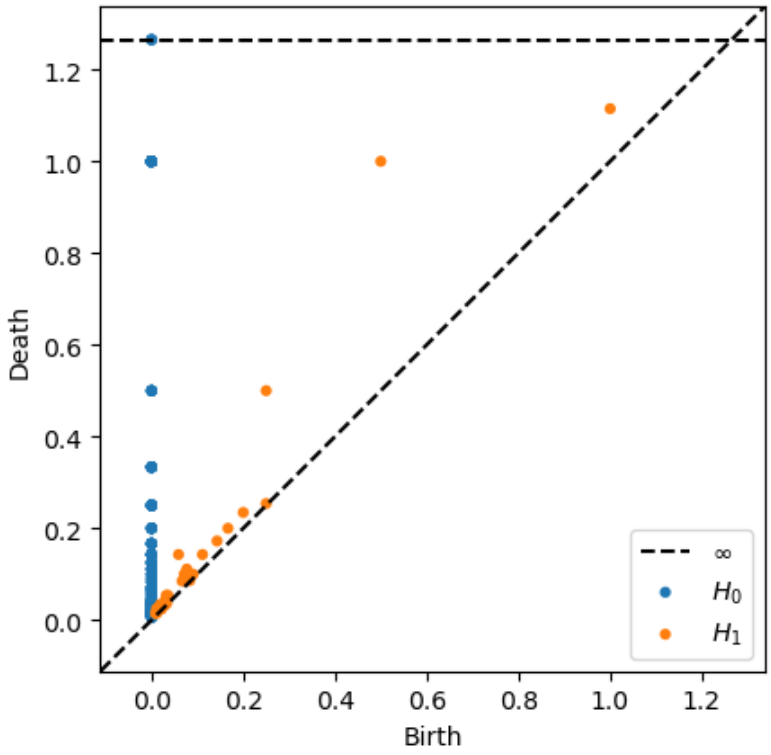

Figure 4: Persistence Diagram using top 200 Subreddits

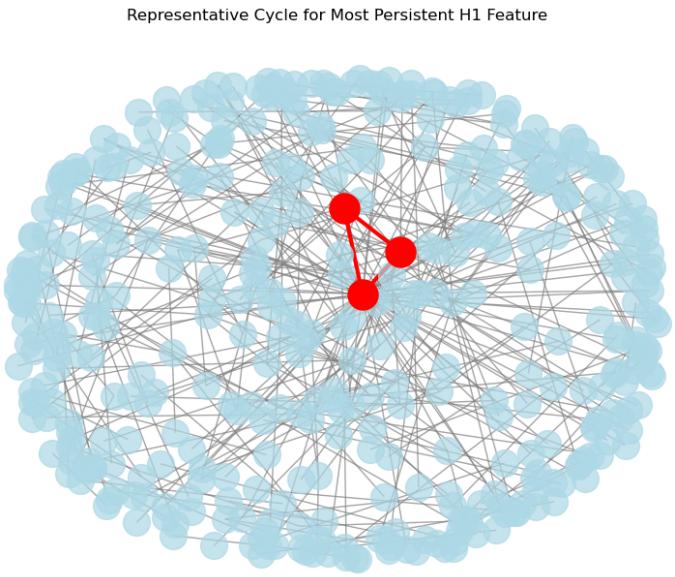

Figure 5: Visualizing H1 features from persistent Homology

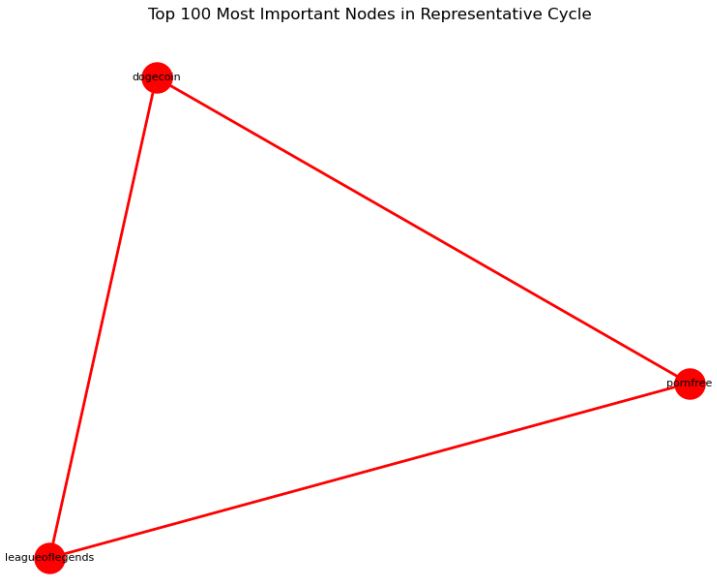

Figure 6: Strongest Connection Within Top Representative H1 Cycle

4.4 Visualization

NetworkX, Matplotlib, and Gudhi were used for all visualizations, including:

- Random subgraph layouts (spring embedding)
- Edge-weight filtered graphs
- Persistence diagrams
- Simplex (Rips) complex overlays

5 Results and Analysis

5.1 Graph Motif Analysis Results

The primary goal of this analysis was to identify echo chambers and pipelines within the Reddit subreddit network. While clique analysis is a natural approach for finding strongly connected communities, it quickly became apparent that it may not fully capture the directional flow of users or information from one community to another—a key aspect of pipeline dynamics. Nevertheless, I hypothesized that clique analysis could still be valuable for identifying insular groups that might function as echo chambers. Upon implementation, the clique motif analysis proved most effective at highlighting large, densely connected community structures within the graph, rather than pinpointing classic echo chambers. For example, as shown in Figure 2, the largest detected clique consisted primarily of football-related subreddits. This suggests that clique analysis tends to surface broad topical communities, where mutual connections are strong, but does not necessarily reveal the insulation or feedback characteristic of echo chambers.

To address this, I extended the analysis to search for the most insulated communities—those with the largest difference between incoming and outgoing links. The rationale was that a community with many incoming links but few outgoing ones might act as a "sink," potentially trapping users and reinforcing shared perspectives (i.e., an echo chamber). This approach yielded more interesting results, surfacing communities that are structurally isolated from the broader network.

However, a key limitation emerged: very few communities exhibited more incoming than outgoing links. This reflects a fundamental constraint of the dataset and methodology. Since edges represent hyperlinks posted from one subreddit to another, they do not always correspond to user migration or recommendation. Hyperlinks may be posted for a variety of reasons—such as referencing content, cross-promotion, or even criticism—many of which do not imply a user is being funneled into a new community. As a result, the directionality of hyperlinks is an imperfect proxy for user flow or the formation of pipelines. This limitation is consistent with recent research questioning the prevalence of echo chambers on Reddit.

While echo chambers are often theorized as insular, self-reinforcing communities, empirical studies have found substantial cross-community interaction, especially around controversial or high-profile topics. My findings reinforce this perspective: while tightly-knit communities exist, true echo chambers—defined by both strong internal cohesion and insulation from the rest of the network—are rare when using hyperlink data as the sole indicator.

5.2 Persistent Homology Analysis Results

The motivation for employing persistent homology was to move beyond classical graph methods and gain a topological view of interconnected echo chambers—specifically, to identify cycles of tightly linked subreddits that could serve as echo chambers, and to detect evidence of user flow into these structures with minimal outflow.

Computational & Methodological Challenges

However, two practical issues arose immediately:

- Computational Scalability: Persistent homology is computationally intensive, making it infeasible to run on the entire Reddit graph (tens of thousands of nodes and edges) with limited computational resources.

- Interpretability and Noise: Applying persistent homology to very large, dense graphs introduces significant noise; the abundance of connected features can obscure interesting or meaningful topological structures.

This presents a tradeoff: while studying the whole graph is essential for drawing conclusions about user flow across all of Reddit, we are forced to analyze smaller subgraphs due to resource constraints. These subsamples can reveal local topological patterns, but they cannot fully capture global user flow or the macro-scale emergence of pipelines and echo chambers.

Findings from Subgraph Analysis

Despite these limitations, persistent homology provided valuable insights when applied to subsamples of the Reddit graph. Notably:

Detection of Subtle Echo Chambers: Persistent homology surfaced cycles (H1 features) that represent potential echo chambers or focus points—clusters of subreddits with strong mutual connectivity. These structures were not easily detectable with traditional clique-based analysis, which tends to highlight only the most densely connected groups.

Representative Cycles: By extracting the most persistent H1 features and identifying the subreddits involved in the corresponding cycles, I uncovered strong user-flow loops that traditional graph analysis missed.

Workflow and Visualization: This process is illustrated in Figure 4 (persistence diagram), Figure 5 (representative cycle), and Figure 6 (cycle of most structurally important nodes).

A key limitation of this approach is its reliance on subsamples rather than the full graph. While subsamples can reveal interesting local structures, they may miss global echo chambers or pipelines that span larger portions of the network. Furthermore, the use of hyperlinks as a proxy for user flow is inherently noisy: hyperlinks are posted for many reasons, not all of which reflect genuine community transitions or recommendations.

6 Conclusion

This study set out to investigate the structural patterns underlying echo chambers and pipelines in Reddit's subreddit network using both classical graph analysis and topological data analysis (TDA). Through clique analysis, I was able to identify large, densely connected communities—such as clusters of football-related subreddits—that represent strong topical cohesion. However, while clique analysis effectively highlights these cohesive groups, it proved less effective at detecting the nuanced, insulated structures characteristic of echo chambers or the directional flow patterns indicative of pipelines.

To address these limitations, I employed persistent homology, a TDA technique capable of capturing higher-order connectivity and cyclic structures within the network. Although computational constraints limited this analysis to subgraphs rather than the entire Reddit network, persistent homology revealed subtle, tightly interconnected cycles—potential echo chambers—that were not apparent through motif or clique analysis alone. By focusing on the most persistent H1 features, I was able to extract representative cycles and identify clusters of subreddits exhibiting strong mutual connectivity and potential feedback loops.

Despite these advances, the analysis also highlighted important limitations. The reliance on hyperlink data as a proxy for user flow introduces noise, as hyperlinks may be posted for reasons unrelated to genuine community transitions. Additionally, the inability to scale persistent homology to the full graph restricts the scope of conclusions about global user flow and the formation of large-scale pipelines or echo chambers.

In summary, while clique analysis is valuable for uncovering large, cohesive communities, persistent homology provides a more nuanced lens for detecting cyclic, potentially echo-chamber-like structures within local regions of the network. Together, these approaches offer complementary insights into the topology of Reddit's community interactions. Future work could address current limitations by leveraging node embeddings, parallel computation, or hierarchical sampling to scale TDA methods,

and by integrating behavioral or content-based data to more accurately trace user flow and the emergence of echo chambers and pipelines across the broader social media landscape.

References

[1] G. Carlsson and A. Zomorodian. Topological Data Analysis for Social Network Analysis. *Discrete & Computational Geometry*, 42(1):71–93, 2009.

[2] C. Sunstein. Echo Chambers in Social Media: A Study of the Political Polarization in Twitter. *Journal of Political Communication*, 2018.

[3] S. Kumar, W. L. Hamilton, J. Leskovec, and D. Jurafsky. Reddit Hyperlink Network Dataset. Available at `https://snap.stanford.edu/data/soc-redditHyperlinks.html`, 2018.

[4] A. Hagberg, D. Schult, and P. Swart. NetworkX. Available at `https://networkx.org/`, 2008.

[5] U. Bauer. Ripser.py: A Lean Persistent Homology Library for Python. Available at `https://github.com/scikit-tda/ripser.py`, 2021.

[6] The GUDHI Project. GUDHI: Geometry Understanding in Higher Dimensions. Available at `https://gudhi.inria.fr/`, 2021.

[7] The pandas development team. pandas: powerful Python data analysis toolkit. Available at `https://pandas.pydata.org/`, 2023.

[8] J.D. Hunter. Matplotlib: Python plotting package. Available at `https://matplotlib.org/`, 2007.