Filter Bubbles

Indu Ramesh and R. Teal Witter

November 27, 2020

Proposal

Motivation and Previous Work

On the surface, the prominence of social media seems like it could make the world more connected, and expose users to a dizzying variety of diverse ideas. However, recent studies have suggested social media encourages the opposite, separating individuals into groups unable to find agreement, thereby polarizing society. A popular explanation for this phenomenon has been coined the filter bubble– the idea that the content displayed on users' feeds is social media networks is simply an echo chamber that prevents people from accessing a diverse array of viewpoints. Since user feed content is constrained by metrics that aim to increase user engagement and ad revenue (i.e. friends' and followers' views, internet search history, user location, etc.), social media companies explicitly incentivize users to pay preferential attention to like-minded content. In this manner, users end up living in a 'filter bubble' of their own ideas. Filter bubbles have been blamed for the spread of misinformation in Brexit, the 2016 U.S. presidential election, and increased distrust in democracy.

Chitra and Musco's paper aimed to provide a rigorous mathematical theory solidifying filter bubbles' emergence, studying user opinion dynamics within social networks [1]. Each network is modeled as a weighted graph per a *stochastic block model*. Each node corresponds to an individual, and edges between nodes correspond to social relationships. The "stochastic" aspect is as follows: the probability of two nodes being connected is higher when the nodes are in the same community, and lower in different communities. Relationships between users with increased interactions (i.e. their respective stories pop up in their news feeds more often) have higher edge weights; relationships where users barely interact are assigned low edge weights. Chitra and Musco adapted the *Friedkin-Johnsen model* to address opinion dynamics, augmenting each node with an opinion, a real number in the interval [-1,1]. As time passes, each node's opinion value is updated based on the average opinion of their neighbors and connections in the social network. Polarization is represented by the variance of opinions within the network, where opinions are represented as a n-dimensional vector.

To formalize social media companies' roles in creating filter bubbles, Chitra and Musco studied the influence of an important outside actor on the Friedkin-Johnsen model: the *network administrator*. The network administrator's job is to minimize disagreement among users by modifying the edge weights of the graph such that users interact with more content from users with similar opinions. The network administrator's modifications are subject to certain constraints; for instance, they cannot change the degree of any vertex, and they can only modify edge weights by a small amount. Chitra and Musco ran experiments on the social networks Twitter and Reddit to simulate the effect of a network administrator on polarization.

Their experimental results confirm filter bubble theory. That is, a social network modeled by the stochastic block model is, with high probability, already in a state of *fragile consensus*. Though a network may exhibit low polarization, a minor change in edge weights can cause a shift to high polarization. For example, when the network administrator changed only 40% of the total edge weight in the graph, polarization increased by a 40-fold factor. Chitra and Musco also demonstrated a simple fix to this phenomenon. Recall that the network administrator's object was to minimize disagreement: $\min_G D_z$. They proposed a solution to add a regularization term to the objective (ridge regression): $\min_G D_z + \lambda w^T w$ Their solution constrained the increase in polarization by a network administrator to 4%.

Research Questions

We propose the following further research directions to extend Chitra and Musco's work.

1. *Formalizing the network administrator's role*: Chitra and Musco informally assert that the network administrator's role is to minimize disagreement. Can we prove, formally, that this is the case– i.e. solve the optimization problem? Could we prove that a "simpler" or "natural" action by the network administrator– i.e. maintaining the overall weight of the graph, but increasing the weights between highly agreeable neighbors and decreasing the weights between disagreeing neighbors– still converges to a more highly polarized vector of opinions? Can we prove this in a stochastic block model? What about other models?

2. *Dealing with outside attacks*: Could we model some sort of attack by an outside actor on increasing polarization? An example is Cambridge Analytica's role in the 2016 presidential election to sway undecided voters toward electing Trump, popularized by the Netflix documentary "The Great Hack." Here's a simple idea: at each time step, introduce a new 'fake' node with an opinion. How many fake nodes, and how strong of opinions, would we have to introduce to converge to a final set of opinions with higher polarization, and how many time steps would that take?

3. *Investigating Bimodality*: Some suggest that polarization, defined by Chitra and Musco as the variance of the n opinions of the network's users, might not actually be what's increasing as much as Chitra and Musco experimentally demonstrate. What's actually going on may be more accurately explained by increasing *bimodality*, where opinions converge in clusters around two centers, as opposed to opinions being further away from where we started. In other words, we can modify Chitra and Musco's question to: "How sensitive are social networks to bimodality?" We propose formally defining bimodality and seeing if we can obtain a similar result to the "fragile consensus" linear algebraic results by Chitra and Musco.

Network Administrator Actions

Algorithm 1: Opinion formation with network administrator.

```
input: adjacency matrix A_0 of dimension n × n,
       initial set of opinions s with s_i ∈ (−1, 1) for i ∈ [n],
       number of iterations T to run
output: a set of equalibrium opinions z^(T)
z^(0) ← s
foreach t in {0, ..., T} do
    D ← degree matrix of A_t
    z^(t) ← (D + I)^{-1}(A_t z^{t-1} + s)
    A_t ← network administrator modifications of A_{t-1}
return z^(T)
```

"Natural" network administrator actions:

1. For each node $u \in [n]$, add $\epsilon > 0$ to the weight of the adjacent edge closest in current opinion to u and subtract ϵ from the weights of the adjacent edges furthest in current opinion from u subject to the constraint that each edge has at most $\delta > 0$ weight.

To Do

- ✓ Code up Algorithm 1.
- ✓ Implement a slightly easier version of the first network administrator action.
- Implement the full version of the first network administrator action.
- Understand Equation 14 in [1].

- Find about an administrator modification that can be written in matrix notation.

References

[1] Uthsav Chitra and Christopher Musco. Analyzing the Impact of Filter Bubbles on Social Network Polarization. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, WSDM '20, pages 115–123, New York, NY, USA, January 2020. Association for Computing Machinery.