Learning Opinion Dynamics From Social Traces

Corrado Monti
ISI Foundation, Italy
corrado.monti@isi.it

Gianmarco
De Francisci Morales
ISI Foundation, Italy
gdfm@acm.org

Francesco Bonchi
ISI Foundation, Italy
Eurecat, Spain
francesco.bonchi@isi.it

ABSTRACT

Opinion dynamics –the research field dealing with how people's opinions form and evolve in a social context– traditionally uses agent-based models to validate the implications of sociological theories. These models encode the causal mechanism that drives the opinion formation process, and have the advantage of being easy to interpret. However, as they do not exploit the availability of data, their predictive power is limited. Moreover, parameter calibration and model selection are manual and difficult tasks.

In this work we propose an inference mechanism for fitting a generative, agent-like model of opinion dynamics to real-world social traces. Given a set of observables (e.g., actions and interactions between agents), our model can recover the most-likely latent opinion trajectories that are compatible with the assumptions about the process dynamics. This type of model retains the benefits of agent-based ones (i.e., causal interpretation), while adding the ability to perform model selection and hypothesis testing on real data.

We showcase our proposal by translating a classical agent-based model of opinion dynamics into its generative counterpart. We then design an inference algorithm based on online expectation maximization to learn the latent parameters of the model. Such algorithm can recover the latent opinion trajectories from traces generated by the classical agent-based model. In addition, it can identify the most likely set of macro parameters used to generate a data trace, thus allowing testing of sociological hypotheses. Finally, we apply our model to real-world data from Reddit to explore the long-standing question about the impact of the *backfire effect*. Our results suggest a low prominence of the effect in Reddit's political conversation.

CCS CONCEPTS

• **Computing methodologies** → **Learning in probabilistic graphical models**; **Agent / discrete models**; • **Human-centered computing** → **Social network analysis**;

ACM Reference Format:
Corrado Monti, Gianmarco De Francisci Morales, and Francesco Bonchi. 2020. Learning Opinion Dynamics From Social Traces. In *26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '20), August 23–27, 2020, Virtual Event, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3394486.3403119

1 INTRODUCTION

Opinion dynamics is the study of how people's opinion on a subject matter form and evolve with time [15, 19]. This branch of social psychology has recently received growing attention due to the widespread adoption of social-media platforms. Users of these platforms can easily access and consume an immense amount of content, as well as engage in debate. In doing so, users share publicly their comments and beliefs, what they like and what they do not like, in other words, i.e., they leave *data traces.* Modeling opinion dynamics from this wealth of data is thus a tremendous opportunity for the social scientist. However, traditional opinion dynamics model are *agent-based*, i.e., they are simulations where a set of agents, interconnected by a network, interacts according to pre-determined mechanisms. These interactions modify the internal opinions of the agents, which in turn generate the dynamic of the opinion formation process.

Starting with the classical model by DeGroot [10], a plethora of refinements have been proposed [2, 9, 16], all sharing the fundamental strengths and weaknesses of agent-based models (ABMs) [28]. ABMs offer a framework for theory development, by allowing to explore empirically the implications of a sociological hypothesis formalized as a rule for interaction among agents. As such, ABMs provide a mechanistic model, which is easily interpretable in a causal way. This property is in sharp contrast with other models used in social science, such as statistical models (e.g., regression), for which a causal interpretation is much harder [22]. However, agent-based models also have several shortcomings. First, their predictive power is rather limited [8]. Second, parameter calibration is a considerable challenge, as it needs to be performed largely by hand. Third, agents cannot be directly used to understand any individual-level digital trace (e.g., from the Web or social media). Typically, in fact, *ABMs do not involve any inference from data.*

In this paper, we overcome these shortcomings of ABMs by proposing *an inference mechanism for fitting a generative, agent-like model of opinion dynamics to real-world social traces.* Such a model, dubbed *Learnable Opinion Dynamics Model* (LODM), retains the desirable properties of ABMs (causal interpretation of the mechanism behind opinion dynamics), while at the same time allowing for parameter inference from real data. Consequently, it can be used to explain individual behaviors, for model selection, and even for prediction: in other words, it produces a more *testable hypothesis.*

In particular, we translate a classical agent-based opinion dynamics model by Jager and Amblard [20] into a probabilistic generative framework. This classical model relies on *bounded confidence* with a *backfire* extension, based on social judgment theory [24]. After translating the model, we design an inference algorithm, based on online expectation maximization, that can fit the model *micro-level* parameters, the opinions of the agents, by looking at a data trace.

KDD '20, August 23–27, 2020, Virtual Event, USA

We show how to use our framework for model selection, i.e., to identify the most likely *macro level* parameters of the model, the rules which prescribe the opinion dynamics, from the data.

Finally, we apply our proposed model to a real-world dataset from Reddit, containing 10-years longitudinal cross-section of active users on subreddits related to politics. We show that our model is able to capture several behaviors of online users, such as the popularity of a user within a community, and the emergence of conflicts between users. Moreover, we show how our framework can test concrete sociological hypotheses expressed as agent-based interaction rules. In particular, we use the model to answer the question *"is there evidence of backfire effect in political discussion on Reddit?"*, to which we find a negative answer.

2 RELATED WORK

Opinion dynamics models (ODM) deal with the evolution over time of opinions in groups [6], and study sociological phenomena such as consensus formation [10], attitude change [20, 24] and polarization [11]. One of the most popular [17, 21] continuous-valued models is the bounded confidence model (BCM) by Deffuant et al. [9], which explains the observed differences in opinion through *bounded confidence*: agents ignore what is perceived as too distant from their own beliefs. Several extensions of BCM have been proposed by implementing other observations from sociology [4]. Quattrociocchi et al. [23] employed social impact theory, which emphasize the role of group pressure in attitude change. Jager and Amblard [20] instead built on social judgment theory [24]: the result of persuasion depends critically on the position of the receiver, and could end up with *acceptance* or *contrast* (the latter also known as *backfire effect*). The backfire effect suggests a link between exposition to opposing views (e.g., on social media) and polarization; as such, its importance has recently become a widely debated issue. Both Sippitt [26] and Fletcher and Jenkins [14] noted the need for more empirical tests confirming or disproving the backfire effect.

In fact, a great concern in opinion dynamics is how to validate the results empirically. According to Flache et al. [13], the field suffers from "a proliferation of theoretical studies and a dearth of empirical work"; for Castellano et al. [4], "there is a striking imbalance between empirical evidence and theoretical modelization, in favor of the latter". Therefore, the question of empirical validation has recently started to attract some attention. For instance, Sobkowicz [27] try to calibrate the model in order to reproduce some observations on the distribution of the resulting opinions in the population. This method has been described [13] as a test on *macro-level* predictions. Instead, the connection of *micro-level* (i.e., individual agents) behavior with real-world observations, that we tackle in this paper, is still largely unexplored.

Some of our ideas are also found in recent work, albeit within different conceptual frameworks, and with different techniques. The inclusion of actions as an observable for opinions was proposed also by Tang and Chorus [30], but without taking into account statistical inference nor real-world observations. Estimation through Maximum-a-Posteriori was proposed by Sichani and Jalili [25] for the sole purpose of inferring the most influential nodes; they considered the opinions to be fully observable.

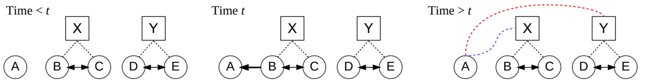

Figure 1: A minimal example of the observed actors (circles) and actions (squares), in the case of a positive (blue) and negative (red) interaction.

De et al. [7] used bayesian inference coupled with an ad-hoc model to predict opinion diffusion trough social influence. They assume that the polarity of messages is given and that followers of a user can be influenced by her messages. In particular, each observable is a triplet (u, m, t), indicating that the user u posted a message with sentiment m at time t, while in our model we observe interactions among users (e.g., discussion) and actions performed by users (e.g., sending a message) without a predefined polarity for the actions. Finally, Grazzini et al. [18] proposed bayesian estimates to calibrate the parameters of other ABMs (not opinion dynamics), but do not consider micro-level predictions.

3 GENERATIVE FRAMEWORK

In agent-based opinion dynamics models, interactions between agents are the driver of opinion change. For instance, the model by Jager and Amblard [20] distinguishes different kinds of interactions (positive and negative) with opposite effects on the opinions of the involved agents. In reality, neither the opinion of a single agent nor the "sign" of the interactions are easily observable. Therefore, it is difficult to use such models to explain individual behavior.

What is observable, instead, is that an interaction between two agents has happened. Moreover, we can often observe some *action* performed by individuals: using a hashtag on Twitter, or participating in a specific Reddit community. Such actions are often a reflection of an individual's personal opinion: hashtags are used as propaganda tools by political campaigns (e.g., `#MAGA`), Reddit communities gather people with similar views (e.g., `r/The_Donald`).

Our proposed probabilistic framework LODM aims exactly at explaining the individual behaviors recoverable from the digital traces found in social media. Our goal is therefore to estimate the micro-level latent variables of interest (i.e., individual opinions, the sign of interactions) given the observed ones (i.e., interactions and actions), under the assumptions of a specific opinion dynamics model. It is thus natural to frame our problem as a probabilistic generative model: the input are observed variables, the output are estimates for the latent ones.

3.1 Observables

Let V be a set of *actors*, who interact and influence each other's opinion. We represent interactions as an arc in a temporal graph, defined over T discrete time steps, where each actor is a node. Actors also perform *actions*. Each action is driven by the latent opinion of the actor: different opinions lead to different actions (think, for instance, of putting a "like" on a politically-charged Facebook page). We consider actions as a noisy proxy for the opinion of an actor. Let A be the set of possible actions. We represent the fact that an actor performs an action as a temporal arc in a bipartite graph, defined by V and A.

Formally, we observe the following two temporal graphs:

- $G = (V, E)$ is the directed interaction graph between actors. Arc $(u, v, t) \in E$ represent that "u interacts with v at time t". The interaction results in u possibly influencing v. Actors can interact multiple times at time step t, so we define E as a multiset, and G as a multigraph.
- $Z = (V, A, F)$ is a bipartite graph of actors and actions. Arc $(v, a, t) \in F$ represents that "actor v performs action a at time t". Similarly to G, each arc can appear multiple times in the same time step. Therefore, Z is a bipartite multigraph.

We depict a minimal example of these observables in Figure 1. We have 5 actors (A,B,C,D,E), and 2 possible actions (X,Y); Actors B and C perform action X at all time steps, while actors D and E perform action Y. These conditions create two clusters of actors who do not interact with each other: a typical instance of polarized opinions. Actor A plays the central role, as its action changes after interacting with the other actors. In the *consensus* scenario, A interacts with B at time t, and then performs action X from time $t + 1$ onwards. In the *backfire* scenario, A interacts with B at time t, but then performs action Y from time $t + 1$ onwards.

3.2 Latent variables

In our setting, each interaction in G is either positive or negative, and it changes the latent opinions of the actors accordingly. The idea that interactions can have different effects is a key concept in several opinion dynamics models [1, 5, 11, 20, 29]. In addition, we need to represent actions in opinion space. Each action is associated to a range of opinions, fixed in time for simplicity. We express these concepts via the following latent variables:

- $x_{t,v} \in [-1, 1]$ represents the latent opinion of actor $v \in V$, on a given subject matter, at time t.
- $\mathcal{S} : E \longrightarrow \{-1, +1\}$ represents the signs of the interaction arcs, which characterize each interaction between actors as either positive or negative.
- $w_a \in [-1, 1]$ and σ_a are the center and half-width of the opinion spectrum $[w_a - \sigma_a, w_a + \sigma_a]$ associated to action $a \in A$.

3.3 Base model

Next, we describe the original, deterministic ABM by Jager and Amblard [20]. This model assumes that interactions are either positive or negative. This is determined by two macro parameters: a *latitude of acceptance* and a *latitude of contrast*, denoted with ϵ^+ and ϵ^- respectively[1] (s.t. $0 \leq \epsilon^+ < \epsilon^- \leq 2$). The sign of an interaction $(u, v, t) \in E$ is determined when u expresses its opinion to v: if it is close (within ϵ^+) v *accepts* it, if it is distant (further than ϵ^-) v *constrasts* it. $\forall (u, v, t) \in E$

$$\begin{aligned} |x_{t,u} - x_{t,v}| < \epsilon^+ &\implies \quad \mathcal{S}(u, v, t) = +1 \text{ (positive arc)} \\ |x_{t,u} - x_{t,v}| > \epsilon^- &\implies \quad \mathcal{S}(u, v, t) = -1 \text{ (negative arc)}, \end{aligned} \tag{1}$$

and the opinions are updated accordingly

$$\begin{aligned} \mathcal{S}(u, v, t) = +1 &\implies \quad x_{t+1,v} = x_{t,v} + \mu^+ \cdot (x_{t,u} - x_{t,v}) \\ \mathcal{S}(u, v, t) = -1 &\implies \quad x_{t+1,v} = x_{t,v} - \mu^- \cdot (x_{t,u} - x_{t,v}), \end{aligned} \tag{2}$$

while clipping $x_{t+1,v}$ in $[-1, 1]$. The parameters $\mu^+, \mu^- > 0$ thus control the speed of the influence due to the interactions.

3.4 Generative process for interactions

Next, we describe how we translate this deterministic ABM into its probabilistic generative counterpart. This change allows us to design an inference procedure for the latent variables, via maximum a posteriori likelihood estimation. The modified model maintains the deterministic update rules for opinions of the agents (Eq. 2).

To make the generative model realistic, there are a few technical concerns to address. An actor might have more interactions in some time steps and fewer in others, for exogenous reasons. In addition, some actors might, in general, interact more than others. We wish for our model to keep these concerns into account, but without modeling them explicitly. Therefore, in our model (1) a node u at time t generates a given, fixed number $\gamma_{t,u}$ of arcs; (2) at each time step t, only a subset $V_t^* \subset V$ of the nodes is considered active and eligible to receive an arc. In other words, we do not explicitly model the probability of directly drawing an arc from all possible pairs given the opinions of agents $P\big((u, v, t) \in E \mid \mathbf{x}_t\big)$. This design choice allows the model to accept any real interaction graphs, with any observed empirical degree distribution, similarly to the configuration model [3].

In order to make interactions stochastic, we first need to determine the a priori probability of an interaction being positive at time t. Considering the opinions $\mathbf{x}_t$, we can use a summary statistic: the fraction of possible positive interactions

$$\alpha_t = \frac{\sum\limits_{(u,v)\in E_t} \mathbb{1}\left(|x_{t,u} - x_{t,v}| < \epsilon^+\right)}{\sum\limits_{(u,v)\in E_t} \mathbb{1}\left(|x_{t,u} - x_{t,v}| < \epsilon^+\right) + \mathbb{1}\left(|x_{t,u} - x_{t,v}| > \epsilon^-\right).} \tag{3}$$

Given u, to draw one of the $\gamma(u, t)$ arcs, we first draw a sign for the arc, positive with probability α_t, and then we pick the target v among the available nodes within the latitude for the given sign.

Now, to turn the agent-based model into a probabilistic generative one, we wish to add stochastic behavior into Equation 1. In particular, to account for noise in the data, we relax the boundaries on the latitudes. We define the probability of an interaction (u, v, t) as a function of the opinions of u and v, and of the sign of the arc. Let $f_G(x) = 1/(1 + e^{-\rho_G \cdot x})$ be a sigmoid function with a certain steepness ρ_G. Then, we define the probability of an interaction so that it depends on two functions

$$\begin{aligned} \kappa^+(x_{t,u}, x_{t,v}) &:= f_G\left(\epsilon^+ - |x_{t,u} - x_{t,v}|\right) \quad \text{(positive)} \\ \kappa^-(x_{t,u}, x_{t,v}) &:= f_G\left(|x_{t,u} - x_{t,v}| - \epsilon^-\right) \quad \text{(negative).} \end{aligned} \tag{4}$$

We can now use these functions to define a probabilistic generative process for the observed temporal graph, such that

$$\begin{aligned} P\big((u, v, t) \in E \mid \mathcal{S}(u, v, t) = +1\big) &\propto \kappa^+(x_{t,u}, x_{t,v}) \\ P\big((u, v, t) \in E \mid \mathcal{S}(u, v, t) = -1\big) &\propto \kappa^-(x_{t,u}, x_{t,v}). \end{aligned} \tag{5}$$

As the steepness of the sigmoid goes to infinity, Equations 4 and 5 turn into the original opinion dynamics model (Equation 1) [20], where every node within the latitude is equally likely to interact with the originating node, and all the nodes outside the latitude have zero probability of interacting with it.

[1] The original paper uses the notation T and U.

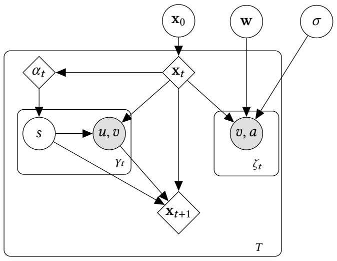

Figure 2: Plate diagram of our model.

In summary, we define the following overall generative process. For each time step t:

(i) Determine α_t from $\mathbf{x}_t$.
(ii) For each actor u, for $\gamma_{t,u}$ times:
  (iii) Extract a sign $s \in \{-1, +1\}$ with probability α_t.
  (iv) Choose an actor $v \in V_t^*$ with probability:

$$P((u, v, t) \in E \mid u, \mathbf{x}_t, s) = \frac{\kappa^s(x_{t,u}, x_{t,v})}{\sum_{v' \in V_t^*} \kappa^s(x_{t,u}, x_{t,v'})} \tag{6}$$

  (v) Add the interaction (u, v, t) to E.
(vi) Finally, update $x_{t+1,v}$ according to Equation 2.

3.5 Generative process for actions

We now define a similar process to account for the *actions* performed by each actor. Let $\zeta_{t,u}$ to be the exogenous, given number of actions that node u performs at time t. We define $P\big((u, a, t) \in F\big)$ to be proportional to

$$\kappa_\sigma(x_{t,u}, w_a) := f_Z\big(\sigma_a - |x_{t,u} - w_a|\big), \tag{7}$$

where f_Z is a sigmoid function with steepness ρ_Z, and σ_a represents a latent concentration in opinion space for action a.

Then, we assume that actions are performed at time step t according to the following process:

(i) For each actor u, for $\zeta_{t,u}$ times:
  (ii) Choose an action a with probability:

$$P((u, a, t) \in F \mid v, \mathbf{x}_t, \mathbf{w}, \boldsymbol{\sigma}) = \frac{\kappa_\sigma(x_{t,u}, w_a)}{\sum_{a' \in A} \kappa_\sigma(x_{t,u}, w_a)} \tag{8}$$

  (iii) Add performed action (v, a, t) to F.

The described model for actions and interactions is represented via plate notation in Figure 2. We provide in Appendix C a reference table outlining the notation we used.

Example. In the minimal example discussed at the end of Section 3.1 and depicted in Figure 1, this model assumes that the observed behavior is a result of a positive or a negative interaction. In the positive example, the opinion of the two nodes A and B are likely within the latitude of acceptance (i.e., $|x_A - x_B| < \epsilon^+$). The interaction between them is therefore positive and brings them closer together by a factor μ^+. Thus, the actor-action arc (A, X, t) that we observe becomes more likely, since $x_A \sim w_X \sim x_B$.

In the negative example, the opinion of the two nodes A and B are likely to be separated at least by the latitude of contrast (i.e., $|x_A - x_B| > \epsilon^-$). The interaction between them is negative and pushes them apart by a factor μ^-. Thus, the observed actor-actor arc (A, Y, t) is more likely, since the action Y is probably far from A and B, who never performed it.

4 LEARNING

Next, we present an algorithm to maximize the complete-data likelihood of the model, which estimates the latent variables given the observables and the macro parameters.

4.1 Complete-data likelihood

We can write the complete likelihood of a given dataset (under the knowledge of all the latent variables) as $P(E, F) = P(E) \cdot P(F)$, thus factoring the likelihood into the interaction likelihood $P(E)$ and the action likelihood $P(F)$. Note that the process described in Section 3.4 implies that, by total probability, the interaction likelihood $P(E)$ can be decomposed into the two mutually exclusive cases of positive and negative interaction,

$$P((u, v, t) \in E \mid u, \mathbf{x}_t) = \sum_{s \in \{+1, -1\}} P(s \mid u, \mathbf{x}_t) P((u, v, t) \in E \mid u, \mathbf{x}_t, s).$$

Therefore, by using the definition from Equation 6, the complete likelihood of interactions is

$$P(E \mid \mathbf{x}) = \prod_{(u,v,t) \in E} \sum_{s \in \{-1,+1\}} P(s) \frac{\kappa^s(x_{t,u}, x_{t,v})}{\sum_{v' \in V_t^*} \kappa^s(x_{t,u}, x_{t,v'})}, \tag{9}$$

where $P(s)$ is α_t for $s = 1$ and $(1 - \alpha_t)$ otherwise.

Similarly, the action likelihood is

$$P(F \mid \mathbf{x}, \mathbf{w}, \sigma) = \prod_{(v,a,t) \in F} \frac{\kappa_\sigma(x_{t,v}, w_a)}{\sum_{a' \in A} \kappa_\sigma(x_{t,v}, w'_a)} \tag{10}$$

by virtue of the probability defined in Equation 8.

We can use recursive Equation 2 to substitute each occurence of x_t in these formulas with a deterministic function of x_0 and $\mathcal{S}$. Therefore, instead of writing $P(E \mid \mathbf{x})$, we write $P(E \mid \mathbf{x}_0, \mathcal{S})$. The details of this function are explored in Appendix B.

Now, we wish to maximize the log likelihood with respect to the latent variables $\Theta = (\mathbf{x}_0, \mathcal{S}, w, \sigma)$ given the observed ones $\Omega = (E, F)$:

$$\widehat{\Theta} = \arg\max_{\Theta} \log P(E \mid \Theta) + \log P(F \mid \Theta) \tag{11}$$

Optimizing this function is not straightforward as the expression for the latent variables Θ contains $\mathcal{S}$ –the sign of each arc in the interaction graph– which is a discrete variable, thus leading to an integer programming problem. We cannot solve this problem via a standard linear relaxation of the sign, since it would mean to define cases “in between” acceptance and contrast. Such an approximation would defeat our purpose of translating a classic opinion dynamics model as faithfully as possible.

We therefore choose to employ the expectation-maximization (EM) technique. In addition, to make the problem tractable, we resort to incremental learning approach in designing the algorithm.

4.2 Online EM

To apply EM, we choose a set of parameters $\theta = (\mathbf{x}_0, \mathbf{w}, \boldsymbol{\sigma})$ from our latent variables Θ. We thus wish to maximize the joint distribution $P(E, F \mid \mathbf{x}_0, \mathbf{w}, \boldsymbol{\sigma})$ given observed variables $\Omega = (E, F)$, the latent variables $\mathcal{S}$, and the parameters θ. Recall that solving this problem requires finding an assignment of the latent variables $\mathcal{S}$ such that for every observed arc $(u, v, t) \in E$ we have a sign $s \in \{-1, +1\}$.

Alas, this formulation would require the summation of the M step to consider all possible $\mathcal{S} : E \longrightarrow \{-1, +1\}$, which are $2^{|E|}$.

To simplify this problem, let us consider our process as an online task. At each time step, our algorithm is presented with the new interactions E_t. The algorithm needs to decide their sign, i.e., whether each interaction is positive or negative. Then, it needs to update its estimate for the opinions of the actors accordingly. While solving the assignment problem for interactions in time step τ, the algorithm can therefore consider interactions and actions exclusively from the past time steps $t \leq \tau$.

Formally, let us consider a time step τ. Then, $E_\tau = \{(\cdot, \cdot, \tau) \in E\}$ are the actor-actor arcs at time τ and F_τ are the actor-actions arcs. Similarly, $\mathcal{S}_\tau : E_\tau \longrightarrow \{-1, +1\}$ are the signs of the interactions at the same time τ. Let also $\mathcal{S}_{<\tau} := \bigcup_{t<\tau} \mathcal{S}_t$. We wish for our algorithm to take inputs $(E_\tau, F_\tau, \mathcal{S}_{<\tau})$ together with a previous estimate for $\widehat{\theta} = (\mathbf{x}_0, \mathbf{w}, \boldsymbol{\sigma})$, and to return as output the maximum a posteriori estimate for $\mathcal{S}_\tau$ and θ. The probabilities of all signs and of the presence of all the links are conditionally independent: $\mathcal{S}(u, v, \tau) \perp\!\!\!\perp \mathcal{S}(u', v', \tau) \mid (\theta, \mathcal{S}_{<\tau})$ and $\big((u, v, \tau) \in E_\tau\big) \perp\!\!\!\perp \big((u', v', \tau) \in E_\tau\big) \mid (\theta, \mathcal{S}_{<\tau})$. As a consequence, we can express the likelihood of the signs as a product of independent likelihoods

$$P(\mathcal{S}_\tau \mid E_\tau, F_\tau, \theta) = \prod_{u,v:(u,v,\tau)\in E} P(\mathcal{S}(u, v, \tau) \mid E_\tau, F_\tau, \theta), \qquad (12)$$

which allows the algorithm to treat each separately. Note that this result requires the online assumption. Without it, since the opinions $\mathbf{x}_{t+1}$ depend on $\mathcal{S}_t$, in the general case $\mathcal{S}(u, v, t) \not\perp\!\!\!\perp \mathcal{S}(u', v', t')$.

Therefore, thanks to the online assumption and Equation 12, we can define the following expectation-maximization steps:

- *E Step.* For each arc $(u, v, \tau) \in E_\tau$ we evaluate

$$q_{s,u,\tau} := P(s \mid (u, v, \tau) \in E_\tau, \widehat{\theta})$$
$$= \frac{P((u, v, \tau) \in E_\tau \mid u, s, \widehat{\theta}) \cdot P(s \mid u, \widehat{\theta})}{\sum_{s' \in \{-1,+1\}} P((u, v, \tau) \in E_\tau \mid u, s', \widehat{\theta})} \qquad (13)$$

  where

$$P(s \mid u, \widehat{\theta}) = \begin{cases} \alpha_t & \text{if } s = 1 \\ 1 - \alpha_t & \text{otherwise} \end{cases}$$

  and $P((u, v, \tau) \in E_\tau \mid u, s, \widehat{\theta})$ is defined as in Equation 6.

- *M Step.* We update parameters θ in order to increase the following function $Q(\theta)$:

$$\sum_{(u,v,\tau)\in E_\tau} \sum_{s\in\{-1,+1\}} q_{s,u,\tau}$$
$$\log\Big(P((u, v, \tau) \in E_\tau \mid s, u, \theta) P(F_{\tau,u} \mid \theta)\Big) = \log P(F_\tau \mid \theta) +$$
$$\sum_{(u,v,\tau)\in E_\tau} \sum_{s\in\{-1,+1\}} q_{s,u,\tau} \log P((u, v, \tau) \in E_\tau \mid s, u, \theta) \qquad (14)$$

  where

$$\log P(F_\tau \mid \theta) = \sum_{(v,a,\tau)\in F} \log\left(\frac{\kappa_\sigma(x_{\tau,v}, w_a)}{\sum_{a'\in A} \kappa_\sigma(x_{\tau,v}, w_{a'})}\right) \qquad (15)$$

  and $P((u, v, \tau) \in E_\tau, s \mid u, \theta)$ is again defined by Eq. 6.

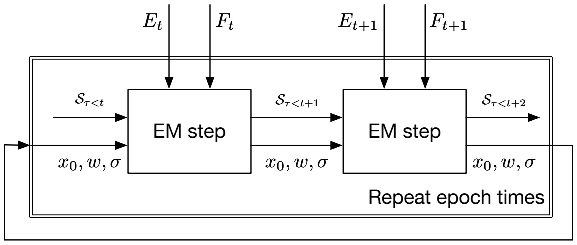

Figure 3: Schema of the proposed online learning process. Starting from $t = 0$, for each time step t, the EM algorithm is presented with the estimate problem for the given time step; it updates the parameters and emits the estimate for $\mathcal{S}_t$. This whole process, from $t = 0$ to $t = T$, can be repeated for a fixed number of epochs to improve the final estimates.

To increase the function in Eq. 14, we employ gradient descent, and maximize it w.r.t. $\mathbf{x}_0, \mathbf{w}, \boldsymbol{\sigma}$. The EM algorithm we have thus defined is summarized in Appendix A (Algorithm 1). It can be applied to one time step at a time, and considers only information coming from the previous time steps to update its parameters. Starting from $t = 0$, at each time step the algorithm is initialized with the current best estimate for its parameters, it updates them with new information, and emits the results for $\mathcal{S}_t$. The resulting $\mathcal{S}_t$ and the updated parameters are then in turn used for the next time step estimate. This schema is depicted in Figure 3 and summarized in Appendix A (Algorithm 2). This process can also be re-iterated: at each epoch, the whole learning process from $t = 0$ to $t = T$ is repeated, starting with the parameter estimates from the previous epoch, and continuously updating the parameters. In practice, we repeat this process for a fixed number of epochs (2 in all the reported experiments). Moreover, as common practice with EM algorithms, we employ a multiple restart approach: for each run, we repeat the learning process a number of times (4 in all the reported experiments) while changing the initial random seed; then, we pick the one with highest likelihood.

What is the complexity of these computations? From Equations 13 and 15, it follows that the complexity for the E step is $O(nm)$ where $n = |V_\tau^*| \leq |V|$ and $m = |E_\tau|$; for the M step, it is $O(nm + n'm')$ where $n' = |A|$ and $m' = |F_\tau|$.Empirically, we report that running our framework on a common laptop for 2 epochs on 10 time steps and 1 000 nodes, takes 140 seconds for each restart; for 3 000 nodes, 1 254 seconds.

5 EMPIRICAL ASSESSMENT

We focus on the following three research questions:

RQ1: Can we recover the micro parameters of the opinion dynamics model? (Section 5.1)

RQ2: Given a data trace from the generative process, can we find which macro-level scenario generated it? (Section 5.2)

RQ3: Can the estimated parameters of the opinion dynamics model on real data explain real user behavior? (Section 5.3)

To answer these questions we use a mix of synthetic data and real data; the latter represent a 10-year data set we crawled from the social rating and discussion website Reddit. While RQ1 and RQ2 deal with the internal validity of our proposal (inference algorithm and

Table 1: Distance (mean average error) and classification accuracy (F1 and average precision) on synthetic experiments; we report average and std. dev. across 8 generated data traces.

	MAE $\mathbf{x}_0$	MAE $\mathbf{w}$	$\mathcal{S}$ F1-score	Z Av.Prec.
Non-commitment	0.13 ± 0.02	0.18 ± 0.03	0.99 ± 0.02	0.95 ± 0.01
Balanced	0.16 ± 0.04	0.14 ± 0.01	1.00 ± 0.00	0.96 ± 0.02
High contrast	0.13 ± 0.02	0.16 ± 0.03	0.98 ± 0.01	0.97 ± 0.01
High acceptance	0.34 ± 0.16	0.26 ± 0.12	0.90 ± 0.08	0.93 ± 0.03

model selection framework, respectively), RQ3 tests the external validity of the inferred model parameters. The results of model selection on real data are quite interesting: Section 6 discusses some possible interpretations. We publicly release our implementation and data set to facilitate reproducibility.[2]

5.1 Recovering opinion micro parameters

RQ1 deals with the micro parameters of our models: the opinions of agents and actions, and the signs of the interactions. To test our inference algorithm, we generate synthetic data traces according to the model by Jager and Amblard [20]. The set of macro parameters (ϵ^+, ϵ^-) for the given trace, which define a *scenario*, are taken from the same work, and reported in Figure 4.

Each scenario represents different assumptions about the behavior of the actors. A *high acceptance* scenario is characterized by a high latitude of acceptance ϵ^+, which results in consensus among the actors. Conversely, a *high contrast* scenario, generated by a low latitude of contrast ϵ^-, results in frequent backfires and a polarized system. A low ϵ^+ and a high ϵ^- generate a scenario of *non-commitment*, where the opinions are stable and fragmented. Finally, in a *balanced* scenario, the distance in opinion space is equally divided among acceptance, neutral, and contrast zones, and opinions cluster into a small number of attraction points.

Actions are not part of the original model, so we generate them according to the stochastic process described in Section 3.5 (initialized uniformly in $[-1, 1]$). For each scenario we generate 8 different data traces. Then, we fit the model with the set of macro parameters corresponding to the specified scenario. Finally, we measure how close the inferred micro parameters (opinions and interaction signs) are to the generated ones.

Table 1 shows four measures of the quality of our predictions in the four scenarios, on average across 8 experiments. First, we show the mean absolute error between the original $\mathbf{x}_0$ and its estimate,[3] and the same for the action opinions $\mathbf{w}$. Then, we treat assigning a positive or a negative sign to each interaction in G as a binary classification problem, and compute the F1 score with respect to the original signs. Finally, we measure how well our model captures the actor-action graph by taking all its edges F, a sample of non-existing edges $(v, a, t) \notin F$ of the same cardinality $|F|$, and we compute the average precision of our model in separating the two.

In most scenarios, the inferred opinions are very close to the generated ones, the signs of the interactions are almost perfectly recovered, and the actions are well fit by the model. The high-acceptance scenario proves to be more challenging, as the final consensus equilibrium blurs the individual opinion of each actor.

5.2 Discriminating macro-level scenarios

We now ask whether our framework is able to discriminate which scenario generated a given data trace. If our framework can accomplish this task, we can use it to assess the plausibility of assumptions of opinion formation models by testing them on real data (as we show in Section 5.3). Operationally, we run our algorithm against the data trace with different sets of macro parameters, one for each scenario hypothesis we wish to test. Finally, we look at the likelihood obtained under each hypothesis.

In our experiments, we generate 8 synthetic data traces for each scenario. Then, for each data trace, we run our algorithm with the four different macro parameters encoding each scenario hypothesis.

Figure 5 shows the likelihood of each different generated data trace under each tested scenario. In all cases, the most likely set of macro parameters found by our framework is the true one that generated the data trace itself. Specifically, it is close to a perfect accuracy for every scenario except “high-acceptance”, for which the results are still mostly positive.

5.3 Opinion dynamics on real data

In this section we apply the framework to real data from Reddit to explore the prominence of the backfire effect, i.e., to see whether a scenario with large latitude of contrast is likely.

Dataset. We gather Reddit data from 2008 to 2017, and bucket it so that one time step corresponds to a month (120 time steps in total). Reddit users are actors, while posting in a subreddit corresponds to an action. User v replying in a comment thread to user u at time step t corresponds to an interaction (u, v, t). We sample from both users and subreddits to create our dataset. In order to study US political discussion, we choose `r/politics` as our seed subreddit and pick the 50 most similar subreddits to `r/politics` according to cosine similarity over a vector representation of the subreddits based on latent semantic analysis, which captures subreddits whose user base is similar to the seed one.[4] Resulting subreddits include political ones such as `r/democrats`, ones dedicated to specific politicians such as `r/hillaryclinton` and `r/The_Donald`, and ideological ones such as `r/Libertarian`. We then sample users posting a minimum of 10 comments per month on `r/politics` for at least half of the months, which gives us 375 users. The resulting action graph has approximatively 144k actor-action arcs, while the interaction graph has approximatively 90k actor-actor arcs.

Reddit allows to up/down-vote posts, which represents the social feedback of the community. The score of a post on a subreddit is a function of the up- and down-votes received by it from other users in that subreddit. It represents how well-received the post is by the specific subreddit community. A negative score means that the post has been disapproved by the community, possibly because it expresses a point of view that is far from the norm of the subreddit. A high absolute score indicates a high attention for the post, i.e., it has been read and voted by a large number of users in the subreddit.

We consider two different application settings for the framework: with or without an *anchored axis*. An anchored axis refers to fixing the position in opinion space of a set of actions. In particular, we fix two actions as the extremes of the opinion space. This way, we create an axis along which all other actions (and actors) lay.

[2] https://github.com/corradomonti/learnable-opinion-dynamics

[3] Since the estimate is symmetric, we take the best between $\mathbf{x}_0$ and $-\mathbf{x}_0$.

[4] https://www.shorttails.io/interactive-map-of-reddit-and-subreddit-similarity-calculator

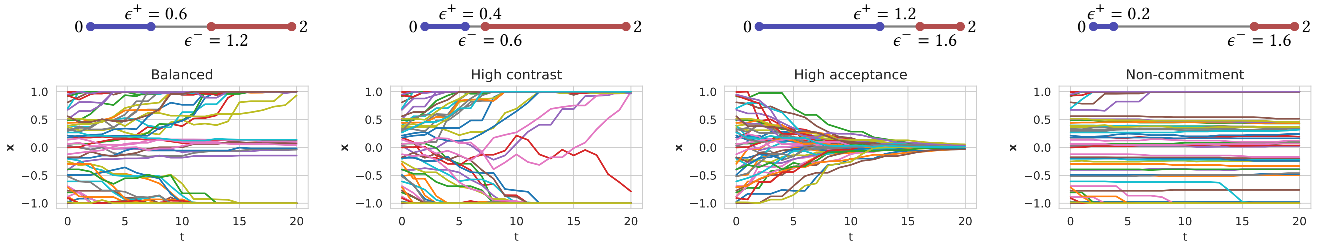

Figure 4: Examples of synthetic data traces generated in each scenario. Plots represent the opinion trajectories along time.

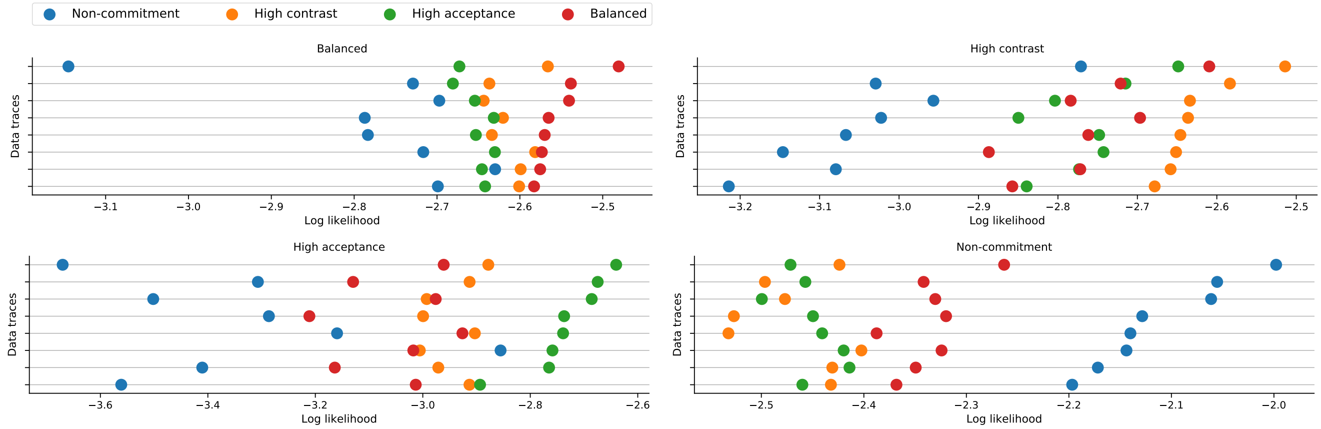

Figure 5: Each panel represents a different macro parameters scenario, and each line on the Y axis a different data trace generated according to that scenario. On the X axis we report, for each data trace, the log likelihood obtained by the the best estimate of our generative model, initialized with a given set of macro parameters, one per color. Rightmost corresponds to highest likelihoods. We see that the estimate corresponding to the true macro parameters has the highest likelihood.

By changing the definition of the axis, we can explore different semantics for the latent opinion space.

In the experiments, we explore two anchors for the axis: one between `r/democrats` or `r/Republican`, by fixing their latent opinion point to $w_a = \{-1, 1\}$, respectively, and one with respect to `r/The_Donald`, by fixing its latent opinion point to $w_a = 1$. The first anchoring represents the traditional political spectrum in US, the second one represents the closeness to Donald Trump supporters. In summary, we have three different axes: a free one (`None`), a left-right one (`r/democrats` — `r/Republican`), and a unipolar one (`r/The_Donald`). For each of the two possible cases w.r.t. anchored axis, we test a set of parameters as the ones presented in Figure 4.

To verify that the model is capturing the underlying behavior from the data, we employ two external validation metrics. These metrics are completely hidden from the framework, and try to capture the user behavior on Reddit:

- **User-subreddit score:** For each user, subreddit, and time step, we compute the average score of all the posts the user has submitted to the given subreddit in the specific month. This score is a proxy for how well-received the opinions of the user are in the specific subreddit.
- **User-user conflicts:** We identify set of user-user interactions that exhibit conflictual behavior. The intuition is that when a reply to a positive-score comment has a negative score (or vice versa), the two authors are probably expressing conflicting points of view. To capture this behavior, we define a conflictual interaction of comment x to comment y when x and y have scores with opposite signs. We restrict our attention to comments that have attracted some attention in the community, i.e., with a minimum absolute score of 10.

We now present results for these two external evaluation metrics by using the best estimate according to our *internal* validation metric, the log likelihood. The best-fitting scenario is a *high-acceptance* one anchored on `r/The_Donald`, but results are qualitatively similar for a *non-commit* scenario. For this experiment, the average precision on the real user-subreddit links F (as defined in Section 5.1) is 0.934. We measure the Pearson correlation coefficient between the user-subreddit score and the distance between user and subreddit in opinion space, as inferred by our model. Our hypothesis is that a higher score corresponds to a lower distance between the two, and therefore the correlation should be negative. This behavior is consistent with the idea that opinions close to the norm of the subreddit are the ones that get the most appreciation, as can be explained by cognitive dissonance theory [12].

Figure 6 shows the regression of the user-subreddit score as a function of their inferred distance. The relationship between the variables is negative as predicted by our hypothesis. In other words, users that are more popular within a community are the closest to that community in our opinion space. This result confirms that the parameters inferred by our model, in particular the opinions of the

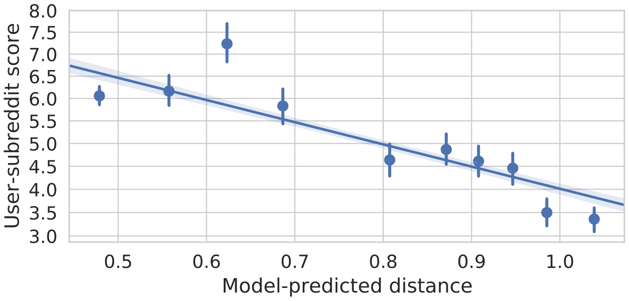

Figure 6: Univariate regression between user-subreddit score and distance between user and subreddit in opinion space as inferred by our model. The correlation coefficient is negative with a value of -0.127 **and highly significant (**$p < 10^{-6}$**). Error bars represent the** 95% **confidence intervals.**

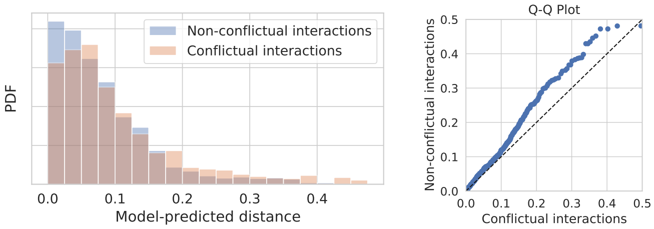

Figure 7: Distributions of distances in opinion space between users during conflictual interactions (orange) and non-conflictual interactions (blue). Conflictual interactions have a significantly larger distance on average ($p < 10^{-6}$**).**

users and the subreddits, capture some of the drivers behind user voting behavior.

For the second validation metric, for each interaction, we measure the distances between pairs of users in opinion space at the time of the interaction. We compute these distances for the interactions specified above, and also for a control group of non-conflictual interactions (i.e., both scores are positive). We also apply the same minimum score threshold as above to select the non-conflictual interactions. Our hypothesis is that conflictual interactions are more likely to happen between users that are further apart in opinion space.

Figure 7 shows the distributions for both kind of interactions. The conflictual interactions present a higher average distance than the control group, which is consistent with our hypothesis. A one-sided non-parametric Mann-Whitney U test confirms the hypothesis that a randomly selected conflictual interactions has a larger distance than a non-conflicting one ($p < 10^{-6}$). The median distances differ by 0.06. This result shows that our model is able to capture some of the mechanisms behind the emergence of conflicts. The small effect size is to be expected, as conflicts might happen for a number of reasons not directly related to the ideological positions of the users interacting. Nevertheless, the strong statistical association between the parameters inferred by our model and real-world user behavior as measured from noisy data is a clear signal that our algorithm is able to capture some latent user characteristic.

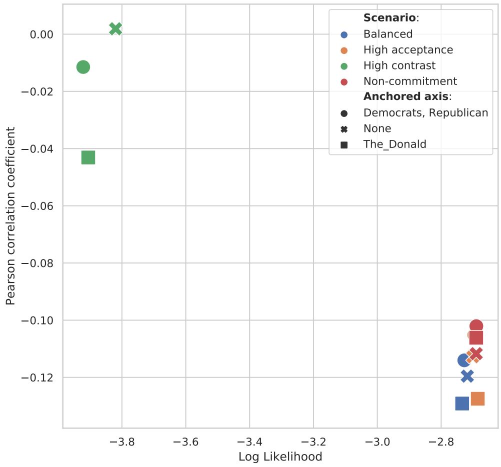

Figure 8: Scatter plot of the results obtained by different hyper-parameters on estimates on Reddit data, with respect to an external and an internal evaluation metric. On the X axis, we report the internal objective function of the model, its log likelihood. On the Y axis, the external evaluation metric: the Pearson correlation between the user-subreddit latent opinion distance and the average user-subreddit scores (bottom right is better).

Finally, we explore the selection of the macro parameters from Figure 4 with respect to the external validation metrics. We use the user-subreddit score as it is numerical, and thus can offer a higher granularity for better presentation. Figure 8 shows the relationship between the likelihood of the model given the set of macro parameters, and the correlation of user-subreddit scores with user-subreddit distances (model parameters). As such, this graph shows the relationship between an internal evaluation metric (the likelihood) and an external validation one (the score-distance correlation coefficient). The two metrics agree for the most part, thus suggesting that we can use the likelihood to identify the most fitting model that explains real-world behaviors.

6 DISCUSSION AND FUTURE WORK

We have proposed LODM: a learnable generalization of an opinion dynamics model. It retains the explainability and causal interpretation of agent-based models, by describing the underlying data generation process via latent and observed stochastic variables. We have shown how to cast a classic agent-based opinion dynamics model into our framework, and designed an algorithm infer its parameters from data. Clearly, this model is a proof-of-concept, and a similar process can be applied to other opinion dynamics models to make them testable and learnable. Since our work is based on a generalization of BCM, it should be easily applicable to other BCM extensions [11, 23]. Thanks to recent efforts in unifying different opinion dynamics model under a common formalism [6], it might be possible to build *general* learnable opinion dynamics model. This framework could leverage social traces to validate empirically several assumptions on opinion dynamics, with the final goal of improving our understanding of how the human mind shapes ideas through social interactions.

Table 2: Position in latent opinion space of the top-20 most popular subreddits in our data. See Section 6 for a discussion.

Subreddit	w_a	σ_a	Subreddit	w_a	σ_a
r/The_Donald	1.00	0.69	r/worldnews	-0.53	0.59
r/Republican	1.00	0.38	r/todayilearned	-0.65	0.60
r/progressive	0.99	0.58	r/atheism	-0.83	0.60
r/Economics	0.89	0.65	r/EnoughTrumpSpam	-0.84	0.53
r/Libertarian	0.88	0.61	r/SandersForPresident	-0.89	0.55
r/TrueReddit	0.87	0.60	r/PoliticalDiscussion	-0.91	0.59
r/Futurology	0.84	0.67	r/worldpolitics	-0.95	0.62
r/conspiracy	0.84	0.60	r/changemyview	-0.97	0.57
r/news	0.52	0.61	r/Conservative	-1.00	0.47
r/politics	-0.34	0.60	r/economy	-1.00	0.52

Our experiments have shown that the framework is able to learn the micro-level parameters of the single actors. For instance, we are able to distinguish positive interaction from negative ones. We are also able to recover the latent opinion of actors, and their trajectory in time. This feature allows fine-grained analysis of real individuals with the same techniques used to describe opinion dynamic models. In other words, we are able to empirically quantify and verify the assumptions of opinion dynamics model at an individual level.

Moreover, we have shown the capabilities of our proposal for model selection. The framework is able to identify the correct scenario (i.e., the set of macro-level parameters that encode the interaction rules) that generated a given data trace in synthetic experiments. This capability is extremely useful for testing sociological assumptions, which can still be expressed as deterministic update rules for agents' internal states.

We have applied our framework to a real-world dataset coming from Reddit, and have shown that the best-fitting model is able to explain user-level behavior. In particular, we are able to explain a trend in voting behavior of users on subreddits by looking at the learned micro parameters (the opinions of users and subreddits).

By using our framework for model selection on Reddit data, we find that the "high acceptance" and "non-commitment" scenarios are the most likely, and the "high contrast" one is the least likely by far. Our model thus rejects the presence of a low latitude of contrast. These results suggest that the backfire effect is negligible among active participants in Reddit's political conversation.

A possible explanation for our results is that a community such as Reddit, over a time span of a decade, tends to evolve more according to a consensus-creation mechanism than an internal polarization one. For example, the social feedback inherent in the platform may stifle extreme opinions, and create more pressure towards mainstream attitudes. Following new trends might be more appealing than the rejection created by polarization mechanisms.

The proposed framework has numerous possible applications. As an example, Table 2 reports the inferred positions in opinion space for the top-20 most popular subreddits in our dataset. Here the latent opinion space is anchored so that `r/The_Donald` (a community of Donald Trump supporters) is fixed at one extreme (1.0). The position of many subreddit in opinion space seems reasonable and follows intuition. The community of Bernie Sanders supporters (`r/SandersForPresident`) is correctly positioned near the other end of the spectrum. A conspiracy group (`r/conspiracy`), which has been described as taking "a pro-Trump bent",[5] is placed very close to Donald Trump supporters. This example shows how our model could be used to analyze trajectories estimated under a specific set of hypothesis.

[5] https://www.vox.com/2018/8/8/17657800/qanon-reddit-conspiracy-data

REFERENCES

[1] A. E. Allahverdyan and A. Galstyan. 2014. Opinion Dynamics with Confirmation Bias. *PLOS ONE* 9, 7 (2014), e99557.

[2] R. Axelrod. 1997. The Dissemination of Culture: A Model with Local Convergence and Global Polarization. *Journal of Conflict Resolution* 41, 2 (1997), 203–226.

[3] E. A. Bender and E. R. Canfield. 1978. The asymptotic number of labeled graphs with given degree sequences. *Journal of Combinatorial Theory, Series A* 24, 3 (1978), 296–307.

[4] C. Castellano, S. Fortunato, and V. Loreto. 2007. Statistical Physics of Social Dynamics. *Reviews of Modern Physics* 81, 2 (2007), 591.

[5] X. Chen, P. Tsaparas, J. Lijffijt, and T. De Bie. 2019. Opinion Dynamics with Backfire Effect and Biased Assimilation. *arXiv:1903.11535* (2019).

[6] A. Coates, L. Han, and A. Kleerekoper. 2018. A Unified Framework for Opinion Dynamics. In *17th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS'18)*. 1079–1086.

[7] A. De, I. Valera, N. Ganguly, S. Bhattacharya, and M. G. Rodriguez. 2016. Learning and Forecasting Opinion Dynamics in Social Networks. In *Advances in Neural Information Processing Systems (NIPS'16)*. 397–405.

[8] G. Deffuant, S. Huet, and S. Skerratt. 2008. An agent based model of agri-environmental measure diffusion: What for? *Agent Based Modelling in Natural Resource Management* (2008), 55–73.

[9] G. Deffuant, D. Neau, F. Amblard, and G. Weisbuch. 2000. Mixing Beliefs among Interacting Agents. *Advances in Complex Systems* 3 (2000), 87–98.

[10] M. H. DeGroot. 1974. Reaching a Consensus. *JASA* 69, 345 (1974), 118–121.

[11] M. Del Vicario, A. Scala, G. Caldarelli, H. E. Stanley, and W. Quattrociocchi. 2017. Modeling Confirmation Bias and Polarization. *Scientific Reports* 7, 40391 (2017).

[12] L. Festinger. 1957. *A Theory Of Cognitive Dissonance.* Stanford University Press.

[13] A. Flache, M. Mäs, T. Feliciani, E. Chattoe-Brown, G. Deffuant, S. Huet, and J. Lorenz. 2017. Models of Social Influence: Towards the Next Frontiers. *Journal of Artificial Societies and Social Simulation* 20, 4 (2017).

[14] R. Fletcher and J. Jenkins. 2019. *Polarisation and the News Media in Europe: A Literature Review of the Effect of News Use on Polarisation across Europe.* Technical Report. European Parliamentary Research Service.

[15] J. R. P. French. 1956. A Formal Theory of Social Power. *Psychological Review* 63, 3 (1956), 181–194. https://doi.org/10.1037/h0046123

[16] Noah E. Friedkin and E. C. Johnsen. 1990. Social influence and opinions. *The Journal of Mathematical Sociology* 15, 3-4 (1990), 193–206.

[17] J. Gómez-Serrano, C. Graham, and J. Boudec. 2010. The Bounded Confidence Model Of Opinion Dynamics. *Mathematical Models and Methods in Applied Sciences* 22, 2 (2010), 1150007.

[18] J. Grazzini, M. G. Richiardi, and M. Tsionas. 2017. Bayesian Estimation of Agent-Based Models. *Journal of Economic Dynamics and Control* 77 (2017), 26–47.

[19] F. Harary. 1959. A criterion for unanimity in French's theory of social power. In *Studies in social power.* 168–182.

[20] W. Jager and F. Amblard. 2005. Uniformity, Bipolarization and Pluriformity Captured as Generic Stylized Behavior with an Agent-Based Simulation Model of Attitude Change. *Computational & Mathematical Organization Theory* 10 (2005), 295–303.

[21] J. Mathias, S. Huet, and G. Deffuant. 2016. Bounded Confidence Model with Fixed Uncertainties and Extremists: The Opinions Can Keep Fluctuating Indefinitely. *Journal of Artificial Societies and Social Simulation* 19, 1 (2016), 6.

[22] J. Pearl. 2009. Causal Inference in Statistics: An Overview. *Statistics Surveys* 3 (2009), 96–146.

[23] W. Quattrociocchi, R. Conte, and E. Lodi. 2011. Opinions within Media, Power and Gossip. *arXiv:1102.2336* (2011).

[24] M. Sherif and C. Hovland. 1961. *Social Judgment: Assimilation and Contrast Effects in Communication and Attitude Change.* Yale University Press.

[25] O. A. Sichani and M. Jalili. 2017. Inference of Hidden Social Power Through Opinion Formation in Complex Networks. *IEEE Transactions on Network Science and Engineering* 4, 3 (2017), 154–164.

[26] A. Sippitt. 2019. *The Backfire Effect: Does It Exist? And Does It Matter for Factcheckers?* Technical Report. Full Fact.

[27] P. Sobkowicz. 2016. Quantitative Agent Based Model of Opinion Dynamics: Polish Elections of 2015. *PLOS ONE* 11, 5 (2016), e0155098.

[28] F. Squazzoni. 2012. *Agent-Based Computational Sociology.*

[29] A. Stefanelli and R. Seidl. 2014. Moderate and polarized opinions. Using empirical data for an agent-based simulation. In *Social Simulation Conference.*

[30] T. Tang and C. G. Chorus. 2019. Learning Opinions by Observing Actions: Simulation of Opinion Dynamics Using an Action-Opinion Inference Model. *Journal of Artificial Societies and Social Simulation* 22, 3 (2019), 1–2.

Algorithm 1 EM Step for time step t

```
Input: Graph $G_t = (V, E_t)$; actions $Z_t = (V, A, F_t)$; $\mathcal{U}^t(\mathbf{x}_0)$; $\alpha_t$.
Output: opinions $\mathbf{x}_0$, actions $(\mathbf{w}, \sigma)$, signs $\mathcal{S}_t : E_t \to \{-1, +1\}$
If not given, initialize $\mathbf{x}_0, \mathbf{w}, \sigma$
$V_t^* := \{v \in V : (\cdot, v, t) \in E_t\}$
repeat                                                        ▷ E Step
    Compute $\mathbf{x}_t := \mathcal{U}^t(\mathbf{x}_0)$
    for $(u, v, t) \in E_t$ do
        $p^+(u, v) := \kappa^+(x_{t,u}, x_{t,v}) / \sum_{v' \in V_t^*} \kappa^+(x_{t,u}, x_{t,v'})$
        $p^-(u, v) := \kappa^-(x_{t,u}, x_{t,v}) / \sum_{v' \in V_t^*} \kappa^-(x_{t,u}, x_{t,v'})$
        $q^+(u, v) := \frac{\alpha_t \cdot p^+(u, v)}{p^+(u, v) + p^-(u, v)}$
                                                              ▷ Eq. 13
        $q^-(u, v) := \frac{(1 - \alpha_t) \cdot p^-(u, v)}{p^+(u, v) + p^-(u, v)}$
    end for
                                                              ▷ M Step
    Using $\mathbf{x}_t = \mathcal{U}^t(\mathbf{x}_0)$, do:
        Update $\mathbf{x}_0$ by ascending the gradient:
            $\nabla_{\mathbf{x}_0} \sum_{(u,v,t) \in E_t} \sum_{s \in \{-1,+1\}} q^s(u, v) \log\left(\frac{\kappa^s(x_{t,u}, x_{t,v})}{\sum_{v' \in V_t^*} \kappa^s(x_{t,u}, x_{t,v'})}\right)$
        Update $\mathbf{x}_0, \mathbf{w}, \sigma$ by ascending the gradient:
            $\nabla_{\mathbf{x}_0, \mathbf{w}, \sigma} \sum_{(v,a,t) \in F_t} \log\left(\frac{\kappa_\sigma(x_{t,v}, w_a)}{\sum_{a' \in A} \kappa_\sigma(x_{t,v}, w_{a'})}\right)$
until convergence
$\forall (u, v, t) \in E,\ \mathcal{S}_t(u, v, t) := \text{sign}\left(q^+(u, v) - q^-(u, v)\right)$
```

Algorithm 2 Complete online learning process

```
Input: Interaction graph $G = (V, E)$; action graph $Z = (V, A, F)$
Output: opinions $\mathbf{x}_0$, actions $(\mathbf{w}, \sigma)$, signs $\mathcal{S} : E \to \{-1, +1\}$
for number of multiple restarts do
    Initialize $\mathbf{x}_0, \mathbf{w}, \sigma$ randomly
    for number of epochs do
        for $t$ in $1, \ldots, T$ do
            Define $\mathcal{U}^t$ with $(E_{<t}, \mathcal{S}_{<t})$                ▷ Eq. 17
            Define $\alpha_t$ with $\mathbf{x}_t = \mathcal{U}^t(\mathbf{x}_0)$                ▷ Eq. 3
            $\mathbf{x}_0, \mathbf{w}, \mathcal{S}_t, := \text{EmStep}(G_t, Z_t, \mathcal{U}^t(\mathbf{x}_0), \mathbf{w}, \sigma, \alpha_t)$
        end for
    end for
    Keep $(\mathbf{x}_0, \mathbf{w}, \sigma, \mathcal{S})$ if $P(E, F | \mathbf{x}_0, \mathbf{w}, \sigma, \mathcal{S})$ is higher than last restart
end for
```

A REPRODUCIBILITY

Algorithm 1 provides the pseudocode of the EM method for each time step t, as introduced in Section 4; while Algorithm 2 the pseudocode of the complete learning process. Our implementation of the proposed framework, alongside all the resources needed to reproduce our experiments are available at:
https://github.com/corradomonti/learnable-opinion-dynamics

Parameter settings. The main parameters for the evaluation are the latitudes of acceptance and contrast (ϵ^+, ϵ^-). We fix the other parameters heuristically via grid search by optimizing the likelihood of the model. Specifically, we use action learning rate 10^{-3}, interaction learning rate 10^{-4}. In this way, we fix the the steepness of the sigmoid functions f_G used in Eq. 4 and f_Z used in Eq. 8 to the values of $\rho_G = 8$ and $\rho_Z = 16$, respectively.

In the synthetic data generation, we use 30 nodes, 20 actions, 10 time steps, 3 interactions per time step per node and 15 actions per time step per node. For the Reddit application, we fix $\mu^+ = 10^{-3}$ and $\mu^- = 10^{-4}$. We also add to the loss function in Equation 15 a prior on $\boldsymbol{\sigma}$ (the half-width of each w_a), so that it follows a $\beta(8, 8)$ distribution (centered in 0.5, with support on $[0, 1]$).

B LINKING BACK TO x_0

At each time step, the EM Algorithm updates an estimate of the same parameters: x_0, w, σ. Thus, we need to express every x_t appearing in the formulas in terms of the same parameters x_0, so that the gradient descent can update x_0. We need therefore an efficient way to define x_t in terms of x_0. The opinion vector $\mathbf{x}_\tau$ is a deterministic function of $\mathbf{x}_0$ and of the signed arcs at previous time steps $(E_{<\tau}, \mathcal{S}_{<\tau})$, that we consider to be fixed.

To find a computationally efficient way to compute $\mathbf{x}_t$, we define the following $n \times n$ matrix M for the signed arcs at time t:

$$M_{u,v}(t) = \begin{cases} -\mu^- \cdot \#_E(u, v, t) & \text{if } \mathcal{S}(u, v, t) = -1 \\ \mu^+ \cdot \#_E(u, v, t) & \text{if } \mathcal{S}(u, v, t) = +1 \\ 0 & \text{if } (u, v, t) \notin E \end{cases} \tag{16}$$

where $\#_E(e)$ is the multiplicity of the arc of e in the multiset E. Then, the opinion update (Equation 2) can be written as

$$\begin{aligned} x_{t,v} + \sum_{u \in N} M_{u,v} \cdot (x_{t,u} - x_{t,v}) &= x_{t,v} + \sum_{u \in N} \Big(M_{u,v} \cdot x_{t,u}\Big) - \Big(\sum_{u \in N} M_{u,v}\Big) \cdot x_{t,v} \\ &= \Big(1 - \sum_{u \in N} M_{u,v}\Big) x_{t,v} + \Big(\sum_{u \in N} M_{u,v} \cdot x_{t,u}\Big). \end{aligned} \tag{17}$$

Therefore, the update equation for $\mathbf{x}_t$ can be conveniently written as a matrix operation $\mathbf{x}_{t+1} = (\mathbf{1} - M^\top \mathbf{1}) \circ \mathbf{x}_t + M^\top \mathbf{x}_t$, where $\mathbf{1}$ is a vector of $|V|$ elements, $\circ$ is the Hadamard product. Let us call $\mathcal{U}^t$ the repeated application of this operation, for the sequence $M(0), \ldots, M(t-1)$, applying also the clipping $\min(1, \max(0, \mathbf{x}_t))$ at each step. This is a deterministic function, computed from $\mathcal{S}_{<t}, E_{<t}$, μ^+, μ^-, that gives $\mathbf{x}_t = \mathcal{U}^t(\mathbf{x}_0)$.

C NOTATION REFERENCE

For readers' convenience we provide a reference table summarizing all the notation used in the paper.

Variable	Meaning
V	Set of actors
E	Interactions: $(u, v, t) \in E$ means u influenced v at time t
G	Temporal graph (V, E)
A	Set of actions
F	Actor-action arcs: $(v, a, t) \in F$ means v performed a at time t
Z	Temporal bipartite graph (V, A, F)
E_t	Subset of E considering only arcs at time t
$E_{<t}$	Subset of E considering only arcs before time t
G_t	Graph (V, E_t)
F_t	Subset of F considering only arcs at time t
Z_t	Graph (V, F_t)
$\mathbf{x}_{t,v}$	Opinion of actor v at time t
$\mathbf{w}_a, \sigma_a$	Center and half-width of action a in opinion space
$\mathcal{S}$	Sign $\{1, -1\}$ of each interaction $(u, v, t) \in E$
$\mathcal{S}_t$	Restriction of $\mathcal{S}$ to E_t
$\mathcal{S}_{<t}$	Restriction of $\mathcal{S}$ to $E_{<t}$
α_t	Probability of an interaction being positive at time t
$\mathcal{U}^t$	Function s.t. $\mathbf{x}_t = \mathcal{U}^t(\mathbf{x}_0)$
ϵ^+	Latitude of acceptance, i.e. threshold for pos. interactions
ϵ^-	Latitude of contrast, i.e. threshold for neg. interactions
μ^+, μ^-	Speed of positive and negative influence
κ^+, κ^-	Sigmoid function for probability of pos. and neg. interactions
κ_σ	Sigmoid function for probability of an actor performing an action