Modeling echo chambers and polarization dynamics in social networks

Fabian Baumann,[1,*] Philipp Lorenz-Spreen,[2] Igor M. Sokolov,[1] and Michele Starnini[3,†]

[1] *Institute for Physics, Humboldt-University of Berlin, Newtonstraße 15, 12489 Berlin, Germany*
[2] *Center for Adaptive Rationality, Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany*
[3] *ISI Foundation, via Chisola 5, 10126 Torino, Italy*

Echo chambers and opinion polarization have been recently quantified in several sociopolitical contexts, across different social media, raising concerns for the potential impact on the spread of misinformation and the openness of debates. Despite increasing efforts, the dynamics leading to the emergence of these phenomena remain unclear. Here, we propose a model that introduces the phenomenon of radicalization, as a reinforcing mechanism driving the evolution to extreme opinions from moderate initial conditions. Empirically inspired by the dynamics of social interaction, we consider agents characterized by heterogeneous activities and homophily. We analytically characterize the transition between a global consensus and emerging radicalization dynamics in the population, as a function of social influence and the controversialness of the topic discussed. We contrast the model's behavior against empirical data of polarized debates on Twitter, qualitatively reproducing the observed relation between users' engagement and opinions, as well as opinion segregation based on the interaction network. Our findings shed light on the dynamics that may lie at the core of the emergence of echo chambers and polarization in social media.

The participatory character of political debates on online social media leads to a high degree of self-organisation in public opinion formation [1]. The low cost for engagement and the distributed architecture of those communication infrastructures did not only increase interaction rates, but also decrease barriers given by geographical distance or social status. In the view of traditional constructive opinion dynamics approaches that model social influence exclusively as opinion averaging [2–5], such unrestricted modes of interaction would ultimately lead to local or global consensus, even in the case of highly controversial issues.

However, this behavior is not always observed empirically. Instead, there is increasingly quantitative evidence that in certain sociopolitical contexts opinions are far from consensus, better described by heterogeneous opinion distributions. People often can be separated in two groups holding qualitatively different opinions - a state of the system referred to as *polarization*. Polarized opinion states have been quantified in political surveys of polling institutes [6, 7], as well as in several debates on online social media, ranging from political orientation [8–10], US and French presidential elections [11], or street protests [12–14]. When segregation in the space of opinions, or polarization, is reflected in the network of interactions among users, *echo chambers* may emerge: situations in which one's opinion is reinforced due to repeated interactions with like-minded individuals [2, 15]. Echo-chambers might be related to the spread of misinformation [17, 18] and may pose a threat for the openness of the democratic debate.

But what relates the dynamics of opinion polarization to the emergence of echo chambers in social networks? Previous modeling approaches mainly described polarization as the result of repulsive interactions, in which users reject opinions that strongly differ from their own [19], or caused by external driving factors, such as propaganda, media influence [20] and disinformation campaigns [21]. More endogenously, a polarized opinion distribution has been shown to be driven by homophily, the preference of agents to interact with similar individuals [22–24], even in the absence of negative influence [25–27]. In a population of interacting agents, homophily was also used to model the emergence of echo chambers [28, 29]. However, several empirical features of social interaction networks characterized by echo chambers have not been adressed within a unified modeling framework [2, 10, 18, 29].

In this Letter, we propose a simple model of opinion dynamics able to capture two frequently observed phenomena of polarized empirical social networks: i) more active users, i.e. those more prone to engage in social interactions, tend to show more extreme opinions, and ii) the similarity between the opinion expressed by a user and those expressed by his/her neighbors in the social interaction network. The model introduces a mechanism by which agents sharing similar opinions can mutually reinforce each other and move towards more extreme views, thus describing *radicalization dynamics*. Alongside, opinion states are coupled to the evolving network of social interactions by homophily. While the convergence toward a global consensus is retained in the model, the introduction of opinion reinforcement and homophily may lead to the emergence of meta-stable polarized states. To gain further insight into the dynamics towards extreme opinions we analytically characterize the transition between consensus and radicalization focusing on social influence and the controversialness of the topic discussed.

Let us consider a system of N agents, each agent i characterized by a dynamic opinion variable $x_i(t)$. For the sake of simplicity, we consider opinions to be one-dimensional, with $x_i \in [-\infty, +\infty]$. The sign of the opinion x_i, $\sigma(x_i)$, describes the agent's qualitative stance to-

wards a binary issue of choice, such as the preference between two candidates or a pro/con attitude in a controversial topic. The absolute value of x_i, $|x_i|$, describes the opinion's strength, or conviction, with respect to one of the sides: the larger $|x_i|$, the more extreme the opinion of agent i. Assuming that the opinion dynamics is solely driven by the interactions among agents, we formulate the model as N coupled ordinary differential equations,

$$\dot{x}_i = -x_i + K \sum_{j=1}^{N} A_{ij}(t) \tanh(\alpha x_j), \tag{1}$$

where $K > 0$ denotes the social interaction strength among agents and α determines the sigmoidal shape of the hyperbolic tangent. The opinion of an agent i follows the aggregated social input from the set of his/her neighbors at time t, determined by the symmetric adjacency matrix of the temporal network $A_{ij}(t)$, where $A_{ij}(t) = 1$ if agents i and j are interacting at time t, $A_{ij}(t) = 0$ otherwise. A similar model with static connectivity has previously been used to describe the dynamics of neural networks showing a transition from stationary to chaotic phase [30].

In a minimal scenario of an interacting pair of agents i and j, we distinguish two fundamentally different scenarios, which depend on the signs $\sigma(x)$ of participating opinions. If the agents share the same general attitude ($\sigma(x_i) = \sigma(x_j)$), the interaction will cause an increase of both of their convictions and hence reinforce their opinions, a mechanism we refer to as *radicalization* dynamics. On the contrary for opposing attitudes ($\sigma(x_i) = -\sigma(x_j)$) the involved opinions move in the opposite direction which potentially leads to sign change(s) for a single or both considered opinions, flipping the general attitude of the respective agent(s). Note that we model opinion dynamics as a purely collective, self-organized process without any intrinsic individual preferences. Agents lacking social interactions will therefore decay towards the neutral state and "forget" their current opinion.

The parameter $\alpha > 0$ tunes the degree of non-linearity between an agent's opinion and the social influence s/he exerts on others. In general, the stronger an agent's conviction, the larger the social influence s/he can exert on other agents. For small α, however, the model gives rise to large opinion ranges around the neutral consensus, where the social influence of weakly convicted individuals ($|x_i| \sim 0$) on others is heavily reduced. For large α, by contrast, agents with weak opinions can already exert a strong social influence on others. In the limit of $\alpha \to \infty$, the hyperbolic tangent in Eq. (1) approaches a Heaviside step function, which gives even agents holding infinitesimally small opinion values maximum social influence. Therefore, the parameter α is interpreted as the *controversialness* of the issue, directly controlling the relationship between an agent's conviction and its social influence on others. Empirically, it has been shown that controversy is an important factor driving the emergence of polarization and echo chambers in debates on online social media [4].

The contact pattern among agents, which sustain the opinion dynamics, represent social interactions and have found to evolve in time [32–34]. Following these empirical observations we model the interaction dynamics as an activity-driven (AD) network [34–37]. Each agent i is characterized by an activity $a_i \in [\varepsilon, 1]$, representing his/her propensity to contact m distinct and randomly sampled other agents. Activities are extracted from a distribution $F(a)$ typically assumed to follow a power-law $F(a) \sim a^{-\gamma}$, as measured in empirical data of real systems [34, 35, 38, 39]. The set of parameters (ε, γ, m) fully encodes the basic AD dynamics. While in the original AD formulation agents establish connections by random uniform selection, we assume here that interactions are ruled by homophily [23, 24]. To this end, the probability p_{ij} that an active agent i will connect to a peer j is modeled as a decreasing function of the absolute distance of their opinions,

$$p_{ij} = \frac{|x_i - x_j|^{-\beta}}{\sum_j |x_i - x_j|^{-\beta}}, \tag{2}$$

where the exponent β controls the power law decay of the connection probability with opinion distance.

It is important to remark that the dynamics of the AD network and the opinion dynamics clearly separate with respect to their time scales. Focusing on a regime, in which social interactions evolve much faster than opinions, we model short lived interactions on social online media, like Twitter. Specifically, we choose to update the network structure at increments of $dt = 0.01$ in time units of the opinion evolution (see SM for details on the numerical approach). The time scale separation allows us to work out an analytical approximation of the model valid in the limit of fast switching interactions. To explore the model numerically we use a system size of $N = 1000$ agents. For each simulation we initialize the opinions uniformly spaced on the interval $x_i \in [-1, 1]$ and fix the basic AD parameters to $m = 10$, $\epsilon = 10^{-2}$ and $\gamma = 2.1$. The behavior of the model is then discussed as a function of the social interaction strength K, the controversialness α and the homophily exponent β.

We identify three qualitatively different dynamical regimes, shown in Fig. 1. The first scenario shows a *neutral consensus*, depicted in Fig. 1(a), in which the opinions of all agents converge towards zero, obtained for small values of controversialness α and social influence strength K. Larger values of α and/or K destabilize the neutral consensus state and give rise to radicalization dynamics, i.e. situations in which agents' opinions do not converge to the single consensus value, but are widespread and reach values far outside of the initial opinion bounds. In these cases, the dynamics of the system strongly depends on how active agents choose their

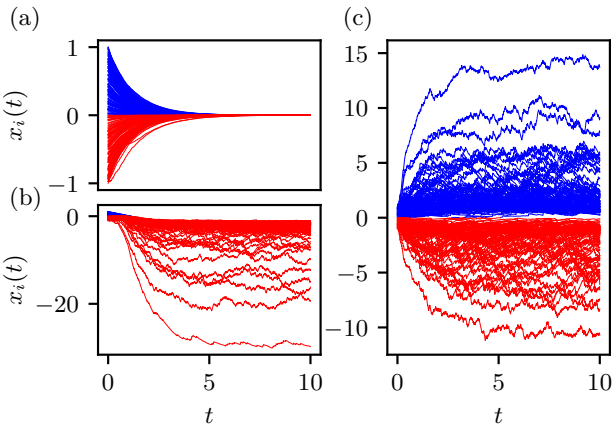

FIG. 1. **Temporal evolution of the agents' opinions.** (a) Neutral consensus for which all opinions converge to zero ($K = 3$, $\alpha = 0.05$, $\beta = 2$). (b) (One-sided) radicalization ($K = 3$, $\alpha = 3$, $\beta = 0$). (c) Opinion polarization, in which opinions split into two opposite sides ($K = 3$, $\alpha = 3$, $\beta = 3$). Positive (negative) opinions $\sigma(x_i) > 0$ ($\sigma(x_i) < 0$) are colored in blue (red). Note different scales on the y-axis.

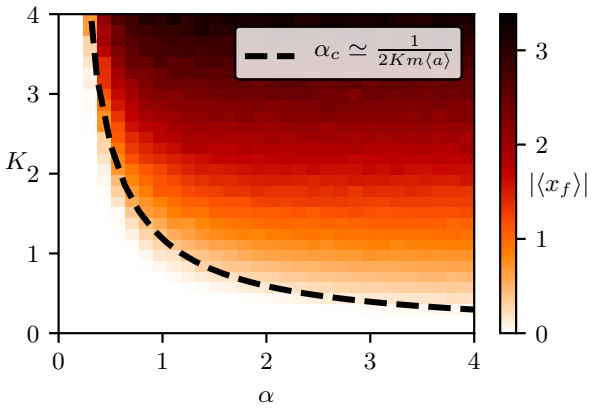

FIG. 2. **Transition from consensus to radicalization dynamics.** Absolute values of the average final opinions $|\overline{x}_f|$ in K-α phase space for $\beta = 0.5$. In the white region, the system approaches a neutral consensus, while in the colored areas the population undergoes radicalization dynamics which become more pronounced for increasing values of K and/or α (color code).

interaction partners. In the absence of homophily bias ($\beta = 0$), where agents pick their interaction partners uniformly at random, all opinions will be directly absorbed by one of the sides, as shown in Fig. 1(b). The introduction of homophily ($\beta > 0$), can drastically change this situation: driven by repeated interactions with like-minded individuals, agents reinforce their opinions and segregate into two groups holding opposite opinions, as shown in Fig. 1(c). In this scenario, a *polarized state* characterized by a bimodal distribution of opinions emerges (see Fig. S1(b) in the SM), which is in line with previous empirical findings [7–12]. It is important to note that the polarized state Fig. 1(c) is not stable in the limit $t \to \infty$ and may eventually decay into a one-side radicalized state, cf. Fig. 1(b). The lifetime of the meta-stable polarized state increases over-exponentially with the strength of homophily β, up to a point where its destabilization becomes numerically inaccessible. In Fig. S3 of the SM, we show this behavior exemplarily for some parameterizations of the model.

The transition from neutral consensus to radicalized states in K-α space, is depicted in Fig. 2, where the color encodes the absolute value of the final average opinion, $|\langle x_f\rangle| \equiv |1/N \sum_i x_i(t_{\text{final}})|$. In the long term regime, the value of $|\langle x_f\rangle|$ identifies the transition between regions exhibiting a stable neutral consensus, $|\langle x_f\rangle| = 0$ (white), characterized by small values of K and α, and regions where radicalization emerges and becomes stronger, $|\langle x_f\rangle| > 0$ (color coded), obtained for increasing K and/or α. It is possible to analytically capture this transition using a mean-field approximation. With a reduced version of the model that neglects homophily ($\beta = 0$), we derive an analytical expression for the critical value of controversialness (see SM for details),

$$\alpha_c \simeq \frac{1}{2Km\,\langle a\rangle}, \tag{3}$$

for which the neutral consensus becomes unstable and radicalized states emerge. The critical controversialness α_c depends inversely on the social influence strength K, the number of contacts per active agent m and the average activity $\langle a\rangle$. Fig. 2 demonstrates that Eq. (3) is able to approximate the critical transition also for the full system assuming moderate values of homophily.

The rich behavior of the model allows us to contrast it with empirical data of polarized debates on social online media. We focus on three different data sets collected from Twitter and analyzed in Ref. [2]. Each contains a set of tweets on a specific topic of discussion, known to be politically controversial: guncontrol, obamacare, and abortion. The data sets have been built along two main features: i) the political orientation of users and ii) their social interaction network. Each user, indeed, is characterized by his/her political leaning on the basis of the content produced, by using a ground truth of political leaning scores of various news organizations (e.g., nytimes.com, foxnews.com)[40], ranging from very conservative to very liberal. Specifically, the political leaning score $x_i \in [-1, +1]$ of user i (equivalent to the opinion variable x_i of agent i in the model) is obtained by considering the set of tweets posted by user i that contain links to news organizations of known political leaning. Moreover, for each data set, the social network of interactions among the users is reconstructed, so that there

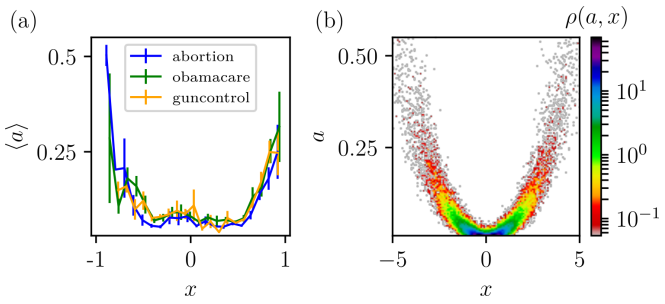

FIG. 3. **Activity vs. opinion.** (a) Average activity $\langle a \rangle$ of users as a function of their political leaning x, for three empirical data sets. (b) Activity-opinion density plot of 10^4 polarized opinion states for $K = 2$, $\alpha = 3$ and $\beta = 1$. The color code encodes the value of $\rho(a, x)$ which is normalized with respect to N.

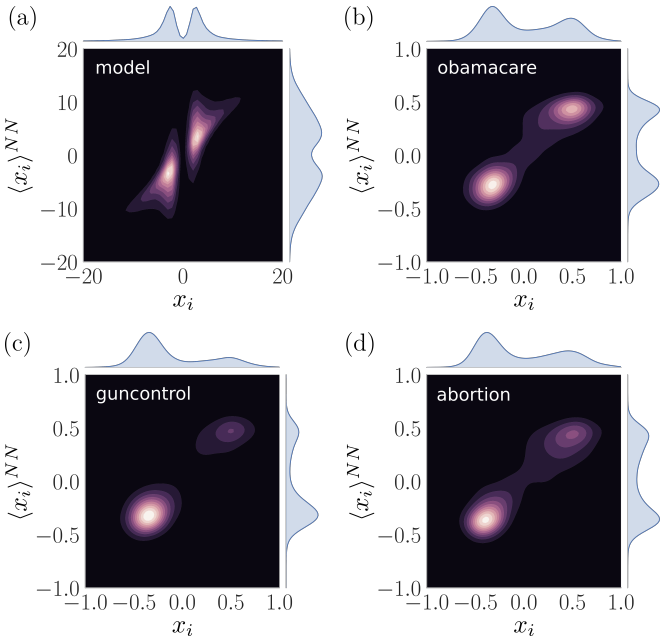

FIG. 4. **Echo chambers.** Contour maps for the average opinions of the nearest-neighbor $\langle x \rangle^{NN}$ against a user's opinion x, for 200 simulations of the radicalization model (a) and three different data sets (b-d). Colors represent the density of users: the lighter the larger the number of users. The marginal distribution of opinions, $P(x)$, and average opinions of the nearest-neighbor $P_{NN}(x)$ are plotted on the x- and y-axis, respectively.

exists a direct link from node i to node j if user i follows user j. These data sets have been collected and validated in previous works, used to show the presence of political polarization and echo chambers in online social media. See SM for further details on the data sets.

The immediate observable – to determine if the system is in a polarized state – is the distribution of expressed opinions $P(x)$ [41]. In considered all data sets, $P(x)$ is characterized by a bimodular shape (see Fig. S1(a) in the SM). Note that even though the method used to infer users' opinions can differ (e.g. likes to Facebook pages [17], Twitter hashtags [10], upvotes to Youtube videos [42] or political leaning of media linked in tweets messages [4]), a similar bimodal shape of the opinion distributions is common in polarized systems, across diverse topics and different online social media platforms. For sufficiently large values of K and α the overall qualitative features of the polarized empirical opinion distributions are reproduced by the model (see Fig. S1(b) in the SM).

Online social media made the engagement of users easily measurable, while the low cost in effort for participating in the discussion allows users to vary strongly in their activity, including highly active individuals. A striking feature emerging in different empirical data sets of polarized debates is a clear relation between the engagement of users in the discussion and their convictions: more active users tend to show more extreme opinions. For the empirical analysis of the Twitter data, we asses the activity of a user as the ratio of tweets containing links to news organizations of known political leaning, a rationale derived from the original activity potential definition [35].

Fig. 3(a) shows the average engagement, or activity a, of users as a function of their opinions x. For all three topics under consideration, one can see rising engagement towards the extremes of opinion space. It is important to note that different definitions of user activity and polarity, such as the number of Likes to Facebook Pages tagged in different classes, shares of political content on Facebook [40], or tweet rates of users classified according to the hashtags they use [10], give rise to the same functional form of association between users' activity and political leaning, or opinions. As depicted in Fig. 3(b) this characteristic U-shaped function is reproduced by our model, which suggests the following interpretation of the empirical finding: While most users have low activities and opinions close to the neutral consensus, some very active users take on more extreme opinions, as their opinions are reinforced by interactions with sufficiently like-minded peers. In a feedback loop, they radicalize themselves and their environment, making it decreasingly likely to listen to opposing opinions. This generic feature of the model is preserved also for different parameter sets, which give rise to similar plots, as shown exemplarily in Fig. S2 of the SM.

Echo chambers, indeed, are identified by the correspondence between the distribution of opinions in the population and the topology of the interaction network. Hence, users are more likely connected to peers sharing similar opinions fostering information exchange among like-minded individuals. On a network level, this translates into a correlation between the opinion of a user i, x_i, and the average opinions of her nearest neighbors, $\langle x_i \rangle^{NN} \equiv 1/k_i \sum_j a_{ij} x_j$, where a_{ij} represents the (static) adjacency matrix of the interaction network and $k_i \equiv \sum_j a_{ij}$ defines the degree of node i. Fig. 4 shows

color-coded contour maps of the density of users in the phase space $(x, \langle x \rangle^{NN})$, for both empirical data and the model. The interaction network in Fig. 4(a) is obtained by aggregating the dynamical contacts of the model for 15 timesteps which corresponds to the same number of temporal network snapshots; a time span for the opinion profile remained stable. Both our model (Fig. 4(a)) and all data sets under investigation (Figs. 4(b) - (d)) clearly show two bright areas characterized by a high density of users with like-minded neighbors, identifying two echo chambers corresponding to opposite opinion groups.

In conclusion, we showed that a simple model of opinion dynamics is able to reproduce several features of empirical social networks characterized by polarization and echo chambers. Our model is based on three main ingredients, inspired by empirical evidence of human interaction dynamics: i) social influence, ii) heterogenous activity of users, and iii) homophily in the interactions. We show that, in the case of controversial issues, a social re-inforcement mechanism leads to radicalization dynamics and may drive groups of agents away from the global consensus. To probe this insight, we analytically characterize the transition from consensus to radicalization dynamics in terms of the social influence strength and the controversialness of the topic, which is in good agreement to numerical simulations.

Our work opens several directions for further research, in both theoretical and empirical domains. On the theoretical side, the heterogeneity of agents' activities and the presence of homophily in the connection probability (given by Eq. (2)) needs to be incorporated in the analytical treatment of the model. Furthermore, the stability of the polarized state remains to be theoretically understood in terms of homophily. On the empirical side, our model identifies controversy as one of the main features driving the transition between global consensus and polarization. While the role of social influence has been extensively studied in the formation of polarized social systems, the effect of topic's controversialness remains poorly understood, and only recently has started to be addressed [4]. Further research should also be devoted to empirically measure the dynamics towards polarized states, to capture the transition between consensus and polarization and shed light on when and how this happens.

This work was developed within the scope of the IRTG 1740/TRP 2015/50122-0 and funded by the DFG/FAPESP. We thank K. Garimella, G. De Francisci Morales, A. Gionis, and M. Mathioudakis for sharing Twitter data with us, and G. De Francisci Morales, L. Cerekwicki and F. Sagues for helpful comments and discussions.

* Corresponding author: fabian.olit@gmail.com
† Corresponding author: michele.starnini@gmail.com

[1] H. Gil de Zúñiga and S. Valenzuela, Communication Research **38**, 397 (2011).
[2] R. Hegselmann and U. Krause, J. Artif. Soc. Simul. **5** (2002).
[3] G. Deffuant, D. Neau, F. Amblard, and G. Weisbuch, Advances in Complex Systems **3**, 87 (2000).
[4] D. Baldassarri and P. Bearman, Am. Sociol. Rev. **72**, 784 (2007).
[5] C. Castellano, S. Fortunato, and V. Loreto, Rev. Mod. Phys. **81**, 591 (2009).
[6] A. J. Morales, J. Borondo, J. C. Losada, and R. M. Benito, Chaos **25**, 033114 (2015).
[7] W. J. Brady, J. A. Wills, J. T. Jost, J. A. Tucker, and J. J. Van Bavel, Proc. Natl. Acad. Sci. **114**, 7313 (2017).
[8] M. Conover, J. Ratkiewicz, M. R. Francisco, B. Gonçalves, F. Menczer, and A. Flammini, ICWSM **133**, 89 (2011).
[9] M. D. Conover, B. Gonçalves, A. Flammini, and F. Menczer, Eur. Phys. J. Data Sci. **1**, 6 (2012).
[10] W. Cota, S. C. Ferreira, R. Pastor-Satorras, and M. Starnini, arXiv preprint arXiv:1901.03688 (2019).
[11] A. Hanna, C. Wells, P. Maurer, L. Friedland, D. Shah, and J. Matthes, in *Proceedings of the 2Nd Workshop on Politics, Elections and Data*, PLEAD '13 (ACM, New York, NY, USA, 2013) pp. 15–22.
[12] I. Weber, V. R. K. Garimella, and A. Batayneh, in *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining - ASONAM '13* (ACM Press, New York, New York, USA, 2013) pp. 290–297.
[13] J. Borge-Holthoefer, W. Magdy, K. Darwish, and I. Weber, in *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing, CSCW 2015, Vancouver, BC, Canada, March 14 - 18, 2015* (2015) pp. 700–711.
[14] S. González-Bailón, J. Borge-Holthoefer, A. Rivero, and Y. Moreno, Scientific Reports **1** (2011), 10.1038/srep00197.
[15] R. K. Garrett, Journal of Computer-Mediated Communication **14**, 265 (2009).
[2] K. Garimella, G. De Francisci Morales, A. Gionis, and M. Mathioudakis, in *Proceedings of the 2018 World Wide Web Conference*, WWW '18 (International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 2018) pp. 913–922.
[17] M. Del Vicario, A. Bessi, F. Zollo, F. Petroni, A. Scala, G. Caldarelli, H. E. Stanley, and W. Quattrociocchi, Proceedings of the National Academy of Sciences **113**, 554 (2016).
[18] M. D. Vicario, G. Vivaldo, A. Bessi, F. Zollo, A. Scala, G. Caldarelli, and W. Quattrociocchi, Scientific Reports **6** (2016), 10.1038/srep37825.
[19] T. V. Martins, M. Pineda, and R. Toral, EPL (Europhysics Letters) **91**, 48003 (2010).
[20] H. Z. Brooks and M. A. Porter, arXiv preprint arXiv:1904.09238 (2019).
[21] J. J. Timothy, arXiv preprint arXiv:1703.10138 (2017).
[22] J. K. Lee, J. Choi, C. Kim, and Y. Kim, Journal of Communication **64**, 702 (2014).
[23] M. McPherson, L. Smith-Lovin, and J. M. Cook, Annual

review of sociology **27**, 415 (2001).
[24] A. Bessi, F. Petroni, M. Del Vicario, F. Zollo, A. Anagnostopoulos, A. Scala, G. Caldarelli, and W. Quattrociocchi, The European Physical Journal Special Topics **225**, 2047 (2016).
[25] M. Mäs and A. Flache, PLOS ONE **8**, 1 (2013).
[26] P. Duggins, Journal of Artificial Societies and Social Simulation **20** (2017).
[27] S. Banisch and E. Olbrich, The Journal of Mathematical Sociology **43**, 76 (2019).
[28] M. Starnini, M. Frasca, and A. Baronchelli, Scientific reports **6**, 31834 (2016).
[29] K. Sasahara, W. Chen, H. Peng, G. L. Ciampaglia, A. Flammini, and F. Menczer, CoRR **abs/1905.03919** (2019), arXiv:1905.03919.
[30] H. Sompolinsky, A. Crisanti, and H.-J. Sommers, Physical review letters **61**, 259 (1988).
[4] K. Garimella, G. D. F. Morales, A. Gionis, and M. Mathioudakis, Trans. Soc. Comput. **1**, 3:1 (2018).
[32] A.-L. Barabási, *Bursts: The Hidden Patterns Behind Everything We Do, from Your E-mail to Bloody Crusades* (Plume, 2010).
[33] A.-L. Barabasi, Nature **435**, 207 (2005).
[34] A. Moinet, M. Starnini, and R. Pastor-Satorras, Phys. Rev. Lett. **114**, 108701 (2015).
[35] N. Perra, B. Gonçalves, R. Pastor-Satorras, and A. Vespignani, Scientific reports **2**, 469 (2012).
[36] M. Starnini and R. Pastor-Satorras, Phys. Rev. E **87**, 62807 (2013).
[37] S. Liu, N. Perra, M. Karsai, and A. Vespignani, Phys. Rev. Lett. **112**, 118702 (2014).
[38] D. Mocanu, A. Baronchelli, N. Perra, B. Gonçalves, Q. Zhang, and A. Vespignani, PloS one **8**, e61981 (2013).
[39] M. Starnini, A. Baronchelli, and R. Pastor-Satorras, Scientific Reports **7**, 8597 (2017).
[40] E. Bakshy, S. Messing, and L. A. Adamic, Science **348**, 1130 (2015).
[41] Y. Lelkes, Public Opinion Quarterly **80**, 392 (2016).
[42] A. Bessi, F. Zollo, M. Del Vicario, M. Puliga, A. Scala, G. Caldarelli, B. Uzzi, and W. Quattrociocchi, PLOS ONE **11**, e0159641 (2016)
.

Supplementary Material

NUMERICAL SIMULATIONS

For each simulation run of the model N activities a_i, one for each agent in the system, are randomly drawn from the distribution $F(a) \sim a^{-\gamma}$. The individual values of a_i are constant over time and give the probability to find agent i in the active state. If agent i is activated, it fires connections to m random distinct nodes in the network. In the case of homophily ($\beta > 0$) those m nodes are not sampled uniformly. Instead, active agents choose their interaction partners based on the probability p_{ij}, cf. Eq. (2). We assume that the opinion exchange between two agents i and j is a symmetric process. Therefore we do not discriminate between cases in which agent i contacts agent j, or vice versa. This leads to a symmetric social interaction matrix $A_{ij}(t)$ and undirected opinion formation processes between agents. For each discrete timestep a separate matrix $A_{ij}(t)$ is generated. Then we integrate the system of Eqs. (1) based on this specific matrix, for a single time step dt, using an explict fourth-order Runge-Kutta method [1]. Unless otherwise stated, the timestep is chosen to be $dt = 0.01$, which leads to a timescale separation between AD dynamics and opinion evolution of a factor of 100.

POLARIZED OPINION DISTRIBUTIONS

The opinion distributions $P(x)$ of all three investigated datasets (**obamacare**, **guncontrol** and **abortion**) show two pronounced maxima on both sides of the neutral consensus. For sufficiently high values of K and/or α the bimodular shape of the empirical distributions is reproduced by the model. This is, however, only the case if additionally homophily is introduced ($\beta > 0$).

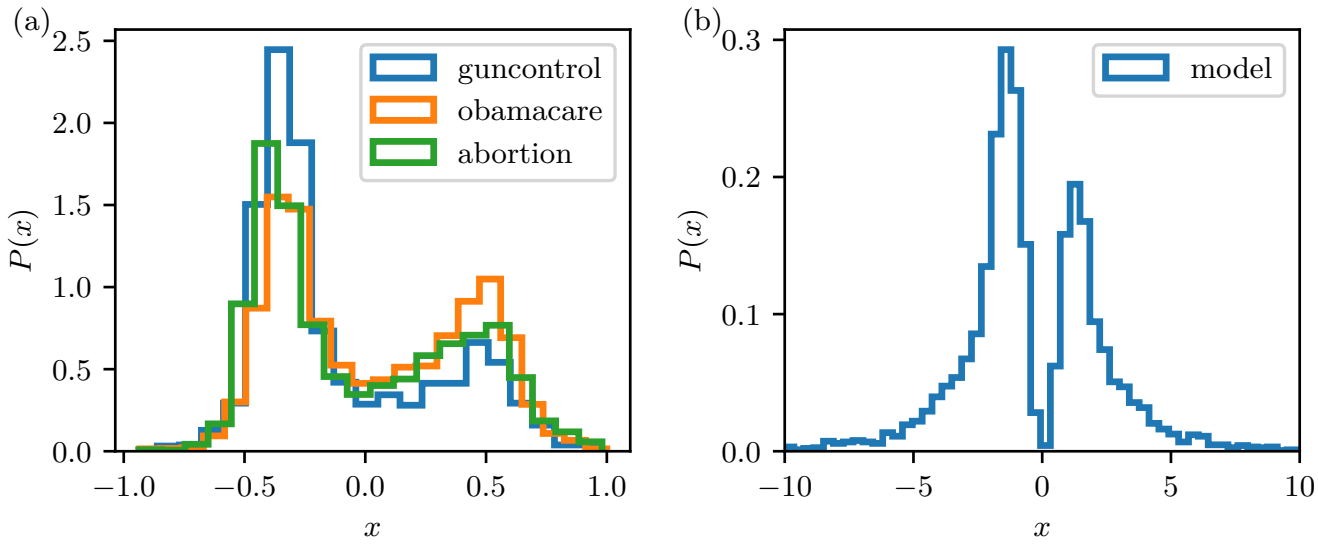

FIG. 5. Normalized opinion distributions as obtained from three different empirical data sets (a) and by simulating the model (b). For sufficiently high values of the parameters K, α and β (here $K = 3$, $\alpha = 3$, $\beta = 0.5$) the model enters a polarized state and gives rise to a bimodular opinion distribution, which is in qualitative agreement with the investigated Twitter data.

LIFETIME OF POLARIZED STATES

Polarized opinion states (cf. Fig. 1(b)), will eventually decay into one-sided radicalized states, cf. Fig. 1(c). However, their lifetimes τ strongly increases with the value of β. In Fig. 6 we depict the mean lifetime, $\langle\tau\rangle$, as a function of β for two different values of the controversialness α. Note the logscale on the y-axis, i.e. the strong dependence of the mean lifetimes on β, which even exceed an exponential growth for higher values of the homophily.

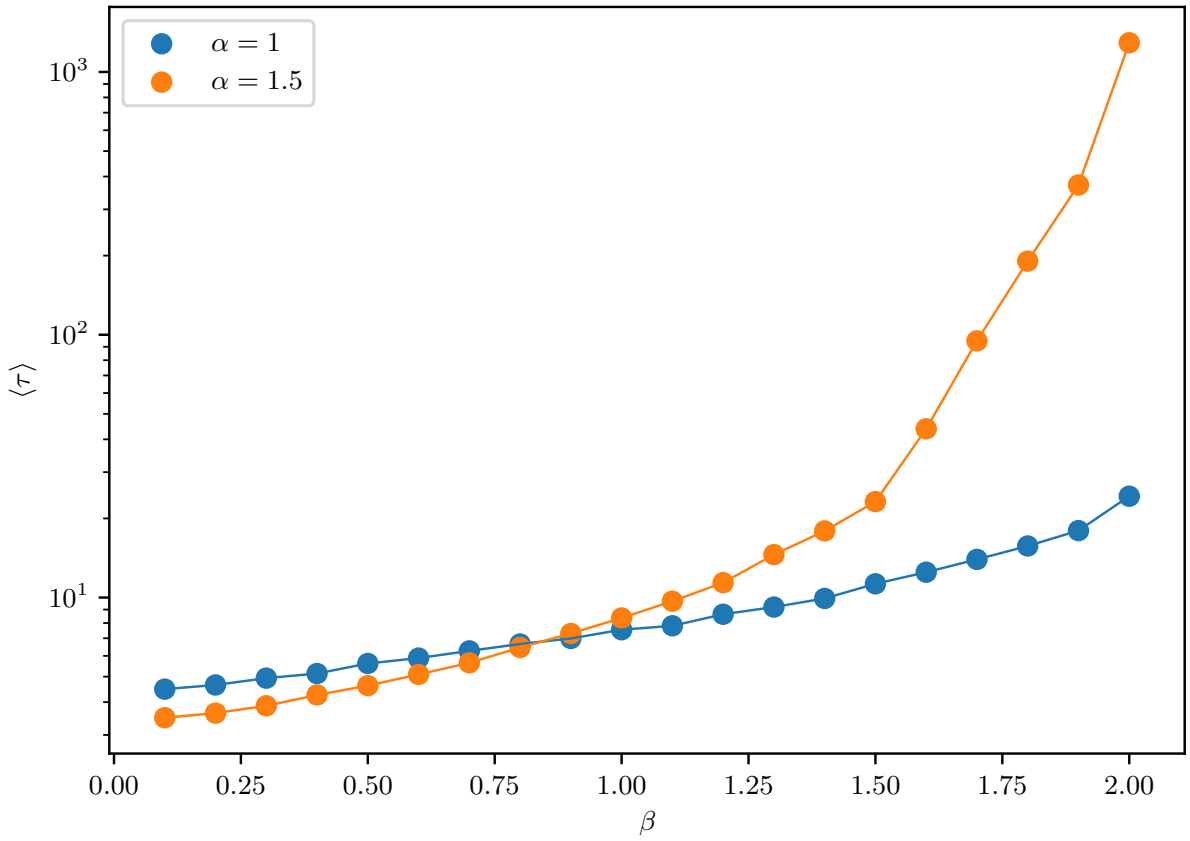

FIG. 6. The mean life-time of the polarized state strongly increases with the value of β. Each dot corresponds to the average of 100 simulations of populations of $N = 250$ agents with $dt = 0.05$. The colors correspond to different values of the controversialness, while $K = 1$ for both depicted curves.

APPROXIMATION OF THE CRITICAL CONTROVERSIALNESS α_c

Fast network dynamics allow the adjacency matrix $A_{ij}(t)$ (cf. Eq. (1) of the main text) to be approximated by its time average, yielding

$$\dot{x}_i = -x_i + K \sum_{j=1}^{N} \langle A_{ij}(t) \rangle_t \tanh(\alpha x_j) \,. \tag{4}$$

Neglecting homophily ($\beta = 0$) the probability that two agents i and j are connected at time t does not depend on their respective opinions x_i and x_j and reduces to

$$\langle A_{ij}(t) \rangle_t = \frac{m}{N}(a_i + a_j) \,, \tag{5}$$

which, averaged over all activities in the system, becomes

$$\Lambda = \frac{2m}{N} \langle a \rangle \,. \tag{6}$$

For activities distributed according to a power law distribution $F(a) = \frac{1-\gamma}{1-\epsilon^{1-\gamma}} a^{-\gamma}$, normalized on the interval $a \in [\epsilon, 1]$, we have $\langle a \rangle = \frac{1-\gamma}{2-\gamma} \frac{1-\epsilon^{2-\gamma}}{1-\epsilon^{1-\gamma}}$. Using Eq. (6) to simplify Eq. (4) we get

$$\dot{x}_i = -x_i + K\Lambda \sum_{j=1}^{N} \tanh(\alpha x_j) \,. \tag{7}$$

To study the transition from neutral consensus to radicalization dynamics within this mean-field approach we compute the Jacobian matrix of the system of Eqs. (7), yielding

$$\mathbb{J}|_{\mathbf{x}=0} = \begin{bmatrix} -1 & K\Lambda\alpha & \dots & K\Lambda\alpha \\ K\Lambda\alpha & -1 & \dots & K\Lambda\alpha \\ \vdots & \vdots & \vdots & \vdots \\ K\Lambda\alpha & K\Lambda\alpha & \dots & -1 \end{bmatrix}, \tag{8}$$

where all off-diagonal elements equal $K\Lambda\alpha$. The largest eigenvalue of $\mathbb{J}$ reads

$$\tilde{\lambda} = (N-1)K\alpha\Lambda - 1 = (N-1)\frac{2m\langle a\rangle K\alpha}{N} - 1 \tag{9}$$

and determines the stability of the fixed point $\mathbf{x} = 0$ with respect to small perturbations. For $\tilde{\lambda} > 0$ the neutral consensus destabilizes, hence, $\tilde{\lambda} = 0$ defined the critical value of controversialness α_c, i.e.

$$\alpha_c = \frac{N}{(N-1)}\frac{1}{2mK\langle a\rangle}. \tag{10}$$

In the limit of $N \to \infty$, Eq. (3) of the main text is recovered.

TWITTER DATA

The datasets used in this work have been collected, analyzed and validated in previous works [2–4]. We use three different datasets from Twitter, each of which contains a set of tweets on a given controversial topic of discussion: **abortion**, **obamacare**, **guncontrol**. In order to keep the three datasets independent, we exclude users present in more than one dataset. In Ref. [2], the authors performed simple checks to remove bots, using minimum and maximum thresholds for the number of tweets per day, followers, friends, and ensure that the account is at least one year old at the time of data collection.

Each dataset is built by collecting tweets posted during specific events that led to an increased interest in the respective topic, during a time period of one week around the event (3 days before and 3 days after the event). Users with less than 5 tweets about the issue during this time window were discarded. The final numbers of users for each data set are: **abortion**: 4130, **obamacare**: 4828, **guncontrol**: 1838.

In Ref. [2], for each dataset, the directed follower network among users has been reconstructed: a directed link from node u to node v indicates that user u follows user v. For each user, a political leaning score is inferred on the basis of the content s/he produces, by using a ground truth of political leaning scores of various news organizations (e.g., nytimes.com) obtained from Bakshy et al. [4]. Specifically, each news organization is classified by a score which takes values between 0 and 1. A value close to 1 (0) indicates that the domain has a conservative (liberal) bent in their coverage. From this classification, the political leaning score, or opinion, of each user i is reconstructed by considering all tweets posted by user i that contain a link to an online news organization with a known political leaning. Each tweet is thus associated with an opinion, corresponding to the political leaning of the news organization linked. The political leaning of the user i is defined as the average of the opinions expressed in his/her tweets. Note that we transformed the original political leaning inferred in Ref. [2], from 0 to 1, into a score from -1 to 1, for coherence with the model.

RELATION BETWEEN USER OPINIONS AND ACTIVITIES

The U-shaped relation between opinions x and activities a is a generic feature of the radicalization dynamics and occurs as soon as the system is in a polarized state. As an example in Fig. 7 we vary the social interaction strength K from top to bottom, while leaving all other model parameters constant. For increasing values of K the convictions of agents of similar activities are increased. This results in a flattening of the U-shaped relation between activities and opinions.

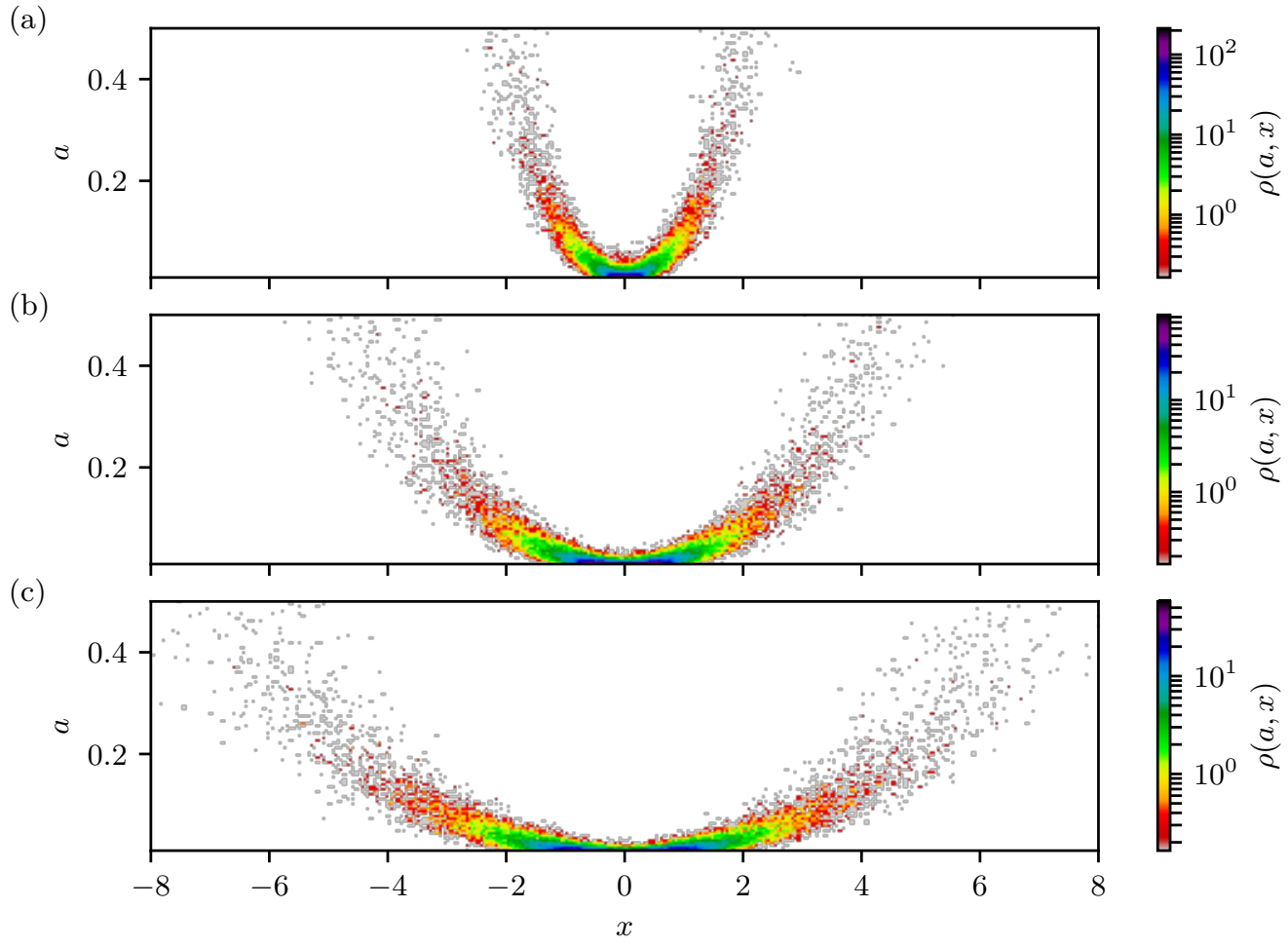

FIG. 7. Normalized histograms of simulation results in x-a space which depict the relation between opinions x and activities a of agents in a polarized state. The color encodes the density of agents. U-shapes for increasing values of the social interaction strength are depicted from top to bottom ($K = 1, 2, 3$), while we fixed the remaining parameters $\alpha = 3$ and $\beta = 1$.

* Corresponding author: fabian.olit@gmail.com
† Corresponding author: michele.starnini@gmail.com

[1] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical recipes 3rd edition: The art of scientific computing* (Cambridge university press, 2007).
[2] K. Garimella, G. De Francisci Morales, A. Gionis, and M. Mathioudakis, in *Proceedings of the 2018 World Wide Web Conference*, WWW '18 (International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 2018) pp. 913–922.
[3] V. Garimella and I. Weber, in *Proceedings of the 11th International Conference on Web and Social Media, ICWSM 2017* (AAAI press, 2017) pp. 528–531.

[4] K. Garimella, G. D. F. Morales, A. Gionis, and M. Mathioudakis, Trans. Soc. Comput. **1**, 3:1 (2018).