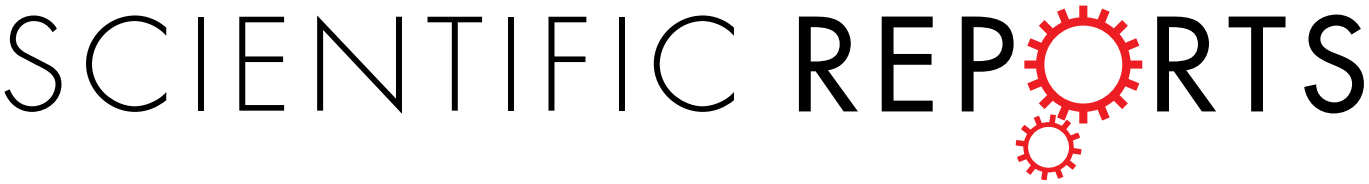

OPEN

Modeling confirmation bias and polarization

Michela Del Vicario[1], Antonio Scala[1,2], Guido Caldarelli[1], H. Eugene Stanley[3] & Walter Quattrociocchi[1]

Received: 12 July 2016
Accepted: 06 December 2016
Published: 11 January 2017

Online users tend to select claims that adhere to their system of beliefs and to ignore dissenting information. Confirmation bias, indeed, plays a pivotal role in viral phenomena. Furthermore, the wide availability of content on the web fosters the aggregation of likeminded people where debates tend to enforce group polarization. Such a configuration might alter the public debate and thus the formation of the public opinion. In this paper we provide a mathematical model to study online social debates and the related polarization dynamics. We assume the basic updating rule of the *Bounded Confidence Model* (BCM) and we develop two variations a) the *Rewire with Bounded Confidence Model* (RBCM), in which discordant links are broken until convergence is reached; and b) the *Unbounded Confidence Model*, under which the interaction among discordant pairs of users is allowed even with a negative feedback, either with the rewiring step (RUCM) or without it (UCM). From numerical simulations we find that the new models (UCM and RUCM), unlike the BCM, are able to explain the coexistence of two stable final opinions, often observed in reality. Lastly, we present a mean field approximation of the newly introduced models.

Online users tend to select claims that adhere to their system of beliefs and to ignore dissenting information[1–5]. The wide availability of content on the web fosters the aggregation of likeminded people where the discussion tends to enforce group polarization[6,7]. Confirmation bias, indeed, plays a pivotal role in viral phenomena[8]. Under such conditions public debates, in particular on social relevant issues, tend to further fragment and polarize the public opinion[9,10].

To better understand this process, in this paper we provide a mathematical model mimicking polarization in online social dynamics.

Opinion dynamics, have been widely investigated in recent years, using different approaches from statistical physics and network science[11]. Classical examples of opinion dynamics models include the Sznajd model[12], the voter model[13–15], the majority rule model[16,17], and the bounded confidence model (BCM)[18–20]. Besides the different assumptions and dynamics rules, for all the cited models the consensus state, in which all agents share the same opinion, is reached for a value of the tolerance parameter big enough.

However, consensus in far from common in real world and Internet based opinion exchanges. A recent study showed the emergence of *polarized communities*, i.e., *echo chambers*, in online social networks[8]. Inside these communities, homogeneity appears to be the primary driver for the diffusion of contents. Both polarization and homogeneity might be the result of the conjugate effect of *confirmation bias* and *social influence*. Confirmation bias is the tendency to acquire or process new information in a way that confirms one's preconceptions and avoids contradiction with prior belief[21]. Social influence is the process under which one's emotions, opinions, or behaviors are affected by others. In particular, *informational influence* occurs when individuals accept information from others as evidence about reality[22,23].

Previous studies[24,25] proposed a non consensus opinion model (NCO) that allowed for the stable coexistence of two opinions by also considering the opinion of the user herself when applying the majority rule update[24], while in ref. 25 the competition between two groups is investigated by the introduction of a set of contrarians in one of the two. The survival of a two-opinions state is studied in ref. 26 from a different point of view, considering the emergence of spontaneous recovery of failed nodes and the majority rule update. Both these models assume only two opinion states (± 1) and a majority rule update, with the novelty of accounting for the individual opinion[24,25] and for an external source of influence[26].

[1]Laboratory of Computational Social Science, Networks Dept, IMT School for Advanced Studies, 55100 Lucca, Italy. [2]ISC-CNR Uos "Sapienza", 00185 Roma, Italy. [3]Boston University, Center for Polymer Studies, Department of Physics, Boston, Massachusetts 02215, USA. Correspondence and requests for materials should be addressed to W.Q. (email: walter.quattrociocchi@gmail.com)

Authors in ref. 27 investigate the emergence of extreme opinion trends in society by employing statistical physics modeling and analysis on polls. By developing an activation model of opinion dynamics with interaction rules based on the existence of individual "stubbornness", they discover a sharp statistical predictor of the rise of extreme opinion trends in society in terms of a nonlinear behavior of the number of individuals holding a certain extreme view and the number of individuals with a moderate opinion and extreme opinion. A model grounded on the BCM and accounting for the interconnection and complexity of the online environment as well as the competition among sources of information is presented in ref. 28. In a recent study[29], authors analyze the effects of the interplay between homophily, social influence, and confirmation bias in the emergence of segregation and echo chambers.

People shape their opinions on the basis of both confirmation bias and social influence, a combination of these two forces generates the observed polarization of communities and homogeneous links[8]. Accounting for this phenomenon, we build a model of opinion dynamics and network's evolution that considers both mechanisms and expands itself from the classical *Bounded Confidence Model* (BCM)[18]. We consider two variations of the model: the *Rewire with Bounded Confidence Model* (RBCM), in which discordant links are broken until convergence is reached; and the *Unbounded Confidence Model*, under which interaction among discordant pairs of users is allowed and a negative updating rule is introduced, either with the rewiring step (RUCM) or without it (UCM). As for the BCM, our models assume a continuous interval of opinions.

The paper is structured as follows. In the first section, *Results and Discussion*, we first present the new models and give an account of the simulation results, then we present a mean field approximation of the newly introduced models. In the last section, *Methods*, we provide references to the methods employed and give a brief overview of the BCM and its convergence results.

Results and Discussion

Models. The paper is a model study derived from the paper[8] on which we provide evidence of the polarizing effect of different narratives and the echo chamber structure of cascades. Hence, here we exploit the bounded confidence proviso (i.e., interacting with an information/opinion iff this is close enough to the agent state) that well mimics the confirmation bias (i.e., acquiring information that adhere to a specific system of beliefs) process.

The Bounded Confidence Model (BCM)[18,20] is a well known opinion dynamics model that takes into account a set of N agents arranged on a complex network G, each of which holds an opinion x_i, $i \in \{1, \ldots, N\}$, uniformly distributed in [0, 1]. Two agents interact if and only if they are connected in G and their present opinions are close enough, i.e. iff $j \in N_G(i)$ and $|x_i - x_j| < \varepsilon$, for $\varepsilon \in [0, 1]$. If these conditions hold, the two agents change their opinions according to Eq. (1), otherwise they do not interact at all:

$$\begin{cases} x_i = x_i + \mu(x_j - x_i) \\ x_j = x_j + \mu(x_i - x_j) \end{cases}, \tag{1}$$

Refer to the section *Methods* for further information on the BCM and for convergence results.

Starting from the BCM we introduce three new models of opinion dynamics and network evolution. The first model we consider is the Rewire with Bounded Confidence Model (RBCM) that considers the same framework as in BCM and involves two phases. In phase one we run the *rewiring steps* in which each agent i interacts with a randomly chosen neighbor j and, if the distance between the two opinions is above the tolerance ε, then their link is broken and i is rewired to a randomly chosen agent $k \in \{1, \ldots, N\}/(N_G(i) \cup \{i\})$. To be specific, we introduce a new distance $|.|_\tau : [0, 1] \times [0, 1] \to [0, 0.5]$ defined as:

$$|x_i - x_j|_\tau = |x_i - x_j - \rho(x_i - x_j)|, \tag{2}$$

where $i, j \in \{1 \ldots, N\}$ and the adjustment ρ ensure the *Periodic Boundary Conditions* (PBC) (refer to the section *Methods* for further details). The condition for the random rewire becomes: $|x_i - x_j|_\tau \geq \varepsilon$, for $\varepsilon \in [0, 0.5]$. Note that we restrict our attention to $\varepsilon \in [0, 0.5]$ after noticing that $\forall\, x, y \in \{1, \ldots, N\}$ we get $|x - y|_\tau \in [0, 0.5]$. We will assume $\varepsilon \in [0, 0.5]$ throughout the paper. Phase one ends when all links have an opinion distance below the tolerance ε.

In phase two we run the BCM on the rewired network. The BCM allows the interaction only for those pairs whose opinion distance is below the tolerance ε, as all the couples in the rewired network are concordant, all the randomly chosen pairs will interact and readjust their opinion according to the rule in Eq. (1), where μ is taken in the interval (0, 0.5).

The Unbounded Confidence Model (UCM) is the second of the models that we introduce and its novelty is to allow the interaction for every randomly chosen pair of neighbors (i, j). To be specific, if two agents have concordant opinions, i.e. if $|x_i - x_j|_\tau < \varepsilon$, as for the previous model, we adjust x_i and x_j by Eq. (1). However, if their opinions are discordant, i.e. if $|x_i - x_j|_\tau \geq \varepsilon$, we use a new updating rule, see Eq. (3), that enables us to replicate the empirically observed repulsion of discordant opinions:

$$\begin{cases} x_i = x_i - \mu[x_j - x_i - \rho(x_j - x_i)] \\ x_j = x_j - \mu[x_i - x_j - \rho(x_i - x_j)] \end{cases}, \tag{3}$$

where μ is taken in the interval (0, 0.5) and $\rho(.)$ is defined in Eq. (14), in the *Methods* section. The adjustment $\rho(.)$ ensures the PBC by maintaining the opinions inside the interval [0, 1].

The last model that we introduce is the Rewire with Unbounded Confidence Model (RUCM) that again allows the interaction for every randomly chosen pair of users (i, j) but at the same time allows for the random rewiring

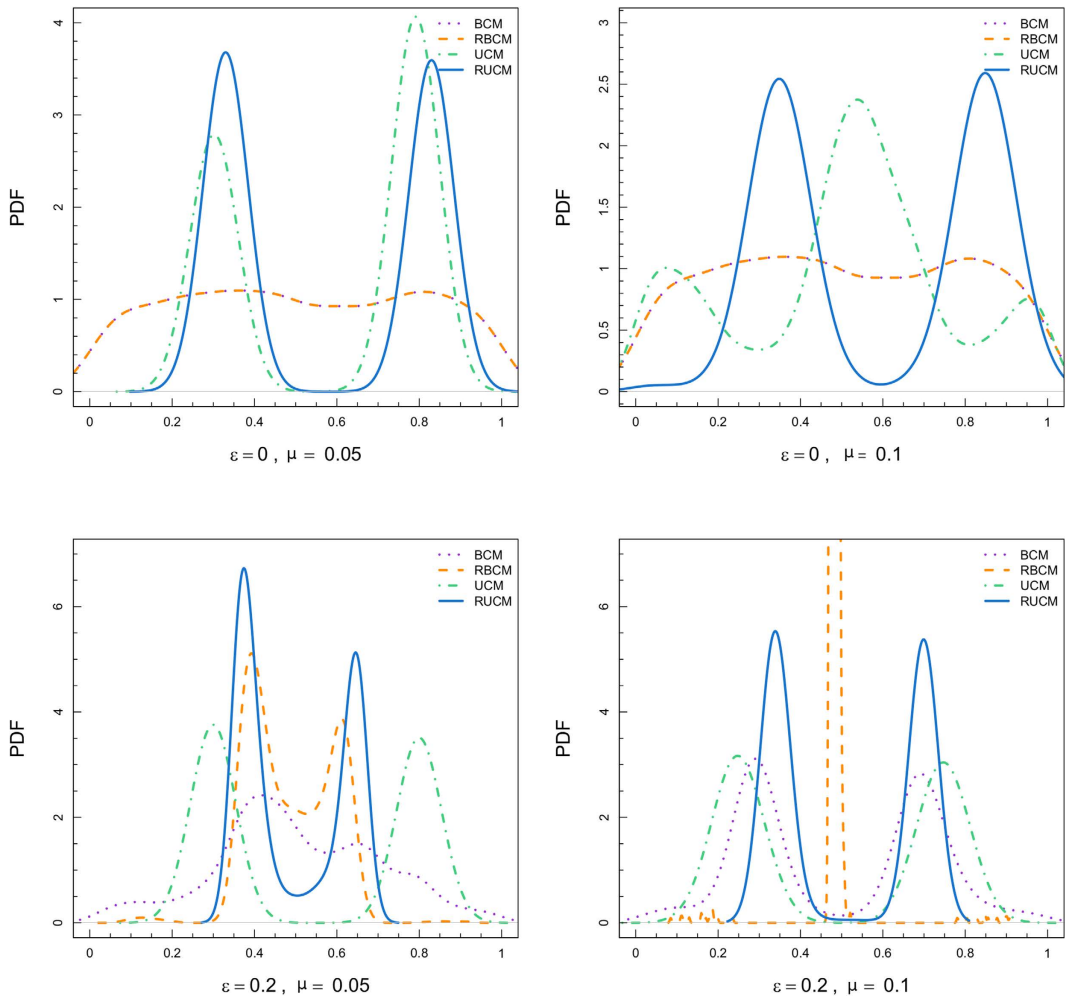

Figure 1. Probability density functions (PDFs) of final opinion, after a maximum of 10^5 time steps or until convergence is reached, for four different combinations of the parameters (ε, μ). In the upper left figure we have $(\varepsilon, \mu) = (0, 0.05)$, in the upper right $(\varepsilon, \mu) = (0, 0.1)$, in the lower left $(\varepsilon, \mu) = (0.2, 0.05)$, and in the lower right $(\varepsilon, \mu) = (0.2, 0.1)$. In all figures the blue solid curve is for RUCM, the green dot-dashed one for UCM, the violet dotted one for BCM, and the pale orange dashed one for RBCM. We observe a bimodal distribution for RUCM and UCM, representing the coexistence of two polarized stable opinions.

of discordant pairs. Specifically, if $|x_i - x_j|_\tau < \varepsilon$, then we adjust x_i and x_j by Eq. (1). If $|x_i - x_j|_\tau \geq \varepsilon$, then we adjust x_i and x_j by Eq. (3), the link between nodes i and j is broken, and a new link between i and a randomly chosen user $k \in \{1, \ldots, N\}/(N_G(i) \cup \{i, j\})$ is created.

Simulation Results. We consider different types of complex networks in the simulations: The Erdös-Rényi random network (ER)[30] characterized by a Poisson degree distribution with average degree $\langle 2 \rangle$, the scale-free network (SF)[31] characterized by a power-law degree distribution $P(k) \sim k^{-\gamma}$ (SF networks are created by using the classic implementation of the Barabási and Albert model, hence $\gamma = 3$), and the small-world network (SW)[32] with rewiring probability equal to 0.2 and neighborhood dimension equal to 2. We restrict our attention to SF networks and report the results for the Erdös-Rényi random network and the small-word network in Supplementary Figs S4 and S5.

Hence, we show the results of Monte Carlo simulations of the BCM and the three new models on a SF network of 2000 nodes with the parameters (ε, μ) varying in the parameter space $[0, 0.5] \times [0, 0.5]$, for $T = 10^5$ times steps and we averaged our results over 5 repetitions. Note that the final state, under the different parameters combinations, is always reached before $T = 10^5$. Refer to Supplementary Fig. S3 for further details. Figure 1 shows the probability density functions (PDFs) of final opinion, after a maximum of 10^5 time steps, for four different combinations of the pair of parameters (ε, μ): $(\varepsilon, \mu) \in \{(0, 0.05), (0, 0.1), (0.2, 0.05), (0.2, 0.1)\}$. The blue solid and the green dot-dashed curves refer to the newly introduced RUCM and UCM respectively, while the violet dotted curve is for BCM and the orange dashed for RBCM. For all the parameter choices we observe a bimodal opinion distribution in the cases of RUCM and UCM (note that we assume periodic boundary conditions). It is interesting to note that for UCM and RUCM there are two polarized opinions also for $\varepsilon = 0$, while in that case BCM and RBCM show no changes with respect to the initial uniform distribution.

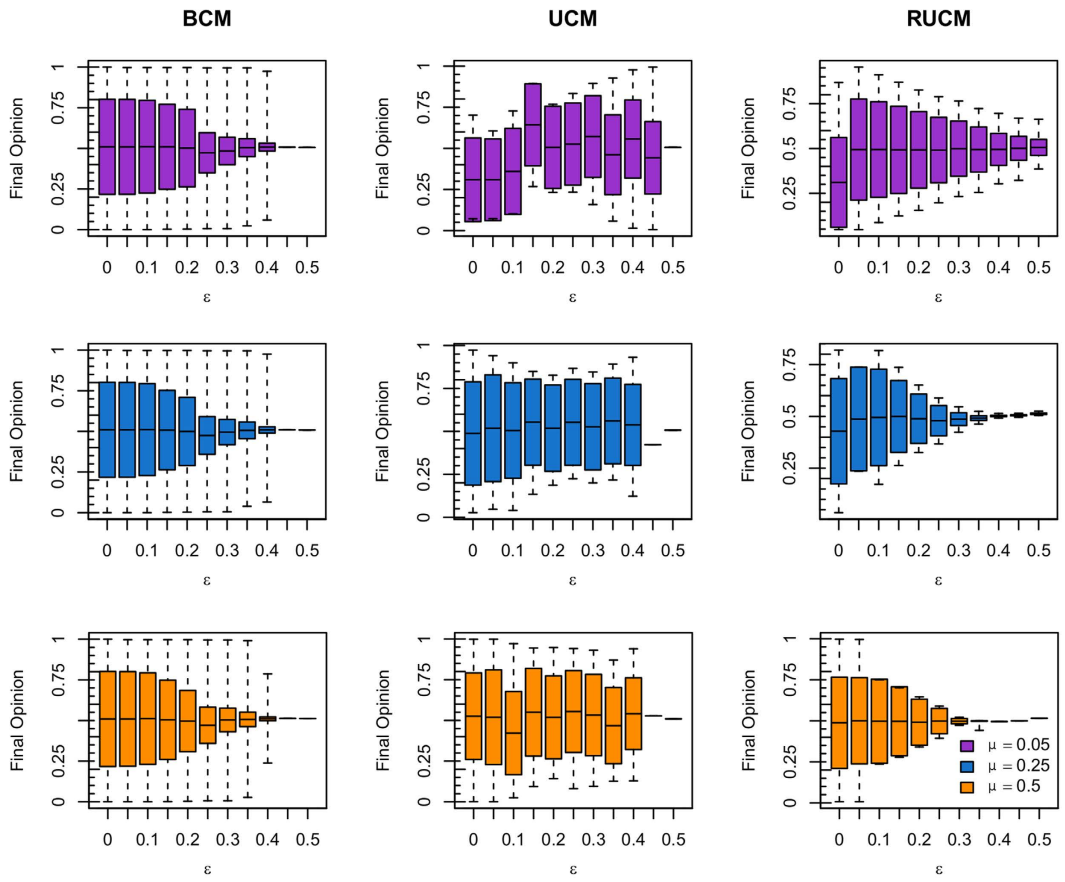

Figure 2. Summary statistics (mean, standard deviation, 1st quantile, and 3rd quantile) of the final opinion distributions for varying ε and three different values of μ: violet denotes $\mu = 0.05$, blue denotes $\mu = 0.25$, and orange denotes $\mu = 0.5$. The left column is for BCM, the central one for UCM, and the right one for RUCM.

Figure 2 reports a collection of summary statistics (mean, standard deviation, 1st quantile, and 3rd quantile) of the final opinion distributions for varying ε and three different values of μ (violet is for $\mu = 0.05$, blue for $\mu = 0.25$, and orange for $\mu = 0.5$). The left column is for BCM, the central one for UCM, and the right one for RUCM. We omit the results for RBCM as we observe from the simulations that, after the rewiring steps, the dynamics are similar to the BCM case but with a faster convergence, refer to the Supplementary Information for an in depth analysis of the RBCM model. We observe different mechanisms for the two newly introduced models, such as a faster convergence to the consensus state for RUCM. However, we need to study the final number of peaks to better characterize the differences between UCM and RUCM, and to relate them with the results for the classical BCM.

Final Distribution of Peaks. We perform Monte Carlo simulations of the BCM, UCM, and RUCM on a scale-free network of 2000 nodes with $(\varepsilon, \mu) \in [0, 0.5] \times [0, 0.5]$, for $T = 10^5$ times steps, that are sufficient to reach the final state of the system under the different parameters combinations (the results are averaged over 5 repetitions). Given the final distributions of opinions obtained by the simulations, we compute the number of peaks of opinions as the local maxima in the distribution of frequencies of opinions. To be specific, we divide the interval [0, 1] in 100 bins of length 0.01 and consider the frequencies of values falling in each interval. We regard two peaks to be separate if the distance between the middle points of the respective bins is smaller than 0.1. All the results are averaged over 5 repetitions.

Figure 3 shows the final distribution of peaks of BCM for varying $(\varepsilon, \mu) \in [0, 0.5] \times [0, 0.5]$. The corresponding result for the RBCM model is shown in Supplementary Fig. S2. The final peaks distribution complies with theoretical[33,34] and simulation[18] results from previous work. Figure 4 shows the final peaks distribution of UCM (left) and RUCM (right) for varying $(\varepsilon, \mu) \in [0, 0.5] \times [0, 0.5]$. For both models we observe a large area of the parameter space for which two final opinions coexist. We register a faster convergence to the consensus state for the RUCM (w.r.t. UCM), that is due to the rewiring rule. Also, we observe that for the RUCM there is a direct transition from many opinions to two opinions, as well as from two opinions to consensus, while for the UCM there is an intermediate area where 3 or 4 opinions emerge, respectively shown in yellow and pale orange.

Comparing Figs 3 and 4, we see that the new models, unlike the BCM, are able to explain the coexistence of two stable final opinions, often observed in reality. Another important difference with respect to the BCM is that

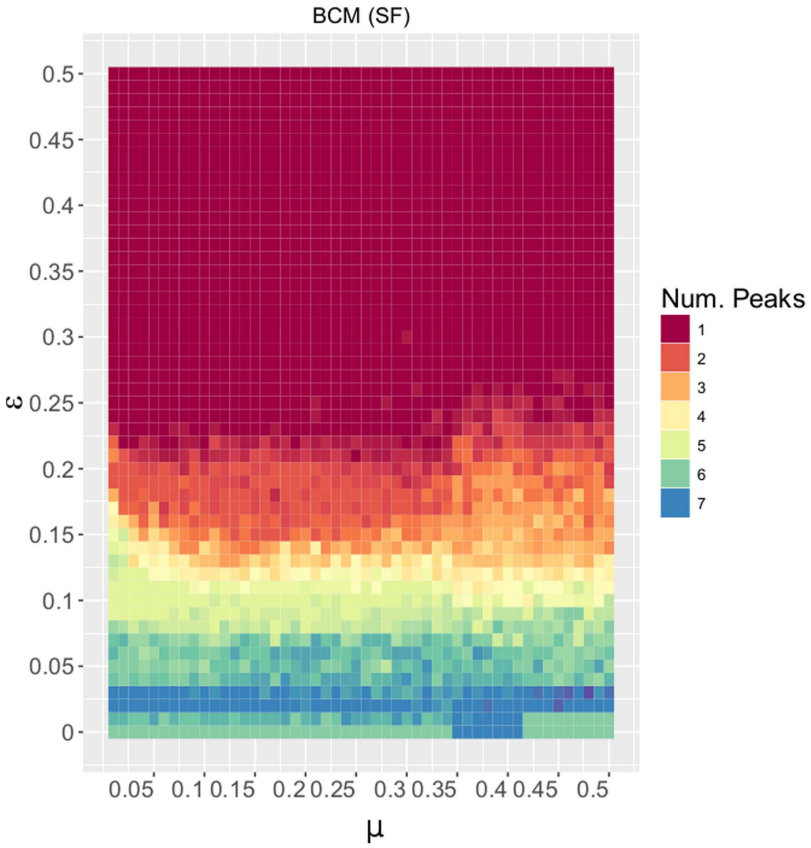

Figure 3. Final distribution of peaks for the BCM, with varying $(\varepsilon, \mu) \in [0, 0.5] \times [0, 0.5]$. The Monte Carlo simulations are carried on a Scale-Free network with 2000 nodes for $T = 10^5$ times steps, that are sufficient to reach the final state of the system under the different parameters combinations (all results are averaged over 5 repetitions).

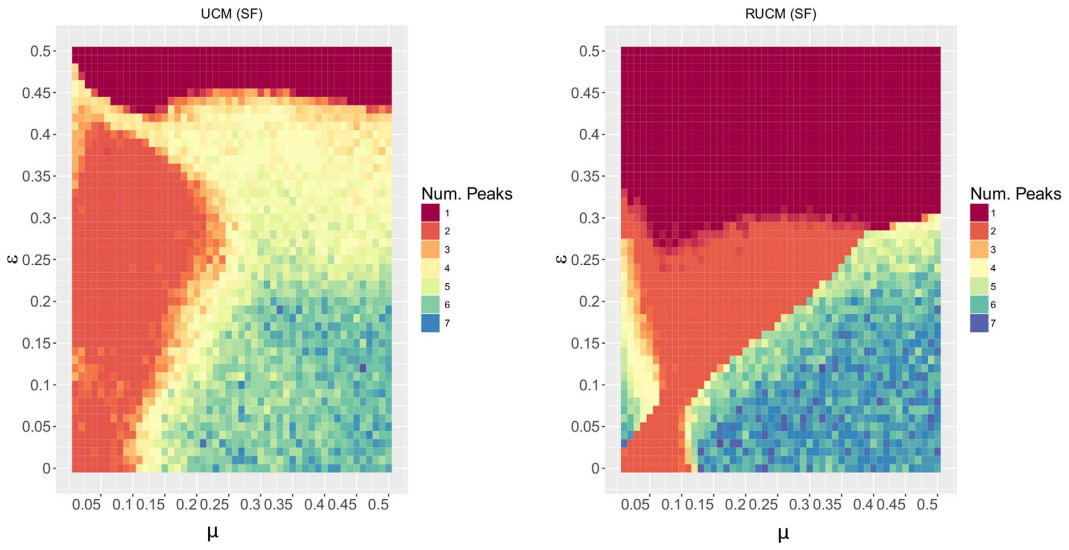

Figure 4. Final distribution of peaks for the UCM (left) and RUCM (right), with varying $(\varepsilon, \mu) \in [0, 0.5] \times [0, 0.5]$. The Monte Carlo simulations are carried on a Scale-Free network with 2000 nodes for $T = 10^5$ times steps, that are sufficient to reach the final state of the system under the different parameters combinations (all results are averaged over 5 repetitions).

the μ parameter assumes an important role in tuning the number of final opinions peaks. The dependence of the number of final peaks on the μ parameter is stronger for the RUCM, where we observe a clear transition from many opinions to exactly two on the diagonal.

Mean Field Approximation. For the RBCM, after the rewiring steps, all connected agents have an opinion distance below ε, meaning that they will always interact. The time rate of change of $\mathbb{P}(x, t)$ is equal to:

$$\frac{\partial \mathbb{P}(x, t)}{\partial t} = -\mathbb{P}(x, t)\int_{-1}^{1}\mathbb{P}(x+y, t)dy + \frac{1}{(1-\mu)}\int_{-1-2x}^{1-2x}\mathbb{P}(x+y, t)\mathbb{P}\left(x - \frac{\mu}{1-\mu}y, t\right)dy. \tag{4}$$

Considerations analogous to the BCM case hold (see the Section *Material and Methods*). A faster convergence scale is also observed in the simulations.

In the UCM and RUCM case we consider two updating rules: the one in Eq. (1) if the opinions (x_i, x_j) of the agents are close enough ($|x_i - x_j|_\tau < \varepsilon$) and the one in Eq. (3) if they are not ($|x_i - x_j|_\tau \geq \varepsilon$). Thus the opinions will change according to $(x_i, x_j) \rightarrow (\hat{x}_i, \hat{x}_j)$:

$$\begin{pmatrix}\hat{x}_i\\ \hat{x}_j\end{pmatrix} = \begin{pmatrix}1-\vartheta_\varepsilon\mu + (1-\vartheta_\varepsilon)\mu & \vartheta_\varepsilon\mu - (1-\vartheta_\varepsilon)\mu\\ \vartheta_\varepsilon\mu - (1-\vartheta_\varepsilon)\mu & 1-\vartheta_\varepsilon\mu + (1-\vartheta_\varepsilon)\mu\end{pmatrix}\begin{pmatrix}x_i\\ x_j\end{pmatrix} + (1-\vartheta_\varepsilon)\mu\begin{pmatrix}\rho(x_j - x_i)\\ \rho(x_i - x_j)\end{pmatrix}, \tag{5}$$

where $\vartheta_\varepsilon = \vartheta(\varepsilon - |x_i - x_j|_\tau)$ is the Heaviside theta function that equals 1 if $\varepsilon - |x_i - x_j|_\tau < 0$, 0 otherwise, and ρ is defined in Eq. (14). There are two ways in which the density of opinion x changes at every time step t: either an agent moves away from x after an interaction (I^-) or she arrives in x after an interaction (I^+). Let $\mathbb{P}(x, t)dx$ be the fraction of agents whose opinion at time t lies in the interval $[x, x+dx]$, then its time rate of change is:

$$\frac{\partial \mathbb{P}(x, t)}{\partial t} = I^-(x, t) + I^+(x, t). \tag{6}$$

The negative part is defined as in the BCM case but for a wider interval:

$$I^-(x, t) = -\mathbb{P}(x, t)\int_{-1}^{1}\mathbb{P}(x+y, t)dy, \tag{7}$$

as $I^-(x, t)$ is simply the probability that an agent with opinion x interacts with some other agent and thus moves away from x. For $I^+(x, t)$ we have two terms depending on the distance of the initial opinions:

$$I^+(x, t) = I_1^+(x, t) + I_2^+(x, t), \tag{8}$$

for the first term we get the same expression as in the BCM case:

$$I_1^+(x, t) = \frac{1}{(1-\mu)}\int_{-\varepsilon-2x}^{\varepsilon-2x}\mathbb{P}(x+y, t)\mathbb{P}\left(x - \frac{\mu}{1-\mu}y, t\right)dy. \tag{9}$$

For $I_2^+(x, t)$ we have to consider the negative update in Eq. (3), and the integrals are over the interval for which $|x_1 - x_2|_\tau \geq \varepsilon$:

$$\begin{aligned} I_2^+(x, t) &= \iint \mathbb{P}(x_1, t)\mathbb{P}(x_2, t)\delta(x + \mu x_1 - (1+\mu)x_2 - \mu\rho))dx_1dx_2\\ &= \frac{1}{(1+\mu)}\int dx_1\mathbb{P}(x_1, t)\int\mathbb{P}(x_2, t)\delta\left(x_2 - \frac{x + \mu x_1 - \mu\rho}{1+\mu}\right)dx_2\\ &= \frac{1}{(1+\mu)}\int_{|x_1-x_2|_\tau\geq\varepsilon}\mathbb{P}(x_1, t)\mathbb{P}\left(\frac{x + \mu(y-\rho)}{1+\mu}\right)dx_1\\ &= \frac{1}{(1+\mu)}\int_{[-1,-\varepsilon-2x]\bigcup[\varepsilon-2x,1]}\mathbb{P}(x+y, t)\mathbb{P}\left(x + \frac{\mu}{1+\mu}(y-\rho)\right)dy, \end{aligned} \tag{10}$$

where $\rho = \rho_{x_2 - x_1}$. Hence we obtain:

$$\begin{aligned} \frac{\partial \mathbb{P}(x, t)}{\partial t} &= -\mathbb{P}(x, t)\int_{-1}^{1}\mathbb{P}(x+y, t)dy\\ &\quad + \frac{1}{(1-\mu)}\int_{-\varepsilon-2x}^{\varepsilon-2x}\mathbb{P}(x+y, t)\mathbb{P}\left(x - \frac{\mu}{1-\mu}y, t\right)dy\\ &\quad + \frac{1}{(1+\mu)}\int_{[-1,-\varepsilon-2x]\bigcup[\varepsilon-2x,1]}\mathbb{P}(x+y, t)\mathbb{P}\left(x + \frac{\mu}{1+\mu}(y-\rho)\right)dy. \end{aligned} \tag{11}$$

When all agents interact positively, i.e. when $\varepsilon \geq 1/2$, the third term of the rate equation disappears and we are again in the BCM case, where consensus is reached asymptotically and:

$$\mathbb{P}_\infty(x) = M_0\delta(x). \tag{12}$$

For smaller values of ε, we rely on simulations results. We notice that the final state is a single peak as long as $\varepsilon \in (0.45, 0.5)$ for the UCM, or $\varepsilon \in (0.3, 0.5)$ for the RUCM (with the exception of those points for which μ is near to zero).

Unlike for BCM, in the new models the parameter μ plays an important role in the evolution of the distribution of opinions. For both UCM and RUCM we have the coexistence of two opinions in the final state for a wide region of the (ε, μ)-plane, this region varies for the two models, in particular the faster convergence to the consensus state for the RUCM is due to the rewiring rule. For smaller values of ε, and outside the two opinions region, we showed by numerical simulations that consensus is not reached, and many opinions at distance larger than ε coexist.

Conclusions

In recent years opinion dynamics has attracted much interest from the fields of both statistical physics and social science. In classical models such as the Sznajd model, the voter model, the majority rule model, and the bounded confidence model, consensus is eventually reached, for values of the tolerance parameter big enough. However, in face-to-face and online opinion exchanges, consensus is not commonly achieved, and classical models fail to explain this empirically observed fact.

We propose a model of opinion dynamics capable of reproducing the empirically observed coexistence of two stable opinions. We assume the basic updating rule of the BCM and we develop two variations of the model: the *Rewire with Bounded Confidence Model* (RBCM), in which discordant links are broken until convergence is reached; and the *Unbounded Confidence Model*, under which the interaction among discordant pairs of users is allowed and a negative updating rule is introduced, either with the rewiring step (RUCM) or without it (UCM).

From numerical simulations we find that the new models (UCM and RUCM), unlike the BCM, are able to explain the coexistence of two stable final opinions, often observed in reality. Another important difference with respect to the BCM is that the convergence parameter μ assumes an important role in tuning the number of final opinions peaks; hence, in our model the speed at which opinions converge/diverge allows to change the final opinion landscape. Lastly, we derive a mean field approximation of all the three new models.

Methods

Periodic Boundary Conditions. We consider N agents and a set of initial opinions x_i, $i \in \{1, \ldots, N\}$, uniformly distributed in [0, 1]. If we compare two agents' opinions by the absolute value distance $|x_i - x_j|$, those agents with near boundary opinions will have less concordant peers by definitions. We can overcome this problem by using the *Periodic Boundary Conditions* (PBC) and the alternative opinions' distance $|.|_\tau : [0, 1] \times [0, 1] \to [0, 0.5]$ defined as:

$$|x_i - x_j|_\tau = |x_i - x_j - \rho(x_i - x_j)|, \tag{13}$$

for $i, j \in \{1 \ldots, N\}$. The $\rho(.) : [-1, 1] \to \{-1, 0, 1\}$ adjustment ensures PBC and it is defined as:

$$\rho(x) = \begin{cases} -1, & \text{if } x \in [-1, -0.5) \\ 0, & \text{if } x \in [-0.5, 0.5]. \\ 1, & \text{if } x \in (0.5, 1] \end{cases} \tag{14}$$

The Bounded Confidence Model (BCM). The *Bounded Confidence Model* (BCM)[18,20] considers a set of N agents arranged on a complex network G. Each agent holds an opinion x_i, $i \in \{1, \ldots, N\}$, uniformly distributed in [0, 1]. Two agents interact if and only if they are connected in G and their present opinions are close enough, i.e. iff $j \in N_G(i)$ and $|x_i - x_j| < \varepsilon$, for $\varepsilon \in [0, 1]$. Note that, as we apply periodic boundary conditions in the simulations, two users will actually interact if: $|x_i - x_j|_\tau < \varepsilon$, for $\varepsilon \in [0, 0.5]$. If these conditions hold, the two agents change their opinions according to Eq. (1), otherwise they do not interact at all.

It is known from previous studies[33,34] that for ε big enough consensus is reached. The time rate change of $\mathbb{P}(x, t)\,dx$, the fraction of agents whose opinion at time t lies in the interval $[x, x + dx]$, is given by:

$$\begin{aligned} \frac{\partial \mathbb{P}(x, t)}{\partial t} &= -\mathbb{P}(x, t) \int_{-\varepsilon}^{\varepsilon} \mathbb{P}(x + y, t)\,dy \\ &\quad + \frac{1}{(1 - \mu)} \int_{-\varepsilon - 2x}^{\varepsilon - 2x} \mathbb{P}(x + y, t)\,\mathbb{P}\left(x - \frac{\mu}{1 - \mu} y, t\right) dy. \end{aligned} \tag{15}$$

The first two moments are given by $M_0 = \int \mathbb{P}(x, t)\,dx = 1$ and $M_1 = \int x \mathbb{P}(x, t)\,dx = 0$, i.e. the total mass and the mean opinion, are conserved[33]. Let $\mathbb{P}(x, 0) = 1$ be a flat initial condition, with $x \in [0, 1]$. We are interested in the final state of the system $\mathbb{P}(x, \infty)$.

When all agents interact, i.e., when $\varepsilon \geq 1$ the rate equation is integrable (as we assume PBC, $\varepsilon \geq 1/2$ is enough for our simulations). The second moment obeys $\dot{M}_2 + M_0 M_2/2 = M_1^2$, and using $M_1 = 0$ and $M_0 = 1$ we find that $M_2(t) = M_2(0) e^{-t/2}$, hence the second moment vanishes exponentially in time, all agents approach the center opinion and the system eventually reaches consensus[33]:

$$\mathbb{P}_\infty(x) = M_0 \delta(x). \tag{16}$$

When $\varepsilon \geq 1$ the final state is a single peak located in the middle and, as long as $\varepsilon \geq 1/2$, this situation persists (again, thanks to the PBC we get $\varepsilon \geq 1/4$ in the simulations). For smaller values of the threshold ε, it has been shown, by numerical simulations, that consensus is not reached and the opinion evolves into clusters that are

separated by a distance larger than ε. Once each cluster is isolated it evolves into a Dirac delta function as in the case $\varepsilon \geq 1$. The final distribution consists of a series of non interacting clusters at locations x_i with masses m_i:

$$\mathbb{P}_\infty(x) = \sum_{i=1}^{r} m_i \delta(x - x_i), \tag{17}$$

where r is the number of evolving opinion clusters[33]. All clusters must fulfill the conservation laws $\sum m_i = M_0 = 1$, and $\sum x_i m_i = M_1 = 0$ is equal to the conserved mean opinion. All different clusters $i \neq j$ must also fulfill $|x_i - x_j| > \varepsilon$.

References

1. Quattrociocchi, W., Scala, A. & Sunstein, C. R. Echo chambers on facebook. *Available at SSRN*, https://ssrn.com/abstract=2795110 (2016).
2. Bessi, A. *et al.* Science vs conspiracy: Collective narratives in the age of misinformation. *PloS one* **10,** e0118093 (2015).
3. Bessi, A. *et al.* Viral misinformation: The role of homophily and polarization. In *Proceedings of the 24th International Conference on World Wide Web Companion*, 355–356 (International World Wide Web Conferences Steering Committee, 2015).
4. Zollo, F. *et al.* Debunking in a world of tribes URL http://arxiv.org/abs/1510.04267 (2015).
5. Jøsang, A., Quattrociocchi, W. & Karabeg, D. Taste and trust. In *IFIP International Conference on Trust Management*, 312–322 (Springer, 2011).
6. Zollo, F. *et al.* Emotional dynamics in the age of misinformation. *PloS one* **10,** e0138740 (2015).
7. Sunstein, C. R. The law of group polarization. *Journal of political philosophy* **10,** 175–195 (2002).
8. Del Vicario, M. *et al.* The spreading of misinformation online. *Proceedings of the National Academy of Sciences* **113,** 554–559 (2016).
9. König, S. *et al.* On the effects of reputation in the internet of services. In *Proceedings of the 1st Int. Conference on Reputation* (*ICORE 2009*), 200–214 (2009).
10. Paolucci, M. *et al.* Social knowledge for e-governance: Theory and technology of reputation. *Roma*: *ISTC-CNR*, https://issuu.com/mario.paolucci/docs/erep_booklet (2009).
11. Castellano, C., Fortunato, S. & Loreto, V. Statistical physics of social dynamics. *Reviews of modern physics* **81,** 591 (2009).
12. Sznajd-Weron, K. & Sznajd, J. Opinion evolution in closed community. *International Journal of Modern Physics C* **11,** 1157–1165 (2000).
13. Holley, R. A. & Liggett, T. M. Ergodic theorems for weakly interacting infinite systems and the voter model. *The annals of probability*, 643–663 (1975).
14. Liggett, T. M. Stochastic models of interacting systems. *The Annals of Probability* **25,** 1–29 (1997).
15. Lambiotte, R. & Redner, S. Dynamics of non-conservative voters. *EPL* (*Europhysics Letters*) **82,** 18007 (2008).
16. Krapivsky, P. L. & Redner, S. Dynamics of majority rule in two-state interacting spin system. *Physical Review Letters* **90** (2003).
17. Galam, S. Sociophysics: a review of galam models. *International Journal of Modern Physics C* **19,** 409–440 (2008).
18. Deffuant, G., Neau, D., Amblard, F. & Weisbuch, G. Mixing beliefs among interacting agents. *Advances in Complex Systems* **3,** 87–98 (2000).
19. Hegselmann, R. *et al.* Opinion dynamics and bounded confidence models, analysis, and simulation. *Journal of Artificial Societies and Social Simulation* **5** (2002).
20. Lorenz, J. Continuous opinion dynamics under bounded confidence: A survey. *International Journal of Modern Physics C* **18,** 1819–1838 (2007).
21. Nickerson, R. S. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology* **2,** 175 (1998).
22. Centola, D. The spread of behavior in an online social network experiment. *Science* **329,** 1194–1197 (2010).
23. Centola, D. & Baronchelli, A. The spontaneous emergence of conventions: An experimental study of cultural evolution. *Proceedings of the National Academy of Sciences* **112,** 1989–1994 (2015).
24. Shao, J., Havlin, S. & Stanley, H. E. Dynamic opinion model and invasion percolation. *Physical review letters* **103,** 018701 (2009).
25. Li, Q., Braunstein, L. A., Havlin, S. & Stanley, H. E. Strategy of competition between two groups based on an inflexible contrarian opinion model. *Physical Review E* **84,** 066101 (2011).
26. Majdandzic, A. *et al.* Spontaneous recovery in dynamical networks. *Nature Physics* **10,** 34–38 (2014).
27. Ramos, M. *et al.* How does public opinion become extreme? *Scientific reports* **5** (2015).
28. Quattrociocchi, W., Caldarelli, G. & Scala, A. Opinion dynamics on interacting networks: media competition and social influence. *Scientific reports* **4** (2014).
29. Starnini, M., Frasca, M. & Baronchelli, A. Emergence of metapopulations and echo chambers in mobile agents. *Scientific reports* **6** (2016).
30. Erdös, P. & Rényi, A. On random graphs. *Publicationes Mathematicae Debrecen* **6,** 90–297 (1969).
31. Barabási, A. L. & Albert, R. Emergence of scaling in random networks. *Science* **286,** 509–512 (1999).
32. Watts, D. J. & Strogatz, S. H. Collective dynamics of 'small-world' networks. *Nature* **393,** 440–442 (1998).
33. Ben-Naim, E., Krapivsky, P. L. & Redner, S. Bifurcations and patterns in compromise processes. *Physica D*: *Nonlinear Phenomena* **183,** 190–204 (2003).
34. Ben-Naim, E. & Krapivsky, P. L. Multiscaling in inelastic collisions. *Physical Review E* **61** (2000).

Acknowledgements

Funding for this work was provided by EU FET project MULTIPLEX nr. 317532, SIMPOL nr. 610704, DOLFINS nr. 640772, SOBIGDATA nr. 654024. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author Contributions

M.D.V., A.S., and W.Q. outlined the research question. M.D.V. performed the simulations. M.D.V. and A.S. interpreted the results. M.D.V., A.S., G.C., H.E.S., and W.Q. contributed equally to the writing and reviewing of the manuscript.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Del Vicario, M. *et al.* Modeling confirmation bias and polarization. *Sci. Rep.* **7**, 40391; doi: 10.1038/srep40391 (2017).