

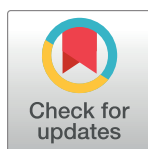
RESEARCH ARTICLE

Algorithmic bias amplifies opinion fragmentation and polarization: A bounded confidence model

Alina Sîrbu^{1,2*}, Dino Pedreschi¹, Fosca Giannotti³, János Kertész^{4,5}

1 Department of Computer Science, University of Pisa, Pisa, Italy, **2** Science Division, New York University Abu Dhabi, Abu Dhabi, United Arab Emirates, **3** Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo" - CNR, Pisa, Italy, **4** Center for Network Science, Central European University, Budapest, Hungary, **5** Department of Theoretical Physics, Budapest University of Technology and Economics, Budapest, Hungary

* alina.sirbu@unipi.it



Abstract

The flow of information reaching us via the online media platforms is optimized not by the information content or relevance but by popularity and proximity to the target. This is typically performed in order to maximise platform usage. As a side effect, this introduces an algorithmic bias that is believed to enhance fragmentation and polarization of the societal debate. To study this phenomenon, we modify the well-known continuous opinion dynamics model of bounded confidence in order to account for the algorithmic bias and investigate its consequences. In the simplest version of the original model the pairs of discussion participants are chosen at random and their opinions get closer to each other if they are within a fixed tolerance level. We modify the selection rule of the discussion partners: there is an enhanced probability to choose individuals whose opinions are already close to each other, thus mimicking the behavior of online media which suggest interaction with similar peers. As a result we observe: a) an increased tendency towards opinion fragmentation, which emerges also in conditions where the original model would predict consensus, b) increased polarisation of opinions and c) a dramatic slowing down of the speed at which the convergence at the asymptotic state is reached, which makes the system highly unstable. Fragmentation and polarization are augmented by a fragmented initial population.

OPEN ACCESS

Citation: Sîrbu A, Pedreschi D, Giannotti F, Kertész J (2019) Algorithmic bias amplifies opinion fragmentation and polarization: A bounded confidence model. PLoS ONE 14(3): e0213246. <https://doi.org/10.1371/journal.pone.0213246>

Editor: Floriana Gargiulo, Centre National de la Recherche Scientifique, FRANCE

Received: March 1, 2018

Accepted: February 19, 2019

Published: March 5, 2019

Copyright: © 2019 Sîrbu et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All results from simulations are included in the manuscript.

Funding: This work was supported by the European Community's H2020 Program under the funding scheme "FETPROACT-1-2014: Global Systems Science (GSS)", grant agreement # 641191 to AS, DP, FG and JK, CIMPLEX "Bringing Citizens, Models and Data together in Participatory, Interactive Social Explorations." The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Introduction

Political polarization and opinion fragmentation is a generally observed, ingravescence negative trend in modern western societies [1–4] with such concomitants as "alternative realities", "filter bubbles", "echo chambers", and "fake news". Several causes have been identified (see, e.g., [5]) but there is increasing evidence that new online media is one of them [6–8]. Earlier it was assumed that traditional mass media mostly influence the politically active elite of the society and they only indirectly affect polarization of the entire population. The recent dramatic changes with the occurrence of online media, the ubiquity of the Internet with all the

Competing interests: The authors have declared that no competing interests exist.

information within reach of a few clicks and the general usage of online social networks have increased the number of communication channels by which political information can reach citizens. Somewhat counterintuitively, this has not lead to a more balanced information acquisition and a stronger tendency towards consensus, as argued in [9], on the contrary. One of the reasons may be that the new media have enhanced the reachability of people, which can be used for transmitting simplified political answers to complex questions and thus act toward polarization [10]. Moreover, the stream of news is not organized in the new media in a balanced way, but by algorithms, which are built to maximise platform usage. It is conjectured that this generates an “algorithmic bias”, which artificially creates opinion fragmentation and enhances polarization. This is an artefact of online platforms, also called “algorithmic segregation” [11].

A considerable and rapidly increasing part of the population does not use traditional media (printed press, radio, TV or even online journals) for obtaining news [12–14] but turns to the new media like online social networks or blogs. However, the flow of news in the new media is not selected by the information value but rather by popularity, by “likes” [15]. As people tend to identify themselves with views similar to their own and will like corresponding news with higher probability, it is in the interest of the service providers to channel the information already in a targeted way. This means users do not even get confronted with narratives different from their favorite ones. Large effort is paid to develop efficient algorithms for the appropriate channeling of the news to provide the stream that has the highest chance to collect maximum number of likes.

Another important factor pointing in the same direction is related to the “share” function, which is largely responsible for the fast spreading of news and thus enhancing popularity. This spreading takes place on the social network, where links are formed mostly as a consequence of homophily, i.e., exchange of information takes place between people with similar views. The variety of interactions of a person (family, school, work, hobby, etc.) may contribute to diversification of information sources [16], nevertheless, regarding political views homophily seems particularly strong [17]. The sharing among users with similar beliefs was also shown to be true in the context of spreading of misinformation on Facebook, leading to an echo-chamber effect [18], as well as on Twitter where partisan users were demonstrated to have an important role in polarization [19]. Additionally, it was observed that consumption of news on Facebook generates very sharp media clusters that users do not leave [20]. While consuming these news, users become increasingly polarised, especially around certain media outlets and countries [21]. Through opinion dynamics modelling, it was conjecture that it is the trust in the media outlets that can counteract polarization. Recently, algorithmic bias in the evolution of the social network (i.e. by social recommendation systems), combined with homophily, was studied [22]. It was demonstrated to enhance the glass ceiling effect, where some user groups (e.g. female users) are excluded from the top layers of the social network.

The link between opinion fragmentation and polarization, and algorithmic bias from online platforms has not been proven to date, with conflicting conclusions from different studies [23]. The aim of the present paper is to study this effect within a simple opinion dynamics model based on bounded confidence. The study of fragmentation and polarisation together is very common in the literature on public opinions, news, media, etc. [24–26]. We measure fragmentation through the number of opinion clusters that arise in a population. Regarding the concept of polarization, a large number of definitions exist, with little consensus on what is the best choice. In general, criteria to measure it encompass various aspects such as appearance of opinion clusters but also distance between opinions [27]. Obviously, the first step of polarization is what we call fragmentation, when an earlier consensual opinion splits into two non-mixing parts as a result of the change of the conditions (e.g., occurrence of algorithmic bias),

or when the number of clusters grows. At the same time, the dispersion of opinions in the population can change. This can be measured through various criteria (e.g. range of opinions, distance between clusters, distance between opinions, standard deviation of opinions) [27]. Here, we opted to use the average pairwise distance between opinions, because this is robust to appearance of isolated small clusters which are common in the bounded confidence model. This criterion does not distinguish well between a situation where opinions are very spread along the opinion spectrum (e.g. uniform distribution) and some clustered situation, however this is not an issue because our model generally produces clusters. Also, we use the measure in conjunction with the number of clusters themselves, so the structure of the population is always clear in our results.

Network effects are obviously important in the spreading of news and opinions, however, in our present study we will ignore the network structure of the system and will exclusively focus on the consequences of the algorithmic bias for selecting the content presented to the users. This corresponds to a mean field approach, which is widely used as a first approximation to spreading problems [28].

The task is therefore to model the evolution of the distribution of individual attitudes in society, provided there is a bias in selecting partners whose opinions are confronted. Recent years has seen the introduction of several models of opinion dynamics [29], however none of them includes algorithmic bias in the sense described here. Hence we provide an extension of one of the existing models to study this effect. We use a bounded confidence opinion dynamics model [30, 31], which is known to be able to describe both consensus and polarization through clustering of opinions, depending on the tolerance level of the agents. We adopt the definition by [30], where agents interact in a pairwise manner rather than taking the average neighbour opinion like in [31]. This due to the need to change the pairwise interaction rates to account for algorithmic bias. In the mean field version two agents are selected at random and their opinions, represented by real numbers, get closer if they are within the tolerance level. The bias is introduced such that already at the selection of the partners their opinions are considered and agents with closer opinions are selected with higher probability. With this simple modification we see that the tendency toward consensus is hindered, i.e., larger tolerance level is needed to achieve consensus. Additionally, the approach to the asymptotic state is slowed down tremendously. Therefore, the selection bias affects the opinion formation process in three profound ways: a) by inducing fragmentation, also in conditions where the original model would predict consensus b) by exacerbating polarization increasing the average distance between opinions and c) by slowing down the spreading process, which makes the system highly unstable.

The paper is organized as follows. In the next section we introduce the model in detail. Section Results contains the results of the simulations. We close the paper with a discussion and an account of further research.

The model

The original bounded confidence model [30] considers a population of N individuals, where each individual i holds a continuous opinion $x_i \in [0, 1]$. This opinion can be considered the degree by which an individual agrees or not to a certain position. Individuals are connected by a complete social network, and interact pairwise at discrete time steps. The interacting pair (i, j) is selected *randomly* from the population at each time point t . After interaction, the two opinions, x_i and x_j may change, depending on a so called *bounded confidence parameter* $\epsilon \in [0, 1]$. This can be seen as a measure of the open-mindedness of individuals in a population. It

defines a threshold on the distance between the opinion of the two individuals, beyond which communication between individuals is not possible due to conflicting views.

If we define the distance between two opinions x_i and x_j as $d_{ij} = |x_i - x_j|$, then information is exchanged between the two individuals only if $d_{ij} \leq \varepsilon$, otherwise nothing happens. The model is based on attractive dynamics, i.e. the exchange of information results in the two opinions becoming closer to one another, modulated by a convergence parameter $\mu \in (0, 0.5]$

$$\begin{aligned}x_i(t+1) &= x_i(t) + \mu(x_j(t) - x_i(t)) \\x_j(t+1) &= x_j(t) + \mu(x_i(t) - x_j(t))\end{aligned}$$

In the following we consider only the case of $\mu = 0.5$, hence when both individuals take the average opinion.

To introduce algorithmic bias in the interaction between individuals, we modify the procedure by which the pair i, j is selected. In the original model, i and j are selected uniformly at random from the population. Instead, algorithmic bias makes encounters of similar individuals more probable. To account for this, we first select i , then the selection of j is performed with a probability that depends on d_{ij} :

$$p_i(j) = \frac{d_{ij}^{-\gamma}}{\sum_{k \neq i} d_{ik}^{-\gamma}} \quad (1)$$

In this way, the probability to select j once i was selected is larger if d_{ij} is smaller, i.e. alike individuals interact more. The parameter γ is the strength of the algorithmic bias: the larger γ , the more rare will be the encounters among individuals with distant opinions. For $\gamma = 0$ we obtain the original bounded confidence model, $p_i(j) = \frac{1}{N-1}$.

In the following we will analyse the new model numerically, through simulation of the opinion formation process. To avoid undefined operations in Eq 1, when $d_{ik} = 0$ ($k \neq i$), we use a lower bound for d_{ik} , d_ε . So, if $d_{ik} < d_\varepsilon$ then d_{ik} is replaced with d_ε in Eq 1. We use $d_\varepsilon = 0.0001$ in the following. Our simulations are design to stop when the population converges to a stable cluster configuration. To achieve this, we analyse the population at every iteration, i.e. every N pairwise interactions. We compute the maximum opinion change that occurred in an individual within an iteration. If the maximum decreases below a threshold (fixed at 0.00001), for at least 1,000 iterations, then we stop the simulation. We also set an overall maximum number of iterations at 10,000,000, but this was never actually reached in the simulations presented below. The model implementation is also available in the NDlib Python library [32, 33].

Results

In order to understand how the introduction of the algorithmic bias affects model performance, we study the model under multiple criteria for various combinations of parameters ε and γ . We are interested in whether the population converges to consensus or to multiple opinion clusters, whether polarization emerges and how fast convergence appears. We also consider the influence of the size of the population on the behavior observed, both for the original and extended model. Furthermore, the effect of a fragmented initial population is studied. For each analysis we repeat simulations multiple times to account for the stochastic nature of the model, and show average values obtained for each criterion above.

Consensus versus opinion fragmentation and polarization

The behaviour of the original bounded confidence model is defined by the parameter ε [30]. When this is large enough, an initially uniformly random population converges to consensus,

while as ε decreases, clusters emerge in the population. It was shown that the number of major clusters can be approximated by $\lfloor \frac{1}{2\varepsilon} \rfloor$. This approximation ignores minor clusters that may emerge in some situations [34].

In the following we analyse fragmentation by studying the number of clusters obtained for our extended model, starting from an uniform initial distribution of opinions, for a population of size $N = 500$. We concentrate on the area in the (ε, γ) space where, in the original model, the number of clusters is smaller or equal to 2, i.e. $\varepsilon \geq 0.2$. In order to quantify the number of clusters, given the existence of major and minor clusters, we use the *cluster participation ratio* as a criterion. This takes into account not only the number of clusters, but also the fraction of the population in each, measuring thus the *effective number of clusters*. Hence two perfectly equal clusters will result in a cluster participation ratio of 2, one single cluster will yield a value of 1, while if one cluster is larger than the other, the measure will take a value in $(1, 2)$. In general, for n clusters, the maximum value of the participation ratio is n and is achieved when all clusters have the same size, while the minimum can be close to 1, if one cluster contains most of the population and a very small fraction is distributed among the other $n - 1$. The effective number of clusters measured in this section is thus computed as:

$$C = \frac{(\sum_i c_i)^2}{\sum_i c_i^2} \quad (2)$$

where c_i is the size of cluster i .

Fig 1 displays the effective number of clusters for various ε and γ values, using averages over multiple runs. We show results for ε between 0.2 and 0.4, since for larger ε the number of clusters was always 1. Note that in these simulations, the population converges to either one, two or three major clusters of similar mass, plus some minor negligible clusters. For the same parameter setting, the population may converge to one cluster in some simulations, or to two clusters in others. We present the average values as obtained for 100 independent runs.

The plot shows that our simulations reproduce perfectly the behaviour of the original model ($\gamma = 0$), where a transition between 2 and 1 cluster takes place for $\varepsilon \in [0.25, 0.3]$. It is clear that the introduction of $\gamma > 0$ causes an increase in the effective number of clusters, compared to the original model. For instance, for $\varepsilon = 0.35$, the original model results in one cluster, while for our model, new clusters start to emerge for $\gamma \geq 1.3$. For $\varepsilon = 0.32$, the transition starts even earlier before $\gamma = 1$, with the average number of clusters close to 2 at $\gamma = 1.6$. For the case of $\varepsilon = 0.2$, when the original number of clusters was already 2, γ increases C towards 3 opinion clusters. Hence, it appears from our simulations that algorithmic bias causes opinion splitting and fragmentation in the bounded confidence model, by increasing the number of clusters with increasing bias.

We also study polarization of the population due to algorithmic bias, showing in Fig 2 the average pairwise distance between opinions. We observe that the introduction of algorithmic bias does increase the distance between opinions. In many cases this appears together with the increase in the number of clusters, i.e. when moving from 1 to 2 or from 2 to 3 clusters. However, a smaller increase in the distance is also observed when the number of clusters is more stable, i.e. the case $\varepsilon = 0.25$.

While the asymptotic number of opinion clusters is important, the time to obtain these clusters is equally so. In a real setting, available time is finite, and so if consensus forms only after a very long period of time, it may never actually emerge in the real population. Thus, we measure the time needed for convergence (to either one or more opinion clusters) in our extended model. This can be counted as total number of pairwise interactions required to obtain a stable configuration, divided by N , the population size.

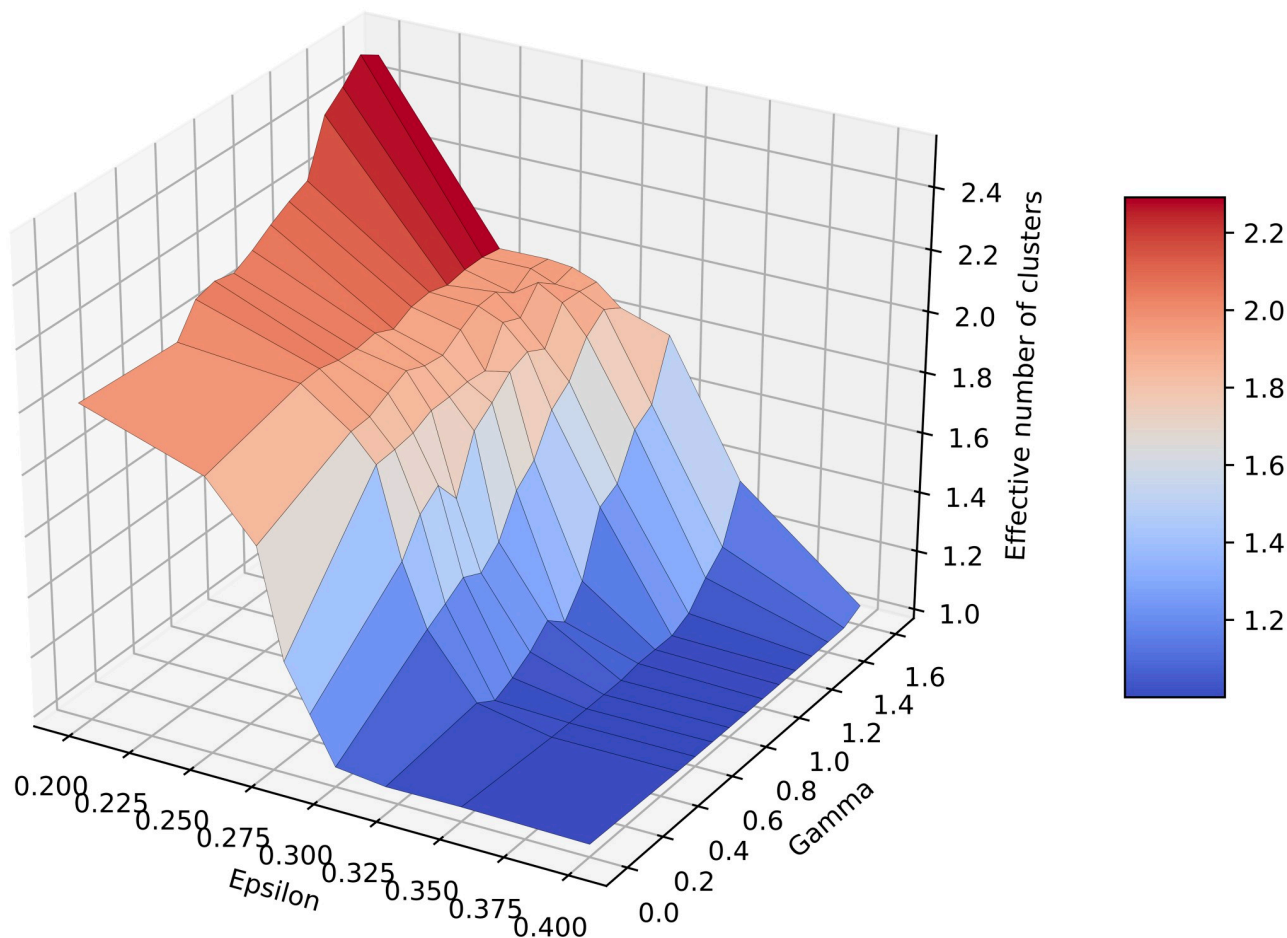


Fig 1. Number of clusters obtained for various ε and γ . For $\varepsilon > 0.4$ the population always converges to 1 cluster, so we omitted the range from the plot, for better visualisation. Values are averaged over 100 runs.

<https://doi.org/10.1371/journal.pone.0213246.g001>

Fig 3 shows how the total number of interactions required depends on both ε and γ . It is clear that the time to convergence grows very fast with γ , for all ε values, including the situation when the population converges to consensus. This indicates that, in a real setting with finite observation time this consensus may not emerge at all. Hence γ has a double fragmenting effect: not only the number of clusters grows, but even in the case of consensus the convergence becomes extremely slow as well, hence leading for reasonable observation times to apparent fragmentation.

Another observation emerging from Fig 3 is that, besides the general fast growth of the time to convergence, we also observe a smaller peak in the convergence time for an ε between 0.25 and 0.35, for all values of γ . This corresponds to a slowing down of convergence around the phase transition (from one to two clusters), which is a physical phenomenon known as *critical slowing down*, observed in many physical systems (e.g. [35, 36]). It relates to the fact that the system is less stable around the phase transition, taking thus longer to reach an equilibrium.

One may argue, however, that measuring the time as the total number of interactions may inflate the figures. Each interaction can have 3 outcomes: (1) nothing happens because a pair of individuals with identical opinions ($x_i = x_j$) were selected, (2) nothing happens because of

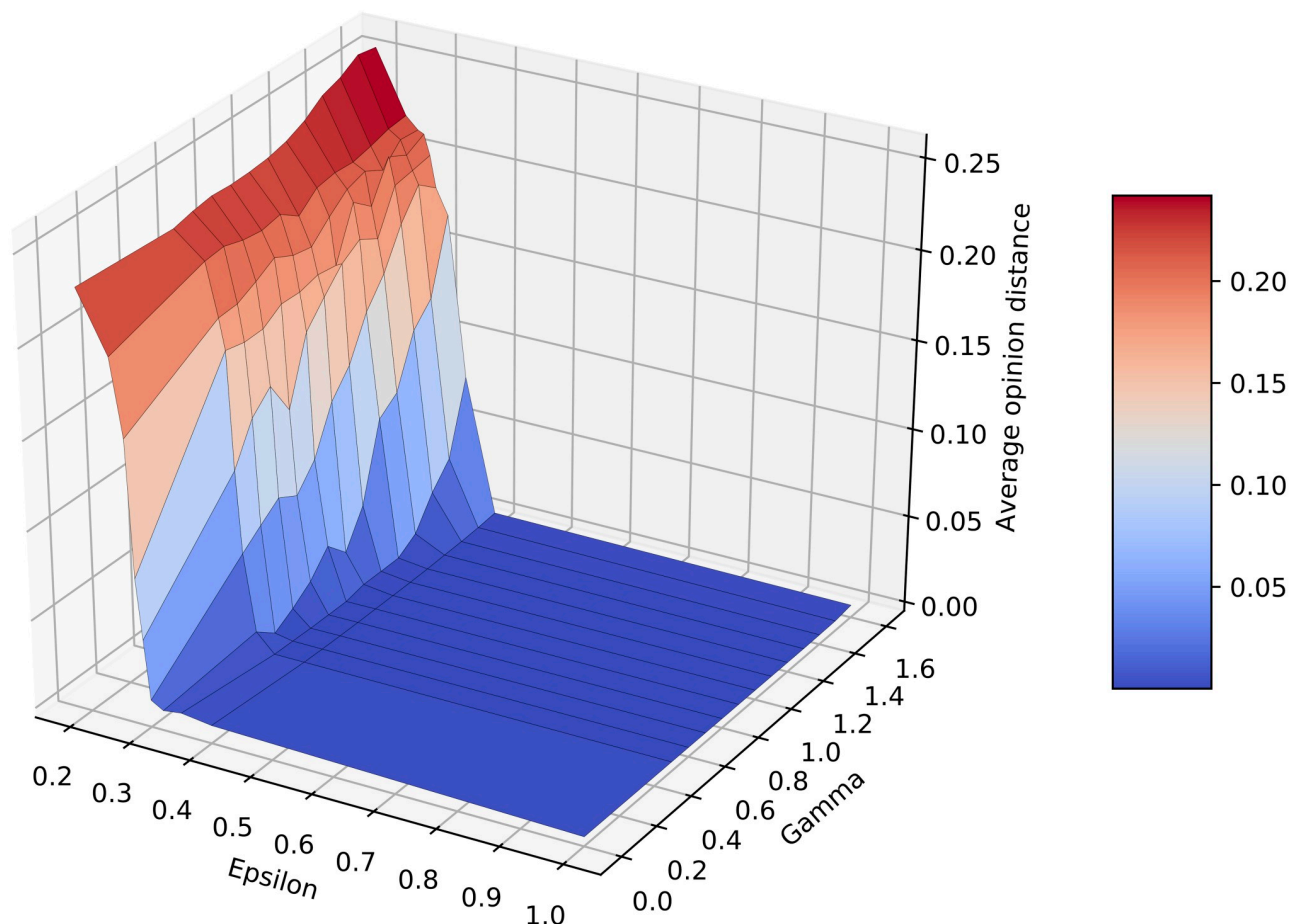


Fig 2. Mean opinion distance obtained for various ε and γ . Values are averaged over 100 runs.

<https://doi.org/10.1371/journal.pone.0213246.g002>

bounded confidence $d_{ij} > \varepsilon$ or (3) the two opinions actually change. In the following we denominate the third type of interaction as ‘active’ interactions.

Fig 4 details the total and active number of interactions for the example case of $\varepsilon = 0.4$. Both measures grow like $\exp(\gamma^\sigma)$ where $\sigma \sim 3.4$, with a small difference visible between active and total interactions. This extremely fast growth of the convergence time means that, in practice, consensus is hindered even by weak algorithmic bias, since consensus is slow to form, hence the population stays in a disordered state for a long time.

This observed slowdown can be explained by considering the time evolution of the sum of pairwise distances between individuals in the population: $SD = \sum_{i < j} d_{ij}$, i.e. a termination function. At each interaction within the bounded confidence limit, SD decreases, until it reaches a minimum corresponding to the steady state. The decrease in SD when individuals i and j interact is directly proportional to d_{ij} . A larger algorithmic bias means on average smaller d_{ij} at each interaction, hence reaching the minimum requires a larger number of iterations.

To prove that SD decreases at each interaction, consider the pair of individuals i and j with $0 < d_{ij} \leq \varepsilon$. After the interaction, the distance among all pairs of individuals different from i and j remains unchanged. What changes is (1) d_{ij} , (2) the distances between i and all other individuals and (3) the distances between j and all other individuals. For (1), after interaction

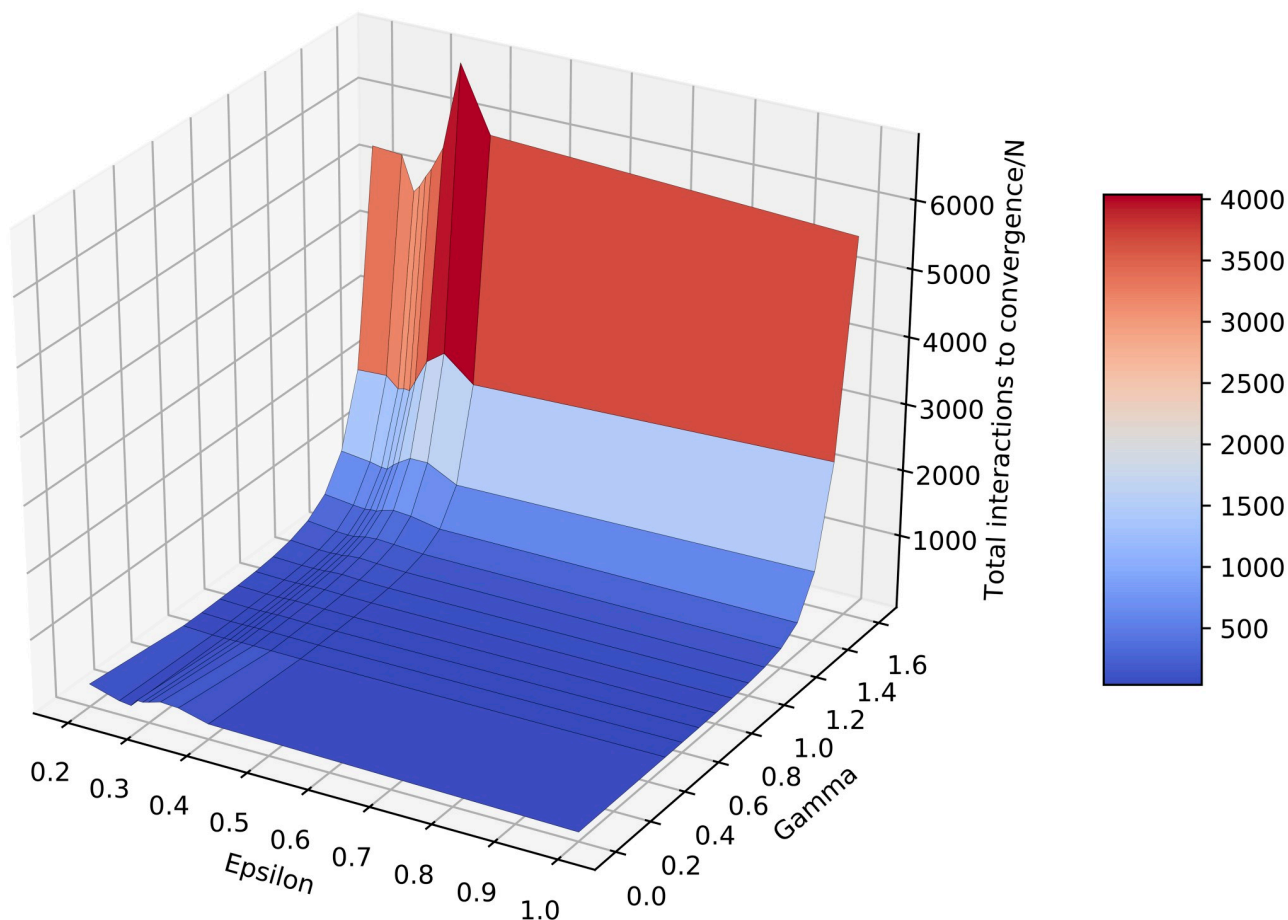


Fig 3. Time to convergence. Total number of interactions required for convergence normalized by the number of individuals, averaged over 100 runs.

<https://doi.org/10.1371/journal.pone.0213246.g003>

$d_{ij} = 0$, hence it decreases. For (2) and (3), consider k an individual different from i and j . If $k < i$ and $k < j$ or if $k > i$ and $k > j$, $d_{ik}(t+1) + d_{jk}(t+1) = d_{ik}(t) + d_{jk}(t)$, because one of i and j will move away from k but the other will move closer by the same amount. If $i < k < j$ (and similarly if $j < k < i$), $d_{ik}(t+1) + d_{jk}(t+1) < d_{ik}(t) + d_{jk}(t)$, because $d_{ik}(t) + d_{jk}(t) = d_{ij}(t)$, while $d_{ik}(t+1) + d_{jk}(t+1) = d_{ij}(t) - 2\min(d_{jk}, d_{ik})(t)$.

To support the points above, we show in Fig 5 the evolution of the population for various values of γ , when $\varepsilon \in 1, 0.32, 0.2$. For $\varepsilon = 1$, the population always converges to one cluster, however we can notice the increase in the number of iterations required, with many clusters coexisting for a long period of time before convergence. When $\varepsilon = 0.32$, the original Deffuant model results in one cluster, with fast convergence. As γ grows, convergence slows down first, and when γ reaches a certain threshold two clusters emerge, resulting also in an increase in the distance between opinions. The case where $\gamma = 1.1$ is close to the transition. It is particularly interesting, because we can notice that initially two clusters coexist for a while, but they eventually merge into one cluster. In other simulation instances with the same parameter values, the two clusters never merge. The last example shows the case $\varepsilon = 0.2$ when the original model results in 2 clusters. By including algorithmic bias the clusters take longer to form, until a third cluster emerges—increased fragmentation—, again causing also an increase in the differences between opinions—increased polarization.

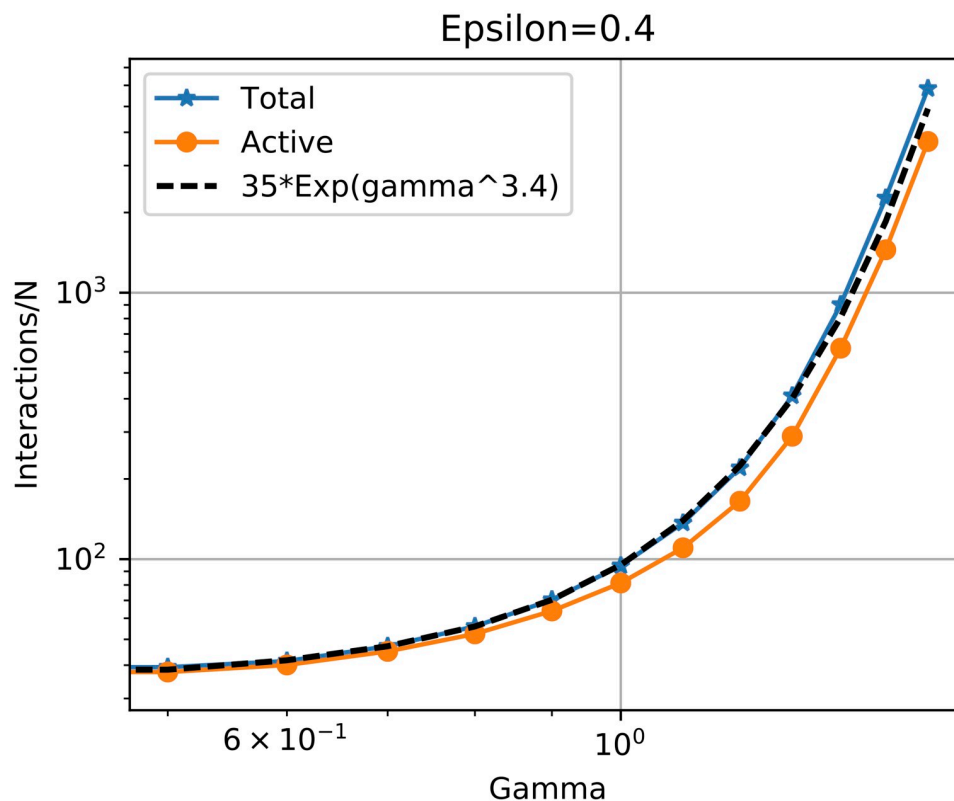


Fig 4. Time to convergence. Normalized total, non-null difference and active number of interactions required for convergence for $\epsilon = 0.4$. The reference $35\exp(\gamma^{3.4})$ is shown as a visual aid only.

<https://doi.org/10.1371/journal.pone.0213246.g004>

Finite size effects

A third analysis that we performed aimed at understanding whether the size of the population plays a role in the effect of the algorithmic bias. Again, this is important for realistic scenarios, since opinion formation may happen both at small and at large scale. Hence we look at the transition between consensus and two opinion clusters for variable population sizes, both for the original and for the extended model.

In the original model, the transition between c and $c + 1$ clusters is shown to be continuous, with an interval for ϵ where both cases can appear in different simulation instances (see Fig 4 in [30]). In particular, the transition between one and two clusters happens for $\epsilon \in (0.25, 0.3)$. We performed numerical simulations to test whether this interval changes with N , given that, to the authors' knowledge, a detailed study in this direction does not exist for the bounded confidence model. Fig 6 shows the mean effective number of clusters over multiple runs obtained with $N \in \{100, 250, 500, 750, 1000\}$. Error bars represent one standard deviation from the mean. It is clear that, as N increases, the transition becomes more abrupt, i.e. the transition interval decreases in length, and gets closer to $\epsilon = 0.25$. For very small populations, in particular, the number of clusters in the transition area is larger, probably due to a lower density of opinions in the starting population, which facilitates formation of clusters even when larger populations would converge to consensus. Hence we can conclude that a small N can also favour opinion splitting. The error bars are almost invisible outside the phase transition interval, but quite large inside this interval. This is due to the fact that within the transition

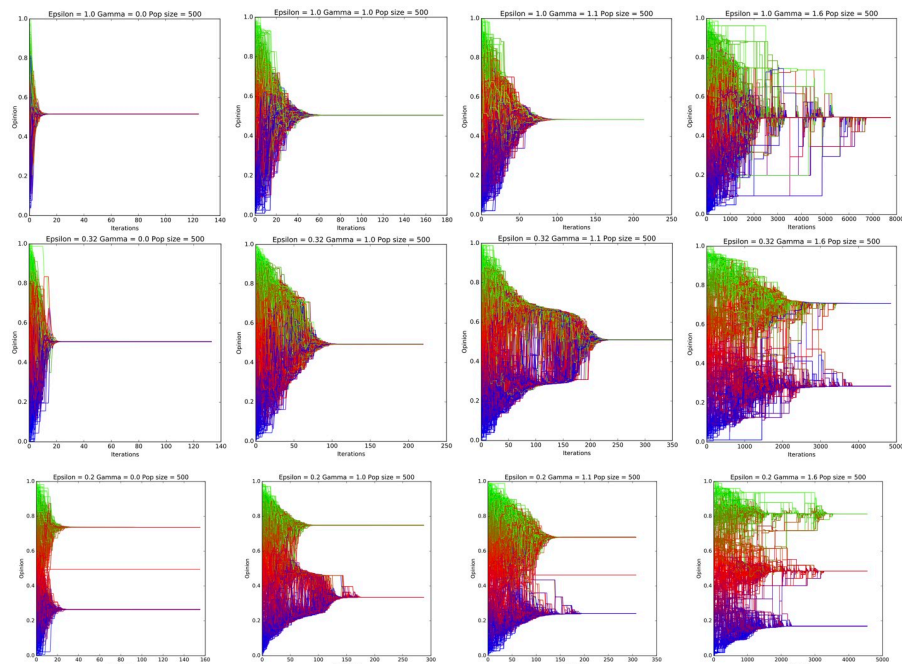


Fig 5. Evolution of the population of opinions for various γ and ϵ values. The first row corresponds to the case where $\epsilon = 1$, the second row corresponds to $\epsilon = 0.32$ while the last row corresponds to $\epsilon = 0.2$. In all cases $\gamma \in \{0, 1, 1.1, 1.6\}$ (left to right).

<https://doi.org/10.1371/journal.pone.0213246.g005>

interval some simulations converge to one cluster, while other to two clusters, obtaining thus large standard deviations from the mean.

To analyse the effect of the population size when $\gamma > 0$, we consider two different ϵ values (0.3 and 0.32). These were chosen because for $\gamma = 0$ they yield one cluster, while opinion splitting emerges as γ grows. Fig 7 shows the effective number of clusters obtained for various population sizes. Again, the transition from one to two clusters is more steep in γ as N increases, with small population sizes favouring opinion splitting. In these conditions, it seems that algorithmic bias can actually be more efficient in hindering consensus for smaller groups. Fig 8 shows average opinion distance for $\epsilon \in \{0.3, 0.32\}$. We observe again that for the same γ smaller populations display higher polarisation levels and larger fluctuations from one simulation to another.

Effect of the initial condition

Previous results were obtained for the case where numerical simulations assumed uniformly random initial opinions in the population. However, in reality, opinion formation may start from slightly fragmented initial conditions. To simulate this, we introduced artificially a symmetric gap around the opinion value 0.5, to simulate a population where opinions are already forming. The width of the gap was varied to understand the effect of various initial fragmentation levels both for the original bounded confidence model ($\gamma = 0$) and for our extension.

Fig 9 shows the mean effective number of clusters obtained for various gap sizes, with error bars showing one standard deviation from the mean. For the original model, the gap shifts the transition from two to one cluster towards larger values of ϵ . Hence a fragmented initial condition favors fragmentation even later during the evolution of opinions. However, as ϵ grows, fragmentation disappears, hence a higher tolerance level can overcome a fragmented initial

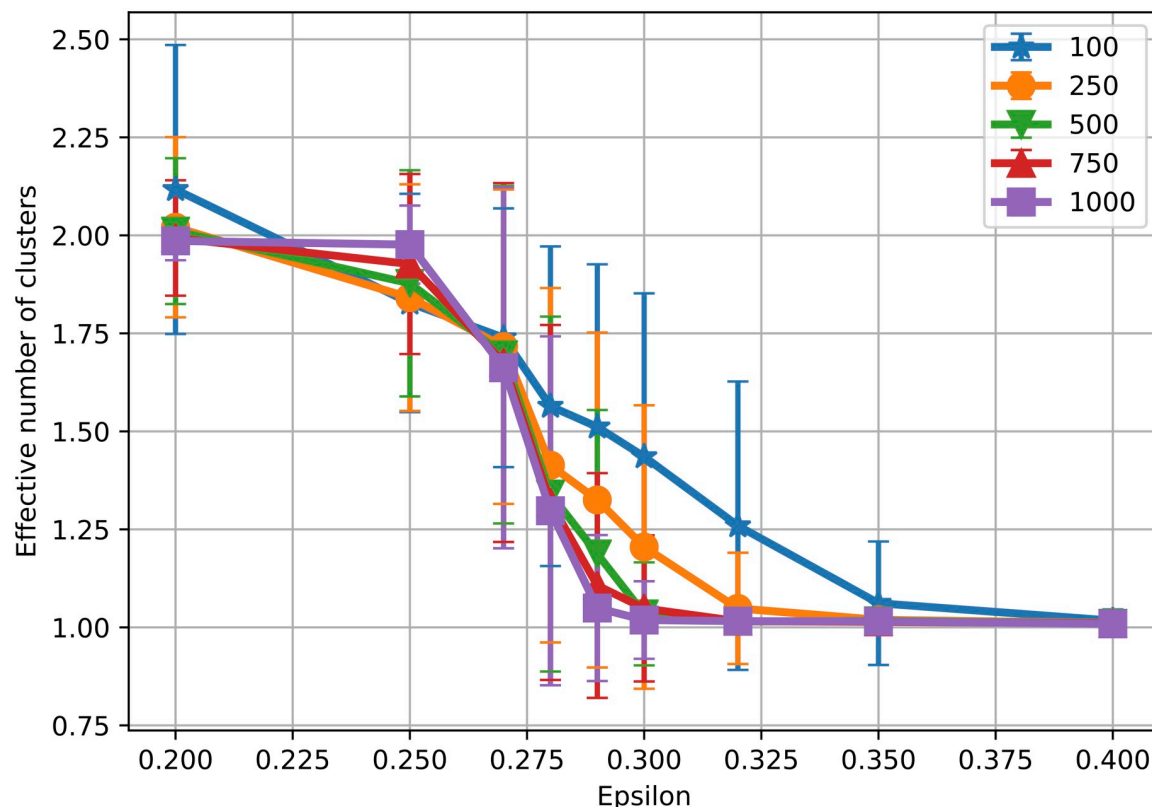


Fig 6. Finite size effects: Number of clusters without algorithmic bias. Effective number of clusters obtained for various ϵ and N , when $\gamma = 0$. Values are averaged over 200, 150, 100, 100 and 100 runs for $N \in \{100, 250, 500, 750, 1000\}$, respectively. Error bars show one standard deviation from the mean.

<https://doi.org/10.1371/journal.pone.0213246.g006>

condition. We also note that as γ grows, the two effects from algorithmic bias and initial fragmentation add up to push the transition to consensus ever closer to $\epsilon = 1$. Error bars show again how in the transition interval the population converges to one cluster in some simulations and to two clusters in others, resulting in relatively large deviations from the mean. However, outside the transition interval error bars are almost invisible, hence the number of clusters is stable from one simulation to another.

Similar behaviour can be observed when analysing the average opinion distance, displayed in Fig 10. Polarisation in the final population is larger when the initial population is fragmented, which can be seen not only in the shift in the transition point towards larger ϵ values, but also in the absolute value of the mean opinion distance when polarisation is observed.

In terms of time for convergence, Fig 11 plots the number of *active interactions* for the case of a fragmented initial condition. It appears that initial fragmentation speeds up convergence when the final population is also fragmented. However, when the final population reaches consensus (one cluster), the effect is reversed, i.e. initial fragmentation slows down convergence. This is due to the fact that the population needs to make up for the initial gap in opinions in order to achieve consensus, increasing thus consensus time. Referring back to the *SD* function introduced previously, the initial fragmentation of the population means a larger value of *SD* at the beginning of the opinion evolution. This means additional iterations are required to reach the value *SD* = 0, corresponding to one cluster. The time to convergence continues to grow very fast with γ in all situations, as seen previously for a uniform initial condition.

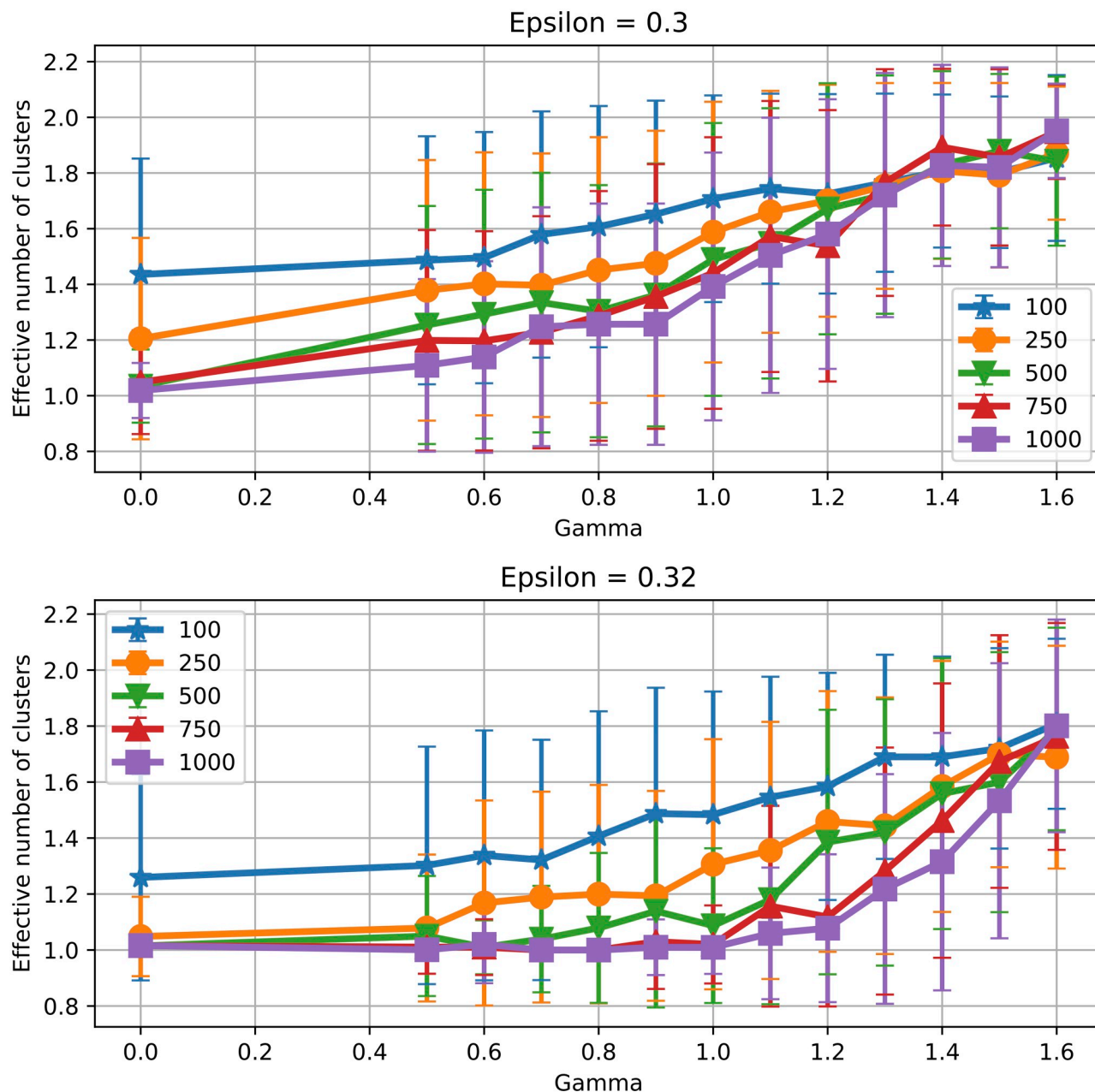


Fig 7. Finite size effects: Number of clusters with algorithmic bias. Effective number of clusters obtained for various ϵ , γ and N . Values are averaged over 200, 150, 100, 100 and 100 runs for $N \in \{100, 250, 500, 750, 1000\}$, respectively. Error bars show one standard deviation from the mean.

<https://doi.org/10.1371/journal.pone.0213246.g007>

Hence, again, the two effects appear to work together against reaching consensus, either by favoring the appearance of additional clusters or by slowing down consensus when this could, in principle, emerge.

Discussion and conclusions

A model of algorithmic bias in the framework of bounded confidence was presented, and its behavior analyzed. Algorithmic bias is a mechanism that encourages interaction among like-minded individuals, similar to patterns observed in real social network data. We found that,

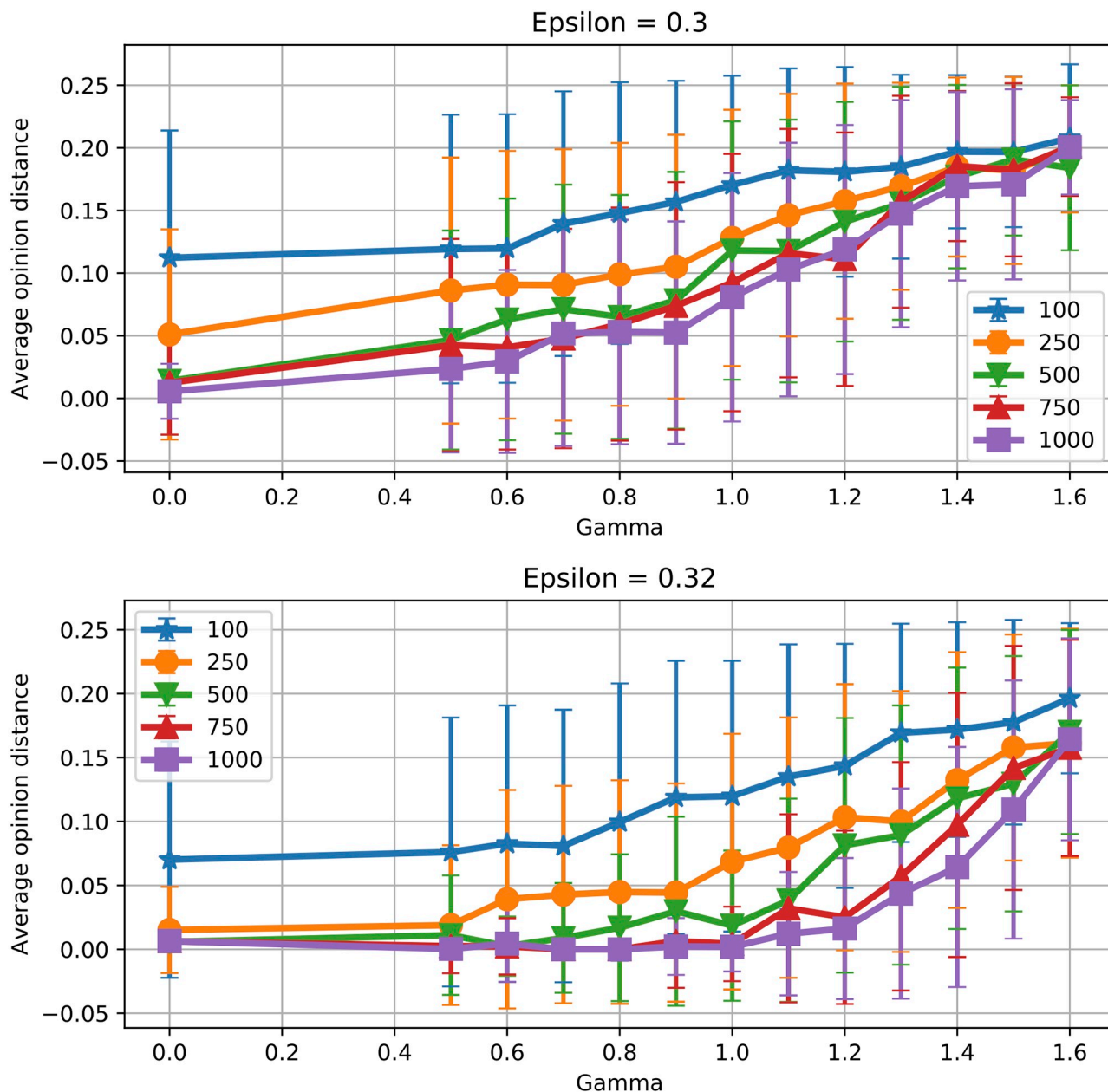


Fig 8. Finite size effects: Opinion distance with algorithmic bias. Mean opinion distance obtained for various ϵ , γ and N . Values are averaged over 200, 150, 100, 100 and 100 runs for $N \in \{100, 250, 500, 750, 1000\}$, respectively. Error bars show one standard deviation from the mean.

<https://doi.org/10.1371/journal.pone.0213246.g008>

for this model, algorithmic bias hinders consensus and favors opinion fragmentation and polarization through different mechanisms. On one hand, consensus is hindered by a very strong slowdown of convergence, so that even when one cluster is asymptotically obtained, the time to reach it is so long that in practice consensus will never appear. Additionally, we observed fragmentation of the population as the bias grows stronger, with the number of clusters obtained increasing compared to the original model. At the same time, the average opinion distance also grew in the population, indicating emergence of polarization. A fragmented initial condition also enhances the fragmentation and polarization, augmenting the effect of

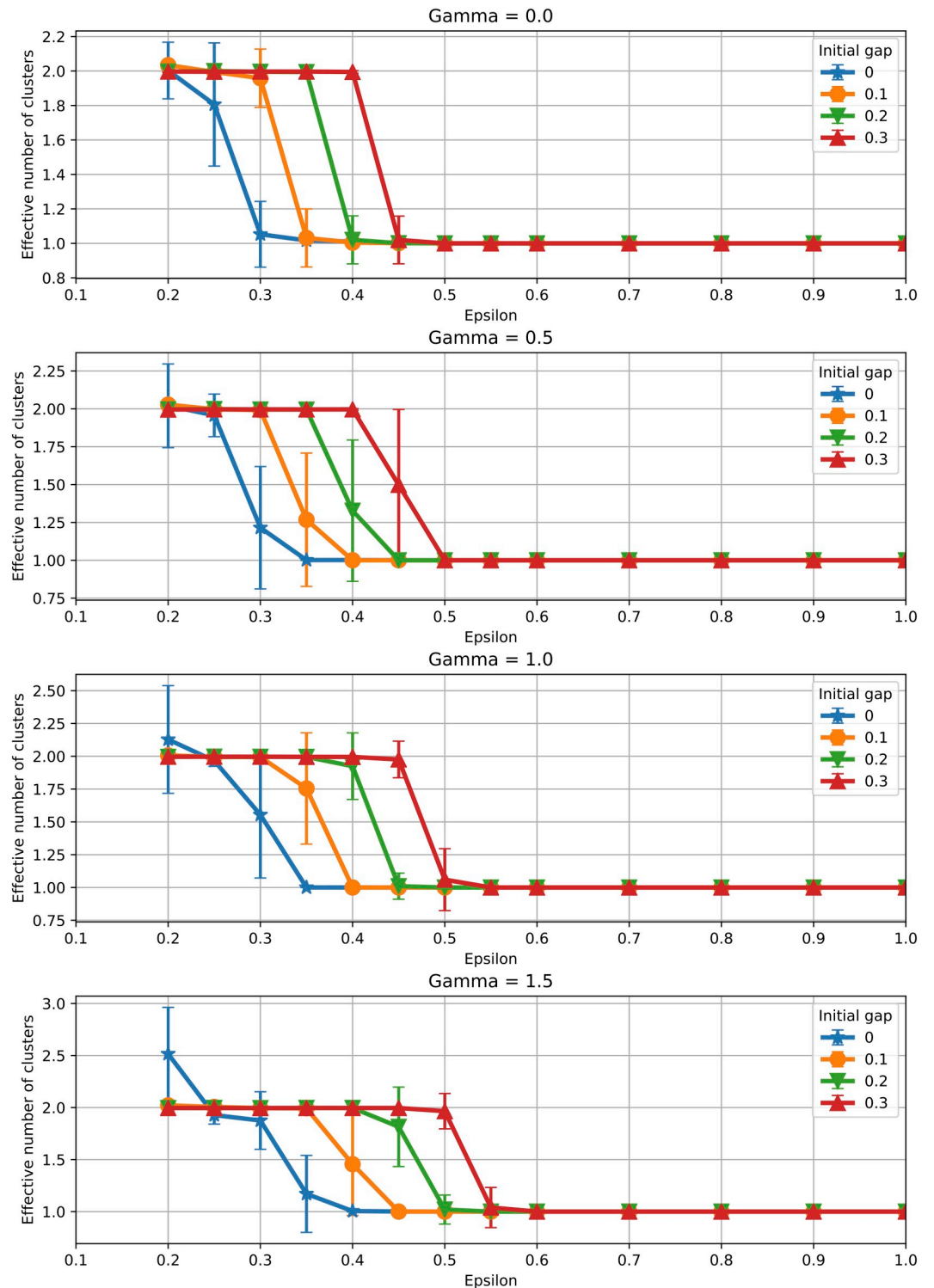


Fig 9. Initial condition: Number of clusters. Effect of the initial condition on the effective number of clusters (averages over 100 runs).

<https://doi.org/10.1371/journal.pone.0213246.g009>

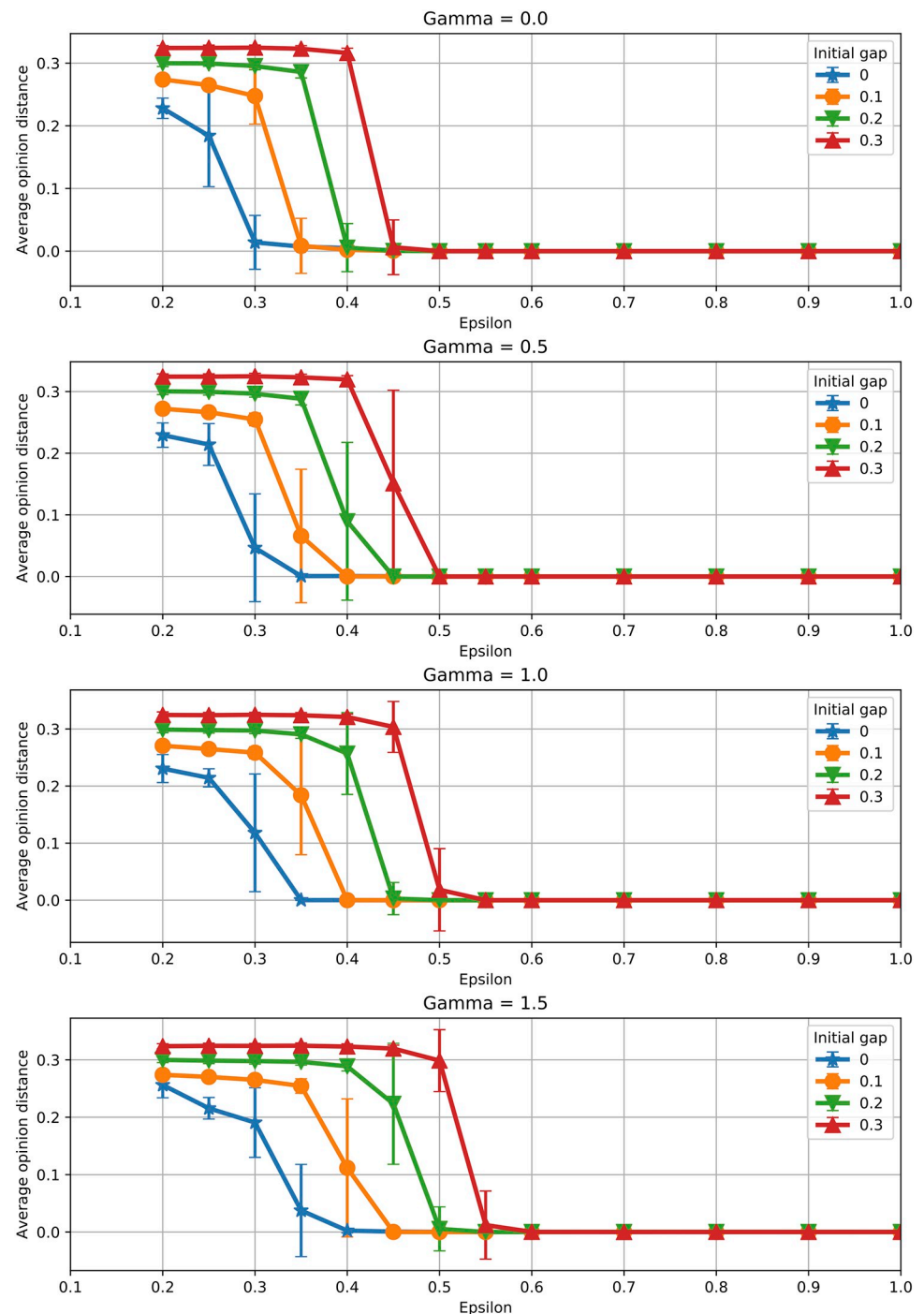


Fig 10. Initial condition: Opinion distance. Effect of the initial condition on the average opinion distance (averages over 100 runs).

<https://doi.org/10.1371/journal.pone.0213246.g010>

the algorithmic bias. Additionally, we observed that small populations may be less resilient to fragmentation and polarization, due to finite size effects.

The results presented here are based on the mean field bounded confidence model, and may be influenced by this choice. A first assumption is that bounded confidence exists, i.e.

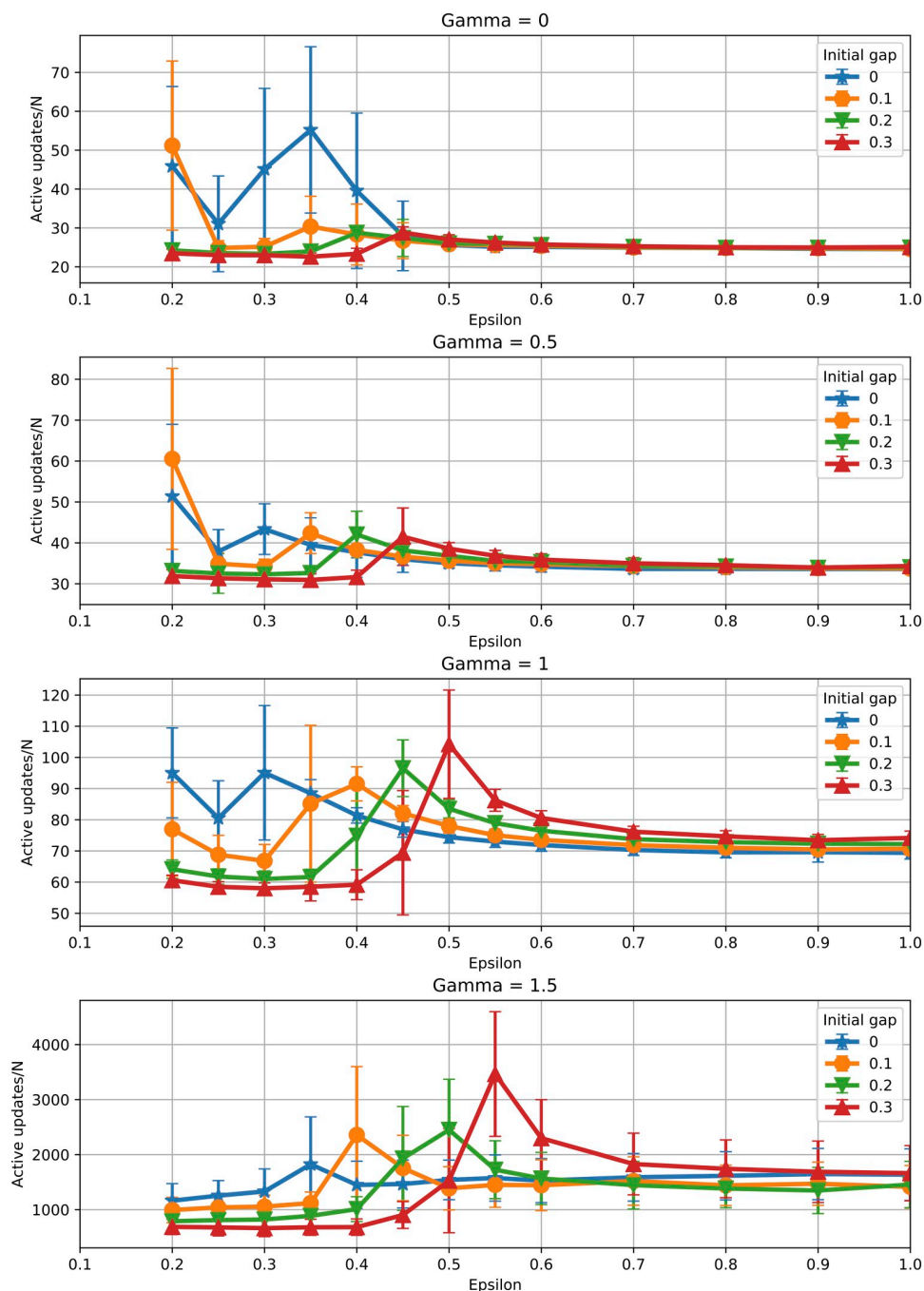


Fig 11. Initial condition: Time to convergence. Effect of the initial condition on the convergence time measured in number of active interactions (averages over 100 runs).

<https://doi.org/10.1371/journal.pone.0213246.g011>

individuals with very distant opinions do not exchange information hence do not influence each other. However, our conclusions regarding the fact that algorithmic bias hinders consensus still stand even when bounded confidence is removed from this model (i.e. $\varepsilon = 1$). In this case, consensus still becomes extremely slow as the bias increases, hence is never achieved in practice, a result that we believe will apply to many other models with attractive dynamics.

Secondly, adding noise to the model could change results, since noise in the bounded confidence model can facilitate consensus [37]. Adaptive noise, on the other hand, could generate metastable clusters [38]. A different issue is the structure of the opinion clusters. In our model clusters form and then they become static, with no further changes possible (neither increase or decrease in opinion distances). However in reality polarization may increase or decrease in time also after clusters form. Adding noise could also be an answer for this issue, however we intend to study alternative mechanisms to embed this type of situation in the model.

Third, it would be interesting to see how taking into account a more realistic social network structure among individuals, instead of a complete graph where anybody may interact with anybody else, would impact the opinion formation process, possibly exacerbating the effects observed in this study. The Deffuant model is known to be affected by other topologies such as scale free networks when the networks do not contain a large number of edges [39–41]: an increase in the number of clusters is visible when the dilution is strong. However as networks become more rich in connections the model returns practically to its original behaviour [39]. We have tested our model on Barabasi-Albert networks [42] and the same effect can be seen: the model behaviour does not change when the network is well connected (number of edges added with each node was 5 in our experiments), while for very diluted networks the number of clusters grows (when each new node comes with one edge only). However, we believe that in order to provide a comprehensive view of the model behaviour on networks, more advanced network models should be studied. We will include evolving social networks based on homophily [43–45], and the possible interaction between algorithmic bias and homophily parameters. Additionally, we are interested in studying weighted networks with community structure [46] and their multilayer version [47], as well as real networks such as an Hungarian online social network [48].

A fourth assumption of our model is in the dynamics of the interaction. Negative interactions can also be important in the dynamics [49–51], and these could have an effect on algorithmic bias. Promoting interaction between similar individuals in such a model could diminish the effect of disagreement, which could enhance consensus. Furthermore, external information could affect the behavior observed, especially when the sources of information are also selected based on a algorithmic bias. Another aspect to be investigated is whether extremists have a louder voice in the population, and whether this would impact in any way cluster formation. Further analysis will be pursued in the future to include all these aspects.

A concept related to algorithmic bias is homophily, i.e. the tendency of people to build friendships with similar others, which is visible in the way the social network is built. This is one factor that could facilitate interactions with like-minded individuals, similar to our model. A recent study shows that homophily enhances consensus in the Deffuant model [43], as opposed to our results. Instead, it seems that external media can have an important effect in opinion polarization and fragmentation. The difference in results among the two studies could be due to the fact that in a strongly homophilic society diverse individuals never have a chance to interact, while in our model this can happen, albeit with a small probability. This has a similar polarizing effect as external information in [43]. A different approach is to take into account homophily to rewire an adaptive social network [45]. Here homophily seemed to make consensus more difficult to achieve, similar to our method. However, for small bounded confidence thresholds, the number of clusters was decreased by the rewiring. As mentioned previously, we plan to analyse our model on adaptive networks as well in future work.

Although there is evidence that many types of social interactions are subject to algorithmic bias, the debate still continues on whether this generates or not opinion polarization in the long term. Our numerical results support the first option, which we plan to analyse in more detail in the future by applying our model to real data from social network processes. Recent

work on how to counteract opinion polarization on social networks has also appeared [52], and initial results suggest that facilitating interaction among chosen individuals in polarized communities can alleviate the issue. We will also investigate this with our model.

Acknowledgments

We thank Guillaume Deffuant for useful discussions on finite size effects in the bounded confidence model. We also thank the IT Center of the University of Pisa (Centro Interdipartimentale di Servizi e Ricerca) for providing access to computing resources for simulations.

Author Contributions

Conceptualization: Alina Sîrbu, Dino Pedreschi, Fosca Giannotti, János Kertész.

Formal analysis: Alina Sîrbu, Dino Pedreschi, Fosca Giannotti, János Kertész.

Funding acquisition: Dino Pedreschi, Fosca Giannotti, János Kertész.

Investigation: Alina Sîrbu, János Kertész.

Methodology: Alina Sîrbu, Dino Pedreschi, Fosca Giannotti, János Kertész.

Software: Alina Sîrbu.

Supervision: Dino Pedreschi, Fosca Giannotti, János Kertész.

Validation: Alina Sîrbu.

Writing – original draft: Alina Sîrbu, Dino Pedreschi, Fosca Giannotti, János Kertész.

Writing – review & editing: Alina Sîrbu, Dino Pedreschi, Fosca Giannotti, János Kertész.

References

1. Pew Research Center. Political Polarization in the American Public. June 12. 2014;.
2. Groskopf C. European Politics Is More Polarized than Ever, and These Numbers Prove It. Quartz com. 2016;.
3. Andris C, Lee D, Hamilton MJ, Martino M, Gunning CE, Selden JA. The rise of partisanship and super-cooperators in the US House of Representatives. PloS one. 2015; 10(4):e0123507. <https://doi.org/10.1371/journal.pone.0123507> PMID: 25897956
4. Del Vicario M, Gaito S, Quattrocioni W, Zignani M, Zollo F. News consumption during the Italian referendum: A cross-platform analysis on facebook and twitter. In: 2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA). IEEE; 2017. p. 648–657.
5. Campbell JE. Polarized: Making sense of a divided America. Princeton University Press; 2018.
6. Prior M. Post-broadcast democracy: How media choice increases inequality in political involvement and polarizes elections. Cambridge University Press; 2007.
7. Baum MA, Groeling T. New media and the polarization of American political discourse. Political Communication. 2008; 25(4):345–365. <https://doi.org/10.1080/10584600802426965>
8. Graber DA, Dunaway J. Mass media and American politics. Sage; 2012, 9th edition.
9. Messing S, Westwood SJ. Selective exposure in the age of social media: Endorsements trump partisan source affiliation when selecting news online. Communication Research. 2014; 41(8):1042–1063. <https://doi.org/10.1177/0093650212466406>
10. Charles A. Media and/or Democracy. In: Charles A, editor. Media/Democracy: A Comparative Study. Cambridge Scholars Publishing; 2013.
11. Ignatieff M. Political Polarization in the American Public. Annual Colloquium on Fundamental Rights, Brussels. 2016;.
12. Purcell K, Rainie L, Mitchell A, Rosenstiel T, Olmstead K. Understanding the participatory news consumer. Pew Internet and American Life Project. 2010; 1:19–21.

13. Hermida A, Fletcher F, Korell D, Logan D. Share, like, recommend: Decoding the social media news consumer. *Journalism studies*. 2012; 13(5-6):815–824. <https://doi.org/10.1080/1461670X.2012.664430>
14. Villi M, Matikainen J, Khaldarova I. Recommend, tweet, share: User-distributed content (udc) and the convergence of news media and social networks. In: *Media Convergence Handbook-Vol. 1*. Springer; 2016. p. 289–306.
15. Pariser E. *The filter bubble: What the Internet is hiding from you*. Penguin UK; 2011.
16. Bakshy E, Messing S, Adamic LA. Exposure to ideologically diverse news and opinion on Facebook. *Science*. 2015; 348(6239):1130–1132. <https://doi.org/10.1126/science.aaa1160> PMID: 25953820
17. Colleoni E, Rozza A, Arvidsson A. Echo chamber or public sphere? Predicting political orientation and measuring political homophily in Twitter using big data. *Journal of communication*. 2014; 64(2):317–332. <https://doi.org/10.1111/jcom.12084>
18. Del Vicario M, Bessi A, Zollo F, Petroni F, Scala A, Caldarelli G, et al. The spreading of misinformation online. *Proceedings of the National Academy of Sciences*. 2016; 113(3):554–559. <https://doi.org/10.1073/pnas.1517441113>
19. Garimella K, De Francisci Morales G, Gionis A, Mathioudakis M. Political discourse on social media: Echo chambers, gatekeepers, and the price of bipartisanship. In: *Proceedings of the 2018 World Wide Web Conference on World Wide Web*. International World Wide Web Conferences Steering Committee; 2018. p. 913–922.
20. Schmidt AL, Zollo F, Del Vicario M, Bessi A, Scala A, Caldarelli G, et al. Anatomy of news consumption on Facebook. *Proceedings of the National Academy of Sciences*. 2017; 114(12):3035–3039. <https://doi.org/10.1073/pnas.1617052114>
21. Schmidt AL, Zollo F, Scala A, Quattrociocchi W. Polarization Rank: A Study on European News Consumption on Facebook. *arXiv preprint arXiv:180508030*. 2018;.
22. Stoica AA, Riederer C, Chaintreau A. Algorithmic Glass Ceiling in Social Networks: The effects of social recommendations on network diversity. In: *Proceedings of the 2018 World Wide Web Conference on World Wide Web*. International World Wide Web Conferences Steering Committee; 2018. p. 923–932.
23. Mäs M, Bischofberger L. Will the personalization of online social networks foster opinion polarization? Available at SSRN 2553436. 2015;.
24. Webster JG. Beneath the veneer of fragmentation: Television audience polarization in a multichannel world. *Journal of Communication*. 2005; 55(2):366–382. <https://doi.org/10.1111/j.1460-2466.2005.tb02677.x>
25. Tewksbury D, Rittenberg J. *News on the Internet: Information and Citizenship in the 21st Century*. Oxford University Press; 2012.
26. Bodrunova SS. Fragmentation and Polarization of the Public Sphere in the 2000s: Evidence from Italy and Russia. *Global Media Journal-German Edition*. 2013; 3(1).
27. Bramson A, Grim P, Singer DJ, Fisher S, Berger W, Sack G, et al. Disambiguation of social polarization concepts and measures. *The Journal of Mathematical Sociology*. 2016; 40(2):80–111. <https://doi.org/10.1080/0022250X.2016.1147443>
28. Pastor-Satorras R, Castellano C, Van Mieghem P, Vespignani A. Epidemic processes in complex networks. *Reviews of modern physics*. 2015; 87(3):925. <https://doi.org/10.1103/RevModPhys.87.925>
29. Sirbu A, Loreto V, Servedio VD, Tria F. Opinion dynamics: models, extensions and external effects. In: *Participatory sensing, opinions and collective awareness*. Springer; 2017. p. 363–401.
30. Deffuant G, Neau D, Amblard F, Weisbuch G. Mixing beliefs among interacting agents. *Advances in Complex Systems*. 2000; 3(4):87–98. <https://doi.org/10.1142/S0219525900000078>
31. Hegselmann R, Krause U, et al. Opinion dynamics and bounded confidence models, analysis, and simulation. *Journal of artificial societies and social simulation*. 2002; 5(3).
32. Rossetti G, Milli L, Rinzivillo S, Sirbu A, Pedreschi D, Giannotti F. NDlib: a python library to model and analyze diffusion processes over complex networks. *International Journal of Data Science and Analytics*. 2018; 5(1):61–79. <https://doi.org/10.1007/s41060-017-0086-6>
33. Rossetti G, Milli L, Rinzivillo S, Sirbu A, Pedreschi D, Giannotti F. Ndlb: Studying network diffusion dynamics. In: *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE; 2017. p. 155–164.
34. Lorenz J. Continuous opinion dynamics under bounded confidence: A survey. *International Journal of Modern Physics C*. 2007; 18(12):1819–1838. <https://doi.org/10.1142/S0129183107011789>
35. Zhang S, Qi S, Klushin LI, Skvortsov AM, Yan D, Schmid F. Anomalous critical slowdown at a first order phase transition in single polymer chains. *The Journal of chemical physics*. 2017; 147(6):064902. <https://doi.org/10.1063/1.4997435> PMID: 28810755

36. Engin D, Orlov S, Segev M, Valley GC, Yariv A. Order-disorder phase transition and critical slowing down in photorefractive self-oscillators. *Physical review letters*. 1995; 74(10):1743. <https://doi.org/10.1103/PhysRevLett.74.1743> PMID: 10057746
37. Carro A, Toral R, San Miguel M. The role of noise and initial conditions in the asymptotic solution of a bounded confidence, continuous-opinion model. *Journal of Statistical Physics*. 2013; 151(1-2):131–149. <https://doi.org/10.1007/s10955-012-0635-2>
38. Mäs M, Flache A, Helbing D. Individualization as driving force of clustering phenomena in humans. *PLoS computational biology*. 2010; 6(10):e1000959. <https://doi.org/10.1371/journal.pcbi.1000959> PMID: 20975937
39. Weisbuch G. Bounded confidence and social networks. *The European Physical Journal B*. 2004; 38(2):339–343. <https://doi.org/10.1140/epjb/e2004-00126-9>
40. Fortunato S. Universality of the Threshold for Complete Consensus for the Opinion Dynamics of Deffuant et al. *International Journal of Modern Physics C*. 2004; 15(09):1301–1307. <https://doi.org/10.1142/S0129183104006728>
41. Stauffer D, Meyer-Ortmanns H. Simulation of consensus model of Deffuant et al. on a Barabasi–Albert network. *International Journal of Modern Physics C*. 2004; 15(02):241–246. <https://doi.org/10.1142/S0129183104005644>
42. Barabási AL, Albert R. Emergence of scaling in random networks. *science*. 1999; 286(5439):509–512. PMID: 10521342
43. Gargiulo F, Gandica Y. The role of homophily in the emergence of opinion controversies. *Journal of Artificial Societies and Social Simulation*. 2017; 20(3):1–8. <https://doi.org/10.18564/jasss.3448>
44. Murase Y, Jo HH, Török J, Kertész J, Kaski K. Structural transition in social networks: The role of homophily. *arXiv preprint arXiv:180805035*. 2018;.
45. Kozma B, Barrat A. Consensus formation on adaptive networks. *Physical Review E*. 2008; 77(1):016102. <https://doi.org/10.1103/PhysRevE.77.016102>
46. Kumpula JM, Onnela JP, Saramäki J, Kaski K, Kertész J. Emergence of communities in weighted networks. *Physical review letters*. 2007; 99(22):228701. <https://doi.org/10.1103/PhysRevLett.99.228701> PMID: 18233339
47. Murase Y, Török J, Jo HH, Kaski K, Kertész J. Multilayer weighted social network model. *Physical Review E*. 2014; 90(5):052810. <https://doi.org/10.1103/PhysRevE.90.052810>
48. Lengyel B, Varga A, Ságvári B, Jakobi Á, Kertész J. Geographies of an online social network. *PloS one*. 2015; 10(9):e0137248. <https://doi.org/10.1371/journal.pone.0137248> PMID: 26359668
49. Macy MW, Kitts JA, Flache A, Benard S. Polarization in dynamic networks: A Hopfield model of emergent structure. *Dynamic social network modeling and analysis*. 2003; p. 162–173.
50. Jager W, Amblard F. Uniformity, bipolarization and pluriformity captured as generic stylized behavior with an agent-based simulation model of attitude change. *Computational & Mathematical Organization Theory*. 2005; 10(4):295–303. <https://doi.org/10.1007/s10588-005-6282-2>
51. Sîrbu A, Loreto V, Servedio VD, Tria F. Opinion dynamics with disagreement and modulated information. *Journal of Statistical Physics*. 2013; 151(1-2):218–237. <https://doi.org/10.1007/s10955-013-0724-x>
52. Garimella K, De Francisci Morales G, Gionis A, Mathioudakis M. Reducing controversy by connecting opposing views. In: *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*. ACM; 2017. p. 81–90.