Information Sciences 609 (2022) 1256–1270

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

# Analyzing the extremization of opinions in a general framework of bounded confidence and repulsion

Jesús Giráldez-Cru [a,c,*], Carmen Zarco [b,c], Oscar Cordón [a,c]

[a] Department of Computer Science and Artificial Intelligence (DECSAI), University of Granada (UGR), Spain
[b] Department of Marketing and Market Research, University of Granada (UGR), Spain
[c] Andalusian Research Institute in Data Science and Computational Intelligence (DaSCI), University of Granada (UGR), Spain

ARTICLE INFO

*Article history:*
Received 14 March 2022
Received in revised form 26 July 2022
Accepted 27 July 2022
Available online 29 July 2022

*Keywords:*
Opinion dynamics
Bounded confidence
Extremization
Repulsion
Social network
Agent-based time-varying rationality

ABSTRACT

In the bounded confidence framework, agents' opinions evolve as a result of interactions with other agents having similar opinions. Thus, consensus or fragmentation of opinions can be reached, but not *extremization* (the evolution of opinions towards an extreme value). In contrast, when repulsion mechanisms are at work, agents with distant opinions interact and repel each other, leading to extremization. This work proposes a general opinion dynamics framework of bounded confidence and repulsion, which includes social network interactions and agent-independent time-varying rationality. We extensively analyze the performance of our model to show that the degree of extremization among a population can be controlled by the repulsion rule, and social networks promote extreme opinions. Agent-based rationality and time-varying adaptation also bear a strong impact on opinion dynamics. The high accuracy of our model is determined in a real-world social network well referenced in the literature, the Zachary Karate Club (with a known ground truth). Finally, we use our model to analyze the extremization of opinions in a real-world scenario, in Spain: a marketing action for the Netflix series "Narcos".

## 1. Introduction

Opinion dynamics (OD) is the field of studies about how opinions spread in a population and how they evolve over time [45,16]. One key aspect of OD is the rule of opinion propagation and fusion, i.e., how the agents of a system interact to spread their opinions, and how opinions are altered as the result of those interactions. Some well-known examples of OD models, which differ in these rules, are the DeGroot model [13], and the bounded confidence (BC) models [11,25]. OD have been successfully applied in many contexts, such as the analysis of *status quo* opinions [3], the identification of leaders in social networks [9], the degree of separation in OD communications [49], and the study of OD in time-dependent communication topologies [42], among others.

In the BC framework, agents' opinions evolve as a result of interactions with other agents sharing similar opinions. One BC framework is the Deffuant-Weisbuch (DW) model [11], according to which agents participate in random pairwise encounters and update their opinions if they are similar, i.e., if the difference between their opinions is smaller than a given confidence threshold. This confidence threshold defines the confidence area of each agent according to its current opinion. When the opinion fusion rule is triggered (that is, the two agents trust on each other), both opinions are updated towards

[*] Corresponding author at: Department of Computer Science and Artificial Intelligence (DECSAI), University of Granada (UGR), Spain.
*E-mail addresses:* jgiraldez@ugr.es (J. Giráldez-Cru), carmen.zarco@ugr.es (C. Zarco), ocordon@decsai.ugr.es (O. Cordón).

https://doi.org/10.1016/j.ins.2022.07.164
0020-0255/© 2022 The Author(s). Published by Elsevier Inc.

an average value. It is well established that either consensus or fragmentation of opinions can be reached in this model, consensus meaning a shared opinion, whereas fragmentation would be the co-existence of distinct clusters of opinions. These results depend on the value of the confidence threshold used in the model: smaller values lead to fragmentation, higher values tend to produce consensus. Still, this model is unable to produce *extremization*, i.e., the phenomenon where some opinions evolve towards an extreme value. Extremization [29] is a well-studied phenomenon in social sciences. There are real-life examples where large populations or large groups of individuals within them modify their opinions towards an extreme value [30].

The lack of extremization in the DW model is one of its main drawbacks in the context of OD systems applied to real-life. To solve it, several repulsion-based extensions of this model have been proposed [12,2,31,28,10]. They are based on the *Social Judgment Theory*, that adopts the notion that any given individual has two attitude thresholds. If the difference in opinions of two agents is smaller than an assimilation threshold, they will shift their opinions towards an average value. Contrariwise, if the difference in opinions is larger than a repulsion threshold, they will shift their opinions away from each other. Unfortunately, all of these models consider the attitude thresholds as global variables, i.e., they are the same for every agent in the population. The motivation behind the present work is to fill this gap by considering the influence of the individual rationality of each agent in the process.

In this work, we propose a general OD framework based on bounded confidence and repulsion. Because our model is a generalization of the DW model, both consensus and fragmentation of opinions can be achieved, but the repulsion mechanism is moreover able to produce the extremization of opinions. Furthermore, we define agent-based rationality in our model, i.e., the confidence and repulsion thresholds are independent for each agent, and they may evolve over time. This allows us to model the evolving behaviors of individuals in a population, and shed light on their affinity towards and against polarization at different points in time.

Through extensive empirical analysis we find that, in effect, an extremization phenomenon is achieved for some agents, yet the model is able to preserve the conditions of consensus or fragmentation of opinions for others. By focusing on the degree of extremization within the system, triggered in the area of repulsion of the agents, we show that the percentage of agents having an extreme opinion depends on the repulsion threshold. We also look into the topology of social networks where agent interaction bear a major impact on the OD. Indeed, the existence of a social network is seen to hinder consensus reaching, since it promotes the emergence of extreme opinions. Social networks with a clear community structure are more prone to preserve clusters of moderate opinions.

Our empirical analysis furthermore shows that agent-based rationality affects the OD. This led us to model both polarized and moderate agents (respectively having high and low repulsion thresholds) in the same population. As rationality "evolves" along time, some scenarios of extremization of opinions arise after a consensus reaching, or *vice versa*. Therefore, our model allows us to represent scenarios that do not tend to a stationary point. Apart from studying synthetic scenarios, we analyze the performance of our model with regard to a real-world social network: the Zachary Karate Club (ZKC) [48]. This network represents the interactions between 34 members of a karate club, split into two factions after a conflict between the administrator and an instructor. Our model served to predict the opinions of these 34 members *a posteriori* with an accuracy of $85.29\%$.

Finally, our model was used to analyze the extremization of opinions in a real-world setting: a viral marketing campaign that the streaming platform Netflix conducted in December of 2016 in Spain to promote the second season of its series "Narcos". A large billboard in downtown Madrid used a very provocative play on words about drugs (the topic of the series) and Christmas (the time the campaign was launched), which generated a clear polarization of opinions in favor and against the campaign. Our model allowed us to reproduce and explore both the early and the late time progressions of the polarization dynamics resulting from this marketing action, and show its versatility to model and explain complex real-world scenarios with polarization and extremization of opinions.

Therefore, the contributions of this work are as follow:

- We present a general OD framework based on bounded confidence and repulsion with agent-based time-varying rationality.
- We extensively analyze the effects of the confidence and repulsion thresholds in the extremization of opinions, including the topology of the social network, and look at how they affect the promotion of extreme opinions.
- We study how several agent-based time-varying rationalities affect the OD. According to our results, the model allows one to represent complex scenarios that do not tend to a stationary state.
- We investigate the performance of our model in a real-world social network, the ZKC [48] (for which the ground truth is known *a priori*), and show that it is able to predict the opinions of a population with high accuracy.
- We use our model to study the extremization of opinions in a real-world setting: the viral marketing campaign for the Netflix series "Narcos" in Spain.

The rest of the work is organized as follows. In Section 2 we summarize some related works. Section 3 contains an overview of the classical DW model and some artificial social networks commonly used in the literature. In Section 4 we present the formal description of our proposed model of bounded confidence and repulsion, while Section 5 is devoted to the empirical analysis of the model. In Section 6, we analyze the polarization and extremization of opinions in a real-world scenario using our model. Finally, we briefly offer some conclusions in Section 7.

## 2. Related works

The DeGroot model [13] is a well-known continuous OD model. Although other extensions of it have been proposed in the literature, none considers opinion similarity under the opinion fusion rule.

Against this drawback, BC models trigger the opinion fusion rule only when agents' opinions are similar. The main BC models in the literature are those of Hegselmann & Krause (HK) [25] plus the DW model [11]. Both rely on the idea of repeatedly averaging the opinions expressed within a bounded confidence area of the agents; they differ in the agents' interactions. With the HK model, each agent considers all the opinions close to his/her own, that is, in their area of confidence. According to the DW model, agents have random pairwise encounters that alter their own opinions whenever both agents are in their confidence area. Therefore, the HK model is more suitable for modeling interactions in large groups, such as formal meetings, whereas the DW is better suited for pairwise interactions within a large population [8]. A theoretical analysis of their similarities and differences can be found in [36], where the two systems are formulated as Markov Chains. The DW model has also been studied considering interactions within a social network [21] or in the presence of external periodic perturbations [7] in order to simulate further intervining factors of opinion influence such as mass communication.

Similarly, several models may explain the prevalence of extremism in a population. Relative Agreement (RA) [12] extends the classical DW model by introducing an uncertainty factor for each agent. Extremism can emerge by assigning a lower uncertainty (high stubbornness) to a subset of agents at the extremes of the opinion distribution. [2] analyze the RA model under a social network topology. Alternatively, in [10] the extremization is viewed as a consequence of biased assimilation (or support) when there is uncertain or inconclusive evidence regarding a complex issue. The Jager-Amblard (JA) model [31] extends DW through a repulsion mechanism, and [28] extends this repulsion to two dimensions. They are based on *Social Judgment Theory*, hence OD are explained by means of two attitude thresholds based on assimilation and repulsion. A generalization of any OD model (including DW and JA) in terms of potential functions is proposed in [39]. Other models that explain the emergence of extreme opinions include the following works: Mathias et al. [37] look into the effects of uncertainty in moderate agents. The Social Judgment Based Opinion (SJBO) of Fan and Pedrycz [20] distinguishes between inner continuous opinions and observable discrete choices, and these authors analyze the emergence of extremist opinions under this model [19].

The representation of agents' opinions is another fundamental question in OD systems. Although the DW model provides a continuous interval of opinions being more expressive than binary values as in [27], it misses the qualitative nature of opinions. To resolve this, some propose the use of fuzzy linguistic variables to represent agents' opinions in a substantially more realistic manner. For instance, [15] propose representing them as 2-tuple fuzzy linguistic variables. In [17], OD is placed in the context of social networks and numerical intervals serve to capture the uncertainty of opinions. Another example can be found in [34], where personalized individual semantics for each agent are proposed.

## 3. Background

In this section we first review the main social networks topologies commonly used in the literature, and then provide a detailed description of the DW model [11] and its adaptation to interactions within a social network following some ideas from [44].

### 3.1. Artificial social networks

A social network can be represented using a graph $G(V,A)$, where $V$ is a set of nodes, and $A$ is its adjacency matrix of size $|V| \times |V|$. For any node $u \in V$, let $N(u)$ be the neighborhood of this node, i.e., $N(u) = \{v \in V \mid A_{uv} = 1\}$. In this graph, each node represents an agent of our model and each edge represents a connection between two agents. These connections are social interactions and must be interpreted in a broad sense, including conversations of any typology (e.g., face-to-face conversations, online communications, etc).

Below we will review some graph topologies commonly applied to model social networks. For the sake of completeness, we also provide the definition of a complete graph to be used alongside this work to study the OD in fully-mixed scenarios.

#### 3.1.1. Complete graphs

A complete graph $K_n$ is a graph $G(V,A)$ with $n = |V|$ nodes and $A_{ij} = 1$ for every node $i,j \in V$. Therefore, this is a fully-connected graph that will allow us to study the OD models in well-mixed systems where every agent can interact with any other agent.

#### 3.1.2. Erdős-Rényi graphs

The Erdős-Rényi (ER) model [18] is the classical example of random graphs. In ER graphs, the probability of any two nodes being connected is uniform. Hence the degree of the nodes exhibits a small variability, following a binomial distribution. To generate them, an iterative process is performed, connecting two randomly chosen nodes at each iteration. For the sake of clarity in this work we will refer these graphs as $ER(n,k)$, where $k = 2m/n$ is the expected node degree, and $m$ is the number of edges.

### 3.1.3. Small-world graphs

The small-world (SW) graph model [43] was devised to preserve the high clustering of regular lattices while reducing the characteristic path lengths of ER graphs. Starting from a regular ring lattice with $n$ nodes and $k$ edges per node, it randomly assigns each edge a new, randomly chosen end-point node with probability $p$. Therefore, $p = 0$ does not alter the regular lattice while $p = 1$ generates ER graphs. In practice a low value of $p$ (e.g., $p = 0.1$) is seen to produce graphs with high local density (i.e., high clustering) and short distances (as in ER graphs). We refer to these graphs as $SW(n, k, p)$.

### 3.1.4. Preferential attachment graphs

Preferential attachment (PA) [6] aims to explain the growth of complex networks, characterized by the existence of highly connected nodes (the so-called hubs). This model starts with a complete graph of size $m_0$, after which it iteratively connects a new node to $m \leqslant m_0$ nodes in the existing graph. Repeating this process, the node degree in the resulting graph follows a power-law distribution $P(i) \sim i^{-a}$. Consequently, most of the nodes have a very low degree whereas just a few receive most of the links. Since this distribution is scale-free (i.e., it does not depend on $n$), this model is usually known as the scale-free model. We will refer to it as $SF(n, m_0, m)$.

## 3.2. The Deffuant-Weisbuch model of bounded confidence

Let us consider a set of $n$ agents $S = \{a_1, a_2, \ldots, a_n\}$. At each discrete time step $t = \{0, 1, 2, \ldots\}$, each agent $a_i \in S$ has an opinion $x_i(t) \in [0, 1]$. Let $X(t)$ be the opinion profile at time step $t$, i.e., the set of opinions $\{x_1(t), \ldots, x_n(t)\}$ for the whole agent population.

At each time step, the DW applies the opinion fusion rule to two distinct agents $a_i$ and $a_j$ (with $i \neq j$) randomly chosen from $S$. By this rule, both agents update their opinions if the difference between $x_i(t)$ and $x_j(t)$ is smaller than a given bounded confidence threshold $\varepsilon$, i.e., when $|x_i(t) - x_j(t)| < \varepsilon$. Otherwise, their opinions are not altered. The precise definition of this model is as follows:

**Definition 1** (*DW model [11]*). Given a bounded confidence threshold $\varepsilon \in [0, 1]$, a set of agents $S = \{a_1, \ldots, a_n\}$, and an initial opinion profile $X(0)$ for these agents, the DW model selects at each time step $t \in \{1, 2, \ldots, T\}$ two distinct agents $a_i$ and $a_j$ (with $i \neq j$), uniformly and independently chosen from $S$, and applies to them the following opinion fusion rule when $|x_i(t) - x_j(t)| < \varepsilon$:

$$\begin{cases} x_i(t+1) = x_i(t) + \mu(x_j(t) - x_i(t)) \\ x_j(t+1) = x_j(t) + \mu(x_i(t) - x_j(t)) \end{cases} \quad t = 1, 2, \ldots \tag{1}$$

where $\mu \in [0, 0.5]$ is the convergence parameter. Otherwise, both opinions remain unaltered.

The convergence speed $\mu$ controls the agent's movement towards a common opinion. In particular, if $\mu = 0$, both agents do not change their opinions; whereas when $\mu = 1/2$ the two agents will move to their average opinion in the next time step. This phenomenon can be observed in Fig. 1, where we depict three executions of the DW model with distinct values of $\mu$. In all three cases, we use a population of 1000 agents, a bounded confidence $\varepsilon = 0.5$, and an initial opinion profile $X(0)$ where every opinion is randomly and independently initialized in the interval $[0, 1]$, while the model is executed for $T = 5 \cdot 10^4$ time steps. In this figure, each point $(x, y)$ means that an agent updated its opinion at time step $x$ with value $y$,[1] except for the points $(0, y)$ and $(T, y)$ which represent the full initial and final opinion profiles $X(0)$ and $X(T)$, respectively. It can be seen that the greater the value of the convergence speed $\mu$, the faster the stationary state of the execution is achieved. Since the three executions use a high value of $\varepsilon$, in the stationary state all the agents reach a consensus with a neutral opinion.

The value of the bounded confidence $\varepsilon$ used in the DW model determines whether a consensus or a fragmentation of opinions can be reached [36]. In Fig. 2 we depict this phenomenon, representing the execution of the DW model with $\varepsilon = \{0.1, 0.2, 0.3, 0.4\}$. In order to measure the distribution of the final opinions among the population, we also present them as histograms, dividing the continuous interval $[0, 1]$ of possible opinions into 20 different bins. For these experiments we use a convergence speed $\mu = 0.1$, and the model is executed for $T = 10^5$ time steps. Also, the final opinion profile $X(T)$ is depicted in red points to emphasize the full set of opinions at the end of the execution.

In the wake of these executions, high values for the confidence threshold (e.g., $\varepsilon = 0.4$) result in a consensus of the whole population, all agents sharing a neutral opinion. Yet as the value of $\varepsilon$ decreases, the number of clusters (i.e., sets of agents with very close opinions) and their volume (i.e., number of agents in those clusters) increase. For instance, when $\varepsilon = 0.3$ there appears a second cluster of very small size with a fairly divergent opinion with respect to the average consensual one, whereas with $\varepsilon = 0.2$ there appear two main clusters of similar size showing two quite different opinions. Finally, for low values of the confidence threshold (e.g., $\varepsilon = 0.1$), there is a clear fragmentation of opinions into many clusters. Let it be said, regarding the number of clusters and their volume, the consensus is very clear for executions with $\varepsilon = 0.4$ and $\varepsilon = 0.3$. However, with $\varepsilon = 0.2$ and $\varepsilon = 0.1$, there exist two or more clusters of opinions of similar size. Note also that if $\varepsilon$ is high enough, the number of opinions contrary to the consensus is negligible (e.g., $\varepsilon = 0.3$).

[1] We recall that at each time step, the DW model only updates the opinions of two agents at most.

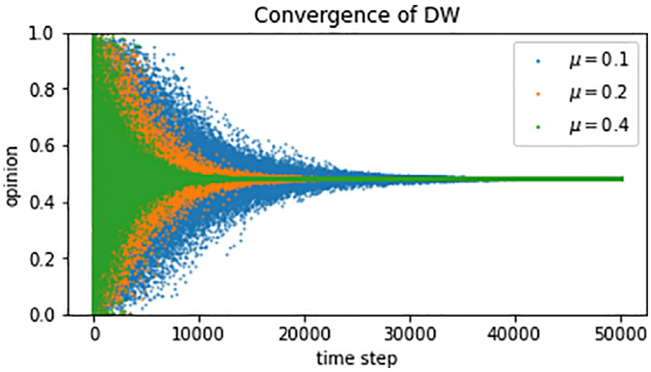

**Fig. 1.** Executions of the DW model on an agent population of size $n = 1000$ with distinct convergence speeds $\mu = \{0.1, 0.2, 0.4\}$ (with a bounded confidence threshold $\epsilon = 0.5$).

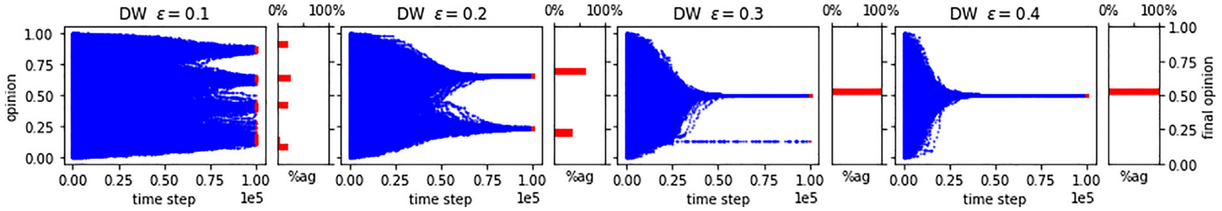

**Fig. 2.** Executions of the DW model and histograms of the distribution of final opinions with distinct bounded confidence thresholds $\varepsilon = \{0.1, 0.2, 0.3, 0.4\}$ (with a convergence speed $\mu = 0.1$).

Therefore, the emergence of consensus or fragmentation of opinions results from the interactions in the confidence area of the participating agents, which in turn depends on the bounded confidence threshold $\varepsilon$. The rationale behind this bounded confidence rule is that agents are only influenced by opinions similar to their own, and this similarity is indeed given by $\varepsilon$. When two agents have opinions that differ to a degree equal to or greater than $\varepsilon$, they do not influence each other, hence their opinions remain unaltered after an encounter. In sum, the bounded confidence threshold $\varepsilon$ determines the dynamics of the agents' opinions.

The classical definition of the DW model (see Definition 1) does not consider agents' interactions within a social network. This would be equivalent to running the model on a complete graph $K_n$, where $n$ is the number of agents in the population. In order to generalize this model to capture the interactions within any social network, we extend it as follows:

**Definition 2** (*DW model with social network (DWSN)* [21]). Given $\varepsilon, S$, and $X(0)$ as in Definition 1, and a graph $G(V, A)$ representing the social network of $S$ such that $|V| = |S|$, the DWSN randomly selects at each time step $t \in \{1, 2, \ldots, T\}$ two agents $a_i$ and $a_j$ such that $i \neq j$ and $A_{ij} = 1$ (i.e., there is an edge between node $i$ and node $j$), and applies them the opinion fusion rule defined in Eq. 1 when $|x_i(t) - x_j(t)| < \varepsilon$. From now on, we refer to this model as $DW(G, X(0), \varepsilon, \mu)$.

Notice that the previous models consider the same confidence threshold $\varepsilon$ for all the agents, and this threshold is, moreover, time-invariant.

## 4. Bounded confidence and repulsion

In this section, we present a new OD model called Agent-independent Time-based Bounded Confidence and Repulsion (ATBCR). This model extends the classical DW model with a repulsion rule. In particular, when two agents with very distant opinions meet in a random pairwise encounter, this repulsion rule results in a reinforcement of each agent's current opinion, which are updated towards a more distant value. Moreover, it incorporates a higher level of rationality in the agents since both the confidence and the repulsion areas are independently defined for each agent and for each timestep, i.e., distinct agents can have distinct thresholds, and even the same agent can modify its thresholds during the execution of the model. This generalization therefore provides a more refined mechanism of adaptation in the agents' rationality. Note that since it is an extension of the DW model, it does not alter the behavior of the agents in the confidence area, i.e., meetings between agents with similar opinions updates them towards a common opinion. Our model is inspired by [12,21,31]. A precise definition of the ATBCR model is provided in Definition 3.

**Definition 3** (*Agent-independent Time-based Bounded Confidence and Repulsion (ATBCR)*). Let $S = \{a_1, \ldots, a_n\}$ be a set of $n$ agents, $t \in \{1, 2, \ldots, T\}$ be a set of time steps, and $X(0)$ be their initial opinion profile. Let $\varepsilon_i(t), \vartheta_i(t) \in [0, 1]$, with $\varepsilon_i(t) < \vartheta_i(t)$ for every agent $a_i$ and time step $t$, be a confidence and a repulsion threshold for agent $a_i$ at time step $t$, respectively. Let $G(V, A)$ be a graph representing the social network of $S$, i.e., it satisfies that $|V| = |S|$. At each time step $t \in \{1, 2, \ldots, T\}$, the ATBCR model randomly selects two agents $a_i$ and $a_j$ uniformly and independently such that $i \neq j$ and $A_{ij} = 1$ (i.e., both agents are neighbors in the social network), and applies them the following opinion fusion rule:

$$|x_i(t) - x_j(t)| < \varepsilon_i(t) \Rightarrow x_i(t+1) = x_i(t) + \mu(x_j(t) - x_i(t)) \tag{2}$$

$$\varepsilon_i(t) \leqslant |x_i(t) - x_j(t)| \leqslant \vartheta_i(t) \Rightarrow x_i(t+1) = x_i(t) \tag{3}$$

$$|x_i(t) - x_j(t)| > \vartheta_i(t) \Rightarrow x_i(t+1) = min(1, max(0, rep(i, j))) \tag{4}$$

and similarly for agent $a_j$, where $rep(a, b) = x_a(t) - \mu(x_b(t) - x_a(t))$ is the repulsion function and $\mu$ is the convergence speed.

We refer to this model as $ATBCR(G, X(0), \varepsilon_i(t), \vartheta_i(t), \mu)$ in the rest of this work.

For the sake of simplicity, when the confidence and repulsion thresholds are time-invariant (i.e., they take the same value during the whole execution), we just call them $\varepsilon_i$ and $\vartheta_i$, for every agent $a_i$. Moreover, for the specific case when these thresholds have the same value for every agent, they are referred to as $\varepsilon$ and $\vartheta$.

It can be seen that ATBCR performs the same confidence rule as the DW model when agents have similar opinions, as per Rule (2). On the contrary, Rule (4) provides a polarization regime where agents reinforce their opinions by an extremization process. Notice that this rule enforces opinions to remain in the interval $[0, 1]$. Finally, outside both regimes (confidence and repulsion areas), the agents do not alter their current opinions (see Rule (3)).

The rationale behind the ATBCR model is the following. When the difference between two opinions is in the interval $[0, \varepsilon]$, both agents apply the confidence rule given by Rule (2), where they mutually influence each other. But when this difference is the interval $[\vartheta, 1]$, both agents apply the repulsion rule (given by Rule (4)), whereby agents reinforce their own opinions updating them towards a more distant value. This is because the two agents participating in this (polarized) encounter have very distant opinions, repelling each other. Finally, the interval $[\varepsilon, \vartheta]$ describes the regime where agents' opinions are not affected, i.e., their opinions are not similar enough to influence each other, but they are not so distant to repel themselves.

The proposed ATBCR model is a generalization of the DW model. In particular, the DW model is equivalent to the ATBCR model with a bounded repulsion threshold $\vartheta = 1$ and agent-uniform time-invariant confidence threshold $\varepsilon$ (i.e., it is the same for every agent $a_i$ at every time step $t$), as stated in Lemma 1.

**Lemma 1.** *The ATBCR model is a generalization of the DW model, where* $DW(G, X(0), \varepsilon, \mu) = ATBCR(G, X(0), \varepsilon, 1, \mu)$.

**Proof.** When the ATBCR model applies Rule (2) or Rule (3), it behaves exactly equal to the DW model for the same value of $\varepsilon$. This is, for any $\varepsilon$, Rules (2) and (3) of the ATBCR model reduce to Rule (1) of the DW model. Also, if $\vartheta = 1$, Rule (4) is never applied. In particular, agents' opinions are defined in the interval $[0, 1]$, thus the difference between any two opinions cannot be greater that $\vartheta$, i.e., $\vartheta = 1 \iff |x_i(t) - x_j(t)| \not> \vartheta$for any $i, j \in \{1, \ldots, n\}$. Therefore, the DW model is reduced to the ATBCR model with $\vartheta = 1$, and hence Lemma 1 holds. □

Therefore, our proposed ATBCR model is an extension of the classical DW model of BC [12]. In particular, the confidence regime of the DW model is preserved in the ATBCR model, whereas a new rule of repulsion affects those agents with distant opinions that meet in a random encounter. We emphasize that these encounters between agents with distant opinions would not alter their opinions in the DW model. It also considers the existence of a social network to model agents' interactions, as in [21]. Finally, it incorporates a repulsion mechanism inspired by [31]. However, none of these models incorporates agent-based time-varying rationality in the agents' behaviors.

## 5. Empirical analysis of the ATBCR model

In this section, we provide an empirical analysis of the ATBCR model. First, we analyze the impact of the repulsion threshold $\vartheta$ on the extremization of agents' opinions. Second, we present a sensitivity analysis to study the degree of extremization achieved in the population under different executions of the model. Third, we analyze the impact of the social network topology on the model. Fourth, we study the influence of incorporating agent-independent and time-based rationality. Finally, we analyze the accuracy of our model in a well-known real-world social network: the ZKC [48]; for which the ground truth is known *a priori*.

### 5.1. Impact of the repulsion threshold

In this section, we analyze the general performance of the ATBCR model. In particular, we study the impact of the repulsion threshold $\vartheta$ on the extremization of agents' opinions. Note that in this first experiment, we consider the agents' thresholds as global variables of the model, shared by all the agents and time-invariant.

Fig. 3 presents some executions of the ATBCR model with distinct values of the confidence and repulsion thresholds, including the histograms of the final opinion profile as before. In particular, we represent a cross-combination of the values $\varepsilon = \{0.2, 0.3, 0.4\}$ and $\vartheta = \{0.7, 0.8, 0.9\}$. In all these executions, the initial opinions $X(0)$ are randomly and independently initialized in the interval $[0, 1]$, and the convergence speed of the ATBCR model is $\mu = 0.1$. All the experiments are executed during $T = 10^5$ time steps, with a fully-mixed population of $|S| = 1000$ agents (i.e., using a complete graph $K_{|S|}$).

In this experiment we observe several remarkable phenomena. When both the confidence and repulsion thresholds $\varepsilon$ and $\vartheta$ are large enough (e.g., $\varepsilon = 0.4, \vartheta = 0.9$, in the bottom right corner of Fig. 3), there is an expected consensus reaching into a neutral opinion. This is because the confidence rule dominates the system and the repulsion rule is hardly triggered. As the confidence threshold $\varepsilon$ decreases (see e.g., $\varepsilon = 0.3$ and $\varepsilon = 0.2$ with $\vartheta = 0.9$, in the right column of Fig. 3), some clusters of opinions, different to the neutral opinion, emerge, including extreme opinions. Their volume increases as $\varepsilon$ decreases, but the cluster with opinions closer to the neutral one still contains the majority of agents due to the large value of $\vartheta$. These (close to) neutral opinions may be slightly biased due to repulsion against the existing extreme opinions (see, e.g., $\varepsilon = 0.2, \vartheta = 0.9$). On the other hand, when the repulsion threshold $\vartheta$ decreases, extreme opinions are promoted, and these extreme opinions become the majority when $\vartheta$ is low enough. See the cases with $\vartheta = 0.8$ and $\vartheta = 0.7$, in the bottom row of Fig. 3. For instance, in the case with ($\varepsilon = 0.2$ and $\vartheta = 0.7$, the repulsion rule dominates the dynamics of the system. Interestingly, the combination of these two phenomena can result into a fragmentation of opinions with some extreme ones. For instance, there are noteworthy differences between the executions in the main diagonal of Fig. 3, i.e., the executions with (a) $\varepsilon = 0.2$ and $\vartheta = 0.7$, (b) $\varepsilon = 0.3$ and $\vartheta = 0.8$, and (c) $\varepsilon = 0.4$ and $\vartheta = 0.9$. In (a), there are three clusters of opinions of similar size. One is a cluster of neutral opinions, while the other two are of extreme opinions. In (b), there is a main cluster of neutral opinions with two small clusters of extreme opinions. Finally, in (c) there is only a cluster of neutral opinions, i.e., a consensus is reached. Thus, the existence of extreme opinions and their volume in the population can be controlled by the confidence and repulsion thresholds $\varepsilon$ and $\vartheta$.

### 5.2. Sensitivity analysis on the extremization of the population

In this section we look into the emergence of extreme opinions and their spread in the population of agents to a deeper extent, in order to measure the volume of extreme opinions that emerges in our model. To this end, we carry out a sensitivity analysis of the two main parameters of the ATBCR model, i.e., the confidence and the repulsion thresholds $\varepsilon$ and $\vartheta$, respectively. Again in this experiment, we only consider the agents' thresholds as global variables of the model shared by all agents and time-invariant.

In this experiment, we execute the ATBCR model with all the combinations of values of $\varepsilon$ and $\vartheta$ in the interval $[0, 1]$, with a step of 0.05. We recall that the ATBCR model enforces $\varepsilon < \vartheta$, hence some combinations of these two thresholds are meaningless and should be forbidden. For each combination of these values, we run 100 independent executions of the ATBCR model, and measure the percentage of agents having an extreme final opinion in $X(T)$. In particular, we consider as extreme opinions those in the interval $[0, 0.1] \cup [0.9, 1]$. Besides $\varepsilon$ and $\vartheta$, each execution differs only in the initialization of the agents' opinions (i.e., $X(0)$), which are always randomly and independently generated in the interval $[0, 1]$. In all the executions, the convergence speed of the model is $\mu = 0.1$ and the system is executed during $T = 10^5$ time steps for a population of $|S| = 1000$ agents using a complete graph $K_{|S|}$ as the social network.

In Fig. 4 we depict the results of this sensitivity analysis. Three distinct regimes can be clearly identified.

First, a consensus is reached when the sum of $\varepsilon$ and $\vartheta$ is large enough (see the upper right quarter of the plot). This phenomenon occurs at approximately $\varepsilon + \vartheta \geqslant 1.1$. In all these executions, the opinion consensus occurs at a neutral opinion (not shown in the plots). This is because the confidence rule (see Rule (2)) has a greater impact on the system, and the repulsion rule is hardly triggered, as mentioned previously.

Second, when the repulsion threshold $\vartheta$ is low (see the left area of the plot), all agents achieve an extreme opinion. In particular, this occurs when $\vartheta \leqslant 0.5$. This is because the repulsion rule (see Rule (4)) dominates the dynamics of the system. In particular, even if the confidence rule is triggered between some agents, it is more likely for a random encounter to result in polarization. As a consequence, all agents' opinions become extremized at some point, and at this stage the agents no longer update their opinions towards a neutral opinion any longer.

Finally, a fragmentation of opinions, including a percentage of extreme opinions, is the result in the remaining executions, i.e., when $\vartheta > 0.5$ and $\varepsilon + \vartheta < 1.1$. See the bottom right area of Fig. 4. This is possibly the most interesting area to analyze for the ATBCR model. As can be observed, the percentage of agents with an extreme opinion can be measured as a function of the sum of both thresholds $\varepsilon$ and $\vartheta$. In particular, the larger the sum $(\varepsilon + \vartheta)$ is, the lower the percentage of agents with an extreme opinion. This function decreases monotonically, showing that the degree of extremization in the ATBCR model can be easily controlled by its two bounded thresholds.

Note that the region where $\vartheta \leqslant \varepsilon$ (upper left quarter of Fig. 4) is not represented in this plot since the ATBCR model does not allow these cases, i.e., the repulsion threshold $\vartheta$ must be strictly greater than the confidence threshold $\varepsilon$.

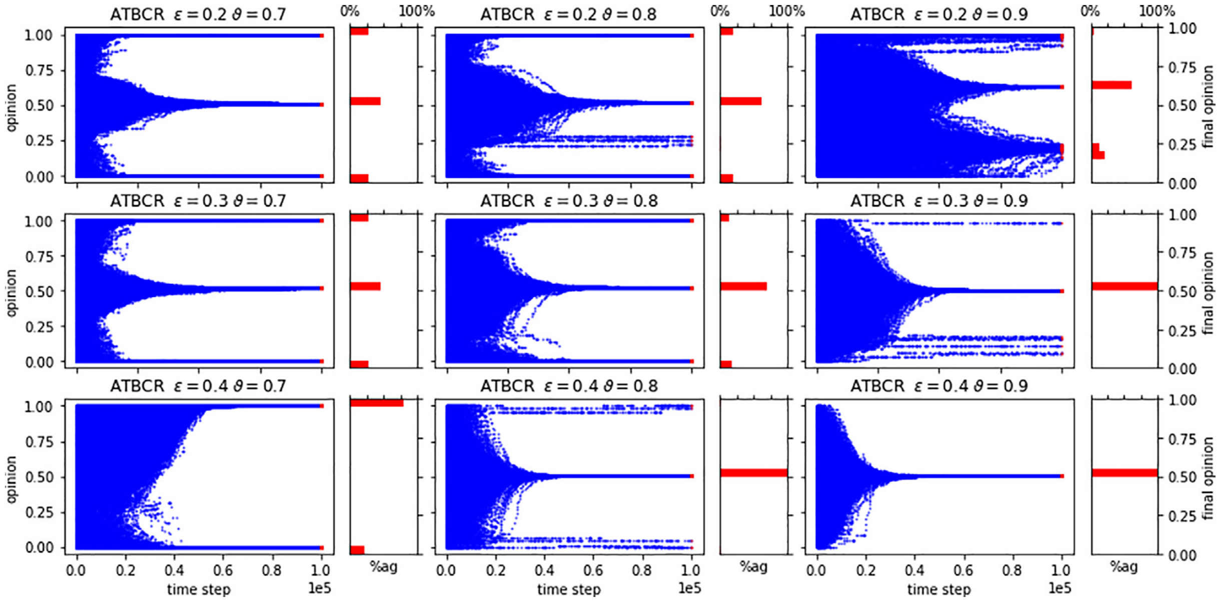

**Fig. 3.** Executions of the ATBCR model and histograms of the distribution of final opinions with distinct values of the confidence threshold $\varepsilon = \{0.2, 0.3, 0.4\}$ (top, center, and bottom, respectively) and repulsion threshold $\vartheta = \{0.7, 0.8, 0.9\}$ (left, center, and right, respectively), with a convergence speed $\mu = 0.1$.

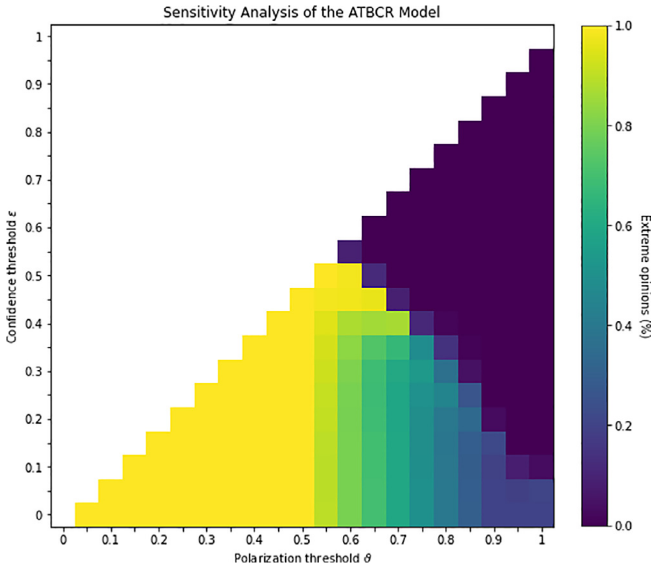

**Fig. 4.** Sensitivity analysis of the ATBCR model.

### 5.3. Impact of the social network topology

In this section we analyze how the topology of the social network affects the dynamics of the ATBCR model. We use three topologies of social networks commonly used in the literature, and study the OD of our model in them, i.e., restricting the interactions of the agents to their neighborhood in those social networks. First, an ER graph is generated with average degree $k = 6$, i.e., meaning on average every agent is connected to six other agents. In this graph the average path length is 4.04 and the average clustering coefficient is $5.86 \cdot 10^{-3}$. Second, a SW graph is produced with average node degree $k = 6$ and probability of rewiring $p = 0.1$. This parametrization gives a graph with average path length 5.92 and average clustering coefficient $5.08 \cdot 10^{-1}$. Finally, an instantiation of the SF model with $m = m_0 = 3$ yields a graph with average node degree $k = 5.98$, average path length 3.45, and average clustering coefficient $3.35 \cdot 10^{-2}$. Although the three graphs have a similar average degree, the graph with the smallest average path length is SF followed by ER, and finally SW, whereas the highest clustering coefficient is found in SW, followed by SF, and finally ER. In the three graphs, the number of nodes is set to $n = |S| = 1000$, i.e., the number of agents in the population.

In Fig. 5 we depict the results of our analysis. In particular, we execute the ATBCR model on these three social network topologies with confidence and repulsion thresholds $(\varepsilon, \vartheta) = (0.2, 0.7)$ (top), $(0.3, 0.8)$ (center), and $(0.4, 0.9)$ (bottom). As in the two previous experiments, these thresholds are time-invariant global variables, the convergence speed is set to $\mu = 0.1$, and the initial opinion profile $X(0)$ is uniformly and independently initialized in the full interval $[0, 1]$ of possible opinions. All the experiments are executed for $T = 10^6$ time steps, i.e., 10 times more time steps than in the previous experiments.

On the one hand, we find that the existence of a social network hinders a consensus reaching and promotes extreme opinions. For instance, in the absence of a social network (i.e., using a fully-mixed population with a complete graph), there is a clear consensus reaching when the thresholds $\varepsilon$ and $\vartheta$ are large enough (e.g., with $\varepsilon = 0.4$ and $\vartheta = 0.9$, see Fig. 3). However this does not completely happen in the analyzed social networks: despite a large fraction of the population reaching such a consensus, there is still a non-negligible set of agents with extreme opinions. Another example can be found when the confidence and repulsion thresholds are low enough (e.g., $\varepsilon = 0.2$ and $\vartheta = 0.7$). In some social networks, the volume of extreme opinions even exceeds the number of neutral opinions (see, for instance, SW networks). This is justified for social network topologies showing a higher clustering coefficient, i.e., being modular networks and thus showing a strong community structure. Therefore, a chance to keep a different opinion in each community arises [5]. We recall that this does not happen in fully-mixed populations.

On the other hand, we note the existence of many intermediate opinions, neither neutral nor extreme. This is also an interesting phenomenon that hardly occurs in the absence of a social network (see Fig. 3). This effect is particularly relevant in SW graphs, with a lower impact in SF graphs, and little impact on ER networks. Again, this is due to the larger clustering coefficient in SW graphs, which promotes the emergence of communities among agents where those moderate opinions can be reinforced.

In sum, consensus reaching appears to be more difficult in the presence of a social network, where extreme opinions are promoted much more easily than in its absence. Furthermore, if the topology of the network eases the emergence of communities, different clusters of intermediate opinions can remain. This is especially relevant in social networks with a high clustering coefficient, where those communities are more clear.

### 5.4. Impact of the agent-based rationality

In this section we analyze the effects of the agent-based rationality on the OD.

In a first experiment, we study how opinions evolve in a population where agents have distinct (time-invariant) confidence and repulsion thresholds. In particular, we initialize $\varepsilon_i \in [0.2, 0.4]$ and $\vartheta_i \in [0.7, 0.9]$ (or sub-intervals of them), and compare the number of extreme (E) and neutral (N) opinions (i.e., opinions in the interval $[0, 0.1] \cup [0.9, 1]$, and $[0.4, 0.6]$, respectively) w.r.t. the vanilla version of our model without agent-based rationality. In Table 1 we report the results of this experiment.

As can be seen, introducing agent-based rationality allows us to model more complex scenarios with agents behaving distinctly, but without dramatically changing the global performance of the system. For instance, when $\varepsilon_i \in [0.2, 0.4]$ and $\vartheta_i \in [0.7, 0.9]$ (see the right bottom cell of Table 1), neutral opinions represent 68.9% of the agent population, while extreme opinions are 31.6%. In this scenario, there are agents with both moderate and polarized behaviors (i.e., low $\vartheta = 0.7$ and high $\vartheta = 0.9$, respectively). If agent-based rationality is not used, the scenario with the most similar results is found in $\varepsilon_i = 0.3$ and $\vartheta_i = 0.8$ (as expected). Notice, however, that in this case all the agents have the same confidence and repulsion thresholds and, hence the model is unable to represent the heterogeneous behavior of a population. For this reason, our model provides a more versatile representation of the complex and heterogeneous behavior of the population.

In a second experiment, we analyze the time-adaptation of agent rationality. In particular, we compare two scenarios to the vanilla time-invariant version of the model. First we study an initially polarized scenario with a temporal tendency to moderation; then we analyze an initially moderate scenario with a temporal tendency to polarization. In Fig. 6 we depict the results of this experiment.

In the initially polarized scenario, when the confidence and repulsion thresholds are time-invariant (i.e., $\varepsilon_i(t) = 0.4$ and $\vartheta = 0.7$), almost all opinions (99.6% of the agents) shift towards an extreme value (see Fig. 6 top left). In contrast, the introduction of temporal behaviors changes the global performance of the system. In particular, we analyze the case with $\varepsilon_i(t) = 0.4 + 10^{-6}t$ and $\vartheta = 0.7 + 10^{-6}t$ (i.e., agents get more moderate during the execution), and find that, although opinions initially move towards an extreme value, these dynamics change after a number of time steps, as a cluster of moderate opinions emerges (see Fig. 6 top right). Consequently, only 16.1% of the agents have an extreme opinion at the end of the execution.

Yet in the initially moderate scenario with time-invariant thresholds ($\varepsilon_i(t) = 0.4$ and $\vartheta = 0.9$), a consensus is reached in a neutral opinion (see Fig. 6 bottom left). In contrast, if the temporal behavior polarizes the agents as time progresses ($\varepsilon_i(t) = 0.4 - 9 \cdot 10^{-6}t$ and $\vartheta = 0.9 - 9 \cdot 10^{-6}t$), although there is a large consensus after a number of time steps, these neutral opinions move towards an extreme value afterwards due to polarization (see Fig. 6 bottom right). In the latter scenarios, extreme opinions represent 99.5% of the population.

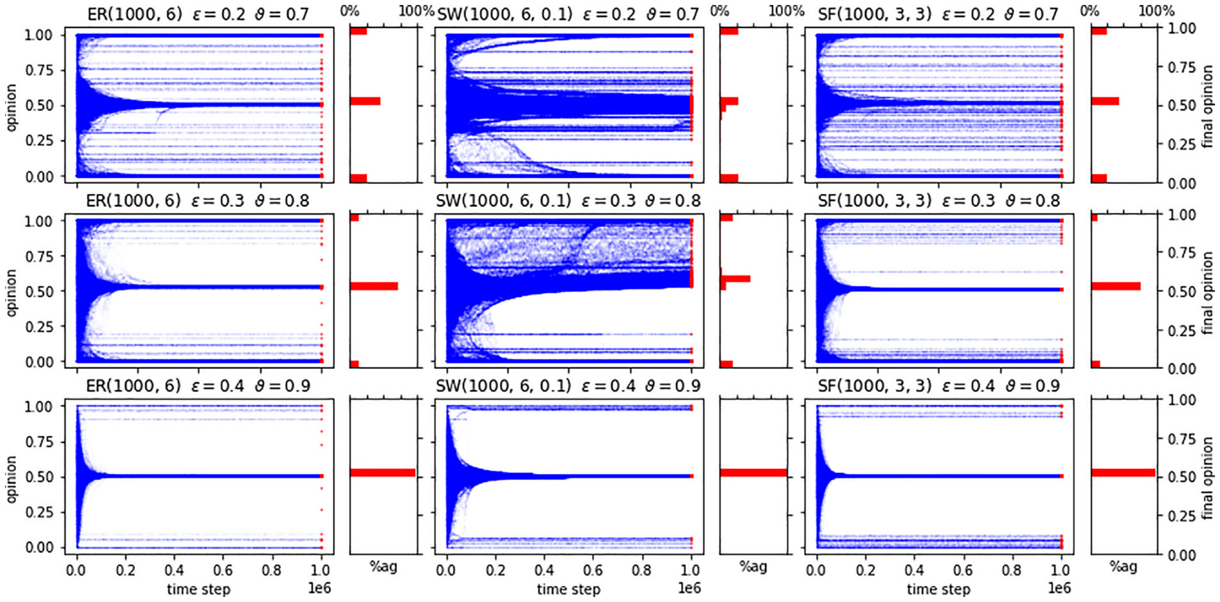

**Fig. 5.** Executions of the ATBCR model and histograms of the distribution of final opinions under distinct social network topologies: ER graphs (left), SW graphs (center), and SF graphs (right). The model is executed with $(\varepsilon, \vartheta) = (0.2, 0.7)$ (top), $(0.3, 0.8)$ (center), and $(0.4, 0.9)$ (bottom), and $\mu = 0.1$.

**Table 1**
Percentage of extreme (E) and neutral (N) opinions after executing the ATBCR model with agent-independent thresholds of confidence $\varepsilon_i$ and repulsion $\vartheta_i$.

| | $\varepsilon_i = 0.2$ | $\varepsilon_i \in [0.2, 0.3]$ | $\varepsilon_i = 0.3$ | $\varepsilon_i \in [0.3, 0.4]$ | $\varepsilon_i = 0.4$ | $\varepsilon_i \in [0.2, 0.4]$ |
|---|---|---|---|---|---|---|
| $\vartheta_i = 0.7$ | $E = 0.550$<br>$N = 0.450$ | $E = 0.540$<br>$N = 0.458$ | $E = 0.547$<br>$N = 0.453$ | $E = 0.591$<br>$N = 0.390$ | $E = 0.996$<br>$N = 0.004$ | $E = 0.568$<br>$N = 0.425$ |
| $\vartheta_i \in [0.7, 0.8]$ | $E = 0.456$<br>$N = 0.526$ | $E = 0.475$<br>$N = 0.525$ | $E = 0.465$<br>$N = 0.530$ | $E = 0.365$<br>$N = 0.603$ | $E = 0.051$<br>$N = 0.941$ | $E = 0.451$<br>$N = 0.562$ |
| $\vartheta_i = 0.8$ | $E = 0.388$<br>$N = 0.610$ | $E = 0.407$<br>$N = 0.591$ | $E = 0.313$<br>$N = 0.687$ | $E = 0.162$<br>$N = 0.840$ | $E = 0.010$<br>$N = 0.990$ | $E = 0.292$<br>$N = 0.703$ |
| $\vartheta_i \in [0.8, 0.9]$ | $E = 0.329$<br>$N = 0.667$ | $E = 0.280$<br>$N = 0.718$ | $E = 0.115$<br>$N = 0.883$ | $E = 0.036$<br>$N = 0.975$ | $E = 0.004$<br>$N = 0.992$ | $E = 0.136$<br>$N = 0.867$ |
| $\vartheta_i = 0.9$ | $E = 0.060$<br>$N = 0.000$ | $E = 0.018$<br>$N = 0.957$ | $E = 0.003$<br>$N = 0.991$ | $E = 0.011$<br>$N = 0.992$ | $E = 0.000$<br>$N = 1.000$ | $E = 0.031$<br>$N = 0.932$ |
| $\vartheta_i \in [0.7, 0.9]$ | $E = 0.414$<br>$N = 0.585$ | $E = 0.415$<br>$N = 0.595$ | $E = 0.325$<br>$N = 0.662$ | $E = 0.162$<br>$N = 0.834$ | $E = 0.009$<br>$N = 0.989$ | $E = 0.316$<br>$N = 0.689$ |

Altogether, our results show that the OD of a complex system are time-dependent, evolving over time, and our model is able to represent this complex behavior.

### 5.5. Analysis of the ATBCR model in a real-world network

In this section we analyze the performance of the ATBCR model in a real-world network: the ZKC [48]. This social network represents the interactions of 34 members of a university karate club. After a conflict between the administrator John A. and instructor Mr. Hi, the club was split into two, and the members chose to join one of them. This social network not only represents a real-world topology of interactions, but it also provides the ground truth of opinions after the division. In Fig. 7 left we depict this ground truth, each color representing one of the two communities (the members who followed either the administrator or the instructor).

In this case to study the OD of this real-world scenario, we randomly initialized each agent's opinion in the interval $[0, 1]$, and executed the ATBCR model with $\varepsilon_i \in [0.2, 0.4]$ and $\vartheta_i \in [0.7, 0.9]$ during 5000 time steps and using time-invariant thresholds. The results of this execution are depicted in Fig. 7 center.

As can be seen, two main clusters of opinions emerge, one composed of moderate opinions, and a second one of extreme values. Fig. 7 right shows the clusters of opinions obtained with the ATBCR, which is very close to the ground truth of this real-world network. The accuracy of this prediction (i.e., the number of agents correctly classified) is 85.29%, meaning that our model is able to capture the behavior of complex systems in an accurate manner.

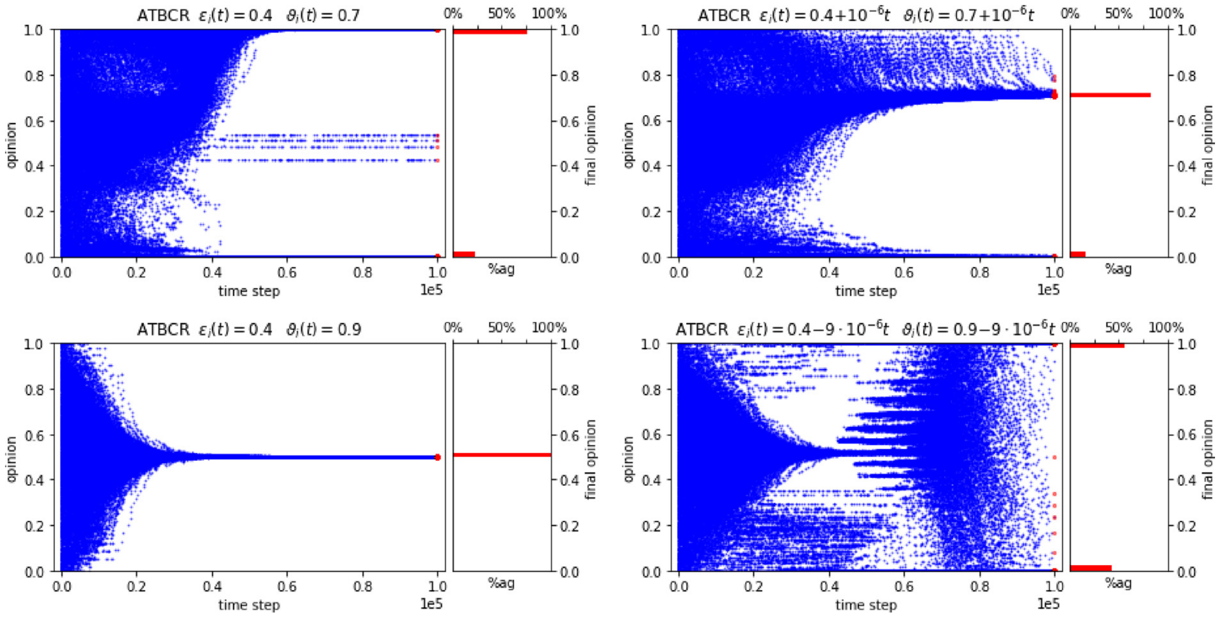

**Fig. 6.** Execution of the ATBCR model without (left) and with (right) time-varying agents' rationality, in a scenario initially polarized with a temporal tendency to moderation (top), and a scenario initially moderate with a temporal tendency to polarization (bottom).

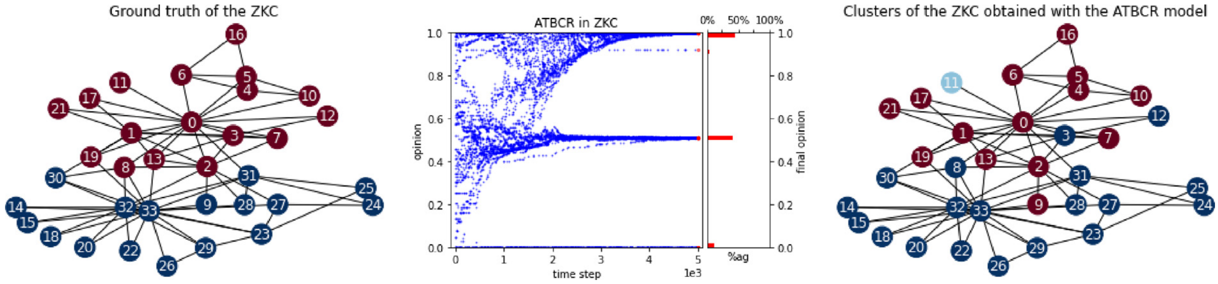

**Fig. 7.** Analysis of the ZKC with the ATBCR model: ground truth (left), OD in the ATBCR model (center), and opinion clusters obtained after executing the ATBCR model (right). The accuracy of the ATBCR model is 85.29%.

## 6. Application to a case study: “Narcos (Netflix)

In this section we describe an application of the proposed methodology in a real case of marketing for the Netflix series “Narcos” in Spain. Although there is no specific data on the modeled phenomenon, it can be viewed as representative of virality in social media and polarization of opinions. The proposed model is intended to achieve realistic results.

Netflix is a streaming service that offers a wide variety of award-winning series, movies, anime titles, documentaries, and other kinds of content on thousands of internet-connected devices.[2] The series “Narcos” chronicles the efforts of the United States, mainly through the DEA, and Colombian authorities and police to corner drug dealer Pablo Escobar and end the Medellí cartel, one of the richest and most ruthless criminal organizations in the history of modern crime.

“Narcos” was very popular during its three seasons, being one of the most watched Netflix series in Spain in 2017.[3] To publicize the series, diverse marketing actions were launched worldwide, generating virality in social media. Examples of these creative campaigns are one in Paris with hourglasses,[4] and in different US cities with word games.[5]

In Spain, “Narcos” was advertised in several places and in different ways starting with the second season of the series (December 2016). The case in point took place at Puerta del Sol, the very center of Madrid. This location was not chosen ran-

[2] https://www.netflix.com.

[3] https://www.highspeedinternet.com/resources/netflix-what-the-world-is-watching.

[4] https://lareclame.fr/ubibene-netflix-narcos-183754.

[5] https://www.businessinsider.com/netflix-is-promoting-the-latest-season-of-narcos-in-bars-and-clubs-2017-9.

domly. It promised maximum visibility as a place well frequented by local residents as well as tourists; if the campaign was striking enough, its virality would be multiplied.[6]

Indeed, the message was very provocative. As shown in Fig. 8, the poster featured the star of the series along with the phrase “Oh, blanca Navidad” (“Oh, white Christmas” in Spanish), a play on words between the topic of the series (cocaine), and the time of year, just before Christmas.

The campaign was a success. In fact, after some days, it provoked reactions even from the Colombian government, whose Ministry of Foreign Affairs requested withdrawal of the promotional poster.[7] What is more important for our study, although the campaign was not designed for social networks, it gave rise to viral action on social media such as Twitter, becoming an electronic Word of Mouth (eWOM) phenomenon affected by the polarization of opinions, as we will detail below.

The current effect of eWOM on consumer behavior is well known. Henning-Thurau and Walshg [26] establishes that “any positive and negative statement made by former clients, current clients or potential clients about a product or company and that is available to a multitude of people and institutions via the Internet acquires vital importance”, especially in controversial marketing campaigns. By definition, eWOM polarizes those opinions expressed in the digital context, placing them in extreme positions, and this result is very useful for certain actions carried out in marketing, such as the case at hand.

Much of the literature on the polarization of opinions has focused on individuals’ political predispositions, such as party affiliation [1,14] and their patterns of media use [38,41] as determinants of polarization. Following those roots from the political sphere, brands are nowadays much more willing to assume controversial roles in their communication, to gain high publicity and many followers. Opinion polarization can prove beneficial for marketing actions, especially within advertising,[8] although on some occasions it can be counterproductive for the brand image [35].

An extreme division of opinions may be sought by marketing campaigns, such as the case of Nike with athlete Colin Kaepernick,[9] Pepsi with Kendall Jenner,[10] Gillette with “We Believe: The Best Men Can Be” [46], meant to be as notorious, not leave indifferent, generate conversation, create controversy, and provoke the audience.

Indeed, the “Narcos” marketing strategy in Spain began harvesting shows of admiration both on the street and on social media, cataloging it as sublime, masterpiece, or excellence. But later it had a negative impact, as previously mentioned. In the end, despite the controversy, the action achieved very positive results for the company: more than 4 million euros in return on investment and more than 1,600,000 impressions were achieved. Likewise, the campaign was recognized by the advertising industry itself and received two silver awards at the 2017 El Sol festival (the Ibero-American Festival of Advertising Communication). The series was immensely popular in Spain at the end of 2016 and during the first half of 2017.[11]

Hence, we take this case of marketing “Narcos” to show how the proposed ATBCR model can capture both the early and the late time progression of dynamics of the case study, something that does not happen in the classic model DW. To this end, we use a population of $|S| = 10^5$ agents distributed along a graph $SF(10^5, 3, 3)$, which represents the social network where agents’ interactions occurred. It should be stressed that eWOM is better modeled by interactions between pairs of agents, rather than broadcasting strategies or other forms of mass communications. Therefore, a social network is the most suitable representation of this kind of reciprocal interaction. In turn, SF graphs are an appropriate representation of real-world interactions, where a few agents have a large number of interactions while most just interact in small groups.

The initial opinions of the population are slightly biased toward positive opinions, as observed during the marketing action. In particular, 90% of the initial opinions are randomly initialized in the interval $[0.3, 1]$; the other 10% is randomly initialized in $[0, 0.3]$. Lacking more precise data, we assume this is a realistic distribution of the original opinion profile. In a first experiment, the ATBCR model is executed with time-invariant agent-based rationality $\varepsilon_i \in [0.15, 0.35]$ and $\vartheta_i \in [0.55, 0.75]$, indicating a high degree of repulsion in the system, as was observed in the studied campaign.

In a second experiment, we introduce agent-independent time-based rationality, in a set-up with $\varepsilon_i(t) = 0.25 - 3 \cdot 10^{-7}t$ and $\vartheta_i(t) = 0.65 - 3 \cdot 10^{-7}t$, whereas the rest of the parameters remain unaltered. This set-up represents an scenario where agents get more repulsive over time, a fairly common situation in marketing campaigns. Consumers occupy a process of assimilation throughout a campaign’s duration [33], during which their awareness of the message is increased. The consequence is that the consumers end up polarizing their opinion, placing themselves at one extreme or the other [32].

In both cases, the number of time steps is set to $T = 10^6$ iterations and the convergence speed is $\mu = 0.05$. Notice that, we use the same simulation time to study both the early and the late time behavior. This decision is based on two reasons. First, to allow comparisons of the dynamics of both scenarios in the same environment. Second, the introduction of time-based rationality variation allows us to simulate a long-term dynamics for the real-world case adjusted to the length of the analyzed period.

In Fig. 9 we depict the OD of this execution, as well as the final opinion profile. In the first experiment (see Fig. 9 left), many opinions are polarized towards an extreme value (either positive or negative). Interestingly, this phenomenon seems more relevant for positive opinions (45%) than negative opinions (15%). This probably reflects the impact of the confidence

[6] https://www.puromarketing.com/7/29175/campanas-netflix-para-narcos-acaban-convirtiendose-virales.html.

[7] https://www.elperiodico.com/es/extra/20161215/colombia-pide-retirar-el-cartel-de-la-serie-narcos-en-la-puerta-del-sol-de-madrid-5692814.

[8] https://elpais.com/economia/2017/01/26/actualidad/1485446729_525959.html.

[9] https://www.elmundo.es/extras/publicidad/2019/01/25/5c49f347fc6c83e54e8b45b9.html.

[10] https://sparkflow.co/blog/pepsis-protest-ad-kendall-jenner-marketing-fail/.

[11] https://www.reasonwhy.es/actualidad/campanas-publicidad-netflix-cinco-anos-espana-legado

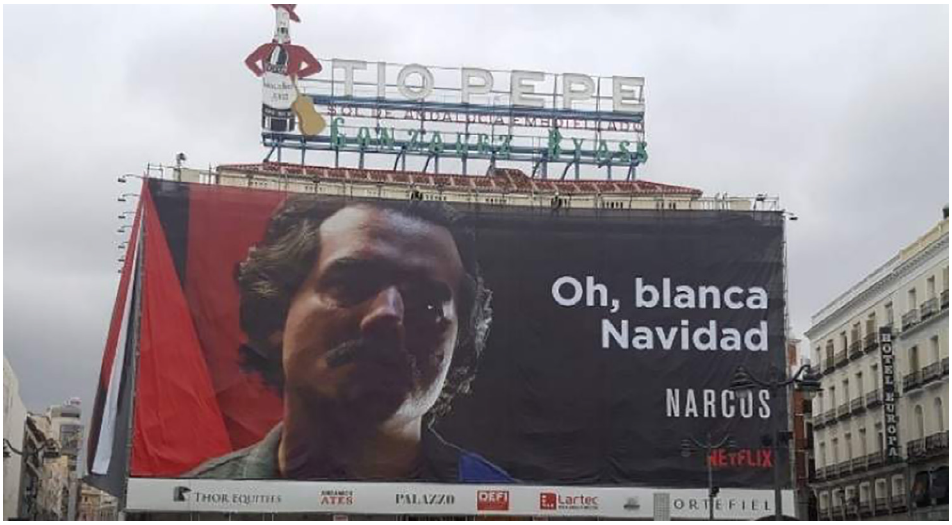

**Fig. 8.** Campaign base ad.

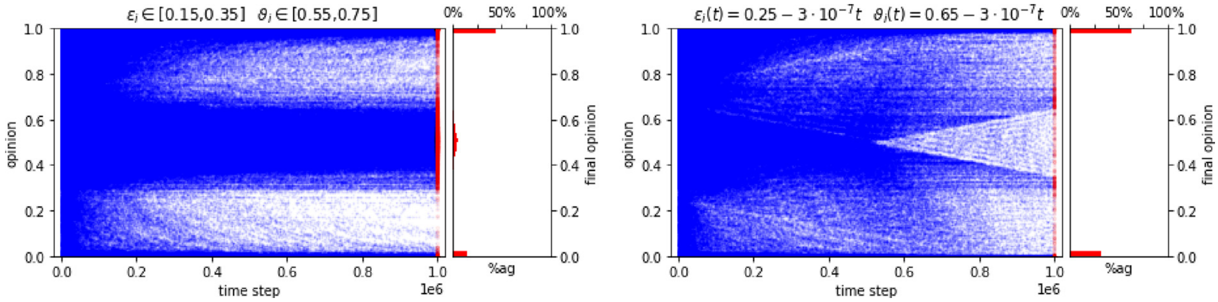

**Fig. 9.** Simulation of the Netflix scenario, with the ATBCR executed on a $SF(10^5, 3, 3)$ graph with $\varepsilon_i \in [0.15, 0.35]$ and $\vartheta_i \in [0.55, 0.75]$ (left), and $\varepsilon_i(t) = 0.25 - 3 \cdot 10^{-7}t$ and $\vartheta_i(t) = 0.65 - 3 \cdot 10^{-7}t$ (right).

regime, marking a greater impact on positive opinions since there is a larger number of them in the initial opinion profile. In addition, the repulsion rule affects both positive and negative opinions, hence polarizing both of them towards extreme values. This is observed throughout the execution. Finally, a large number of agents (around 32%) scarcely modify their opinions, remaining neutral or hardly affected by this marketing campaign.

In the second experiment (see Fig. 9 right), we observe that the time-based rationality of the agents has a major impact on the late time progression of their opinions. In this case, most of the opinions are extremized after the execution (96%), whereas less than 1% would represent neutral opinions. As in the previous experiment, extreme opinions are biased towards a positive (extreme) value. In particular, they represent a 64% of the population, while negative (extreme) opinions are only 32%. This signals the impact of the agents' temporal-varying behavior of the agents, the repulsion regime gaining weight as time progresses. It explains why most opinions, including those that remained neutral in the previous experiment, become extreme when the time-based agents' rationality is in play.

Overall, this behavior makes sense in the light of the effects that advertising creativity has on consumers [40] and their emotional attachment to a brand [47]. These conditions are optimal to arouse a favorable emotion or sympathy towards the advertiser as personal affection is enhanced by frequent exposure and publicity [4], and this is very common with Netflix, whose creativity tends to be outstanding, effectively influencing consumers [24].

In summary, the present study shows that the ATBCR is able to simulate a real-wold scenario entailing a polarization of opinions, even in the case of a non-uniform distribution. Although the confidence rule of the DW model is able to model many of the interactions in the system, it is unable to reproduce the extremization of opinions in the population, which is achieved by the repulsion phase in the ATBCR model. Moreover, the distinct rationality of the agents (including temporal adaptation) provides a more realistic representation of the complex behaviors of individuals in a population. The complex system dynamics at different points in time instants can be analyzed by considering different ATBCR model parameterizations.

## 7. Conclusions and future work

The classical DW model [11] is a system of OD with BC, where agents meet in random pairwise encounters and update their opinions whenever they are similar. This similarity is defined by a BC threshold, whose value determines the dynamics of the system. In particular, for low values of the confidence threshold, the agents reach an opinion consensus in terms of a neutral opinion, whereas for higher values a fragmentation of opinions into different clusters emerges. Nevertheless, this model is unable to explain the emergence of extremization, i.e., the process whereby opinions evolve towards extreme values.

Although there exist some extensions of the DW model that may be applied to analyze the extremization of a population [12,2,31,28,10], they fail to model the heterogeneous and time-dependent rationality of agents. Alternatively, in this work we propose an extension of the DW model with a repulsion procedure, able to capture the mechanism of extremization in a population in a more versatile and general way. Our model, called ATBCR, behaves as the classical DW model in the confidence area of the agents (i.e., when they have very similar opinions), but it also implements a repulsion rule through which agents reinforce their own opinions when they encounter other agents having very distant opinions. This repulsion mechanism is triggered by a repulsion threshold, and in conjunction with the confidence rule proposed in the DW model, the ATBCR model can result in an extremization of certain agents (due to the repulsion mechanism), while preserving the consensus and fragmentation of opinions in the rest (due to the confidence rule). Moreover, it incorporates agent-based rationality, i.e., both the confidence and the repulsion thresholds may be distinct for each agent, and they may evolve over time.

Our model is empirically evaluated, showing that both the confidence and the repulsion thresholds have a direct impact on the extremization of the agents. In particular we show that, as under the classical DW model, a low value for the confidence threshold leads to a fragmentation of opinions, whereas the model eventually reaches a consensus when a high threshold value is used. At the same time, the value of the repulsion threshold affects the extremization of the agents, lower values promoting the emergence of extreme opinions and increasing their volume in the population. Therefore, the degree of extremization in the system can be controlled with both parameters. Social networks can promote extreme opinions, where a consensus is more difficult to reach, but they also promote moderate opinions (neither neutral nor extreme) that would not remain in the absence of a social network. Having analyzed the effects of agents' time-based adaptation on the OD, we may state that these dynamics can be dramatically affected by temporal changes in agents' rationality. In addition, we checked the performance of our model in the ZKC, a real-world social network with a known ground truth, finding that the ATBCR model exhibits an accuracy of $85.29\%$. These results suggest that our model provides a versatile and general framework to represent many complex OD scenarios.

Lastly, we explored a real-world marketing action initiated by Netflix in Spain in December, 2016 to promote its series "Narcos". This campaign provoked a massive polarization of opinions, with a large number of extreme opinions (both positive and negative). In our work, we use the ATBCR model to simulate the dynamics of this opinion polarization, showing the versatility of the model to handle complex real-world scenarios. It is suitable to reproduce both the early and the late time progression of opinion polarization in viral guerrilla marketing actions like the advertising campaign described here.

As future work, we plan to extend this model in several different directions. First, we plan to analyze other factors of opinion influence, such as advertising, media, and marketing campaigns, among others [7]. They can be modeled considering that mass communication campaigns spread an opinion (either positive or negative) within a population according to a level of influence and reach, and agents update their opinions considering those *interactions* [23]. This line of study can be further extended by analyzing OD in contexts of multiple coexisting opinions, as in [22]. Second, we plan to incorporate a more realistic representation of opinion to our model, based on fuzzy linguistic variables [15,34]. Finally, our proposed model can serve as the basis of a population-based optimization algorithm, where agents' opinions are the values of the objective function, and interactions between agents (i.e., solutions) can give rise to individuals with extreme opinions, which would allow the algorithm to better explore the search space in order to look for high quality solutions.

## CRediT authorship contribution statement

**Jesús Giráldez-Cru:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Writing – original draft, Writing – review & editing. **Carmen Zarco:** Conceptualization, Validation, Resources, Writing – review & editing. **Oscar Cordón:** Conceptualization, Validation, Resources, Writing – review & editing, Funding acquisition.

## Data availability

Data will be made available on request.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgments

This work was supported by the Spanish Ministry of Science and Innovation (MCIN), Agencia Estatal de Investigación (AEI), under grant CONFIA (PID2021-122916NB-I00), and by the Andalusian Government (Spain) and the University of Granada (UGR, Spain) under grants AIMAR (A-TIC-284-UGR18) and SIMARK (P18-TP-4475), including European Regional Development Funds (ERDF). J. Giráldez-Cru is also supported through the Juan de la Cierva program, fellowship IJC2019-040489-I, funded by MCIN and AEI. Funding for open access charge: Universidad de Granada / CBUA.

## References

[1] A. Abramowitz, K. Saunders, Is polarization a myth?, J Politics 70 (2) (2008) 542–555.
[2] F. Amblard, G. Deffuant, The role of network topology on extremism propagation with the relative agreement opinion dynamics, Phys. A 343 (2004) 725–738.
[3] A. Anagnostopoulos, L. Becchetti, E. Cruciani, F. Pasquale, S. Rizzo, Biased opinion dynamics: when the devil is in the details, Inf. Sci. 593 (2022) 49–63.
[4] C. Atkin, G. Heald, Effects of political advertising, Public Opin. Q. 40 (2) (1976) 216–228.
[5] A.L. Barabási, Network Science, Cambridge University Press, 2016.
[6] A.L. Barabási, R. Albert, Emergence of scaling in random networks, Science 286 (1999) 509–512.
[7] T. Carletti, D. Fanelli, S. Grolli, A. Guarino, How to make an efficient propaganda, Europhys. Lett. 74 (2) (2006) 222–228.
[8] C. Castellano, S. Fortunato, V. Loreto, Statistical physics of social dynamics, Rev. Mod. Phys. 81 (2009) 591–646.
[9] J. Chen, G. Kou, H. Wang, Y. Zhao, Influence identification of opinion leaders in social networks: an agent-based simulation on competing advertisements, Inf. Fusion 76 (2021) 227–242.
[10] P. Dandekar, A. Goel, D.T. Lee, Biased assimilation, homophily, and the dynamics of polarization, Proc. Nat. Acad. Sci. 110 (15) (2013) 5791–5796.
[11] G. Deffuant, D. Neau, F. Amblard, G. Weisbuch, Mixing beliefs among interacting agents, Advances in Complex Systems 3 (01n04) (2000) 87–98.
[12] G. Deffuant, F. Amblard, G. Weisbuch, T. Faure, How can extremism prevail? a study based on the relative agreement interaction model, J. Artif. Soc. Soc. Simul. 5 (4) (2002).
[13] M.H. Degroot, Reaching a consensus, J. Am. Stat. Assoc. 69 (345) (1974) 118–121.
[14] P. DiMaggio, J. Evans, B. Bryson, Have american's social attitudes become more polarized?, Am J. Sociol. 102 (3) (1996) 690–697.
[15] Y. Dong, X. Chen, H. Liang, C.-C. Li, Dynamics of linguistic opinion formation in bounded confidence model, Inf. Fusion 32 (2016) 52–61.
[16] Y. Dong, M. Zhan, G. Kou, Z. Ding, H. Liang, A survey on the fusion process in opinion dynamics, Inf. Fusion 43 (2018) 57–65.
[17] Y. Dong, M. Zhan, Z. Ding, H. Liang, F. Herrera, Numerical interval opinion dynamics in social networks: Stable state and consensus, IEEE Trans. Fuzzy Syst. 29 (3) (2021) 584–598.
[18] P. Erdós, A. Rényi, On random graphs, Publicationes Mathematicae 6 (1959) 290–297.
[19] K. Fan, W. Pedrycz, Emergence and spread of extremist opinions, Phys. A 436 (2015) 87–97.
[20] K. Fan, W. Pedrycz, Opinion evolution influenced by informed agents, Phys. A 462 (2016) 431–441.
[21] S. Fortunato, Universality of the threshold for complete consensus for the opinion dynamics of Deffuant et al., Int. J. Modern Phys. C 15(09) (2004) 1301–1307.
[22] J. Giráldez-Cru, M. Chica, O. Cordón, A framework of opinion dynamics using fuzzy linguistic 2-tuples, Knowl.-Based Syst. 233 (2021) 107559.
[23] J. Giráldez-Cru, M. Chica, O. Cordón, The effects of mass communication in a fuzzy linguistic framework of opinion dynamics, in: Proceedings of the 2022 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2022), 2022.
[24] J. Hazla, Y. Jin, E. Mossel, and G. Ramnarayan. A geometric model of opinion polarization. CoRR, abs/1910.05274, 2019.
[25] R. Hegselmann, U. Krause, Opinion dynamics and bounded confidence models, analysis, and simulation, J. Artif. Soc. Soc. Simul. 5 (3) (2002).
[26] T. Henning-Thurau, D. Walshg, Electronic word of-mouth via consumer-opinion platforms: what motivates consumers to articulate themselves on the internet, J. Interactive Market. 18 (1) (2004) 38–52.
[27] J.A. Hołyst, K. Kacperski, F. Schweitzer, Phase transitions in social impact models of opinion formation, Phys. A 285 (1) (2000) 199–210.
[28] S. Huet, G. Deffuant, W. Jager, A rejection mechanism in 2D bounded confidence provides more conformity, Adv. Complex Syst. 11 (4) (2008) 529–549.
[29] D.J. Isenberg, Group polarization: A critical review and meta-analysis, J. Pers. Soc. Psychol. 50 (6) (1986) 1141–1151.
[30] G. Iyer, H. Yoganarasimhan, Strategic polarization in group interactions, J. Mark. Res. 58 (4) (2021) 782–800.
[31] W. Jager, F. Amblard, Uniformity, bipolarization and pluriformity captured as generic stylized behavior with an agent-based simulation model of attitude change, Comput. Math. Organization Theory 10 (4) (2005) 295–303.
[32] K. Kułakowski, Opinion polarization in the receipt–accept–sample model, Phys. A 388 (4) (2009) 469–476.
[33] J. Li, R. Xiao, Agent-based modelling approach for multidimensional opinion polarization in collective behaviour, J. Artif. Soc. Soc. Simul. 20 (2) (2017) 4.
[34] H. Liang, C.-C. Li, Y. Dong, F. Herrera, Linguistic opinions dynamics based on personalized individual semantics, IEEE Trans. Fuzzy Syst. 29 (9) (2021) 2453–2466.
[35] L. Lilla, P. Szabolcs, Using controversial values in csr communication – analysing the coca-cola #loveislove campaign, in: Proceedings of the 11th European Marketing Academy, 2020, pp. 83385.
[36] J. Lorenz, Heterogeneous bounds of confidence: Meet, discuss and find consensus!, Complexity 15 (4) (2010) 43–52
[37] J.-D. Mathias, S. Huet, G. Deffuant, Bounded confidence model with fixed uncertainties and extremists: The opinions can keep fluctuating indefinitely, J. Artif. Soc. Soc. Simul. 19 (1) (2016) 6.
[38] N.H. Nie, I. Darwin, W. Miller, S. Golde, D.M. Butler, K. Winneg, The world wide web and the u.s. political news market, Am. J. Polit. Sci. 54 (2) (2010) 428–439.
[39] H. Noorazar, M.J. Sottile, K.R. Vixie, An energy-based interaction model for population opinion dynamics with topic coupling, Int. J. Mod. Phys. C 29 (11) (2018) 1850115.
[40] R. Smith, X. Yang, Toward a general theory of creativity in advertising: Examining the role of divergence, Market. Theory 4 (1/2) (2004) 29–55.
[41] N. Stroud, Media effects, selective exposure, and fahrenheit 9/11, Polit. Commun. 24 (4) (2007) 415–432.
[42] W. Su, X. Wang, G. Chen, Y. Yu, T. Hadzibeganovic, Noise-based synchronization of bounded confidence opinion dynamics in heterogeneous time-varying communication networks, Inf. Sci. 528 (2020) 219–230.
[43] D.J. Watts, S.H. Strogatz, Collective dynamics of 'small-world' networks, Nature 393 (6684) (1998) 440–442.
[44] G. Weisbuch, Bounded confidence and social networks, Eur. Phys. J. B 38 (2004) 339–343.
[45] H. Xia, H. Wang, Z. Xuan, Opinion dynamics: A multidisciplinary review and perspective on future research, Int. J. Knowl. Syst. Sci. 2 (4) (2011) 72–91.
[46] S. Xu, Y. Xiong, Setting socially mediated engagement parameters: A topic modeling and text analytic approach to examining polarized discourses on gillette's campaign, Public Relations Review 46 (5) (2020) 101959.
[47] X. Yang, R.E. Smith, Beyond attention effects: Modeling the persuasive and emotional effects of advertising creativity, Market. Sci. 28 (5) (2009) 935–949.
[48] W.W. Zachary, An information flow model for conflict and fission in small groups, J. Anthropol. Res. 33 (4) (1977) 452–473.
[49] Q. Zhou, Z. Wu, A.H. Altalhi, F. Herrera, A two-step communication opinion dynamics model with self-persistence and influence index for social networks based on the DeGroot model, Inf. Sci. 519 (2020) 363–381.