# Polarization on social media: when group dynamics leads to societal divides

Agnieszka Rychwalska
University of Warsaw
a.rychwalska@uw.edu.pl

Magdalena Roszczyńska-Kurasińska
University of Warsaw
magda.roszczynska@gmail.com

## Abstract

*Polarization of group opinions – a natural mechanism that enables groups to stay intrinsically cohesive – explains why after multiple interactions individual and group opinions shift towards the extremes. Recently, significant polarization of opinions can be witnessed in the public discourse of many Western societies in a range of topics. We argue here that the prevalence of social media together with its specific design may amplify natural group dynamics and strengthen the divisions. We present an agent based model wherein implementation of polarization mechanisms together with social media properties leads to increased segregation and radicalization of opinions. We propose certain design choices for social media platforms that could help ameliorate the problem.*

## 1. Introduction

Recent years have seen major divisions within western societies. Both in the US and Europe public discourse is full of conflicting issues on which constructive dialogue is increasingly more difficult. While such topics as abortion or homosexual marriages have always been divisive, these days a range of seemingly unproblematic issues raise heated disputes (e.g. membership in the EU, vaccinations, health insurance regulations, etc.) More and more often, positions are taken at the extremes of the possible breadth of an opinion. While there are numerous political, economic and societal factors that might have led to the increase of polarization of public opinions, there is one more characteristics of modern societies that may contribute to the problem.

Polarization and radicalization of opinions are nothing new. Back in the 1970-80s seminal work in social psychology has shown that interactions within social groups cause both the group and its constituent individuals to dig in into their positions and – in many circumstances – to shift their opinions towards the extremes [17]. Yet, in real life, the mechanisms that cause group polarization are counterbalanced by processes that lead groups to work in concert [22]. What is different now is that an increasing volume of social interactions take place in social media rather than in face to face contacts.

Drawing from the research in social psychology, we argue here that current social media design distorts group dynamics in such a way that the balance between polarizing and unifying forces is unsettled.

To support our claims we present an agent based model in which we implement some of the basic mechanisms of polarization (action commitment, social desirability, reactance) and one critical property of social media networks – the possibility to rewire social connections in a purposeful manner. We show that under those conditions polarization is amplified and happens in a wider range of situations. Interestingly enough, rewiring – while being generally detrimental – proves to be advantageous in select conditions: it becomes the last resort in situations that would have otherwise led to radicalization. While the presented model is certainly not exhaustive and other social mechanisms and technological specifics are at play in the currently observed processes of polarization, it shows that a prevalent design choice is sufficient to produce increased opinion divides in social systems. We conclude the paper by pointing out certain design solutions that could help ameliorate the problem of unchecked polarization on social media.

## 2. Polarization

Polarization of opinions has first been defined as a ”risky shift” of decisions after a group discussion [20]. Numerous studies have shown that this phenomenon is ubiquitous: in a variety of situations and choices a discussion among group members leads them to shift their opinion toward a more extreme option than the mean individual choice [10]. An early review by Mayers and Lamm [14] outlined three mechanisms underlying the attitude change that leads to opinion polarization: social motivation, action commitment and cognitive foundation. Social motivation refers to the desire to be perceived favorably by relevant others. Socially motivated individuals first test what is the general opinion of others and then shift towards the extreme.

URI: http://hdl.handle.net/10125/50150
ISBN: 978-0-9981331-1-9

Presenting a more extreme opinion creates the image of high self-esteem and therefore is more socially desired.

Once the opinion is formed and expressed on the group forum the second element of attitude change activates – action commitment mechanism. The verbalization of opinion leads to its enhancement, making it less vulnerable to change in future. Interestingly, even the duration of time spend on thinking on an issue might lead to polarization of opinion [21]. The last component – cognitive foundation – relates to the cognitive processing of arguments, cognitive rehearsal and acknowledging the novel information shared between the members [22].

Two branches of research on opinion polarization, one concentrating on the social mechanism – social comparison theory [19] – and the second concentrating on the cognitive aspects of dispute – persuasive arguments theory [22] – both report similar effects of polarization. In more recent studies, segregation and clustering of opinions have been often found in social media. For example, political blogs link to other blogs of the same political ideology [1, 9]. Individuals prefer to read content from authors who present similar political views [11]. Analysis of political disputes on Twitter demonstrates that networks of retweets are clearly segregated into two clusters corresponding to the political left and right [2].

While this polarization seems to reflect a natural group mechanism, it is worth noting that social media design may potentially amplify this process. First, the pervasiveness of information on one's social circle increases the awareness of the opinions of relevant others [8]. This might lead to a shift in the individual's own opinion in pursue of social acceptance, as was described in the social motivation mechanism of polarization.

Further, the prevalent design choice of social media is to broadcast one's content – including opinionated statements – to all social connections at once. This voicing of one's attitudes constitutes both cognitive rehearsal as well as action commitment – not only are the views displayed on the forum of the acquaintance group but they may also be strengthened by any positive feedback, such as likes, follows, retweets, etc.

Finally, social media enables unprecedented exchange of information and opinions on a daily basis, allowing even distant individuals to influence one another – even if this influence is passive (e.g. being exposed to content posted by unknown others). This intensification of information exchange increases the probability of encountering the opposite opinion in its extremes – sometimes even in malicious attempts to spite (e.g. trolling). Exposure to such overstated views may trigger another mechanism that alongside polarization can lead to extremization of one's opinions: *reactance* – an emotional response to the reduction of perceived freedom of choice [13]. In the context of opinion formation reactance would manifest as strengthening of the opposite opinion to the one imposed, i.e., radicalization of opinions.

## 3. Methodology

To investigate the possible effects of social media design on polarization dynamics we have chosen the agent based modelling approach. This methodology is well suited for exploration of macro-level – systemic – effects of individual behavior. While agent based models rarely allow for quantitative prediction, they offer a unique opportunity to test possible qualitative effects of interventions and solutions that are difficult or impossible to implement experimentally [3, 4, 5].

In an agent based model the researcher specifies the behavior of elements and the rules that govern their interactions and observes the behavior of the whole system by describing it with a few aggregate variables (sometimes called order parameters). Agent based models are thus constructed in a bottom-up way – the assumptions concern the individual (micro) level rules of behavior and the hypotheses are tested at the level of the system (macro).

Agent based models are used to analyze various complex systems, but one of the most fruitful and rich areas of their application has been the modelling of social systems. In particular, opinion dynamics models – such as the one described in this paper – have shown, for example, how the natural drive to follow majority choice results in the final distribution of voting [12, 26] or how social influence of opinion leaders allows minority opinion to survive in the sea of majority [16].

While many early opinion dynamics models relied fully on physics based mechanics [12], recent models often draw from social psychological knowledge to inform the construction of the agents representing individuals and their interactions [5]. The challenge in such modelling attempts is to choose the most appropriate psychological variables for the agents' characteristics and most fitting social contexts for them to interact. On the one hand, the model needs to realistically depict social processes but on the other hand inclusion of too many variables might render the model intractable and impossible to interpret.

As an example, in the Weisbuch-Deffuant bounded confidence model [24], which served as the starting point for the model presented here, the authors set out to investigate how peoples' tolerance for different views affects the process of social influence and the resulting opinion variation. They observed that individuals usually are not impacted by opinions that are very

different from their own. Therefore if they meet a person that presents a vastly different view, they will ignore his or her beliefs and instead will seek influence or advice from someone else, whose opinion is within the limits of the individual's tolerance for differences.

In this model the tolerance for differences is the variable under scrutiny and all other possible design choices are kept as simple as possible. Thus, agents have the theoretical possibility to interact with any other agent in the system, provided that their opinions are within the tolerance limits. The opinions are continuous – ranging from 0 to 1 – and the agents do not have any other properties that could differentiate them. At each simulation step, a pair of agents is drawn randomly (i.e. they "meet" to discuss opinions) and if their views are within the tolerance (which is identical for every agent) they shift their opinion slightly towards that of their interaction partner.

The resulting opinions in the social system of this design depend on the value of tolerance. If the tolerance is high (>0.3), all agents converge on one opinion (the middle point, ~0.5); when it is low, a few opinion clusters form. The lower the tolerance, the more clusters appear and this relation scales as 1 over 2 times the tolerance value.

The basic Weisbuch-Deffuant model is the simplest opinion dynamics model that produces clear divisions in the final opinion distribution fully accountable to a single parameter and therefore is well suited for studying polarization.

## 4. The model

To test how social media design might impact polarization process we have designed an agent based model wherein the agents influence each other in a way similar to the Weisbuch-Deffuant bounded confidence model but also behave according to Myers and Lamm's polarization mechanisms conceptual scheme [14]. Therefore, the agents have an opinion that they broadcast to others and are influenced by opinions presented by their social relations – provided that the opinions are within their tolerance for difference. If this is the case, the agents shift their opinions towards that of their social links. However, tolerance for difference is not a static trait of the agents; rather, it changes as the agent's confidence in her opinion changes: the more confident the agent is, the lower her tolerance for different opinions.

Drawing from the Myers & Lamm's concepts, we let the agents be motivated by social comparison – that is, they seek to be perceived favorably by others. Therefore, their confidence is affected by being in majority within their social circle. If the agent is in majority, her confidence grows, if she is in minority, it drops.

Moreover, confidence also changes when the agent is being "heard" by others, following Myers & Lamm's proposal that verbalization of one's opinion affects the actor's attitude. If an agent influences another one whose opinion is within her tolerance for difference, her confidence rises.

The agents in the model are linked into a network of social relations. Since we interpret the system as operating on social media, we allow the agents to cut off their links and create new connections – that is, to rewire their social network. Therefore the agents can, from time to time, severe a relation that is far beyond their tolerance level and instead connect to another person.

It is worth stressing that to keep the model tractable it was intentionally rendered non exhaustive – both with respect to psychological mechanisms of social influence as well as the technological design solutions. Investigating the effects of e.g. tie strength, power structures or algorithmic filtering would require separate modelling studies.

### 4.1 Implementation – the basic model

The modeled social system is composed of $N = 300$ individuals, connected by undirected links into a locally clustered network similar to a small world network (SWN) with an average node degree $d = 20$. The size and connectivity of the network were chosen to simulate a medium sized organization or a large acquaintance clique in which a certain number of meaningful interactions (transmitting sufficient amount of information or opinions) per individual can take place in a day. To a certain degree it is possible to emulate a system of larger size in this model by manipulating the speed constant and probability constants (as introduced later) but the model was not intended to simulate a truly big system (e.g. the whole Facebook network). Polarization has been defined in small group studies and therefore to simulate much bigger systems a careful choice of other mechanisms might be necessary.

The agents interact with their link neighbors. Specifically, each agent in the network is characterized by an opinion (ranging from 0 to 100 and drawn randomly from a flat distribution at the beginning of simulation) and her tolerance for differing opinions that defines the range of opinions that can influence her (i.e. form a range around her opinion +/- tolerance). The model was also run using other opinion distributions at the start (normal and bimodal, results not presented), but the flat distribution resembles a discussion on a topic on which the individuals do not yet have an opinion. Therefore, it is more apt for the study of polarization of opinions than normal distribution (possibly,

characterizing an already established issue, on which people do not differ much in opinions) or bimodal distribution (an established issue that has already polarized public opinion).

In each simulation step, each agent *i* randomly draws one of her social connections *j* and is influenced by his or her opinions (if it is within her tolerance range) – she shifts her opinion *Oi* toward the opinion *Oj* of the selected agent by a fraction of the difference of their opinions:

$$O_{t+1}^{i} = O_t^i + 2s * (O_t^j - O_t^i) \quad \text{Eq. 1}$$

where *s* is the speed constant that affects the volatility of the influence process, set at 0.1 for all simulations.

The process of social influence here is asymmetrical, resembling the Weisbuch-Deffuant model as implemented on scale-free networks [23]. That is, only the agent *i* changes her opinion in this mechanisms, while agent *j* – the source of influence – stays with her previous opinion. This is important for our intended comparisons between real life contacts and interaction on social media. On social media, a user broadcasts information for all to see (friends or public, depending on the settings). Therefore, in the act of "communicating out" the opinions of the user herself are not affected by those of her readers – it is only the recipients that can be affected. Feedback (such as likes, follows, etc.) affects the opinions of the broadcaster in an indirect way, as described later.

The asymmetry of interactions, as has been noted [23], also helps to relate the power law distribution of connections into an asymmetrical influence process – people that are heavily connected (hubs) will have a greater chance of influencing others but will not be themselves more prone to being influenced, which reflects the direction of opinion spread in real social networks [15, 18].

The tolerance value can be interpreted as the characteristics of the topic being discussed in the network. Important issues, e.g. related to the value system, would cause smaller tolerance (i.e. lower acceptance of differing opinions) and trivial issues would mean bigger tolerance (translating into a wider range of opinions that can influence an individual). As has been described in the previous section, in a fully connected network introducing tolerance ranges into the process of influence produces varied numbers of peaks in the final distribution of opinions, depending on the value of tolerance [24]. In a scale-free network with low connectivity (~4 average links per node) the number of peaks and the dispersion of opinions around them are different, but for more connected structures, they resemble the fully connected case [23]. Therefore, we also expected similar effects for a SWN with 20 connections per node on average.

We have used such defined model as the null model that provides a baseline for comparisons.

## 4.2. Introducing confidence dynamics

In the next step we have modified the influence dynamics to add to it the mechanisms of polarizing group processes [14]. To that end, we have added another agent characteristics, confidence *C*, ranging from 0 to 0.99 and drawn randomly at the beginning of the simulations from a normal distribution with mean 0.5 and standard deviation of 0.1. This parameter describes the self-assurance of an individual in her opinion. The polarization mechanisms of social desirability and action commitment (as well as cognitive rehearsal, not modeled), can be thought of as depending on changing the confidence of opinions within the group. This internal trait then can be used as a modifier for the social influence process.

First, since polarization studies show that strengthening of one's confidence depends on assessing what is the social norm in one's group (social desirability mechanism), we assume that confidence of agents depends on their being in majority with regards to their held opinion. Each agent *i* assess how many of her neighbors (*countSO*) voice an opinion that is within the limits of her tolerance – i.e. have similar opinions – and computes what fraction of her total contacts (*countN*) that group constitutes. She changes her confidence *Ci* depending on how much of a majority she is in (her confidence increases if she is in majority, and it decreases if she is in minority):

$$C_{t+1}^{i} = C_t^i + {}^1\!/_3\, s * (\frac{countSO}{counN} - 0.5) \quad \text{Eq. 2}$$

The speed constant is divided by 3 to keep confidence dynamics slower than opinion dynamics.

Second, we let the agents grow confident when they transmit influential opinions, resembling the effects of publicly committing to an opinion by verbalizing it in group discussion, as described by the action commitment mechanism. Each agent *j* that has been used as a source of influence (i.e. was drawn by another agent in that agent's process of social influence), increases her confidence proportionally to the similarity of their opinions (the more similar the opinions are, the more confidence the broadcaster gains; if the follower's opinion is outside the agent's tolerance, her confidence drops):

$$C_{t+1}^{j} = C_t^j + \frac{s*(T*\left(1-C_t^j\right)-\left|O_t^i - O_t^j\right|)}{t*\left(1-C_t^j\right)} \quad \text{Eq. 3}$$

where *T* is tolerance. The speed constant for this confidence dynamics mechanism is bigger than for minority / majority related changes to counteract the fact

that being chosen as influence source is much rarer than assessment of majority opinion.

This behavior in face to face contacts is easy to imagine – a person who is listened to and whose opinions are repeated among group members would grow more confident. In social media this is also visible, albeit differently. All signs of positive responses for one's broadcast content (e.g. likes, follows, shares, reposts, comments, etc.) increase positive emotions and are (sometimes desperately) sought after, and can well predict whether the author will publish again or not.

Finally, to implement the effects of confidence on social influence dynamics, we modify an agent's tolerance range *TR* by multiplying it by the inverse of her confidence. That is, high confidence results in narrower tolerance limits and low confidence broadens those limits. A person who is unsure whether his or her opinion is right, would be more prone to seek influence from individuals with even very different mindset. In contrast, a person who is very sure of his or her correctness will likely shut out those who broadcast different opinions, even if that difference is not really big.

$$TR_t^i = \{O_t^i - t * \left(1 - C_t^i\right), O_t^i + t * \left(1 - C_t^i\right)\} \quad \text{Eq. 4}$$

### 4.2. Reactance implementation

In our next modification we have implemented the process of reactance, i.e. changing one's opinion in the direction contrary to the influence of others when such influence threatens one's freedom of choice [13]. In face to face communication, reactance is visible in rejecting persistent attempts at influence. For example, a person seeing an obtrusive ad for one brand of soda on a vending machine, might choose a different drink – even against her preference – just to maintain a feeling of freedom of choice. The phenomenon of reactance can be traced in social media for example in reactions to trolling attempts (i.e. presenting conflicting and extreme opinions intended to start a dispute or to spite) or when confronted with content of a vastly different viewpoint. A recipient of such content or target of trolling usually gets involved in a heated discussion, voicing opinions that are closer to the extreme just to get the upper hand in the conflict, even though in normal circumstances she would not have voiced them. Since verbalizing opinions is an act of action commitment, reversing opinions to their previous state might not be possible, especially if they receive positive feedback from other involved disputants.

Reactance was implemented in the model as the inverse of social influence. With a small probability (varied from 0 to 0.06) each agent in each simulation step had a chance of "being reached" by content from outside her tolerance range. That is, she randomly drew a link neighbor $k$ with opinion outside her limits (provided she had such neighbors). The agent then adjusts her opinion to move away from the intolerable neighbor, increasing the difference of opinions proportionally to the breadth of the difference:

$$O_{t+1}^i = O_t^i + 2s * (O_t^i - O_t^k) \quad \text{Eq. 5}$$

These two mechanisms of polarization – confidence that limits the tolerance for different opinions and reactance – and their co-occurrence, provided us three models to be tested against the null model: confidence, reactance and both confidence and reactance. To this specification we added models that implement one selected trait of social media - rewiring.

To implement the possibility of purposefully adjusting one's social connections – rewiring – we allow the agents to seek for one social contact per simulation step that is outside the agents tolerance range (if the agent has any such) and whose opinion is the farthest away from the agent's. This connection is severed and instead the agent links to another, randomly selected node to keep the network density stable. While in terms of network connectivity this moves us from an SWN to a more random connection structure (effectively increasing the probability $p$ of weak ties) it is worth noting that the connections are not truly random as they reflect opinion structure.

The model was implemented in the NetLogo agent based modelling platform [25]. The versions of the model described above (the null model and 7 possible combinations of confidence, rewiring and reactance) have been run as separate simulations for 700 time steps and with 50 repetitions for each combination of parameters.

## 5. Results

When comparing the implemented group dynamics mechanisms to the null model we were interested in assessing how segregated are the final opinions, for which tolerance levels the segregation happens and how close to the extremes the opinion peaks are. To assess this, we have computed Shanon's entropy on the final opinion distributions as well as measured the number of modes using a statistics developed by [7]:

$$m = \frac{1}{M}\sum_{i=2}^{n}|x_i - x_{i-1}| \quad \text{Eq. 6}$$

where $M$ is the highest frequency in the histogram (i.e. the maximum value), $n$ is the set of bins in the histogram and $x_i$ is the frequency in the bin $i$.

A value of 2.40 is roughly the threshold for bimodal distributions and a value above 4 describes a distribution with three or more modes.

Additionally, we have inspected opinion histograms from sample simulation runs for increasing values of tolerance and have measured the frequency of extreme

opinions (10% of the lower and upper opinion range) to see the degree of radicalization of opinions.

The null model resembles the properties of a fully connected lattice in Weisbuch-Defuant bounded-confidence model (Fig 1, Fig 2). For values of tolerance above 23% of the opinion range, the opinions of all agents converge on the midpoint; tolerance below 20% threshold produces two peaks of segregated opinions, that grow increasingly distant as tolerance is decreased (Fig 3). Below the tolerance value of 10% of opinion range, the number of peaks grows, and the lower the tolerance, the less distinct they become.

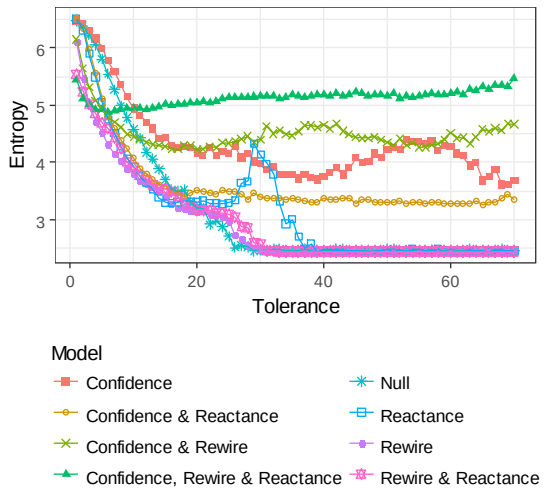


**Figure 1. Entropy values for the different models computed on final distributions of opinions (after 700 simulation steps); each point is an average of 50 simulation runs.**

Adding the possibility to purposefully rewire connections amplifies the segregation process in that the segregation is complete (i.e. all agent's opinions converge on the peaks, with minimal dispersion as reflected in the entropy values, and as can be seen from the single bin peaks in the histograms in Fig 3, right panel) and in that the range of tolerance values that generate divisions is slightly wider. Specifically, the first threshold (that produces two peaks) is at 27% tolerance. In all, rewiring reduces the diversity of opinions and slightly increases polarization for less important issues (i.e. those characterized by higher tolerance for differing opinions).

Introducing the mechanisms that change agents' confidence and therefore make tolerance range dynamical, changes the number of convergence points for agents' opinions. For all the investigated tolerance values, at least two peaks appeared in the final distribution of opinions (m value > 2.5, Fig 2). However, the peaks are generally smaller and there is a lot of dispersion around them (Fig 3, left panel). While they tend to form around the midpoint of the opinion scale (i.e. the agents are not segregated into clearly polarized opinions), there is considerable variation in their size – some opinions are very popular but there are also many opinions that garner some small following. In sum, introducing confidence into the social influence process, while dividing opinions in a wide range of tolerance values, also increases the diversity of opinions and reduces segregation as agents' growing confidence counteracts the force to unify on single opinion peaks.

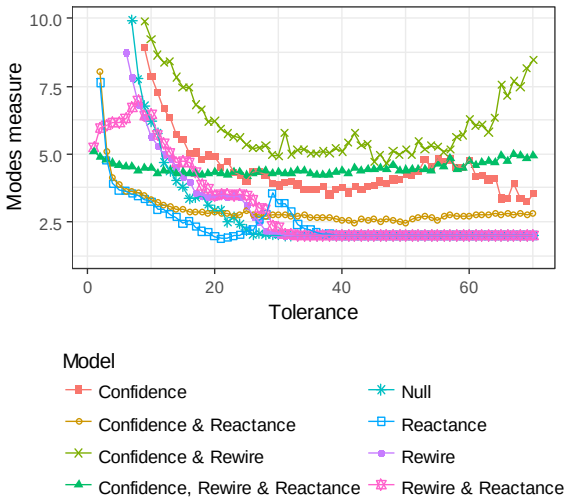


**Figure 2. Modes measure for the different models computed on final distributions of opinions (after 700 simulation steps); each point is an average of 50 simulation runs. The *y* scale has been truncated at 10 to increase readability.**

Implementation of both rewiring and confidence results in typical polarization dynamics. The diversity introduced by agents' confidence all but disappears – even at high tolerance levels (i.e. 50% of the opinion range, Fig 4) clear opinion peaks form and there is little dispersion around them. Rather, all agents converge on two opinion bins that are visibly segregated. This segregation is also visible at very low tolerance – many opinion peaks form, but they are very distinct and separated by gaps in opinion frequencies.

When only confidence mechanisms were at play, low values of tolerance produced divisions in opinions, but they were not so segregated and some midpoint opinions were present (Fig 3). In all, this shows that combining the natural group segregation mechanisms with the possibility to purposefully shape connections amplifies the polarization dynamics in a non-linear way.

Rewiring alone increased segregation and confidence mechanisms introduced segregation into issues whose importance (tolerance) would not normally warrant it. Together, these two mechanisms

produce a social system where opinions are almost perfectly segregated for a very wide range of issues.

The mechanism of reactance, when introduced into the null model by itself strengthens the segregation of opinions at a wider range of tolerance values, i.e. by producing clear two peaks for tolerance between 20% and 30% of opinion range (Fig 2). Moreover, for lower values of tolerance, where in the null model a few more or less distinct peaks form, reactance clearly pushes agents to converge predominantly on two opinions only.

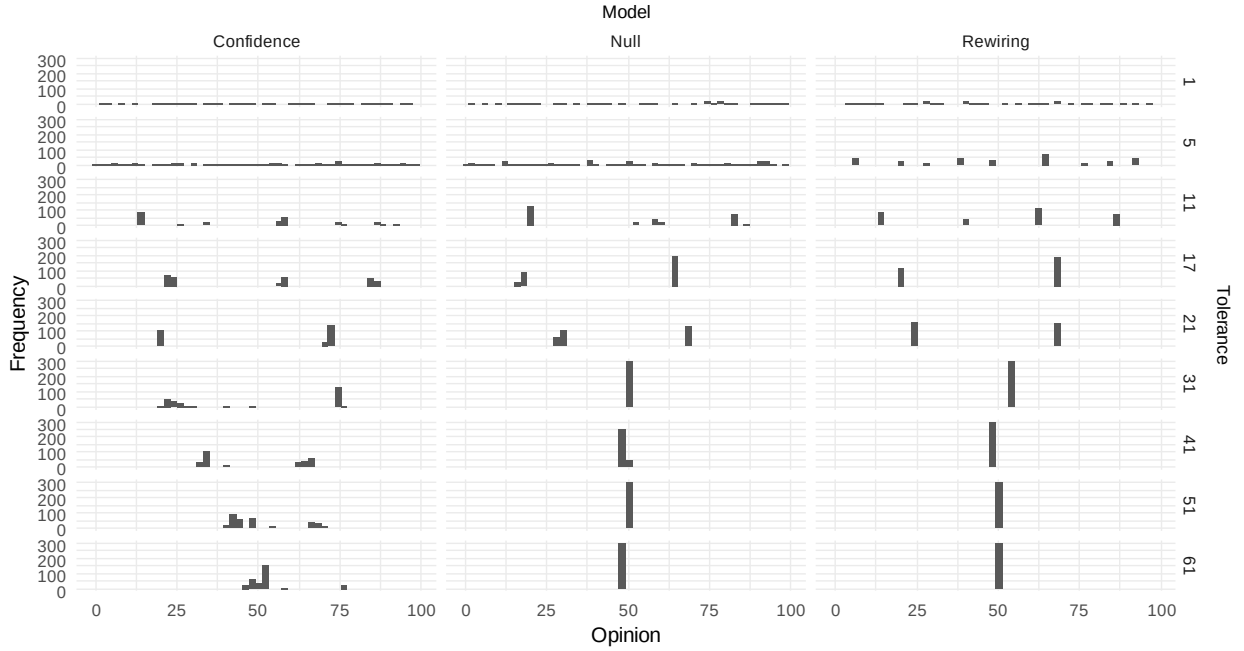


**Figure 3. Histograms of final opinions after 700 simulation steps in sample simulation runs for the null model, the model with confidence and the model with rewiring. For each model 9 simulations were run, for different values of tolerance (labelled on the right) resulting in 9 histograms.**

For really low tolerance levels (<10%) those two peaks are separated by tiny groups of followers of midpoint views, but this dispersion is minimal (Fig 6). Yet, the most visible and striking result is that reactance is the force that pushes the opinion peaks to the extremes. As can be seen from the frequencies for high and low opinion values (Fig 5) most agents converge on the poles of the opinion scale for a wide range of tolerance values (up to around 30% tolerance, above which a single opinion peak is visible).

Adding the confidence mechanisms to the reactance phenomenon produces the most polarized and segregated opinion distribution of all presented here. For all values of tolerance investigated, clear factions form at the end of the opinion spectrum with little to no dispersion around them – there is simply no middle viewpoint present. Even for very low values of tolerance (10% and below), where normally many separate peaks are present, the two poles dominate the view. On the other hand, for higher values of tolerance (> 35%) where reactance alone was not enough to divide opinions, together with the strengthening force of confidence it produces clear divisions. What is more, the opinions are not only perfectly segregated but they also converge at the very extremes of the scale (Fig 6). What happens when the rewiring possibilities of social media are added to the picture?

Interestingly enough, combining all the group mechanisms together with the rewiring possibility of social media gives an opinion distribution that is more diversified than that without the rewiring. A few, dispersed midpoint opinions are present, especially in the lower ranges of tolerance values (Fig 6) and the extreme opinions are not so frequent among agents (Fig 5). It might seem from this that social interactions in physical space, where we cannot "delete" social relations should produce more radicalization and stronger segregation than social media. This puzzling result warrants a closer analysis.

The first thing to notice is that in face to face interactions extreme reactance leading to radicalization of views is rare – much rarer than the 6% probability introduced in the model. Simply, groups and individuals have a wider range of behaviors available that allow them to avoid protracted conflicts. Even if a stubborn party where to pester another group with attempts similar to online trolling this most often would not lead to open fights or intractable disputes resulting in radicalization of opinions, because the target can

employ a variety of strategies to quench the fire before it spreads.

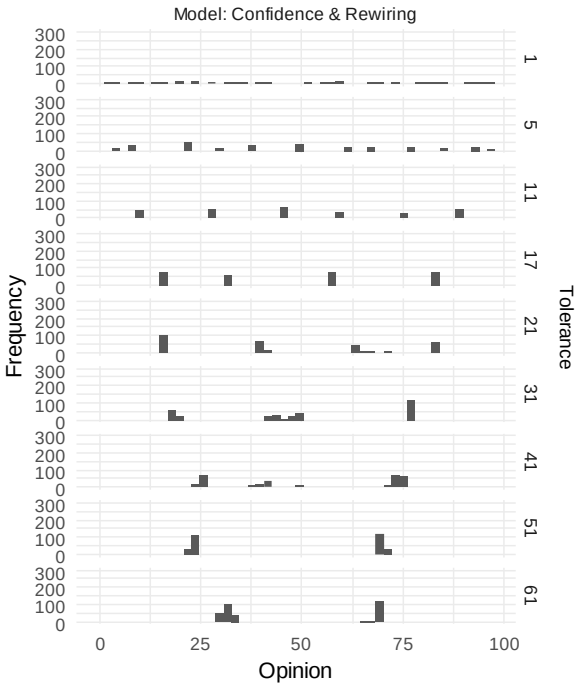


**Figure 4. Histograms of the final opinions after 700 simulation steps in sample simulation runs for the model with both confidence and rewiring. Nine simulations were run, for different values of tolerance (labelled on the right) resulting in 9 histograms.**

For example, in physical contacts people tend to "agree to disagree". If opposing views are identified, it can be quickly established how far the opponent is willing to change his or her opinion. If change is impossible, the interlocutors simply change the topic, avoiding the conflict.

This is especially visible in situations where cutting the contact is impossible – for example among family members or coworkers that naturally need to coexist in the same physical space and often need to collaborate. In those cases, opponents in one issue can find similarities of opinions in another issue to improve the mutual connection that enables cooperation. What is more, very often people are able to highly respect another person for her professional knowledge or skills while knowing that their political or moral viewpoints are irreconcilable. Stressing similarities and diversifying the topical plateaus for interaction is thus the main counterbalance to the polarizing forces such as reactance.

The situation is different in online social life. Trolling is simple and while it can be emotionally draining its costs are far less than what open conflicts in physical space incur. Even without such malicious intents, content of all viewpoints is easily spread and there are high chances of encountering opposing opinions in everyday online functioning.

Moreover, social media design does not allow diversifying social relations along topical domains and therefore the strategy to find similarities in another area is not available. Often, one divisive issue can destroy an otherwise fruitful collaboration. The only solution is to cut the link altogether to reduce the pressing need to react negatively to adverse opinions. It is simply hard to accept the opinions on a trivial matter of a person with whom one is in violent dispute over moral issues. For these reasons, the possibility to rewire (that is: cut) some links in social media is actually reducing polarization and radicalization of conflicting parties.

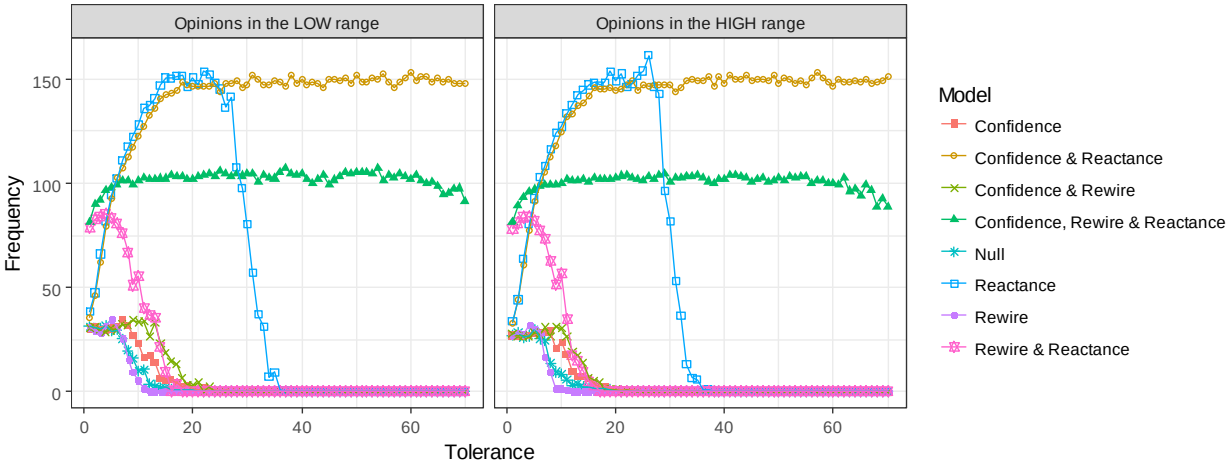


**Figure 5. Frequencies of opinions from the high (10%) and low (10%) ends of the opinion scale at the end of simulations for the various models. Each point is an average of 50 simulation runs.**

This explains the counter intuitive amelioration of polarization in the model resulting from combining all group mechanisms with social media rewirability. The seeming improvement from the "real life" case is due to the fact that the model does not implement other confrontation avoidance techniques that are abundant in physical interaction and are actually more effective than connection severance. Moreover, the model might underestimate the probability of reactance in social media as opinion "wars", trolling attempts and similar actions are probably positively correlated with issue importance (i.e. tolerance). However, this analysis of limitations of the model gives clear clues as to how to improve social media design to counteract the negative phenomena of polarization and radicalization.

## 6. Conclusions

While polarization mechanisms in natural contexts help groups maintain inner cohesiveness, when amplified by social media design they may lead to disproportional divisions. The model presented here shows that dynamics of individuals' confidence which is a result of the polarization mechanisms of social desirability and action commitment causes even non-important issues to become divisive (i.e. issues for which individuals have wider tolerance for different opinions). On the other hand, the possibility to purposefully rewire links, which is a prevalent design choice in social media platforms, increases the degree of segregation of opinions. Together, these two mechanisms lead to more complete segregation of opinions over a much larger range of issues.

When the phenomenon of reactance is added to the picture, the segregated opinions tend to radicalize – not only are the individuals predominantly grouped into distinct opinion peaks, but also the opinions they commit to are close to the extremes of the opinion scale.

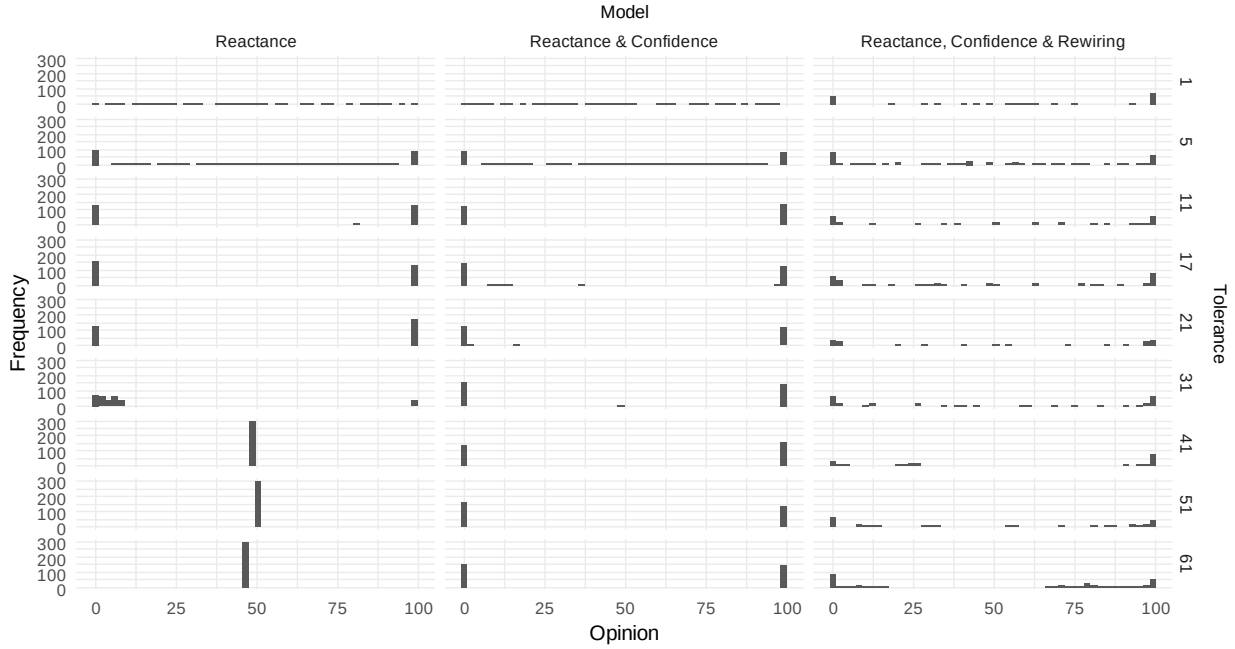


**Figure 6. Histograms of final opinions after 700 simulation steps in sample simulation runs for the three models with reactance. For each model 9 simulations were run, for different values of tolerance (labelled on the right) resulting in 9 histograms.**

Surprisingly, rewiring possibility ameliorates the polarization in models where reactance is present. This is due to the fact that cutting social links prevents protracted conflicts and subsequent radicalization.

In physical space socializing "deleting" relations is often impossible. Therefore groups and individuals have developed certain behaviors and strategies that enable them to avoid the worsts of conflicts. For example, they tend to avoid conflict prone topics and diversify the relations with respect to topic, importance, etc.

This translates into a complex structure of social links wherein individual ego networks are composed of partially overlapping subnetworks that serve different social or professional purposes. This concept – embeddedness of social networks [6] – ensures that a

social system can perform complex functions that are interdependent but not fully dependent on each other.

To help ameliorate the problem of polarization on social media, solutions that would help diversify ego networks and that would strengthen their embeddedness can be designed. Certain social media platforms are already topic specific (e.g. LinkedIn, ResearchGate) but the biggest ones are not (Facebook, Twitter). Moreover, users are encouraged to link their different identities from various platforms, by e.g. logging in everywhere with a single (Facebook) ID and to import all their social contacts from one platform to another. In effect, the ego networks are collapsed and create a huge amalgamate that in physical social context would be unmanageable. With the collapse of the social network comes also the collapse of areas, topics and issues. Content from all platforms is combined and broadcast for all to see. In effect, conflicts from one sphere might be generalized to other spheres and may become a general divide across many areas of social functioning.

Maintaining diversified ego networks could be promoted by tagging connections and keeping content thematically separated and flowing over different links. In effect a complex multiplex network could form. Moreover, strategies for conflict avoidance could be implemented. Instead of simple "thumbs up" and "thumbs down" more complex assessment could be available within one click – for example "agree to disagree".

Social media limits certain behaviors due to the affordances of the technology. Yet, the potential to enrich ICT mediated social interactions is immense and innovative solutions are sprouting in many platforms. However, we advocate to analyze each new concept with respect to how it can affect known social processes and phenomena to avoid turning those processes into distorted reflection of what they are in real space.

## 7. References


[1] Adamic, L.A. and Glance, N. The Political Blogosphere and the 2004 U.S. Election: Divided They Blog. *Proceedings of the 3rd International Workshop on Link Discovery*, ACM (2005), 36–43.

[2] Conover, M.D., Ratkiewicz, J., Francisco, M., Flammini, A., and Menczer, F. Political polarization on Twitter. *In ICWSM*, (2011), 89–96.

[3] Edmonds, B. and Hales, D. Computational Simulation as Theoretical Experiment. *The Journal of Mathematical Sociology 29*, 3 (2005), 209–232.

[4] Epstein, J. Why Model? *Journal of Artificial Societies and Social Simulation 11*, 4 (2008).

[5] Gilbert, N. *Agent-Based Models*. SAGE, 2008.

[6] Granovetter, M. Economic Action and Social Structure: The Problem of Embeddedness. *American Journal of Sociology 91*, 3 (1985), 481–510.

[7] Gregg, B. Frequency Trails: Modes and Modality. http://www.brendangregg.com/FrequencyTrails/modes.html.

[8] Hampton, K.N., Shin, I., and Lu, W. Social media and political discussion: when online presence silences offline conversation. *Information, Communication & Society 20*, 7 (2017), 1090–1107.

[9] Hargittai, E., Gallo, J., and Kane, M. Cross-ideological discussions among conservative and liberal bloggers. *Public Choice 134*, 1–2 (2008), 67–86.

[10] Isenberg, D.J. Group Polarization: A Critical Review and Meta-analysis. *Journal of Personality and Social Psychology 50*, 6 (1986), 1141–1151.

[11] Kohut, A., Keeter, S., Doherty, C., and Dimock, M. Social networking and online videos take off: Internet's broader role in campaign 2008. *TPR Center, The PEW research center*, (2008).

[12] Liggett, T. *Interacting Particle Systems*. Springer Science & Business Media, 2012.

[13] Miron, A.M. and Brehm, J.W. Reactance Theory - 40 Years Later. *Zeitschrift für Sozialpsychologie 37*, 1 (2006), 9–18.

[14] Myers, D.G. and Lamm, H. The group polarization phenomenon. *Psychological Bulletin 83*, 4 (1976), 602–627.

[15] Nowak, A., Bartkowski, W., Samson, K., et al. No need for speed: Modeling trend adoption in a heterogeneous population. *Advances in Complex Systems 16*, 04n05 (2013), 1350025.

[16] Nowak, A., Szamrej, J., and Latané, B. From private attitude to public opinion: A dynamic theory of social impact. *Psychological Review 97*, 3 (1990), 362–376.

[17] Pruitt, D.G. and Teger, A.I. The risky shift in group betting. *Journal of Experimental Social Psychology 5*, 2 (1969), 115–126.

[18] Rogers, E.M. and Shoemaker, F.F. *Communication of Innovations; A Cross-Cultural Approach.* The Free Press, New York, N. Y., 1971.

[19] Sanders, G.S. and Baron, R.S. Is social comparison irrelevant for producing choice shifts? *Journal of Experimental Social Psychology 13*, 4 (1977), 303–314.

[20] Stoner, J. A comparison of individual and group decisions involving risk. 1961.

[21] Tesser, A. and Conlee, M.C. Some effects of time and thought on attitude polarization. *Journal of Personality and Social Psychology 31*, 2 (1975), 262–270.

[22] Vinokur, A. and Burnstein, E. Depolarization of attitudes in groups. *Journal of Personality and Social Psychology 36*, 8 (1978), 872–885.

[23] Weisbuch, G. Bounded confidence and social networks. *The European Physical Journal B 38*, 2 (2004), 339–343.

[24] Weisbuch, G., Deffuant, G., Amblard, F., and Nadal, J.-P. Interacting Agents and Continuous Opinions Dynamics. In P.R. Cowan and D.N. Jonard, eds., *Heterogenous Agents, Interactions and Economic Performance*. Springer Berlin Heidelberg, 2003, 225–242.

[25] Wilenski, U. NetLogo. 1999. http://ccl.northwestern.edu/netlogo/.

[26] Incomplete ordering of the voter model on small-world networks. *EPL (Europhysics Letters) 63*, 1 (2003), 153.