Two Papers on Polarization Metrics

A Summary by C. Lente — 2019-05-12

Introduction

Many countries' recent political shifts alongside growing awareness about how social networks are able to manipulate users' opinions have brought to the public eye many concepts such as political polarization, echo chambers, filter bubbles, and the role of "the algorithm" in social medias' news feeds. Recently, Mariani et al.[1] have even published what they called an "Ideology GPS" to track political polarization on Twitter.

Many of these trackers, however, suffer from three problems: need for subjective input from the researcher, ill defined similarity metrics between nodes of the network, and impossibility to track more than two clusters. As a result, most analyses feel post-hoc, making it hard for the authors to defend their methodology.

A Measure of Polarization

Guerra et al. (2013)[2] discusses an important but often overlooked aspect of research on possibly polarized networks: sometimes what seems like polarized communities are just cohesive but not actually polarized. The online community of football fans might, for example, have few connections to the community of basketball fans, but this doesn't necessarily mean that they are polarized.

After some experiments with Twitter data, they conclude that *modularity*[3] (a very popular community quality metric) alone can't differentiate between non-polarized communities (graduate vs. undergraduate students on Facebook) and polarized ones (liberals vs. conservatives blogs). This motivates them to look for another metric that is able to do that.

One key concept they observe is that, unlike polarized networks, non-polarized networks tent to have their most "popular" nodes in the community boundaries. This makes intuitive sense: in the graduate-undergraduate example, the students with the most friends probably tend to also have friends in both groups, making them highly connected boundary nodes.

With this in mind, the authors propose a new metric for measuring polarization itself instead of modularity: *polarization*[4]. P lies in the range $(-1/2; +1/2)$ and values grater than zero indicate that "on average, nodes on the boundary tend to connect to internal nodes rather than to nodes from the other group, indicating that antagonism is likely to be present".

Using the same data sets they used to test modularity, the researches test their polarization metric. They conclude that, while modularity has no evident threshold at which polarization is present in the network, their metric does.

[1] Mariani, Daniel, et al. "A posição ideológica de mil influenciadores no Twitter." *Folha de S. Paulo*. 2019.

[2] Guerra, Pedro Calais, et al. "A measure of polarization on social media networks based on community boundaries." *Seventh International AAAI Conference on Weblogs and Social Media*. 2013.

[3] $Q = \frac{1}{2m} \sum_{ij} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \frac{s_i s_j + 1}{2}$

[4] $P = \frac{1}{|B|} \sum_{v \in B} \left[\frac{d_i(v)}{d_b(v) + d_i(v)} - 0.5 \right]$

Minimizing Polarization and Disagreement

Musco et al. (2017)[5] have a much more prescriptive approach in their work. They seek out to find a network structure to an existing social graph that minimizes at the same time polarization and disagreement simultaneously.

The authors argue that, currently, social networks' algorithms are trained to maximize user engagement and revenue. This ends up generating recommender systems that minimize challenging users' opinions, suggesting only content and friendships that reinforces each user's echo chamber. Minimizing disagreement, though, has the side-effect of generating greater polarization.

Given this knowledge about social media, they frame the following optimization problem:

> Given n agents, each with its own initial opinion that reflects its core value on a topic, and an opinion dynamics model, what is the structure of a connected social network with a given total edge weight that minimizes polarization and disagreement simultaneously?

With formal definitions of polarization[6] and disagreement[7], the researches conclude that there is always a graph with $O(n/\epsilon^2)$ edges that is within an acceptable range from the optimal solution to the problem above. They also describe how such minimization would be achieved with a convex optimization program which can be solved in polynomial time.

To test this solution, the authors ran the algorithm (publicly available on GitHub) on both synthetic and real data sets. On a network of Reddit users, they claim that their method is able to optimize the graph topology enough to achieve an almost 60,000-fold reduction in polarization and disagreement.

Personal Comments

As someone who has just started researching this topic, many of the jargon was unknown to me (specially graph topology optimization). Nevertheless, I understood enough to gather some strong and weak points of each article.

The first paper has a powerful intuitiveness to it. Most of the concepts can be illustrated with simple examples, unlike the paragraph-like equations of the second paper that seem obscure and unverifiable. But this is also a weak point: the authors of the latter don't talk about any competing metrics or how they came up with theirs.

Guerra et al. also make no prescriptive claims about their polarization index (e.g. how to minimize it). But Musco et al. also show their shortcomings when they make assumptions about the network[8]. It wasn't made clear why all social media graphs must be undirected and unweighted; there is also no relationship between edge weight and nodes' expressed opinions. Finally, they also rely strongly on NLP tools to classify these expressed opinions, which implies that the researchers need to have an *a priori* list of classifications.

[5] Musco, Cameron, et al. "Minimizing polarization and disagreement in social networks." *Proceedings of the 2018 World Wide Web Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2018.

[6] $D_{G,s} = \sum_{(u,v)\in E} d(u,v)$

[7] $P_{G,s} = \sum_{u\in V} \bar{z}_u^2 = \bar{z}^T \bar{z}$

[8] Two research questions that they propose themselves are: could one use another opinion formation model with their framework? Do links that cross communities always improve the polarization-disagreement index, or not?