Estimating Affective Polarization on a Social Network

Keywords: affective polarization, partisan hostility, social networks, network analysis, Twitter

Extended Abstract

Concerns about polarization on social media are widespread. While many studies have analyzed whether social media users are embedded in isolated echo chambers [e.g. 1], there is only little research on the affective dimension of digital user interactions [2]. Affective polarization describes negative affect between political opponents and is thus concerned with in-group versus out-group dynamics [3]. Traditionally, surveys are used to measure affective polarization [4], but this approach has two critical shortcomings.

First, the concept of affective polarization focuses on social identity and group dynamics, but collecting information on social ties in surveys is challenging. Since this information is usually unavailable in survey data, a relevant conceptual component of affective polarization remains undiscovered in these standard survey measures. Second, surveys can only capture general attitudes and intentions, but they are inadequate indicators of actual behavior [3]. Although a respondent might claim to dislike or avoid others they disagree with, a survey cannot conclusively clarify whether the participant indeed acts upon these intentions [4].

To address these shortcomings, we propose a new network-based measure of affective polarization that models the social ties between people and their group dynamics. We apply this measure to a large-scale Twitter dataset on the Covid-19 pandemic, thereby analyzing real-world behavior on social media instead of relying on abstract survey data.

Our measure $\rho_{x,y,W}$ is based on the recently introduced Pearson correlation on complex networks [5]. To begin, we consider a network $G = (V, E)$ with V as the set of individuals and $E \subseteq V \times V$ as the set of connections. We record an opinion value $o_i \in [-1, 1]$ per node in V. For each edge between two nodes i and j, we determine a disagreement value $x_{i,j}$ as the absolute difference of opinion values between the two individuals. Moreover, we document a hostility value $y_{i,j} \in [0, 1]$ per edge. An edge represents an interaction between two individuals, such as a reply or mention on a social media platform, and we assume that the edges are undirected. The Pearson correlation on complex networks can only determine correlations for node attributes which poses a problem for quantifying affective polarization since we model both disagreement and hostility as edge attributes. To solve this problem, we retrieve the line graph G', which translates all edges and their attributes into nodes and associated node attributes. Using the line graph is an established approach to, for instance, find communities among edges rather than nodes [6]. Each edge in G is represented as a node in the line graph G', and each node common to two edges in G is represented by an edge in G'.

Our measure $\rho_{x,y,W}$ takes three inputs: a vector x specifying the disagreement between each pair of users, a vector y recording the hostility between user pairs, and a weight matrix W. W contains the exponentiated effective resistance, a measure of distance in a network, for each node pair in the line graph G'. The measure is then defined as:

$$\rho_{x,y,W} = \frac{sum(W \times (\hat{x} \otimes \hat{y}))}{\sigma_{x,W}\sigma_{y,W}}$$

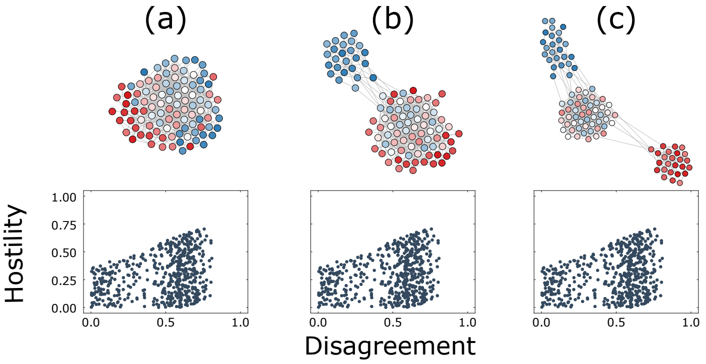
where $\hat{x}$ and $\hat{y}$ are the centered versions of x and y, e.g., $\hat{x} = (x - \bar{x})$, $\times$ is an element-wise product operation, $\otimes$ is the outer product operation, *sum* is the sum of all the cells of a matrix, and $\sigma_{x,W}$ and $\sigma_{y,W}$ are the respective (network) standard deviations of x and y [see 5]. The measure can return values from -1 (negative correlation), passing 0 (no correlation, i.e., no affective polarization), to $+1$ (positive correlation, i.e., high affective polarization). Intuitively, $\rho_{x,y,W}$ captures the correlation between disagreement and hostility, given the correlation of the two variables in the vicinity of each edge in G.

To validate our measure, we conduct experiments with synthetic data which compare $\rho_{x,y,W}$ to the Pearson correlation coefficient $\rho_{x,y}$ – an established measure of affective polarization in survey-based political science studies [4] – and a measure proposed in [7] that relies on the Earth Mover's Distance ($EMD_{x,y,G}$). In the first experiment, we generate three different network structures while keeping the relationship between disagreement and hostility fixed as shown in Figure 1. In (a), the nodes are randomly connected and there is no community structure, whereas in (c), there is both a liberal (blue) and a conservative (red) community as well as a mixed community of moderate nodes. In the latter case, extreme conservative and extreme liberal individuals are surrounded by their non-hostile in-groups and we therefore expect affective polarization to be highest here. The results confirm that our network correlation coefficient $\rho_{x,y,W}$ and $EMD_{x,y,G}$ are sensitive to topological features since their values increase from (a) to (c) as expected, while the Pearson correlation coefficient $\rho_{x,y}$ stays constant for all three network topologies. In further experiments (not displayed here), we test how the measures react when the network structure stays fixed, but the relation between disagreement and hostility changes. The results show that our measure $\rho_{x,y,W}$ and the Pearson correlation coefficient $\rho_{x,y}$ behave as expected, while $EMD_{x,y,G}$ produces inconsistent and counter-intuitive affective polarization values. We conclude that among the three approaches tested, only our measure $\rho_{x,y,W}$ can appropriately quantify affective polarization in a social network.

Additionally, we apply this measure to a large-scale Twitter data set of approximately 47 million tweets discussing Covid-19, posted over six months in 2020. We find that affective polarization was low in early February ($\rho = 0.17$) and then increased to moderately high levels ($\rho = 0.6$) in the following months before reaching very high levels ($\rho = 0.95$) in July 2020.

References

1. Barberá, P., Jost, J. T., Nagler, J., Tucker, J. A. & Bonneau, R. Tweeting from left to right: is online political communication more than an echo chamber? *Psychological Science* **26,** 1531–1542 (2015).
2. Marchal, N. "Be nice or leave me alone": an intergroup perspective on affective polarization in online political discussions. *Communication Research* **49,** 376–398 (2022).
3. Druckman, J. N. & Levendusky, M. S. What do we measure when we measure affective polarization? *Public Opinion Quarterly* **83,** 114–122 (2019).
4. Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N. & Westwood, S. J. The origins and consequences of affective polarization in the United States. *Annual Review of Political Science* **22,** 129–146 (2019).
5. Coscia, M. Pearson correlations on complex networks. *Journal of Complex Networks* **9** (2021).
6. Evans, T. S. & Lambiotte, R. Line graphs, link partitions, and overlapping communities. *Physical Review E* **80** (1 2009).
7. Tyagi, A., Uyheng, J. & Carley, K. M. Heated conversations in a warming world: affective polarization in online climate change discourse follows real-world climate anomalies. *Social Network Analysis and Mining* **11,** 87 (2021).



Network correlation $\rho_{x,y,W}$	**0.66**±0.06	**0.90**±0.02	**0.94**±0.01
Pearson correlation $\rho_{x,y}$	0.43±0.04	0.43±0.03	0.43±0.04
EMD$_{x,y,G}$ (Blue)	0.13±0.02	0.15±0.02	0.15±0.02
EMD$_{x,y,G}$ (Red)	0.13±0.02	0.12±0.02	0.15±0.02

Figure 1: The first row shows the network topology used in this experiment where the node color reflects opinion (blue as liberals, and red as conservatives). The second row shows how (dis)agreement and hostility are related across all node pairs in the network. The rows below report the results for the measures compared here. Since the EMD measure proposed by [7] is calculated per group, we specify a value for the group of blue nodes and red nodes separately. All values reported are averages over 100 iterations of the experiment and each average is followed by its standard deviation.