# In-Group Love, Out-Group Hate: A Framework to Measure Affective Polarization via Contentious Online Discussions

Buddhika Nettasinghe*
University of Iowa
Tippie College of Business
Iowa City, Iowa, USA
buddhika-nettasinghe@uiowa.edu

Ashwin Rao*
USC Information Sciences Institute
Department of Computer Science
Marina del Rey, California, USA
mohanrao@usc.edu

Bohan Jiang
Arizona State University
School of Engineering
Tempe, Arizona, USA
bjiang14@asu.edu

Allon G. Percus
Claremont Graduate University
Institute of Mathematical Sciences
Claremont, California, USA
allon.percus@cgu.edu

Kristina Lerman
USC Information Sciences Institute
Department of Computer Science
Marina del Rey, California, USA
lerman@isi.edu

## Abstract

Affective polarization, the emotional divide between ideological groups marked by in-group love and out-group hate, has intensified in the United States, driving contentious issues like masking and lockdowns during the COVID-19 pandemic. Despite its societal impact, existing models of opinion change fail to account for emotional dynamics nor offer methods to quantify affective polarization robustly and in real-time. In this paper, we introduce a discrete choice model that captures decision-making within affectively polarized social networks and propose a statistical inference method estimate key parameters—in-group love and out-group hate—from social media data. Through empirical validation from online discussions about the COVID-19 pandemic, we demonstrate that our approach accurately captures real-world polarization dynamics and explains the rapid emergence of a partisan gap in attitudes towards masking and lockdowns. This framework allows for tracking affective polarization across contentious issues has broad implications for fostering constructive online dialogues in digital spaces.

## CCS Concepts

• **Information systems** → **Social networks**; • **Applied computing** → *Sociology*; • **Computing methodologies** → Modeling methodologies.

## Keywords

Polarization; Twitter; COVID-19; Social Networks; Exposure

**ACM Reference Format:**
Buddhika Nettasinghe, Ashwin Rao, Bohan Jiang, Allon G. Percus, and Kristina Lerman. 2025. In-Group Love, Out-Group Hate: A Framework to Measure Affective Polarization via Contentious Online Discussions. In *Proceedings of the ACM Web Conference 2025 (WWW '25), April 28-May 2, 2025, Sydney, NSW, Australia.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3696410.3714935

*Both authors contributed equally to this research.


ACM ISBN 979-8-4007-1274-6/25/04
https://doi.org/10.1145/3696410.3714935

## 1 Introduction

Americans have grown increasingly divided along ideological lines. These divides extend beyond mere disagreements on policy issues to open antagonism and hostility between the two parties. This emotional divide, known as *affective polarization* [24, 25], is characterized by two key aspects: (1) the tendency for people to like and trust others from their own political party (*in-group love*) and (2) the tendency to dislike and distrust people from opposing parties (*out-group hate*) [26]. The out-group hate is a corrosive force that often undermines consensus, driving individuals to adopt a position not necessarily on its merits or because it aligns with their own beliefs, but simply to oppose the other side [11, 37]. Thus, when an issue becomes politicized, i.e., associated with a partisan identity, out-group animosity can split society, transforming seemingly mundane choices such as mask wearing and vaccination into contentious partisan divides [25].For these reasons, affective polarization can disrupt governance in unexpected ways [20, 24].

Despite its significance, few robust methods exist to estimate the core components of affective polarization, namely, in-group love and out-group hate, in real-time. Current methods rely heavily on offline surveys [11], which are vulnerable to framing effects and respondent bias [10], and are slow. Data-driven methods usually rely on homophily [4, 42]—the tendency of similar users to form new connections—to explain emergence of polarization in social networks [14, 15]. However, these methods do not generalize well due to challenges of collecting network data, and they fail to explain divergence of partisan opinions in fully connected networks.

To address these challenges, we propose a principled framework for estimating in-group love and out-group hate using social media data. Our key contributions are as follows:

***Discrete Choice Model***: We introduce an intuitive discrete choice model of individual decision-making within an affectively polarized social network. This model includes parameters that capture the two key aspects of affective polarization: in-group love ($\alpha$) and out-group hate ($\beta$). The temporal dynamics of the model demonstrate how emotionally polarized decision-making can either unify or divide a population along party lines, depending on the values of $\alpha$ and $\beta$. We also generalize the model to account for multi-party contexts (e.g., far left, left, moderate, right, far-right), which can explain issue polarization in societies where a spectrum of ideologies exists.

***Parameter Estimation***: We develop a statistical method to estimate model parameters—in-group love ($\alpha$) and out-group hate ($\beta$)—from data based on stances of individuals and their network neighbors (both in-group and out-group). For example, this method can analyze opinions expressed on Twitter about the effectiveness of masks in curbing the spread of COVID-19. The estimation method is derived from the discrete choice model and uses logistic regression. We also discuss variations of the method for settings where only ego nodes' stances (but not their neighbors') are available.

***Empirical Validation***: We first empirically analyze attitudes on masking and lockdowns expressed online during the COVID-19 pandemic to show that a partisan gap emerged rapidly over these issues, likely triggered by key events such as the CDC masking recommendation (or co-occurring events). We the estimate the in-group love $\alpha$ and out-group hate $\beta$ (i.e., parameters of affective polarization) related to these issues via the proposed framework. When the model is calibrated with these estimated parameters, it accurately reproduces the real-world the real-world polarization dynamics leading to the divergence of attitudes along ideological lines. We also demonstrate that the more active partisans show higher levels of out-group hate.

Rapidly identifying emotionally polarized issues on social media through empirical estimates of in-group love and out-group hate enables more constructive dialogue. Journalists can report responsibly to avoid inflaming tensions, while social media platforms can implement real-time strategies to prevent manipulation and mitigate societal divisions, fostering healthier civic discourse. The proposed model and estimation method also provides a framework for network and computational social science researchers to systematically understand manifestations of affective polarization in a wide array of issues.

## 2 Related Work

***Survey-based methods.*** The existing methods rely largely on self-reported surveys to measure affective polarization (i.e., in-group love and out-group hate) [24]. For example, the popular *feelings thermometer* used by the American National Election Study (ANES) asks democrats and republicans how warm or cold they feel towards their own party and the opposing party on a scale of 0 (coldest) to 100 (warmest). Similar approaches ask about traits that people tend associate with the two parties (such as intelligence, patriotism, meanness, hypocrisy) and how comfortable they are associating with a member of the opposite party (as a friend, a neighbor, a relative, etc.) [10, 31]. Despite their widespread use, survey-based methods like the ANES feeling thermometer have limitations, including respondents' subjective interpretations, selection bias favoring politically engaged individuals, and sensitivity to survey mode (e.g., in-person vs. online) [12, 43].

***Online Polarization.*** Research on social media's role in polarization has increasingly focused on exacerbation by emotional divides. Studies show that cross-ideological interactions tend to be more negative and toxic compared to exchanges between same-ideology users [29], consistent with affective polarization. Network-based methods have been proposed to detect polarization and controversy. These methods are inspired by *partisan sorting*, where social media users position themselves near same-ideology others, aligning their beliefs. Garimella et al. [15] use random-walk-based measures to analyze the structure of conversation to identify topics. Bonchi et al. [4] examine polarization in signed networks where positive and negative edges indicate friendly or antagonistic interactions. They use community detection to split the network by antagonistic and friendly interactions to identify issues where disagreement exists. Similar to them, Fraxanet et al. [14] quantify polarization in signed networks of online interactions as the interplay between 1) *antagonism*, which reflects the level of negative interactions or hostility within a community, and 2) *alignment*, measured by how much interactions are structured along a primary division, or fault line, within the community. They identify issues that drive polarization (high antagonism and alignment) and conflicts that do not reinforce societal divisions. Polarization grew during the COVID-19 pandemic, with partisan divides emerging over virus origins, mitigation strategies, and vaccine mandates [27, 39]. Political ideology shaped adherence to health guidelines [18, 20], while political elites fueled early-stage polarization by emphasizing different issues [19].

***How the proposed method differs from existing methods.*** Unlike existing opinion dynamics models, such as DeGroot-type models, our model explicitly parameterizes affective polarization through in-group love and out-group hate, allowing for a direct study of their effects on group dynamics. By leveraging social media data, we estimate the relative importance of these factors in a principled manner. Further, unlike the existing survey-based methods, the proposed approach takes into account both the social network structure (i.e., how people are connected to other individuals in their in-group as well as out-group) as well as the temporal dynamics of the stances. Additionally, it captures binary decision-making, reflecting real-world political choices (e.g., pro- vs. anti-vaccine, pro- vs. anti-masking) [9, 28]. Also, the model is based on a bi-populated society, aligning with real-world partisan divisions, unlike signed network models that focus on friendly and antagonistic ties [14, 41].

Our binary choice model is a generalization of [37] and better captures real-world behaviors. The model in [37] emerges as a special case of our framework. The model proposed in this paper enables logistic regression to estimate in-group love and out-group hate via social media data. This data-driven approach provides statistical guarantees and can be applied to issues such as vaccines and climate change, offering a theoretically tractable tool to complement empirical studies on affective polarization.

We contribute to the state of the art by explaining how novel COVID-19 issues became polarized and the role of affective polarization in the partisan gap. We use a calibration approach to link the model to real-world observations by estimating parameters directly from data (see e.g., [22]), instead of adjusting the model to fit observed behaviors. Solving the calibrated model, we demonstrate how minimal mathematical models can predict complex real-world phenomena.

## 3 Modeling Affective Polarization

This section introduces a dynamical model of individual decisions in an affectively polarized society, separating the effects of in-group love and out-group hate.

### 3.1 The Binary Choice Model

**Context:** We consider an undirected social network $G = (V, E)$ where each node (individual) $v \in V$ has two binary attributes: a static attribute $R(v) \in \{0, 1\}$ representing its identity or aspect of an identity such as ideology or political affiliation, and a dynamic attribute $H_k(v) \in \{0, 1\}$ that evolves over time $k = 0, 1, 2, \ldots$. The node $v$ is labeled as red if $R(v) = 1$ (and blue otherwise). The dynamic attribute $H_k(v)$ reflects the individual $v$'s stance (or choice) from among the two available alternatives at time $k$, such as wearing a mask or not. The initial stances (i.e., $H_k(v)$ for each $v \in V$ at time $k = 0$) may be assigned deterministically or randomly.

Each individual can observe the stances of their network neighbors, regardless of their own or the neighbor's affiliation. We refer to individual's same-affiliation neighbors as their *in-group* and opposite affiliation neighbors as their *out-group*.

**Quantifying the In-Group and Out-Group Influences:** In order to define the evolution of stances of individuals, we first need to formally specify how each individual is influenced by the stances of their in-group and out-group neighbors. For this purpose, let $\Delta_k^{in,1}(v), \Delta_k^{out,1}(v)$ denote measures that quantify the prevalence of stance-1 (with respect to the stance-0) among the in-group neighbors and the out-group neighbors of $v \in V$, respectively. Similarly, $\Delta_k^{in,0}(v), \Delta_k^{out,0}(v)$ denote the prevalence of stance-0 (with respect to the stance-1) among the in-group and out-group neighbors of $v$, respectively. The application context and the available data granularity could dictate the exact definitions of those measures of in-group and out-group influences. Below are two examples.

*Definition 1. Net number of neighbors with a stance normalized by degree:* $\Delta_k^{in,1}(v)$ is defined as the difference between the number of in-group neighbors of $v$ with stance-1 and stance-0, normalized by the degree of $v$ (number of neighbors of the $v$), and $\Delta_k^{in,0}(v) = -\Delta_k^{in,1}(v)$. The quantities for the out-group are defined similarly. For example, consider a blue node $v$ with 70 out of 100 blue neighbors (in-group connections) wearing masks (stance 1) and 7 out of 10 red-neighbors (out-group connections) wearing masks. Under this first definition of influence, we get $\Delta_k^{in,1}(v) = (70 - 30)/110 = 40/110, \Delta_k^{in,0}(v) = -40/110$ and $\Delta_k^{out,1}(v) = (7 - 3)/110 = 4/110, \Delta_k^{out,0}(v) = -4/110$. This approach takes into account the number of individuals with stance-1 and not just the fraction in each group (i.e., a larger number of individuals is likely to exert more influence).

*Definition 2. Net fraction of neighbors with a stance:* Alternatively, $\Delta_k^{in,1}(v)$ is defined as the difference between fraction of in-group neighbors of $v$ with stance-1 and fraction of in-group neighbors of $v$ with stance-0, and $\Delta_k^{in,0}(v) = -\Delta_k^{in,1}(v)$. Quantities for the out-group are defined similarly. For the same example as before, this second approach would yield same magnitude of the in-group and out-group effects from masking i.e., $\Delta_k^{in,1}(v) = \Delta_k^{out,1}(v) = 0.7 - 0.3 = 0.4$ and $\Delta_k^{in,0}(X_k) = \Delta_k^{out,0}(v) = -0.4$. Thus, unlike the first approach, given the fraction of individuals in each group of neighbors who are masking and non-masking, the actual numbers do not matter.

The formal versions of the above definitions are given in the Appendix C. For the remainder of this paper, we use the first definition. Results related to Definition 2 are provided in the Appendix.

**Dynamics of the Stances:** At each time step $k$ (where $k = 0, 1, 2, \ldots$), a node $X_k \in V$ is sampled uniformly to observe the stances of its neighbors and then update its own stance as follows. Let,

$$p_{X_k,k}(1|0) = \mathbb{P}(H_{k+1}(X_k) = 1 | H_k(X_k) = 0) \tag{1}$$

$$p_{X_k,k}(0|1) = \mathbb{P}(H_{k+1}(X_k) = 0 | H_k(X_k) = 1) \tag{2}$$

denote the probabilities with which the node $X_k$ switches its choice at time $k + 1$. Those transition probabilities are modeled using the following logistic functions,

$$\begin{aligned} p_{X_k,k}(1|0) &= \frac{1}{1 + \exp\left[-\left(\alpha\Delta_k^{in,1}(X_k) - \beta\Delta_k^{out,1}(X_k) - \delta\right)\right]} \\ p_{X_k,k}(0|1) &= \frac{1}{1 + \exp\left[-\left(\alpha\Delta_k^{in,0}(X_k) - \beta\Delta_k^{out,0}(X_k) - \delta\right)\right]}, \end{aligned} \tag{3}$$

where $\alpha, \beta, \delta \geq 0$ are fixed model parameters. In the event that the node $X_k$ doesn't switch choices, it stays with the same choice at time $k + 1$, i.e., $H_{k+1}(X_k) = H_k(X_k)$.

**Discussion of the Model:** In the above model, the parameters $\alpha, \beta$ quantify the strengths of in-group love and out-group hate, respectively. For example, consider a scenario where the randomly selected node at time $k$, $X_k$, belongs to the blue group (i.e., $R(X_k) = 0$) and is not wearing a mask ( i.e., holds stance-0 or $H_k(X_k) = 0$). In order to decide whether to adopt stance-1 (wear a mask) at the next time step, $k+1$, $X_k$ focuses on two factors: the prevalence of stance-1 among the in-group (blue) neighbors (quantified by $\Delta_k^{in,1}(X_k)$) and the prevalence of stance-1 among out-group (red) neighbors (quantified by $\Delta_k^{out,1}(X_k)$). In particular, notice from Eq. (3) that the log-ratio of adopting stance-1 to staying with stance-0 can be expressed as,

$$\alpha\Delta_k^{in,1}(X_k) - \beta\Delta_k^{out,1}(X_k) - \delta = \log\left(\frac{p_{X_k,k}(1|0)}{1 - p_{X_k,k}(1|0)}\right)$$

A graphical illustration is given in the Appendix Fig. 9. If the in-group neighbors are more inclined towards stance-1 (i.e., larger $\Delta_k^{in,1}(X_k)$), it encourages $X_k$ to adopt stance-1 at the next time step (i.e., the probability of switching to stance-1, $p_{X_k,k}(1|0)$, is larger). On the other hand, if the in-group neighbors are less inclined to wear masks (i.e., smaller $\Delta_k^{in}(X_k)$), it discourages $X_k$ from masking. The out-group has the opposite effect: a higher tendency among the out-group to wear masks (i.e., larger $\Delta_k^{out,1}(X_k)$) provokes $X_k$ to resist masking (reflecting out-group animosity) and vice-versa. Thus, the decision of $X_k$ is influenced by the interplay of these two social tendencies. The relative strengths of the in-group love and out-group hate are quantified by the parameters $\alpha$ and $\beta$, respectively. For example, $\beta > \alpha$ indicates that people pay more attention to the out-group than the in-group when making decisions. The inertia $\delta$ models the tendency to resist changes in behavior. In particular, larger values of $\delta$ implies that a large collective influence of the in-group and out-group is needed for $X_k$ to switch its stance.

The proposed model decouples the effects of in-group love and out-group hate (e.g., high, equal, or low out-group hate relative to in-group love) and helps to understand how their interplay shapes the dynamics of people's stances. Additionally, it takes into account the fact that people may be exposed to more in-group individuals (due to homophily of the political ideology in the network $G = (V, E)$) as well as the asymmetry of the presence of the two stances within the in-group and out-groups of individuals.

## 3.2 Dynamics of the Discrete Choice Model

The evolution of the model presented in Sec. 3.1 is stochastic since the node that updates its stance at each time instant is chosen at random. However, the stochastic dynamics can be meaningfully approximated in a mean-field manner when the number of nodes are large and the network is fully connected. In particular, let $\theta^{\mathcal{B}}(t), \theta^{\mathcal{R}}(t)$ denote the fraction of blue nodes with stance-1 and fraction of red-nodes with stance-1 at (continuous) time $t$, respectively. Further, let $r \in (0, 1)$ be the fraction of red nodes in the network. The influence measures defined earlier ($\Delta_k^{in,1}(X_k)$, $\Delta_k^{out,1}(v)$, $\Delta_k^{in,0}(v)$, $\Delta_k^{out,0}(v)$) can then be written in terms of $\theta^{\mathcal{B}}(t), \theta^{\mathcal{R}}(t)$ and $r$. For example, for any blue node, $\Delta_t^{in,1}(v) = (1-r)\left(2\theta^{\mathcal{B}}(t) - 1\right)$ under the first influence measure (net number of individuals with a stance normalized by degree) and $\Delta_t^{in,1}(v) = \left(2\theta^{\mathcal{B}}(t) - 1\right)$ under the second influence measure (fraction of neighbors in each group with a stance). Consequently, the transition probabilities in Eq. (3) can be expressed in terms of $\theta^{\mathcal{B}}(t), \theta^{\mathcal{R}}(t)$ and $r$. In particular, let $p_\theta^{\mathcal{B}}(1|0), p_\theta^{\mathcal{B}}(0|1)$ (resp. $p_\theta^{\mathcal{R}}(1|0), p_\theta^{\mathcal{R}}(0|1)$) denote the transition probabilities given in Eq. (3) when $X_t$ is a blue (resp. red) node in a fully connected graph. Under the first influence measure defined earlier (net number of neighbors with a stance normalized by degree), they can be expressed in terms of $\theta^{\mathcal{B}}(t), \theta^{\mathcal{R}}(t)$ and $r$ as,

$$
\begin{aligned}
\text{logit}\left(p_\theta^{\mathcal{B}}(1|0)\right) &= \alpha(1-r)\left(2\theta^{\mathcal{B}}(t) - 1\right) - \beta r\left(2\theta^{\mathcal{R}}(t) - 1\right) - \delta, \\
\text{logit}\left(p_\theta^{\mathcal{B}}(0|1)\right) &= \alpha(1-r)\left(1 - 2\theta^{\mathcal{B}}(t)\right) - \beta r\left(1 - 2\theta^{\mathcal{R}}(t)\right) - \delta, \\
\text{logit}\left(p_\theta^{\mathcal{R}}(1|0)\right) &= \alpha r\left(2\theta^{\mathcal{R}}(t) - 1\right) - \beta(1-r)\left(2\theta^{\mathcal{B}}(t) - 1\right) - \delta, \\
\text{logit}\left(p_\theta^{\mathcal{R}}(0|1)\right) &= \alpha r\left(1 - 2\theta^{\mathcal{R}}(t)\right) - \beta(1-r)\left(1 - 2\theta^{\mathcal{B}}(t)\right) - \delta.
\end{aligned}
$$

The transition probabilities under second measure (fraction of neighbors in each group with a stance) can be obtained by simply removing $r$ from the above expressions (since the group sizes do not matter). Then, the evolution of $\theta(t) = \left[\theta^{\mathcal{B}}(t), \theta^{\mathcal{R}}(t)\right]$ on a fully connected network can be represented using the following differential equation

$$
\begin{bmatrix} \dot{\theta}^{\mathcal{B}}(t) \\ \dot{\theta}^{\mathcal{R}}(t) \end{bmatrix} = \begin{bmatrix} \left(1 - \theta^{\mathcal{B}}(t)\right) p_\theta^{\mathcal{B}}(1|0) - \theta^{\mathcal{B}}(t)\, p_\theta^{\mathcal{B}}(0|1) \\ \left(1 - \theta^{\mathcal{R}}(t)\right) p_\theta^{\mathcal{R}}(1|0) - \theta^{\mathcal{R}}(t)\, p_\theta^{\mathcal{R}}(0|1) \end{bmatrix}, \tag{4}
$$

where $\dot{\theta}^{\mathcal{B}}(t), \dot{\theta}^{\mathcal{R}}(t)$ are the rates of change of $\theta^{\mathcal{B}}(t), \theta^{\mathcal{R}}(t)$, respectively.

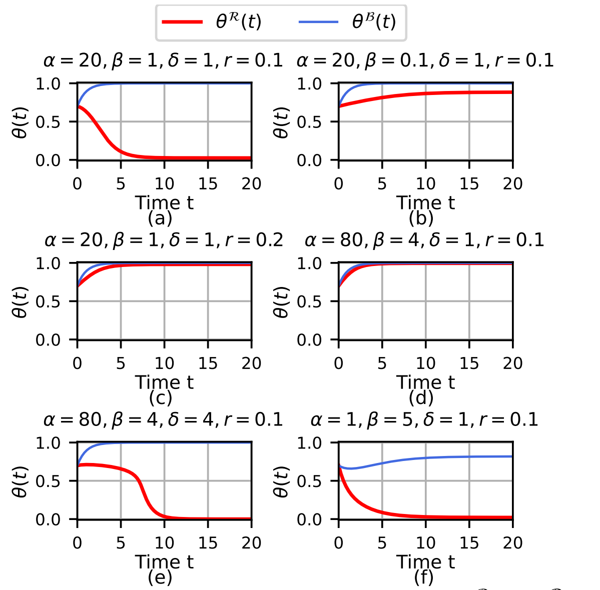



**Figure 1: Example trajectories of the $\theta(t) = \left[\theta^{\mathcal{B}}(t), \theta^{\mathcal{R}}(t)\right]$ on a fully connected graph based on Eq. (4) (under the first definition of peer influence) for various parameter configurations. In each case, it is assumed that $\theta^{\mathcal{B}}(0) = \theta^{\mathcal{R}}(0)$ i.e., both groups initially have the same prevalence of the dynamic attribute.**

The trajectories of the differential equation in Eq. (4) can easily be obtained numerically.[1] Fig. (1) shows example scenarios where both red and blue groups initially have the same prevalences of dynamic attribute (i.e., $\theta^{\mathcal{B}}(0) = \theta^{\mathcal{R}}(0)$). Several observations can be made from Fig. 1.

First, comparing Fig. (1)(a) and Fig. (1)(b) shows that decreasing the out-group hate $\beta$ (compared to in-group love $\alpha$) facilitates consensus where both groups largely adopt the same stance over time (i.e., $\theta^{\mathcal{B}}(t) \approx \theta^{\mathcal{R}}(t) \approx 1$ or $\theta^{\mathcal{B}}(t) \approx \theta^{\mathcal{R}}(t) \approx 0$, for large $t$). Importantly, the values of $\alpha, \beta$ matter and not just their ratio. For example, Fig. (1)(a) and Fig. (1)(d) both have the same $\alpha$ to $\beta$ ratio, and yet correspond to two different outcomes (partisan polarization and consensus).[2]

Second, more balanced group sizes (i.e., $r$ closer to 0.5) facilitate consensus as seen by comparing Fig. (1)(a) and Fig. (1)(c). Comparing Fig. (1)(d) with Fig. (1)(e) shows how inertia (while all other parameters kept same) can affect the dynamics of polarization.

[1]Using the Euler method, we can get the discretized trajectory as,

$$\theta^{\mathcal{B}}(k\epsilon + \epsilon) = \theta^{\mathcal{B}}(k\epsilon) + \epsilon\dot{\theta}^{\mathcal{B}}(k\epsilon) \tag{5}$$
$$\theta^{\mathcal{R}}(k\epsilon + \epsilon) = \theta^{\mathcal{R}}(k\epsilon) + \epsilon\dot{\theta}^{\mathcal{R}}(k\epsilon) \tag{6}$$

at time steps $t = k\epsilon$ where $k = 0, 1, 2, \ldots$ for some small $\epsilon > 0$.

[2]In the model proposed in [37], only $\alpha/\beta$ and $r/(1-r)$ determine the outcomes over a long period of time. The long-term outcomes of the model proposed in [37] is limited to exact consensus (where $\theta^{\mathcal{B}}(t) = \theta^{\mathcal{R}}(t) = 0$ or $\theta^{\mathcal{B}}(t) = \theta^{\mathcal{R}}(t) = 1$, as $t$ goes to infinity), exact partisan polarization (where $\theta^{\mathcal{B}}(t) = 1, \theta^{\mathcal{R}}(t) = 0$ or $\theta^{\mathcal{B}}(t) = 0, \theta^{\mathcal{R}}(t) = 1$, as $t$ goes to infinity) or non-partisan polarization where each group will be split evenly (where $\theta^{\mathcal{B}}(t) = \theta^{\mathcal{R}}(t) = 0.5$ as $t$ goes to infinity). In contrast, the generalized model presented in this paper is able to capture scenarios in between as seen from Fig. (1) and Fig. (2).

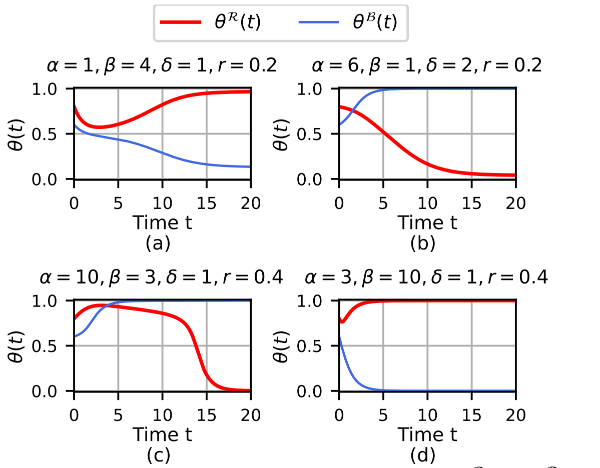



**Figure 2: Example trajectories of the $\theta(t) = \left[\theta^{\mathcal{B}}(t), \theta^{\mathcal{R}}(t)\right]$ on a fully connected graph based on Eq.** (4) **(under the first definition of peer influence) for various parameter configurations. Unlike Fig. 1, the two groups initially have different prevalences of the dynamic attribute.**

Fig. (2) shows several scenarios where the two groups have different initial states (i.e., $\theta^{\mathcal{B}}(0) \neq \theta^{\mathcal{R}}(0)$). In particular, Fig. (2)(b) shows an example of a cross-over where the minority (red) group ends up largely giving up the stance-1 despite that being initially more prevalent compared to the blue group. Fig. (2)(c) shows an example of the minority red group (red) reversing the trend when the two groups are about to reach consensus.

Interestingly, under the second influence measure specified by Definition 2, both groups follow an identical trajectory when the initial states are same for both groups. This can be seen from Fig. 10 in the Appendix. Intuitively, since the initial fractions of stances are the same in each group, both the in-group and out-group effects for each group remains symmetric and therefore the trajectories are identical. When the initial conditions for the two groups are different, the dynamics under Definition 2 can show a rich array of phenomena as seen from Fig. (11) in the Appendix.

As Fig. 1 and Fig. 2 illustrate, the proposed model can replicate a wide-array of phenomena observed in real-world even on a fully connected network. The trajectories for fully connected networks can be obtained directly from Eq. (4) and the model can easily be simulated on any arbitrary network (not necessarily fully connected) as well. An example with a Facebook network dataset is discussed in Appendix C.4.

## 3.3 Extending the Model beyond Two Parties

The model can be extended to systems with more than two parties to account for finer-grained social divisions. For example, one could partition the US population in to five non-overlapping groups based on their political ideological leaning: hard-left, left, moderate, right, hard-right. To model such systems, let $A_{ij}$ be the emotion of group $i$ towards group $j$ (where $i, j \in \{1, 2, \ldots, N\}$ and $N$ is the total number of groups). If $A_{ij} > 0$ (resp. $A_{ij} < 0$), then the individuals of group $i$ have a positive (resp. negative) emotion towards group $j$. Assume that the random node chosen from the population belongs at time $k$, $X_k$, belongs to group $i$. Then, the transition probabilities of the node $X_k$ can be expressed as

$$\begin{aligned} p_{X_k,k}(1|0) &= \frac{1}{1 + \exp\left[-\left(\sum_{j=1}^{N} A_{ij}\Delta_k^{j,1}(X_k) - \delta\right)\right]} \\ p_{X_k,k}(0|1) &= \frac{1}{1 + \exp\left[-\left(\sum_{j=1}^{N} A_{ij}\Delta_k^{j,0}(X_k) - \delta\right)\right]}, \end{aligned} \tag{7}$$

where $\Delta_k^{j,1}(X_k), \Delta_k^{j,0}(X_k)$ denote measures that quantify the prevalence of stance-1 and stance-0, respectively, among the neighbors of $X_k$ that belong to group $j$, and $\delta_i$ is the inertia of individuals belonging to group $i$.

Let $\theta^{(i)}(t)$ denote the fraction of individuals in group-$i$ with stance-1 at time $t$, $r_i$ denote the number of group-$i$ individuals as a fraction of the population, and $\delta_i$ denote the inertia of group-$i$ individuals. Then, the mean-field dynamics of the multi-party system (using the first influence measure) can be expressed as,

$$\frac{d\theta^{(i)}(t)}{dt} = (1-\theta^{(i)}(t)) \cdot p_{01}^{(i)}(t) - \theta^{(i)}(t) \cdot p_{10}^{(i)}(t), \quad i = 1, \ldots, N, \tag{8}$$

where,

$$\begin{aligned} p_{01}^{(i)}(t) &= \frac{1}{1 + \exp\left(-\left(\sum_{j=1}^{N} A_{ij} r_j (2\theta^{(j)}(t) - 1) - \delta_i\right)\right)} \\ p_{10}^{(i)}(t) &= \frac{1}{1 + \exp\left(-\left(\sum_{j=1}^{N} A_{ij} r_j (1 - 2\theta^{(j)}(t)) - \delta_i\right)\right)}. \end{aligned}$$

Similar to the two party system, the differential equation Eq. (8) which governs the dynamics in multi-party populations can be numerically simulated to obtain its trajectory. Two example scenarios under the first influence measure are discussed in Appendix E.

## 4 Estimating Affective Polarization

We illustrate how the parameters of the above model can be estimated from social media data. We present two approaches for two levels of granularity of the available data: stances (dynamic attribute) and parties (static attribute) of a set of sampled nodes as well as their neighbors are known (case 1) and stances and party of a set of randomly sampled nodes available (case 2).

***Case 1: Stances of Sampled Egos and their Neighbors are Known***
When the stance as well as neighbor influence measures are known for a sample of ego nodes $v_i, i = 1, 2, \ldots, n$, model parameters can be estimated without any further assumptions. Specifically, we assume that the peer influences (based on in-group and out-group neighbors with each stance) at some time instant $k_i$, $\Delta_{k_i}^{in,1}(v_i)$, $\Delta_{k_i}^{out,1}(v_i)$, $\Delta_{k_i}^{in,0}(v_i)$, $\Delta_{k_i}^{out,0}(v_i)$, are known for each node $v_i, i = 1, 2, \ldots, n$. Further, the stance of each ego node in the sample $v_i, i = 1, 2, \ldots, n$ at two consecutive time instants, $k_i$ and $k_i + 1$ (i.e., $H_{k_i}(v_i), H_{k_i+1}(v_i)$), are also known. Then, logistic regression can be utilized to estimate the parameters $\alpha, \beta, \delta$ as shown in Algorithm 1 (Refer Appendix C.2). It leverages the logistic model Eq. 3, ensuring that all theoretical properties of logistic regression via maximum likelihood are preserved. A key practical point is that the same node can appear in observations at different time intervals, allowing monitoring of a few nodes' stances and influence over time as sufficient input for

Algorithm 1. Additionally, no assumptions about the network structure are required to estimate $\alpha, \beta, \delta$ with theoretical guarantees.

***Case 2: Only Stances of Sampled Egos are Known***

While collecting stances from a random set of ego nodes over time is feasible in most practical settings, obtaining peer influence measures is difficult due to the need for network structure. In such cases, the random set of ego nodes can be treated as a sample from a fully connected network. For Algorithm 1, this means that the influence measures $\Delta_{k_i}^{in,1}(v_i)$, $\Delta_{k_i}^{out,1}(v_i)$, $\Delta_{k_i}^{in,0}(v_i)$, $\Delta_{k_i}^{out,0}(v_i)$ are calculated by viewing each $v_j$, $j \neq i$ in the sample as a neighbor of $v_i$.

## 5 Empirical Validation

We now estimate the model parameters using real-world discussions about contentious issues on Twitter. We then use the model calibrated with the estimated parameters show that the model dynamics align closely with empirical observations and derive further insights

### 5.1 Data and Issue Detection

We use publicly available data consisting of 1.4 billion tweets related to the COVID-19 pandemic posted between January 21, 2020, and November 4, 2021 [5]. These tweets were collected using pandemic-relevant keywords and geolocated within the US with Carmen [8], a geo-location tool for Twitter data. Carmen leverages tweet metadata, including "place" and "coordinates" objects, as well as mentions of locations in users' bios, to assign tweets to specific locations at state and county level (see Appendix F.3). We restrict our focus to tweets between January 21, 2020 and January 1, 2021. This leaves us with 230M tweets from 8.7M users in the US.

Prophylactic measures like stay-at-home orders and masking recommendations dominated the early online discussions about the pandemic [6, 40], growing contentious [29, 38]. The *masking* issue encompasses posts about the use of face coverings, mask mandates, mask shortages, and anti-mask sentiments. The *lockdowns* issue includes discussions of state and federal mitigation efforts, such as quarantines, stay-at-home orders, business closures, reopening strategies, calls for social distancing and access to transportation. To identify posts relevant to these issues, we employ a weakly-supervised approach which mines Wikipedia pages related to each issue for relevant keywords and phrases [38]. After manually verifying the extracted terms, we filter the tweets for posts that mention these issues. We identify 11.4M tweets as relevant to masking and 8.3M tweets as relevant to lockdowns.

### 5.2 User Ideology Classification

To identify ideology, we use a two-phase approach. In Phase 1, we analyze retweet interactions between users and a curated list of 2,200 political elites with known ideologies (liberal: −1, conservative: 1). Excluding interactions with fewer than 10 occurrences, we map 92,000 users to a latent ideological space using methods from [2]. Retweets are modeled as agreement, with similar accounts positioned closer in the space. The model incorporates elite popularity, user political interest, and retweet behavior, estimating the likelihood of interaction as:

$$p(y_{i,j} = 1 \mid \lambda_j, \eta_i, \gamma, \phi_j, \theta_i) = \text{logit}^{-1}(\lambda_j + \eta_i - \gamma * (\phi_j - \theta_i)^2), \quad (9)$$

where $\lambda_j$, $\eta_i$ represent elite popularity and user interest, $\phi_j$ and $\theta_i$ denote elite and user ideologies, and $\gamma$ is a regularizing constant. Bayesian inference via PyMC3 with GPU acceleration is used to estimate parameters, drawing 1,000 samples across four chains after a 500-iteration warm-up. The resulting ideology scores for 94,000 users strongly agree with follower-based methods (Pearson $r = 0.89$). This enables us to estimate continuous ideology scores for all users in the bipartite network of elite interactions. We call these users *political partisans*. After binarizing scores (liberal $\leq 0$, conservative $> 0$), we achieve an F1-score of 0.95 for classification.

In Phase 2, we extend classification using supervised fine-tuning of the LLaMA 3.1-8B Instruct model. Aggregating tweets from 8.7M users (January–December 2020), we train the model with 94$K$ users' tweets and Phase 1 ideology estimates. The fine-tuned model achieves an F1-score of 0.98 on the test set, classifying 6.6$M$ liberal and 2.1$M$ conservative accounts. Results are validated against state-level Republican vote shares in the 2020 election (Appendix F.1). A detailed description of the ideology classification is provided in Appendix A.

### 5.3 Stance Classification

Pro-masking tweets advocate masks as effective public health measures, while anti-masking tweets argue masks harm well-being and infringe on liberty. Neutral tweets often share news or unrelated content. Similarly, pro-lockdown tweets highlight lockdowns as essential for controlling COVID-19, whereas anti-lockdown tweets emphasize economic and mental health harms and threats to freedom. Neutral tweets share news without a clear stance (see Table 4 in the Appendix).

We fine-tune a LLaMA 3.1-8B Instruct model with Low-Rank Adaptation (LoRA) [23] to classify tweet stances. Training data from [17] includes 1,921 annotated tweets on masking and 1,717 on lockdowns, categorized as favor, against, or neutral/irrelevant. We use 80% for training and 10% each for validation and testing. LoRA enables efficient fine-tuning of large models with minimal parameter adjustment, preserving pre-trained knowledge for faster convergence. Tweets are input with the prompt: `What is the stance expressed towards masking (resp. lockdown mandates) in the following tweet?`, with ground-truth stances from [17].

The fine-tuned model infers stances for 11.4$M$ masking-related tweets and 8.3$M$ lockdown-related tweets. Among masking tweets, 6.7$M$ are pro-masking, 1.2$M$ are anti-masking, and 3.2$M$ are neutral. For lockdown tweets, 2.1$M$ favor lockdowns, 1.4$M$ oppose them, and 4.8$M$ are neutral. Classifier performance on test sets achieves F1-scores of 0.91 (masking) and 0.85 (lockdowns). Masking stances strongly correlate with New York Times state-level masking surveys (Pearson $r = 0.63$, see Appendix F.1).

A user is classified as pro-masking (or pro-lockdowns) if their average stance across all tweets on the issue exceeds 0.5, otherwise as anti-masking. Figs. 3 (a) and (b) show the share of pro-masking users over time by user ideology, for all users and partisans. Before the CDC's April 3, 2020 masking recommendation, liberals and conservatives supported masking at similar rates. Afterward, a partisan gap emerged, with a sharp decline in conservative support,

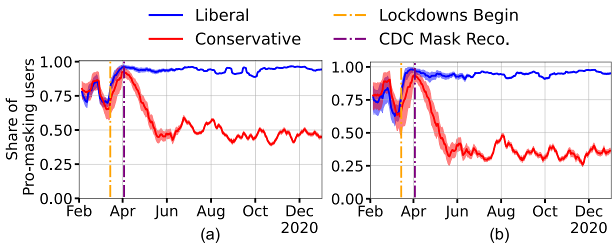



**Figure 3: Masking stances over time. Share of liberal and conservative users expressing pro-masking stance for (a) all users and (b) partisans.**

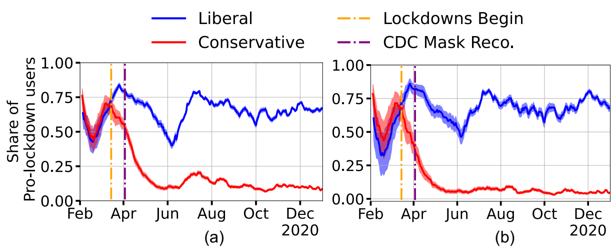



**Figure 4: Lockdown stances over time. Share of liberal and conservative users expressing pro-lockdown stance for (a) all users and (b) partisans.**

**Table 1: Affective Polarization Parameters Estimated via Logistic Regression from Social Media Data.**

| *Issue* | | *All users* | *Pseudo-R2* | *Partisans* | *Pseudo-R2* |
|---|---|---|---|---|---|
| | $\alpha$ | 3.75±0.008 | | 5.11 ± 0.018 | |
| *Masking* | $\beta$ | 0.25±0.005 | 0.31 | 0.63± 0.007 | 0.38 |
| | $\delta$ | 0.63±0.003 | | 0.28 ± 0.005 | |
| | $\alpha$ | 3.76 ± 0.020 | | 5.08± 0.032 | |
| *Lockdowns* | $\beta$ | 0.75± 0.017 | 0.15 | 1.05 ± 0.027 | 0.23 |
| | $\delta$ | 0.91 ± 0.004 | | 0.80 ± 0.007 | |

especially among conservative partisans (Figures 3(b)). Figs. 4(a) and (b) show the share of pro-lockdown users split by ideology for all users and partisans. Like masking, the novel issue of lockdowns quickly became polarized. Before U.S. stay-at-home orders (around March 15, 2020), there were no significant differences in attitudes of liberals and conservatives. Afterward, pro-lockdown sentiment sharply declined among conservatives, especially partisans. Liberals' support for lockdowns gradually declined until June 2020, then spiked sharply.

## 5.4 Model Calibration via Parameter Estimation

We use the method described under Case 2 in Section 4 to estimate affective polarization parameters. which assumes a fully connected network. Under this assumption, the in-group of liberal users includes all other liberals, and their out-group includes all conservatives, and vice versa. The number of active users varies over time, as not all users are consistently engaged on Twitter, resulting in sparse time series. To address this challenge, we divide the timeline into 24 intervals of 15 days. A user is considered active if they post

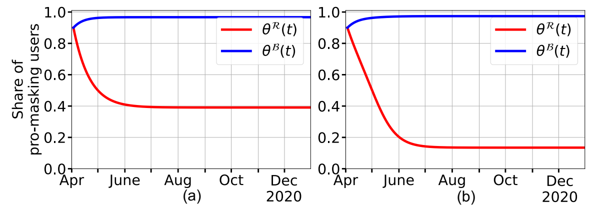



**Figure 5: Plotting trajectories of issue positions on masking for (a) all users and (b) political partisans.**

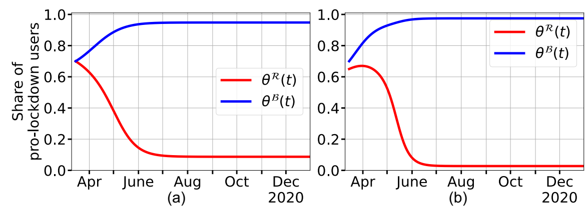



**Figure 6: Plotting trajectories of issue positions on lockdowns for (a) all users and (b) political partisans.**

within that interval. We ensure that the same users are active in two consecutive intervals; e.g., intervals $t = i$ and $t = i + 1$ share the same users, which may be different from users active during intervals $t = i + 1$ and $t = i + 2$. This allows us to assess the changes in user stances between two consecutive intervals based on the stances of the user's in-group and out-group members.

The results of the estimated parameters using the above approach (see Appendix F.2 for more details and Appendix C.4 for a comparison with a baseline method) are shown in Table 1. We observe that for both issues—masking and lockdowns— $\alpha > \beta$, suggesting that opinions within a user's in-group have a greater impact on their own stance than changes in views among the out-group. However, positive $\beta$ values suggest that out-group stances influence users' own positions on masking and lockdowns. We observe that the estimated values of $\alpha$ and $\beta$ are higher for political partisans compared to all users, suggesting that their positions are more strongly influenced by in-group favoritism and out-group hostility. Additionally, the relatively high Pseudo-R2 values for both issues and user groups indicate that these parameters account for a significant portion of the variation in stance transitions over time.

## 5.5 Model Solutions with Estimated Parameters

Given three estimated parameters ($\alpha$, $\beta$, $\delta$), and three parameters measured from data— $r$, $\theta^{\mathcal{B}}(0), \theta^{\mathcal{R}}(0)$, where $r$ denotes the fraction of conservative users, and $\theta^{\mathcal{B}}(0), \theta^{\mathcal{R}}(0)$ are the initial share of liberals ($\mathcal{B}$) and conservatives ($\mathcal{R}$) who are pro-masking (resp. lockdowns)—we solve the ODE in Eq. (4) numerically to obtain the trajectories of the system. We set March 15, 2020 as the initial time ($t = 0$) for the lockdowns issue ($\theta^{\mathcal{B}}(0) = \theta^{\mathcal{R}}(0) = 0.7$) and April 3, 2020 for the masking issue ($\theta^{\mathcal{B}}(0) = \theta^{\mathcal{R}}(0) = 0.9$). We chose these starting points as they denote key events—CDC masking recommendation on April 3, 2020 and lockdown orders on March 15, 2020—following which the attitudes of liberals and conservatives diverged. The share of conservatives among all users at $t = 0$ discussing masking (resp. lockdowns) was 18% (resp. 43%).

The share of conservatives among all partisans was 34% and 49% for masking and lockdowns respectively.

By numerically solving the differential equation in Eq. (4) using Euler's method (with discretized time intervals of 0.01), we compute $\theta^{\mathcal{B}}(t)$ and $\theta^{\mathcal{R}}(t)$ from $t = 0$ to 100, which denote the share of users who are liberal pro-masking (resp. lockdowns) and conservative pro-masking (resp. lockdowns), respectively. Figs. 5 (masking) and 6 (lockdowns) show results with a time axis scaled to empirical data, where each step equals 7 days.

The model trajectories shown in Figs. 5 and 6 correspond to real-world trajectories in Figs. 3 and 4. In particular, the model correctly reproduces a larger gap among active partisans than among all users, and also higher approval of masking among conservatives unlike lockdowns. However, the model overestimates approval of lockdowns among liberals, compared to real-world data Figs. 3 and 4). The discrepancy may arise from simplifying assumptions like a fully connected network and synchronous user transitions.

Prior studies [3, 4, 7, 42] suggest that homophily alone may explain the development of issue polarization. Along with estimating homophily on individual issue stance, captured by in-group love ($\alpha$), we also estimate the impact of the out-group ($\beta$). The parameter $\delta$, as previously discussed, measures the user's resistance to change. We argue that out-group dynamics is necessary for polarization, especially in fully connected networks. To illustrate this, we plot the evolution of issue stances by varying $\beta$ relative to $\alpha$ (using values of $\alpha$ estimated for all users in Table 1). The trajectories are shown in Appendix Figs. 7 and 8. When considering only in-group love (homophily), with $\alpha \neq 0$ and $\beta = 0$, we find that liberals and conservatives reach near-consensus or remain weakly polarized (Figs. 7(d) and 8(d)). When group differences disappear and out-group attitude shifts to out-group love ($\beta \to -\alpha$), both groups converge to consensus, becoming pro-masking (or pro-lockdowns) (Figs. 7(a-c) and 8(a-c)). Conversely, as out-group animosity grows ($\beta \to \alpha$), groups diverge along ideological lines.

In the absence of in-group love ($\alpha = 0$ and $\beta \neq 0$) both groups converge to $\theta^{B}(t) = \theta^{R}(t) \to 0.5$, indicating half the group holds one belief and half the other. Even if both groups started out with a majority of members supporting masking (or lockdowns), without in-group love they evolve into this undecided state. However, if the majority of one group supported masking along with a minority of the other group ($\theta^{B}(0) > 0.5$ & $\theta^{R}(0) < 0.5$), we would still witness polarization under sufficiently high $\beta$.

### 5.6 State-level Variation of Parameters

We uncover systematic geographic differences in the associations between affective polarization and ideology. Given user's inferred state-level location (determined with Carmen [8]), we separately estimate affective polarization parameters for each state across both issues. While user nodes are restricted to their respective states, their exposures are not limited i.e., we assume that they are exposed to users from across the U.S. We then examine the relationship between a state's conservatism (measured by Republican vote share in the 2020 Federal elections) and three parameters—$\alpha$, $\beta$, and $\delta$—for both issues. Appendix Figure 15(a-c) presents the correlation between these parameters and vote share for masking, and Appendix Figure 15(d-f) for lockdowns. For the masking issue, we find that users in more conservatives states are increasingly less influenced by their in-group and more influenced by their out-group ($\alpha$ falls but $\beta$ rises with Republican vote share), and they become more susceptible to social influence ($\delta$ decreases with Republican vote share). In contrast, for lockdowns, users in more conservative states are more influenced by the in-group ($\alpha$ increases) and less susceptible to social influence ($\delta$ increases). Issue-specific geographic differences in out-group hate and in-group love highlight the need to consider both in polarization assessments.

## 6 Conclusion

We model opinion polarization in an emotionally divided society and propose a framework to estimate in-group love and out-group hate. Unlike homophily-based models [3, 4, 7], we show that emotional dynamics alone can polarize even a fully mixed society. Using COVID-19 discussions on masking and lockdowns, we calibrate the model by estimating its three parameters from data. We show that the calibrated models reproduce observed polarization levels, capturing ideological divides and partisan gaps. Results demonstrate the model's robustness and provide a quantitative explanation of societal divisions. Unlike previous findings that emphasize out-group hate [10, 13], our results suggest in-group love is the stronger force—had it been weaker or group differences absent, consensus would have emerged on both issues.

***Future Work and Limitations:*** While the proposed model and estimation framework does not assume any specific graph structure, our theoretical analysis and parameter estimation assumed a fully connected graph due to difficulties in collecting social media data. Future studies could relax this assumption and analyze the polarization dynamics and estimate parameters in more structurally rich, sparse networks. It is worth noting that the estimated trajectories were slightly overestimated, possibly due to this simplifying assumption. Further, previous work has shown that sampling biases such as the friendship paradox [21] and its variants (e.g., [1, 16]) can be exploited to improve the estimation accuracy of various parameters [34–36]. Similar approaches may be devised to sample the network non-uniformly to reduce the variance of the estimates of in-group love and out-group hate. Finally, robustness of parameter estimation to the choice of the time interval also has to be explored. Our findings emphasize the crucial role of interactions between identity (e.g., ideology) and belief formation. Further research should explore intersectional identities, multi-way interactions, and diverse issue positions within a pluralistic society. We focused on two polarizing issues during the Covid-19 pandemic, which were driven by urgent circumstances. However, traditional wedge issues such as abortion rights, gun control, LGBTQ+ rights, immigration, and racial/social justice have polarized over longer periods. Assessing the applicability of our models and empirical frameworks to these issues could be key to ensuring their generalizability. Despite the validation accuracy of our estimates for issue positions and ideology, they may still be inconsistencies. Relying on large language model for these classifications may introduce biases that could contribute to these inconsistencies.

## Acknowledgments

This work was funded in part by the National Science Foundation under grant CCF 2200256.

## References


[1] Nazanin Alipourfard, Buddhika Nettasinghe, Andrés Abeliuk, Vikram Krishnamurthy, and Kristina Lerman. 2020. Friendship paradox biases perceptions in directed networks. *Nature communications* 11, 1 (2020), 707.

[2] Pablo Barberá. 2015. Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data. *Political analysis* 23, 1 (2015), 76–91.

[3] Fabian Baumann, Philipp Lorenz-Spreen, Igor M Sokolov, and Michele Starnini. 2020. Modeling echo chambers and polarization dynamics in social networks. *Physical Review Letters* 124, 4 (2020), 048301.

[4] Francesco Bonchi, Edoardo Galimberti, Aristides Gionis, Bruno Ordozgoiti, and Giancarlo Ruffo. 2019. Discovering Polarized Communities in Signed Networks. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management* (Beijing, China) *(CIKM '19)*. Association for Computing Machinery, New York, NY, USA, 961–970. https://doi.org/10.1145/3357384.3357977

[5] Emily Chen, Kristina Lerman, Emilio Ferrara, et al. 2020. Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set. *JMIR public health and surveillance* 6, 2 (2020), e19273.

[6] Aidan Connaughton. 2021. Those on ideological right favor fewer COVID-19 restrictions in most advanced economies. Pew Research Center. Available at: https://www.pewresearch.org/lockdown-continues-across-sydney-as-nsw-continues-to-record-new-covid-19-cases/. Accessed: 2024-04-11.

[7] Morris H DeGroot. 1974. Reaching a consensus. *Journal of the American Statistical association* 69, 345 (1974), 118–121.

[8] Mark Dredze, Michael J Paul, Shane Bergsma, and Hieu Tran. 2013. Carmen: A twitter geolocation system with applications to public health. In *Workshops at the twenty-seventh AAAI conference on artificial intelligence.* Association for the Advancement of Artificial Intelligence.

[9] James Druckman, Samara Klar, Yanna Kkrupnikov, Matthew Levendusky, and John Barry Ryan. 2020. The political impact of affective polarization: how partisan animus shapes COVID-19 attitudes. (2020).

[10] James N Druckman and Matthew S Levendusky. 2019. What do we measure when we measure affective polarization? *Public opinion quarterly* 83, 1 (2019), 114–122.

[11] James N Druckman and Jeremy Levy. 2022. 18. Affective polarization in the American public. *Handbook on politics and public opinion* (2022), 257.

[12] James N Druckman and Jeremy Levy. 2022. Affective polarization in the American public. In *Handbook on politics and public opinion.* Edward Elgar Publishing, 257–270.

[13] Eli J Finkel, Christopher A Bail, Mina Cikara, Peter H Ditto, Shanto Iyengar, Samara Klar, Lilliana Mason, Mary C McGrath, Brendan Nyhan, David G Rand, et al. 2020. Political sectarianism in America. *Science* 370, 6516 (2020), 533–536.

[14] Emma Fraxanet, Max Pellert, Simon Schweighofer, Vicenç Gómez, and David Garcia. 2024. Unpacking polarization: Antagonism and alignment in signed networks of online interaction. *PNAS Nexus* (2024), pgae276.

[15] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2018. Quantifying controversy on social media. *ACM Transactions on Social Computing* 1, 1 (2018), 1–27.

[16] Amir Ghasemian and Nicholas A Christakis. 2023. The enmity paradox. *Scientific Reports* 13, 1 (2023), 20040.

[17] Kyle Glandt, Sarthak Khanal, Yingjie Li, Doina Caragea, and Cornelia Caragea. 2021. Stance detection in COVID-19 tweets. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (long papers)*, Vol. 1.

[18] Anton Gollwitzer, Cameron Martel, William J Brady, Philip Pärnamets, Isaac G Freedman, Eric D Knowles, and Jay J Van Bavel. 2020. Partisan differences in physical distancing are linked to health outcomes during the COVID-19 pandemic. *Nature human behaviour* 4, 11 (2020), 1186–1197.

[19] Jon Green, Jared Edgerton, Daniel Naftel, Kelsey Shoub, and Skyler J Cranmer. 2020. Elusive consensus: Polarization in elite communication on the COVID-19 pandemic. *Science advances* 6, 28 (2020), eabc2717.

[20] Guy Grossman, Soojong Kim, Jonah M Rexer, and Harsha Thirumurthy. 2020. Political partisanship influences behavioral responses to governors' recommendations for COVID-19 prevention in the United States. *Proceedings of the National Academy of Sciences* 117, 39 (2020), 24144–24153.

[21] Nathan Hodas, Farshad Kooti, and Kristina Lerman. 2013. Friendship paradox redux: Your friends are more interesting than you. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 7. 225–233.

[22] Tad Hogg and Kristina Lerman. 2009. Stochastic models of user-contributory web sites. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 3. 50–57.

[23] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685* (2021).

[24] Shanto Iyengar, Yphtach Lelkes, Matthew Levendusky, Neil Malhotra, and Sean J Westwood. 2019. The origins and consequences of affective polarization in the United States. *Annual review of political science* 22 (2019), 129–146.

[25] Shanto Iyengar, Gaurav Sood, and Yphtach Lelkes. 2012. Affect, not ideology: A social identity perspective on polarization. *Public opinion quarterly* 76, 3 (2012), 405–431.

[26] Shanto Iyengar and Sean J Westwood. 2015. Fear and loathing across party lines: New evidence on group polarization. *American journal of political science* 59, 3 (2015), 690–707.

[27] Julie Jiang, Emily Chen, Shen Yan, Kristina Lerman, and Emilio Ferrara. 2020. Political polarization drives online conversations about COVID-19 in the United States. *Human Behavior and Emerging Technologies* 2, 3 (2020), 200–211.

[28] Jun Lang, Wesley W Erickson, and Zhuo Jing-Schmidt. 2021. # MaskOn!# MaskOff! Digital polarization of mask-wearing in the United States during COVID-19. *PloS one* 16, 4 (2021), e0250817.

[29] Kristina Lerman, Dan Feldman, Zihao He, and Ashwin Rao. 2024. Affective polarization and dynamics of information spread in online networks. *npj Complexity* 1, 1 (2024), 8.

[30] Jure Leskovec and Julian Mcauley. 2012. Learning to discover social circles in ego networks. *Advances in neural information processing systems* 25 (2012).

[31] Matthew S Levendusky. 2018. Americans, not partisans: Can priming American national identity reduce affective polarization? *The Journal of Politics* 80, 1 (2018), 59–70.

[32] Nonna Mayer. 2011. Why extremes don't meet: Le Pen and Besancenot voters in the 2007 French presidential election. *French Politics, Culture & Society* 29, 3 (2011), 101–120.

[33] Stefan McCabe, Jon Green, Pranav Goel, and David Lazer. 2023. Inequalities in Online Representation: Who Follows Their Own Member of Congress on Twitter? *Journal of Quantitative Description: Digital Media* 3 (2023).

[34] Buddhika Nettasinghe, Kowe Kadoma, Mor Naaman, and Vikram Krishnamurthy. 2024. Estimating Exposure to Information on Social Networks. *ACM transactions on social computing* 7, 1-4 (2024), 1–24.

[35] Buddhika Nettasinghe and Vikram Krishnamurthy. 2021. " What Do Your Friends Think?": Efficient Polling Methods for Networks Using Friendship Paradox. *IEEE transactions on knowledge and data engineering* 33, 3 (2021), 1291–1305.

[36] Buddhika Nettasinghe and Vikram Krishnamurthy. 2021. Maximum likelihood estimation of power-law degree distributions via friendship paradox-based sampling. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 15, 6 (2021), 1–28.

[37] Buddhika Nettasinghe, Allon G Percus, and Kristina Lerman. 2024. Dynamics of Affective Polarization: From Consensus to Partisan Divides. *arXiv preprint arXiv:2403.16940* (2024).

[38] Ashwin Rao, Siyi Guo, Sze Yuh Nina Wang, Fred Morstatter, and Kristina Lerman. 2023. Pandemic Culture Wars: Partisan Differences in the Moral Language of COVID-19 Discussions. In *2023 IEEE International Conference on Big Data (BigData).* IEEE, 413–422.

[39] Ashwin Rao, Fred Morstatter, Minda Hu, Emily Chen, Keith Burghardt, Emilio Ferrara, and Kristina Lerman. 2021. Political partisanship and antiscience attitudes in online discussions about COVID-19: Twitter content analysis. *Journal of medical Internet research* 23, 6 (2021), e26692.

[40] Rick Rojas. 2020. Masks Become a Flash Point in the Virus Culture Wars. New York Times. Available at: https://www.nytimes.com/2020/05/03/us/coronavirus-masks-protests.html. Accessed: 2024-04-11.

[41] Yu Tian and Renaud Lambiotte. 2024. Spreading and structural balance on signed networks. *SIAM Journal on Applied Dynamical Systems* 23, 1 (2024), 50–80.

[42] Petter Törnberg. 2022. How digital media drive affective polarization through partisan sorting. *Proceedings of the National Academy of Sciences* 119, 42 (2022), e2207159119.

[43] Matthew Tyler and Shanto Iyengar. 2023. Testing the robustness of the ANES feeling thermometer indicators of affective polarization. *American Political Science Review* (2023), 1–7.


## A Ideology Classification

To identify ideology, we use a two-phase approach. First, we reference a curated list of political elites (e.g., politicians, pundits) with ideological leanings of $-1$ for liberals and 1 for conservatives. We analyze frequent retweet interactions between regular users and these elites, creating a bipartite network and excluding interactions with fewer than 10 occurrences. This yields roughly 3 million interactions between 92,000 users and 2,200 elites. Following [2], we map users to latent ideology space by positioning ideologically similar accounts closer based on retweet behavior, as retweets often reflect agreement.

Barberá [2] leverages two other indicators—political interest and elite popularity— in addition to retweet interactions to quantify user ideology in latent space. Political interest can be estimated as

**Algorithm 1:** Estimating $\alpha$, $\beta$, and $\delta$ via Logistic Regression

**Data:** The influence measures ($\Delta_{k_i}^{in,1}(v_i)$, $\Delta_{k_i}^{out,1}(v_i)$, $\Delta_{k_i}^{in,0}(v_i)$, $\Delta_{k_i}^{out,0}(v_i)$) and stances over two adjacent time instants ($H_{k_i}(v_i), H_{k_i+1}(v_i)$) for nodes $v_i, i = 1, 2, \ldots, n$

**Result:** Estimates $\hat{\alpha}$, $\hat{\beta}$, $\hat{\delta}$

$J \leftarrow []$ $X_{\text{in}} \leftarrow []$ $X_{\text{out}} \leftarrow []$

**for** $i = 1, 2, \ldots, n$ **do**
  **if** $H_{k_i+1}(v_i) \neq H_{k_i}(v_i)$ **then**
    $J[i] = 1$;
  **end**
  **else**
    $J[i] = 0$;
  **end**
  **if** $H_{k_i}(v_i) = 0$ **then**
    $X_{\text{in}}[i] = \Delta_{k_i}^{\text{in},1}(v_i)$;
    $X_{\text{out}}[i] = \Delta_{k_i}^{\text{out},1}(v_i)$;
  **end**
  **else if** $H_{k_i}(v_i) = 1$ **then**
    $X_{\text{in}}[i] = \Delta_{k_i}^{\text{in},0}(v_i)$;
    $X_{\text{out}}[i] = \Delta_{k_i}^{\text{out},0}(v_i)$;
  **end**
**end**

**return** *Logistic regression fit via maximum likelihood:*

$\log\left(\frac{P(J=1)}{1-P(J=1)}\right) = \beta_0 + \beta_1 X_{in} + \beta_2 X_{out}$

Extract parameters: $\hat{\alpha} := \beta_1$, $\hat{\beta} := -\beta_2$, $\hat{\delta} := -\beta_0$;

the total number of tweets, retweets, replies and quoted tweets that non-elites generate over time. Elite popularity is quantified as the in-degree of elite nodes in the bipartite network. This accounts for the variation in activity of non-elites and the popularity of elites (realDonaldTrump maybe more popular than SenatorRomney). After log-transforming both features, we model the probability of user $i$ retweeting elite $j$ as a logit:

$$p(y_{i,j} = 1 \mid \lambda_j, \eta_i, \gamma, \phi_j, \theta_i) = logit^{-1}(\lambda_j + \eta_i - \gamma * (\phi_j - \theta_i)^2),$$

where $\lambda_j$, $\eta_i$ represent the popularity of elite node $j$ and political interest of user $i$ respectively, and $\phi_j$ and $\theta_i$ denote the ideology of the elite $j$ and user $i$ in the latent ideological space. Parameter $\gamma$ is a regularizing constant and $(\phi_j - \theta_i)^2$ quantifies the distance between elite $j$ and user $i$ in the latent space. The goal is to maximize the likelihood of the interaction between elite $j$ and user $i$ while minimizing the distance between them. Maximum likelihood estimation is intractable given that we have to estimate five parameters over $94K$ accounts. Instead, we use Bayesian inference. We use normal priors for $\lambda, \eta, \theta, \phi$ and half-normal prior for $\gamma$ (need for positive regularizing constant). We set initial values for $\lambda, \eta$ to be the log-transformed values for elite popularity and user political interest estimated from our dataset. We set the initial values for ideological leaning of elites $\phi$ to be the ones we obtained from [33]. We then use PyMC3 to run the No-U-Turn Sampler (NUTS) to sample from the posterior distribution of the model parameters, which include $\lambda$, $\eta, \gamma, \theta$, and $\phi$. The sampling process estimates the joint posterior distribution ($p(\lambda, \eta, \gamma, \theta, \phi \mid y)$), given the observed binary outcomes $y$. We draw a total of 1000 samples under four chains, after a warm-up period of 500 iterations for tuning. We achieve significant speedup by running the sampling on a RTXA6000 GPU node by leveraging the Numpyro Python library.

This method enables us to estimate continuous ideology scores for all users in the bipartite network of elite interactions. We call these users *political partisans*. We compare ideology scores to those estimated using the follower network-based approach in [2] for an overlap of $40K$ users and find a strong agreement (Pearson $r = 0.89$). After binarizing scores with a threshold of 0 (liberal $\leq 0$, conservative $> 0$), we achieve an F1-score of 0.95 for ideology classification.

However, most users in our dataset do not interact with political elites. In the second phase, we extend ideology classification to all users via supervised fine-tuning of the LLaMA 3.1-8B Instruct model. We compile all tweets from 8.7 million users between January 21, 2020, and December 31, 2020, creating a document for each user that aggregates all their tweets. We then use the prompt - `What is the political leaning expressed in these tweets?` We randomly select 80% of the $94K$ users as a training set, using their tweets as input to the model and the estimated ideology from the first phase as the output. After fine-tuning the model, we test it on the remaining 20% and achieve an F1-score of 0.98. Using the fine-tuned model, we then infer the ideology of all 8.7 million users in our dataset, resulting in approximately $6.6M$ liberal accounts and $2.1M$ conservative accounts. We perform further validation by comparing the share of conservative Twitter users at the state-level to the 2020 Federal Election Republican vote share for the state (Refer Appendix F.1).

# B Additional Results

Figs. 7 and 8 show the trajectories of issue stances when we vary $\beta$ relative to $\alpha$ while using values of $\alpha$ estimated for all users in Table 1. Fig.15 (a-c) shows the correlation between a state's share of Republican voters in the 2020 Federal elections and the three estimated parameters - $\alpha, \beta, \delta$ for the issue of masking. Fig.15 (d-f) highlights the same for the issue of lockdowns.

# C Additional Details about the Model

## C.1 Formal Definitions of Peer Influence

The formal definitions of the two peer influence measure are given below.

*Definition 1 (Net number of individuals with a stance normalized by degree).* Let,

$$\begin{aligned}
d_k^{in,0}(v) &= \sum_{(v,u)\in E} \mathbb{1}(R(u) = R(v) \wedge H_k(u) = 0) \\
d_k^{in,1}(v) &= \sum_{(v,u)\in E} \mathbb{1}(R(u) = R(v) \wedge H_k(u) = 1) \\
d_k^{out,0}(v) &= \sum_{(v,u)\in E} \mathbb{1}(R(u) \neq R(v) \wedge H_k(u) = 0) \\
d_k^{out,1}(v) &= \sum_{(v,u)\in E} \mathbb{1}(R(u) \neq R(v) \wedge H_k(u) = 1)
\end{aligned} \tag{10}$$

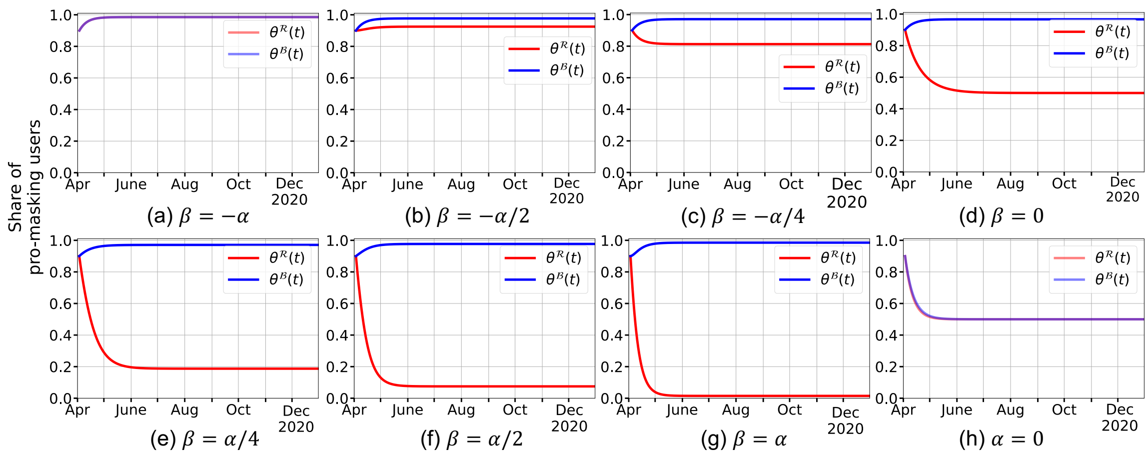



**Figure 7: Estimating trajectories of pro-masking users by ideology for varying values of $\beta$ relative to $\alpha$.**

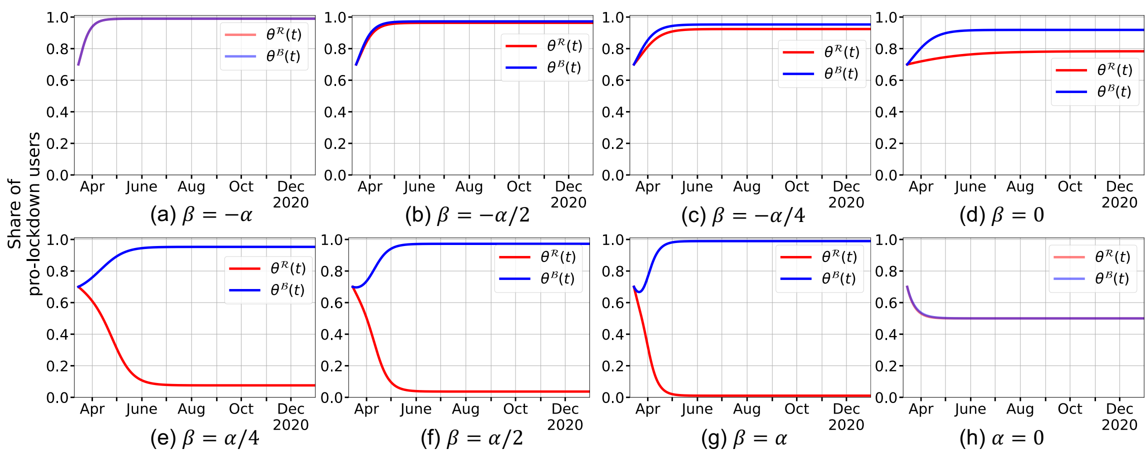



**Figure 8: Estimating trajectories of pro-lockdown users by ideology for varying values of $\beta$ relative to $\alpha$.**

denote the number of in-group and out-group neighbors of $v$ who have stance-0 and stance-1 at time $k$ on graph $G = (V, E)$, and let $d(v) = d_k^{in,0}(v) + d_k^{in,1}(v) + d_k^{out,0}(v) + d_k^{out,1}(v)$ be the degree of node $v$. Then, peer influences are defined as,

$$\Delta_k^{in,1}(v) = \frac{d_k^{in,1}(v) - d_k^{in,0}(v)}{d(v)}, \qquad \Delta_k^{in,0}(v) = -\Delta_k^{in,1}(v, 1)$$

$$\Delta_k^{out,1}(v) = \frac{d_k^{out,1}(v) - d_k^{out,0}(v)}{d(v)}, \quad \Delta_k^{out}(v, 0) = -\Delta_k^{out}(v, 1).$$

*Definition 2 (Net fraction of neighbors in each group with a stance ).* Consider the notation in Eq. 10. The peer influences are defined as,

$$\Delta_k^{in,1}(v) = \frac{d_k^{in,1}(v) - d_k^{in,0}(v)}{d_k^{in,1}(v) + d_k^{in,0}(v)}, \qquad \Delta_k^{in,0}(v) = -\Delta_k^{in,1}(v, 1)$$

$$\Delta_k^{out,1}(v) = \frac{d_k^{out,1}(v) - d_k^{out,0}(v)}{d_k^{out,1}(v) + d_k^{out,0}(v)}, \quad \Delta_k^{out}(v, 0) = -\Delta_k^{out}(v, 1).$$

Fig. 9 shows how the transition probability $p_{X_k,k}(1|0)$ in Eq. (3) varies as a sigmoid function of the linear combination of the peer influence measures. Consequently, the log-odds (log ratio of switching from 0 to 1 vs. not swicthing) varies linearly.

## C.2 Algorithm for Estimating Parameters

Algorithm 1 outlines the logistic regression approach for estimating the parameters of the proposed model. A dummy variable $J$ is used to denote whether the transition occurred or not, and that dummy variable is treated as the dependent variable.

## C.3 Comparison of Dynamics under Two Influence Measures

Fig. 10 and Fig. 11 show how the dynamics differ under the two influence measures (net number of individuals with a stance normalized by degree and net fraction of neighbors in each group with a stance).

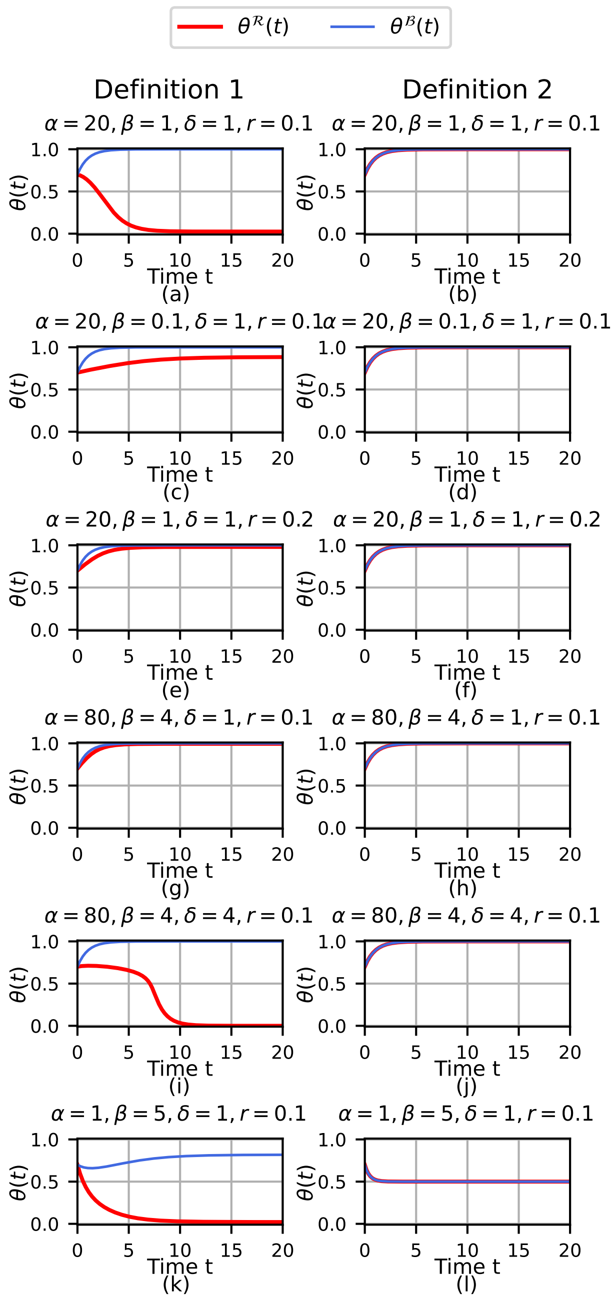



**Figure 10: Example trajectories of the $\theta(t) = \left[\theta^{\mathcal{B}}(t), \theta^{\mathcal{R}}(t)\right]$ on a fully connected graph (based on Eq. (4)) under the two definitions (the two columns) of peer influence for various parameter configurations (the four rows). In each case, it is assumed that $\theta^{\mathcal{B}}(0) = \theta^{\mathcal{R}}(0)$ i.e., both groups initially have the same prevalence of the dynamic attribute. It can be seen that under the second influence measure specified by Definition 2, both groups follow an identical trajectory when the initial states are same for both groups.**

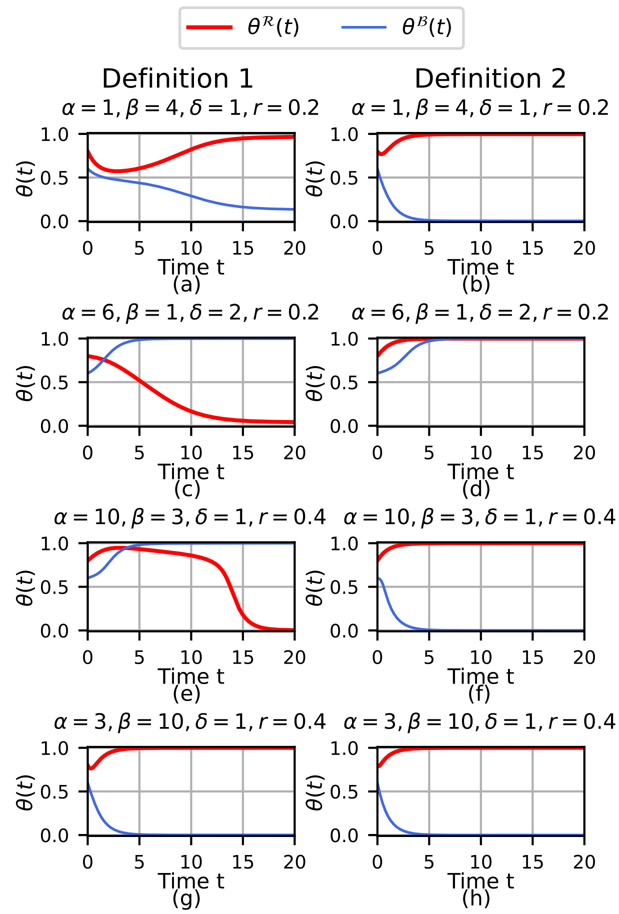



**Figure 11: Example trajectories of the $\theta(t) = \left[\theta^{\mathcal{B}}(t), \theta^{\mathcal{R}}(t)\right]$ on a fully connected graph (based on Eq. (4)) under the two definitions (the two columns) of peer influence for various parameter configurations (the four rows). Unlike Fig. 10, the two groups initially have different prevalences of the dynamic attribute i.e., $\theta^{\mathcal{B}}(0) \neq \theta^{\mathcal{R}}(0)$.**

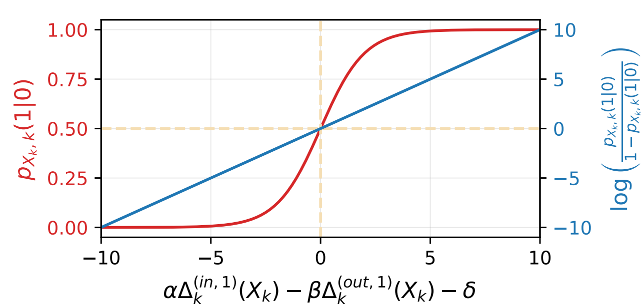



**Figure 9: The probability that a random node with stance-0 switches to stance-1 at time $k$ varies as a logistic function (red line) of the $\alpha\Delta_k^{in,1}(X_k) - \beta\Delta_k^{out,1}(X_k) - \delta$ as specified in Eq. 3. Consequently, the log-odds of switching stances vary linearly with $\alpha\Delta_k^{in,1}(X_k) - \beta\Delta_k^{out,1}(X_k) - \delta$ (blue line).**

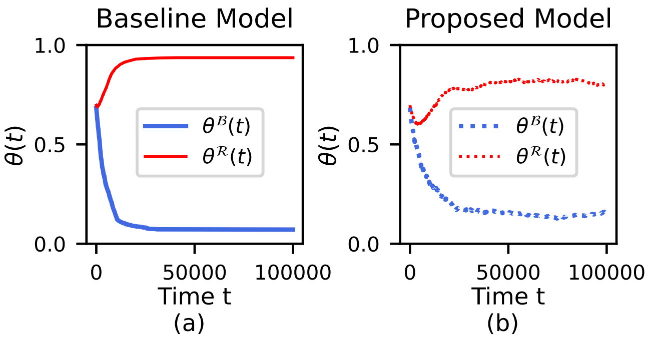



**Figure 12: Comparison of the dynamics of under the proposed model and the model in [37] on a Facebook subgraph (detailed in Sec. C.4).**

| | Ground Truth | Proposed Model | Baseline |
|---|---|---|---|
| $\alpha$ | 2.00 | 2.00 ± 0.035 | 8.37 ± 0.225 |
| $\beta$ | 6.00 | 5.94 ± 0.097 | 19.13 ± 0.757 |
| $\delta$ | 0.10 | 0.11 ± 0.017 | 1.20 ± 0.103 |

**Table 2: Comparison of performance of Algorithm 1 with the proposed model and the baseline model in [37]. While the model in [37] is theoretically tractable, the proposed model is more amenable to logistic regression based inference of the underlying parameters.**

## C.4 Comparison with the Baseline Model

To illustrate the amenability of the proposed model with logistic regression on a real network, we implement the proposed model and the baseline approach in [37] on the Facebook dataset from the SNAP repository [30]. This network contains a subgraph of Facebook containing 4039 nodes and 88234 edges. We first randomly assign 80% of the nodes in the network to be red and the remaining 20% nodes to be blue (i.e., $r = 0.8$). We then assigned 70% of the nodes of each group (chosen randomly) to have the dynamic attribute 1 initially i.e., we set $H_k(v) = 1$ at time $k = 0$ for 70% of randomly chosen nodes from each group. This leads to $\theta^{\mathcal{B}}(0) = \theta^{\mathcal{R}}(0) = 0.7$. The dynamical model described in Sec. 3.1 was then implemented on this network for 100000 iterations with $\alpha = 2, \beta = 6, \delta = 0.1$. We also implemented the model proposed in [37] with those same parameters as a baseline. The dynamics of $\theta(t)$ observed under these two models is shown in Fig. 12. Then, the logistic regression based parameter estimation Algorithm 1 was used to estimate the parameters $\alpha, \beta, \delta$ by using the sequence of sampled nodes sampled over the 1000000 time epochs where the model was implemented. The estimates obtained for the two models are given in Table 2.

It can be seen that the proposed model leads to a smaller absolute error as well as a smaller standard deviation compared the baseline model. While this is not surprising since the proposed model is amenable to logistic regression by construction, the key point is that even though the original model proposed in [37] is theoretically tractable, the model proposed in this paper is more useful for inference purposes. For inferring parameters from experiments, the model proposed in this paper can be used and then for analyzing the implications of those inferred parameters, the theoretically tractable version in [37] can also be utilized.

## D Measuring Exposure Through Tweets

In the main manuscript, we assume that users are exposed to other users, meaning they track how many users hold a particular issue position. Alternatively, one could assume that users track the number of text instances (tweets) associated with a particular group identity and expressing an issue position, rather than the number of users. As an alternative approach to modeling exposures, we rely on *Definition 2* as discussed earlier in Section 3

We leverage this alternative definition to re-estimate the parameters $\alpha, \beta$ and $\delta$. The results of this estimation are shown in Table 3. Furthermore, the trajectories of pro-masking and pro-lockdown users with liberal and conservative ideologies estimated using these parameters are shown in Figs. 7 and 8. We observe that the estimated trajectories closely resemble those in Figs. 5 and 6 suggesting that whether we measure exposure in terms of users or tweets, the resulting trajectories are similar.

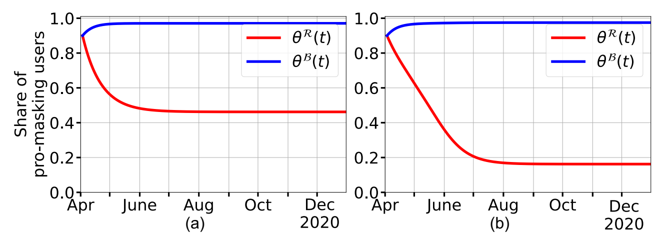



**Figure 13: Plotting trajectories of issue positions on masking for (a) all users and (b) political partisans with exposure measured through tweets.**

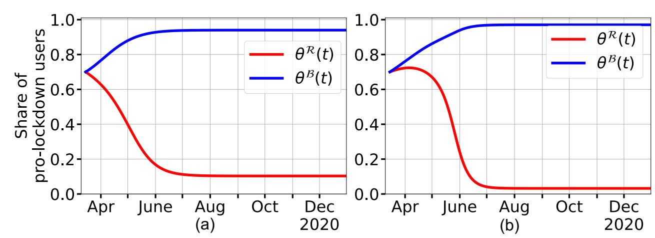



**Figure 14: Plotting trajectories of issue positions on lockdown mandates for (a) all users and (b) political partisans with exposure measured through tweets.**

## E Dynamics of a Multi-party Population

Two example scenarios under the first influence measure (net number of individuals with a stance normalized by degree) are shown in Fig. 16: Fig. 16(a) corresponds to a situation where the two ideologically extreme groups (hard-left and hard-right) are driven more by in-group love and out-group hate compared to the scenario in Fig. 16(b). Additionally, two ideologically extreme groups show some amount of positive emotion towards the moderates in

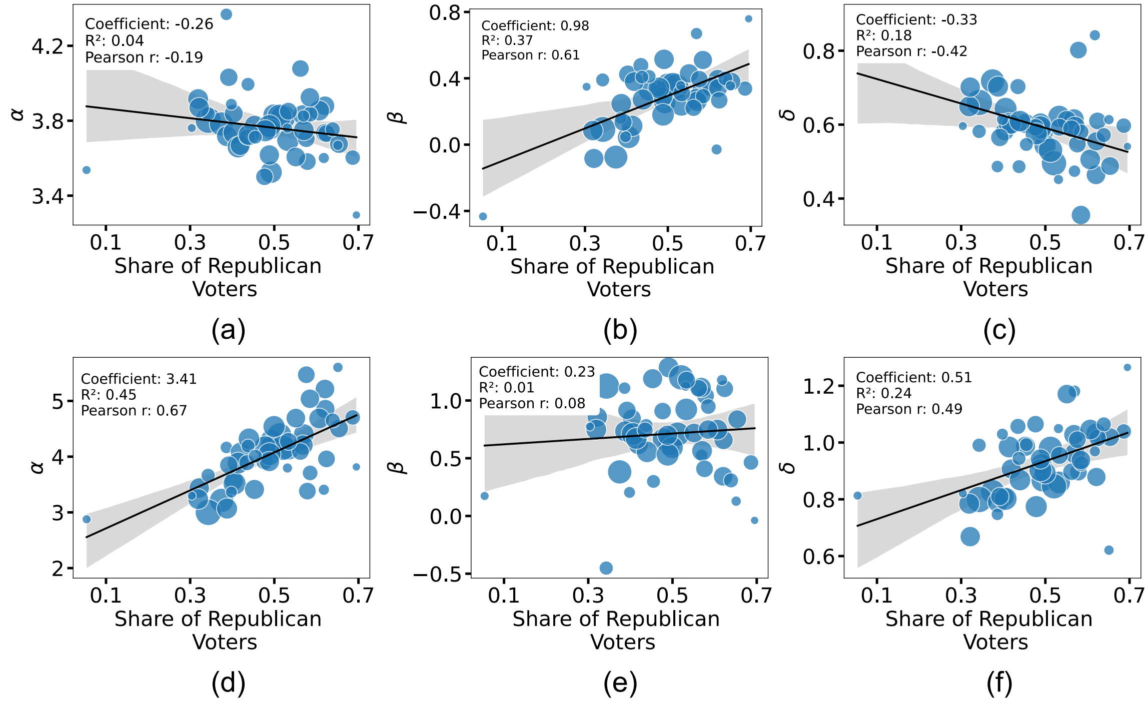
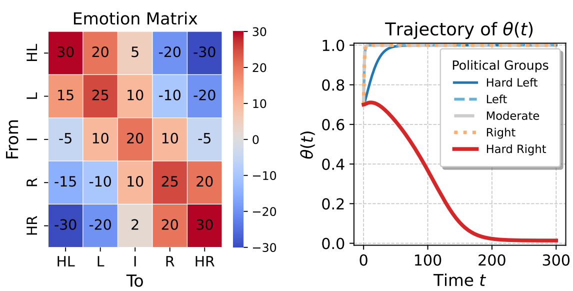
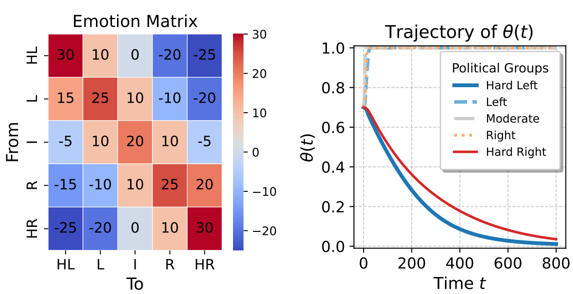



**Figure 15: The model parameters $\alpha, \beta, \delta$ for individuals in each state plotted against the state's 2020 Republican vote share: subplots (a-c) show results for masking and (d-f) for lockdowns.**




**(a) Example trajectories with $\delta_i = 4$ for each group $i$**




**(b) Example trajectories with $\delta_i = 8$ for each group $i$**

**Figure 16: Comparison of two scenarios (under first influence measure) of an affectively polarized multi-party population with group sizes 5% (HL), 25% (L), 40% (I), 25% (R), 5% (HR). Fig. 16(b) corresponds to a larger inertia for each group as well as smaller in-group love for the ideologically extreme groups (HL, HR) compared to the scenario in Fig. 16(b). It can be seen that the parameters in Fig. 16(b) lead to a horseshoe effect where the two extreme groups unite.**

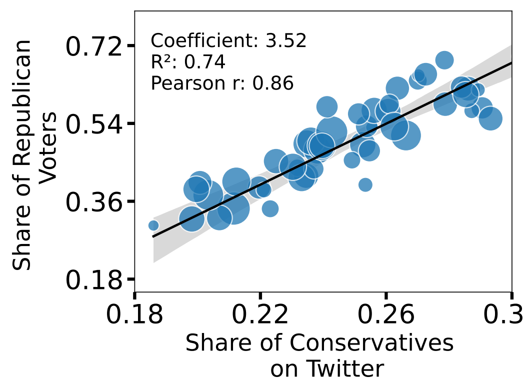



**Figure 17: Correlation between state-level Twitter ideology estimates and 2020 Republican vote share in Federal elections.**

**Table 3: Logistic Regression Parameters with exposure as a measure of tweets.**

| *Issue* | | *All users* | *Pseudo-R2* | *Partisans* | *Pseudo-R2* |
|---|---|---|---|---|---|
| *Masking* | $\alpha$ | 3.85±0.008 | 0.31 | 5.27 ± 0.019 | 0.38 |
| | $\beta$ | 0.08±0.005 | | 0.41± 0.005 | |
| | $\delta$ | 0.62±0.003 | | 0.28 ± 0.001 | |
| *Lockdowns* | $\alpha$ | 3.80 ± 0.017 | 0.18 | 5.03± 0.028 | 0.27 |
| | $\beta$ | 0.70± 0.014 | | 1.22 ± 0.023 | |
| | $\delta$ | 0.78 ± 0.004 | | 0.58 ± 0.008 | |

| Issue | Example Tweet | Stance |
| --- | --- | --- |
| Masking | Wait we're supposed to wear masks? Hell no! I don't wanna smell my own breath all day. what if I had liver and onions for breakfast? | Anti-masking |
| | Masks help stop the spread of coronavirus – the science is simple and I'm one of 100 experts urging governors to require public mask-wearing | Pro-masking |
| | It's hard because I'm not sure whether to trust a rando twitter account versus the cdc director when it comes to the wisdom of masks | Neutral |
| Lockdowns | Every day the left's hypocrisy is on display. Believe all women.... unless they are accusing Democrats. you can't come out of your house. Protesting against the lockdown is wrong and people will die! a protest against racism is ok and it's more important than social distancing. | Anti-lockdowns |
| | Cancel everything. Put us back on lockdown. Cut another check and let try again in a couple months. That's my opinion… | Pro-lockdowns |
| | RT @account: The supreme court just banned covid restrictions on attendance at synagogues and churches. They voted 5-4 | Neutral |

**Table 4: Examples of tweets expressing a stance on masking and lockdowns issues.**

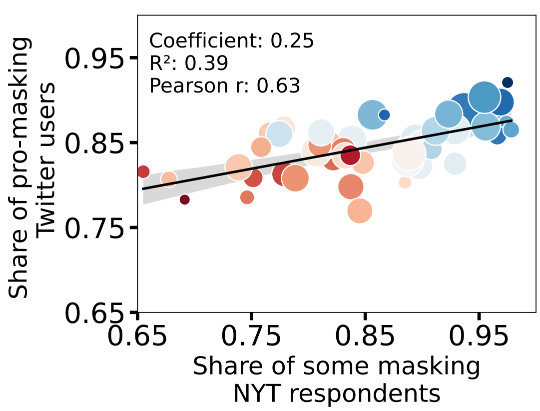



**Figure 18: Correlation between the share of respondents to New York Times masking survey who answered that they mask sometimes, often or always and the share of that state's Twitter users who endorse masking. Color represents the state's Trump vote share.**

Fig. 16(a) whereas they are indifferent to the moderates in Fig. 16(b). Consequentially, the hard left aligns with the remaining groups in Fig. 16(a) due to the positive emotion towards the largest group, moderates, and the highly negative feeling towards hard-right while the hard-right converges to the opposite stance. Fig. 16(b) illustrates a horseshoe effect [32] where the political extremes end up uniting with each other due to their large animosity towards the more moderate groups.

Thus, the multi-party model can capture a richer array of phenomena compared to the two-party systems even on a fully connected network. As we will subsequently see, such phenomena are observed in stances related to real-world issues, and their parameters can be identified by relying on the proposed model.

# F Additional Details

## F.1 Correlation of Stance and Ideology with Survey Data

Table 4 presents example tweets along with their corresponding stances for each issue. Fig. 17 shows the correlation between the state's 2020 Republican vote share in the Federal elections and the share of state's Twitter users who were identified as conservative by the ideology detection classifier. The Pearson's correlation of $r = 0.86$ with a $R^2$ value of 0.74 indicate that the classifier is reliable. Additionally, Fig. 18 shows the correlation between the share of state's New York Times masking survey [3] respondents who said they mask at least some of the time and the share of state's Twitter users who were identified to be pro-masking. The survey was administered from July 1, 2020 to July 14, 2020. We focus on masking tweets from the same period to assess the relationship between the survey responses and Twitter estimates. The Pearson's correlation of $r = 0.63$ with a $R^2$ value of 0.39 validates the stance classifier's performance.

## F.2 Additional Details on Model Calibration

We formalize the estimation of parameters $\alpha, \beta, \delta$, as a supervised learning task. More specifically, we model the shifts in user stances on masking and lockdowns for a user $X$ from $H_t(X)$ to $H_{t+1}(X)$ between two consecutive intervals, $t$ and $t$+1 (where $t$ = 0,1,2,...) as a feature of stances of users in the in- and out-group neighborhoods of $X$. We denote the fraction of users who are liberal and pro-masking (resp. lockdowns) and fraction of users who are conservative and pro-masking (resp. lockdowns) at (continuous) time $t$ as $\theta^B(t)$ and $\theta^R(t)$ respectively. Using *Definition 1* (in Sec. 3.1) we quantify for each time period $t$ $\theta^B(t)$ as the number of active liberal users who support the issue at hand, normalized by the total number of *active* users in that period. Similarly, $\theta^R(t)$ represents the number of active conservatives in period $t$ who support the issue, also normalized by the total number of active users in that period. Similarly, we define $\theta^{B'}(t)$ as the number of liberals who oppose the issue, and $\theta^{R'}(t)$ as the number of conservative users who oppose the issue, both normalized by the total number of active users in period $t$. We can then model the transition probabilities as:

$$p_\theta^B(1|0) = logit^{-1}\left(\alpha\left(\theta^B(t) - \theta^{B'}(t)\right) - \beta\left(\theta^R(t) - \theta^{R'}(t)\right) - \delta\right),$$
$$p_\theta^B(0|1) = logit^{-1}\left(\alpha\left(\theta^{B'}(t) - \theta^B(t)\right) - \beta\left(\theta^{R'}(t) - \theta^R(t)\right) - \delta\right),$$
$$p_\theta^R(1|0) = logit^{-1}\left(\alpha\left(\theta^R(t) - \theta^{R'}(t)\right) - \beta\left(\theta^B(t) - \theta^{B'}(t)\right) - \delta\right),$$
$$p_\theta^R(0|1) = logit^{-1}\left(\alpha\left(\theta^{R'}(t) - \theta^R(t)\right) - \beta\left(\theta^{B'}(t) - \theta^B(t)\right) - \delta\right).$$

We represent $p_\theta^B(1|0), p_\theta^B(0|1), p_\theta^R(1|0), p_\theta^R(0|1)$ using a dummy variable $J$, that indicates transition. We have $J = 1$ when we have a transition from anti to pro-masking (resp. lockdowns) regardless of the group identity ($p_\theta^B(1|0), p_\theta^R(1|0)$). Conversely, we have

[3] https://github.com/nytimes/covid-19-data/tree/master/mask-use

$J = 0$ when we have a transition from pro to anti-masking (resp. lockdowns) regardless of the group ($p_\theta^B(0|1), p_\theta^R(0|1)$).

## F.3 Identifying Geolocation

Location information for tweets is available for a subset of tweets through the coordinates and place fields within the tweet object. However, this data is present in less than 5% of the tweets in our dataset. To address this, we used Carmen [8] [4], a geo-location inference tool for Twitter data, to assign tweets to U.S. locations. Carmen utilizes information present in the user's bio, in addition to the place and coordinates fields of the tweet object, to infer location. The location object provided by Carmen includes details like the country, state, and county of the tweet. A manual review confirmed the method's effectiveness in identifying a user's home state.

[4] https://github.com/mdredze/carmen-python