IEICE TRANSACTIONS

on Communications

VOL. E105-B NO. 8
AUGUST 2022

A PUBLICATION OF THE COMMUNICATIONS SOCIETY

The Institute of Electronics, Information and Communication Engineers
Kikai-Shinko-Kaikan Bldg., 5-8, Shibakoen 3chome, Minato-ku, TOKYO, 105-0011 JAPAN

IEICE TRANS. COMMUN., VOL.E105–B, NO.8 AUGUST 2022

PAPER

Modeling Polarization Caused by Empathetic and Repulsive Reaction in Online Social Network

Naoki HIRAKURA[†a)], ***Student Member***, **Masaki AIDA**[†b)], ***and*** **Konosuke KAWASHIMA**[†c)], ***Fellows***

SUMMARY While social media is now used by many people and plays a role in distributing information, it has recently created an unexpected problem: the actual shrinkage of information sources. This is mainly due to the ease of connecting people with similar opinions and the recommendation system. Biased information distribution promotes polarization that divides people into multiple groups with opposing views. Also, people may receive only the seemingly positive information that they prefer, or may trigger them into holding onto their opinions more strongly when they encounter opposing views. This, combined with the characteristics of social media, is accelerating the polarization of opinions and eventually social division. In this paper, we propose a model of opinion formation on social media to simulate polarization. While based on the idea that opinion neutrality is only relative, this model provides new techniques for dealing with polarization.

key words: *social media, polarization, filter bubble*

1. Introduction

Today, a lot of social media streams exist and play important roles in information distribution. Unlike traditional mass media, social media allows users to collect information on interesting topics more efficiently. However, the increased convenience has been accompanied by the unexpected problem of the diminution of actual information sources. The main reasons for this are users' freedom to choose which streams to follow and the recommendation system, which delivers information thought to be of interest to individual users. Polarization is a concern as only users with similar ideas tend to communicate with each other, which creates a biased information distribution environment [1]. Polarization is the emergence of groups with conflicting opinions on a topic. As polarization strengthens, it leads to an intensification of slander and defamation between the conflicting groups. Polarization was confirmed on several topics. For example, polarization on three major topics: gun control, same-sex marriage, and climate change [2] and secular and Islamist groups [3].

Networks play a significant role in the formation of the individual's opinion, and many models have been proposed [4], [5]. In this paper, we propose a model of opinion formation on social media to simulate polarization. This model incorporates users' empathetic and repulsive reactions to social media posts based on the following ideas. People are more likely to accept others' opinions if they support their preferred opinions, and to ignore the opposing opinions. This is called confirmation bias [6]. In particular, social media promotes this, making it easy to form clusters of users with similar opinions. In a closed community of people with similar opinions, the same opinions are reinforced, and different opinions are actively suppressed. This can create a special kind of common sense that is valid only within a community; it is referred to as the echo chamber phenomenon. For example, biased linkage patterns in Blogs about political topics [7] and social media users' tendency clustering along party lines [8] were found. Besides, quantifying echo chambers on Twitter that emerged with the impeachment of the former Brazilian President was conducted [9]. These studies support the presence of echo chambers in OSNs.

Moreover, people can feel strong repulsion towards opinions that are different from their own. This means that when people see an opinion that differs from their own, they will come to believe more strongly in the opinion they originally had. In psychology this is called the backfire effect [10], and there is a report that confirms the backfire effect in social media users [11].

A opinion formation model incorporates the backfire effect was proposed [12]. This model is called the BEBA model because it incorporates two types of interactions: backfire effect and biased assimilation. In this model, regardless of the actual distribution of opinions, neutral opinions are artificially introduced in advance and are used as a standard to determine which interaction predominates. Based on the assumption of neutral opinions, the BEBA model posits that biased assimilation is established between users who have the same opinions, while the backfire effect emerges if the users have different opinions. Polarization occurs between users who have conflicting opinions, which are opinions on opposite sides of the neutral opinion. Polarization never occurs if all users have opinions on the same side of the neutral opinion. Of importance, our paper proposes a model that reflects the idea that opinion neutrality is determined by relative differences in user opinions.

In this paper, we propose a model of user opinion formation as evidenced by interaction via social media. A unique feature of our model is that it eliminates the need to establish an absolute opinion neutrality point. This feature reflects our idea that polarization arises with relative differences in user opinions, rather than from any predetermined neutral opinion. Our model is designed to change continuously the

Manuscript received September 12, 2021.
Manuscript revised December 24, 2021.
Manuscript publicized February 16, 2022.

[†]The authors are with the Tokyo Metropolitan University, Hinoshi, 191-0065 Japan.
a) E-mail: hirakura-naoki@ed.tmu.ac.jp
b) E-mail: aida@tmu.ac.jp
c) E-mail: k-kawa@cc.tuat.ac.jp

DOI: 10.1587/transcom.2021EBP3150

strength of the interaction based on differences in opinion. Also, two types of reactions to social media posts are considered: empathy and repulsion. We assume that the former is stronger than the latter as opinions are closer, and its strength decays as opinion difference strengthens. Conversely, the latter is stronger than the former when the opinions differ, and its strength decays the closer the opinions are.

To understand the behavior of the proposed model, we conduct simulations. They reveal parameter characteristics and show that the model can form either large groups or multiple small groups.

The rest of this paper is organized as follows. In Sect. 2, we explain related work and clarify the position of this study. Section 3 proposes a model of opinion formation based on empathetic and repulsive reactions to social media posts. In Sect. 4, we evaluate the proposed model and show how different parameters contribute to polarization. Section 5 compares the proposed model with previous study and shows that it can effectively deal with polarization. In Sect. 6, we conclude.

2. Related Work

As mentioned in Sect. 1, features of the proposed model are that it obviates the need to artificially introduce the neutral point of opinions, that the strength of the interaction changes based on the difference in opinions, and that the opinion value is updated based on information diffusion on the OSN. We will explain the key advances over existing study by discussing these three points.

First, we discuss the fact that the proposed model eliminates the need to artificially introduce a neutral point of opinion in advance. In addition to [12], which we discussed in Sect. 1, [13] is an existing study that introduces repulsive interactions after assuming that the objective neutral opinion is 0. In that study, it is rare for agents to completely change their original opinion by crossing over the neutral point of opinion. The interactions of users with opinions of different signs are governed by different rules. For the situation in which the two agents' opinion values are close but their sign is different, [13] proposed three types of interactions: negative interactions, indifference, and repulsion. Regardless of interaction type, the sign of each agent's opinion value does not change. Besides, many other models also artificially introduce a neutral point of opinion in advance at the origin of the axis. Typical models that reflect such thought about opinion neutrality are DeGroot model [14] and Friedkin-Johnsen (FJ) model [15]. Ordinary, these kinds of models describe the user's opinion as continuous real value in $[-1, 1]$ or in $[-\infty, +\infty]$, and 0 represents a neutral opinion. Many models that extend DeGroot model and FJ model have been proposed. For example, there are extended models focusing on polarization: a model based on biased assimilation that extends the eGroot model [16] and a model examining the filter bubble effect to the polarization that extends the FJ model [17]. Since each model extends the DeGroot model or the FJ model, thought of opinion neutrality is the same. There are also models that have the same thought of opinion neutrality other than DeGroot, FJ, and these extending models. For example, [18] use a sign of the opinion value to describe difference of user's stance. On the other hand, since our model sets no objective neutral opinion and instead addresses the difference in opinion strength, it allows for more subtle interactions among users.

Next, we discuss how the strength of the interaction changes based on the difference in opinions. The bounded confidence model is a model that determines the presence or absence of an interaction based on the difference in opinions. This model uses a predetermined threshold value, called the confidence radius, and interaction occurs only between agents whose difference in opinion value is less than or equal to the confidence radius. One type of bounded confidence model is the Deffuant-Weisbuch (DW) model [19]. In this model, if the difference in the opinion values of two randomly selected agents is less than or equal to the confidence radius, their opinions approach each other.

Another type of bounded confidence model, the Hegselmann-Krause (HK) model, updates opinion values synchronously [20]. In this model, if the difference in the opinion values of two agents is less than the confidence radius, it is considered that they are adjacent on the network. The user's opinion value is updated to the average value of the opinion values of itself and neighboring users.

In addition, a model that extends the HK model based on social judgment theory has been proposed [21]. In this model, one agent classifies another agent as being suitable for acceptance, non-commitment, or rejection based on the difference in opinion; the corresponding responses are positive acceptance, ignoring, and negative perception. Two thresholds are needed to establish these three judgement types.

As shown above, in the bounded confidence model and its extended model, threshold values are set in advance to determine the existence and type of interaction. Whereas these previous studies set thresholds in advance and so lack flexibility, this paper proposes a continuous, threshold-less model.

In addition, the proposal differs in terms of how it handles repulsive interactions. In [21], based on the social judgment theory, agents whose opinions are close to each other always develop a positive attitude, and agents who have different opinions always develop a negative attitude. The proposed model, on the other hand, stochastically selects empathetic and repulsive reactions, and the strength of the reaction is determined by a continuous function of the difference in opinion values: if the opinions are close, the empathy is large and the repulsion is small. It should be noted that the amount of change in the opinion value with repulsion when opinions are close is so small that there is virtually no substantially change in opinion. The same can be said for empathy and repulsion when opinions are far apart. The strength of the empathetic and repulsive reactions will be similar for those who are some distance apart in opinion, and both reactions are moderate.

Similar to the proposed model, a model that defines

interaction rules as a continuous function is also proposed, in which the opinions of users approach each other if they are close, and they move away from each other if they are far apart [30]. In [30], the cubic curve is used as the simplest form to represent the reaction, which is attractive when the opinions are close and repulsive when they are far apart. Although the approximation of cubic functions is a main term of the Taylor expansion near the equilibrium point, it is unnatural to define it in the whole parameter range. On the other hand, the proposed model assumes that the strength of the reaction changes exponentially. In the proposed model, the strength of the interaction is reasonable. For example, the rate of decrease in the strength of the interaction when the difference increases by δ is $\exp(-k(|\Delta| + \delta))/\exp(-k|\Delta|) = \exp(-k\delta)$, which is constant and highly universal regardless of the value of δ.

In [21] and [30], opinions are definitely attracted with each other when opinions are close and opinions are definitely repulsed each other when opinions are far away. On the other hand, the proposed model probabilistically chooses empathetic reaction or repulsive reaction. Though the average characteristics of the proposed model are similar to the model [30], social dynamics are not always driven by average characteristics. Individual reactions are diverse and such fluctuation of individuals can influence macro social dynamics. Our proposed model can describe individual reactions with fluctuations. It is necessary to compare how the difference between the two models actually affects the dynamics, but this is future work.

Next, we discuss how to update opinion values to reflect information diffusion over the OSN. This is related to the choice of interact partners. In this paper, opinion formation is considered to occur with information diffusion on OSNs. Existing opinion formation models describe several ways in which interaction partners are selected: all agents update their opinions synchronously [14], [15], or two randomly selected agents update their opinions [19], [22]. In this paper, the multivariate Hawkes process is used to determine the user who posts and the time of the posting, and the posting user's neighbors may update their opinions in response. The multivariate Hawkes process is a point process with mutual excitation in which multiple processes promote the occurrence of each other's events [23]. It is effective in modeling posts on social media, where a post by one user triggers posts by its neighbors [24], [25].

This paper is an extension of our prior study [26]. In a specific advance, we evaluate the behavior of the model when each user has different parameters.

3. Modeling Polarization by Empathy and Repulsion

We model situations in which a user's posts on social media influence the opinions of other users. In this model, two types of interactions are considered: empathy and repulsion. Users strongly empathize with posts that are close to their opinions and dismiss posts that express opinions very different from their own. In the case of repulsion, users are repulsed more

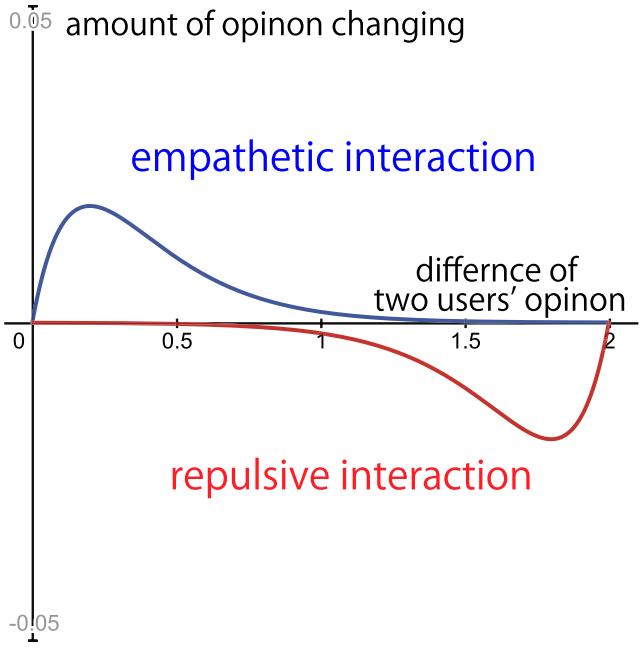
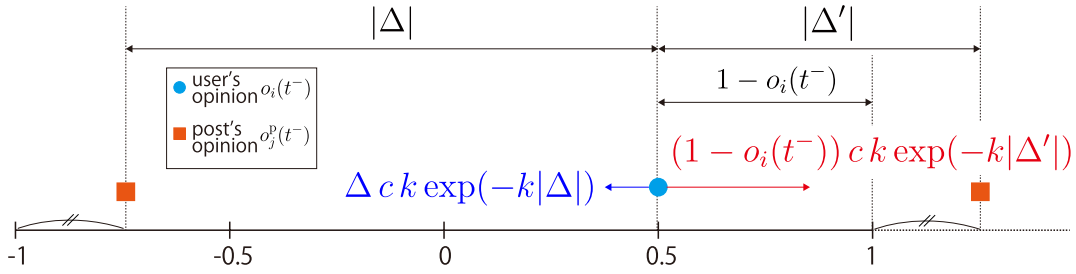


Fig. 1 Amounts of opinion changing by empathetic and repulsive interaction.

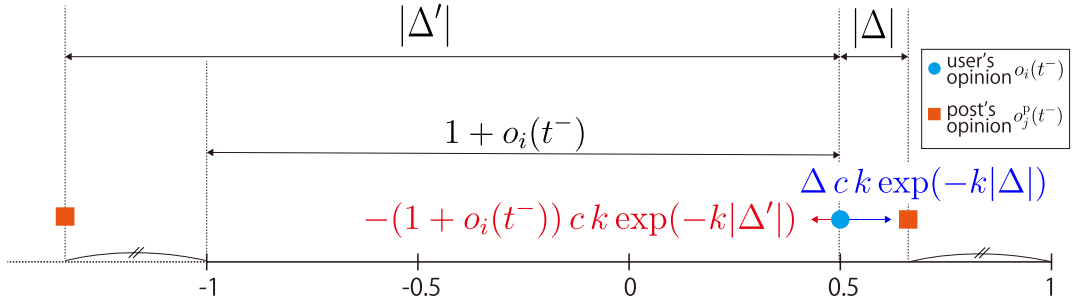
by posts that are more distant from their opinions, and are less likely to be repulsed by posts that are closer to their own opinions. These two kinds of interactions are stochastically chosen, therefore, three types of reactions are introduced: empathy, repulsion, and disregard. Important points of the proposed interaction rule are that it eliminates the need to introduce both a point of neutral opinion and a threshold that determines the type of interaction. Intuitive drawing of interaction rules is shown in Fig. 1. When the difference between two users' opinions is small, the amount of opinion changing by empathetic interaction is large and the amount of opinion changing by repulsive interaction is tiny. Conversely, when the difference between two users' opinions is large, the amount of opinion changing by empathetic interaction is tiny and the amount of opinion changing by repulsive interaction is large. The following rules determine how opinions are changed based on these reactions.

Consider a social network with N users. User i ($i = 1, \ldots, N$) is deemed to have opinion value $o_i(t) \in [-1, 1]$; the opinion value $o_i^{\mathrm{p}}(t)$ of user i towards a post at time t is $o_i^{\mathrm{p}}(t) = o_i(t)$ and reflects the user's opinion at that time. We assume that the opinion of user i is impacted by concurrent posts. If user j is the author of the latest post seen by user i, then the opinion value of user i at time t changes in response to $o_j^{\mathrm{p}}(t^-)$, where t^- represents time right before t. The opinion value of user i is given by $o_i(t) = o_i(t^-) + f(o_i(t^-), o_j^{\mathrm{p}}(t^-))$. The second term on the right side is explained later.

When user i sees the latest post of user j, user i develops either an empathetic or repulsive reaction with probability of p_i or $1 - p_i$, respectively. We propose a model of change in opinion value, assuming that the magnitude of the empathy and repulsion depends on the difference in opinion values.



(a) The amount of opinion change in the case of $\Delta = o_j^{\mathrm{p}}(t^-) - o_i(t^-) \leq 0$.



(b) The amount of opinion change in the case of $\Delta = o_j^{\mathrm{p}}(t^-) - o_i(t^-) > 0$.

Fig. 2 These two illustrations show the amount of opinion change in the cases of $\Delta = o_j^{\mathrm{p}}(t^-) - o_i(t^-) \leq 0$ (Fig. 2(a)) and $\Delta = o_j^{\mathrm{p}}(t^-) - o_i(t^-) > 0$ (Fig. 2(b)). The cyan circles and orange squares are the user's and post's opinion values, respectively. The blue arrow and red arrow indicate the amount of opinion changing by empathetic reaction and repulsive reaction, respectively.

First, consider the case of empathy. User i should show a stronger empathetic reaction to posts whose opinion values are close to their own, and weaker empathy when the opinion values are very different. Based on this, we introduce function $f(o_i(t^-), o_j^{\mathrm{p}}(t^-))$, which represents the change in the opinion value, as follows

$$\Delta\, c\, k\ \exp(-k\,|\Delta|). \tag{1}$$

Note that $\Delta = o_j^{\mathrm{p}}(t^-) - o_i(t^-)$ is the difference between the opinion value of the latest post seen by user i and the user's opinion value. Parameter $k(> 0)$ is a constant that determines the attenuation in the intensity of empathy with respect to the absolute value of the difference of opinion values $|\Delta|$, and c is a parameter that is adjusted to yield $o_i(t) \in [-1, 1]$, which satisfies $ck \leq 1$.

Next, we consider the case of repulsion. In this case, a stronger repulsive reaction should be developed to posts that are very different from own opinion values, and the repulsion should weaken as the opinion values approach each other. In order to express the repulsive reaction in the same form as empathy, we impose a periodic boundary condition on the two ends of $+1$ and -1 to harmonize the difference between the opinion value of a post and the user's opinion, $|\Delta'| = 2 - |\Delta|$; the smaller the difference $|\Delta'|$ is, the larger the repulsive reaction becomes. Based on this, we design function $f(o_i(t^-), o_j^{\mathrm{p}}(t^-))$, which represents the change in opinion value, as

$$(1 - o_i(t^-))\, c\, k\ \exp(-k\,|\Delta'|), \quad (\Delta \leq 0), \tag{2}$$
$$-(1 + o_i(t^-))\, c\, k\ \exp(-k\,|\Delta'|), \quad (\Delta > 0). \tag{3}$$

Figure 2 illustrates the concrete responses of expressions (1), (2), and (3). The blue circle indicates a user's opinion and the orange square indicates the post's opinion. Numbered line of $[-1, 1]$, the solid line, delineates opinion space. We also impose a periodic boundary condition on the two ends of $+1$ and -1 to define the new difference of the opinion values as $|\Delta'|$ and the relationship is $|\Delta'| = 2 - |\Delta|$. It is used to define the opinion change rule of the repulsive reaction in the same form as the empathetic reaction. Note that the orange square on the dashed line indicates the same opinion on the solid line. The green arrow indicates the amount of opinion value change caused by the empathetic reaction. The smaller $|\Delta|$ is, the stronger the empathy is. The purple arrow indicates the amount of opinion value change caused by a repulsive reaction. The smaller $|\Delta'|$ is, the stronger the repulsion is.

Expression (1) shows the change in the opinion value in the case of an empathetic reaction; it indicates that the user's opinion value approaches the opinion value of the post by an amount proportional to the difference, $c\, k \exp(-k\,|\Delta|)$. In this case, the positive or negative value of Δ corresponds to the direction of the movement in the user's opinion to be closer to the opinion of the post as shown in the negative direction of Fig. 2(a) and the positive direction of Fig. 2(b).

Figure 2(a) shows the case of $\Delta \leq 0$. The arrow pointing in the negative direction from the blue circle representing the user's opinion value indicates that the user's opinion value approaches the opinion value of the post by the proprtion $c\, k\ \exp(-k\,|\Delta|)$ of difference $|\Delta|$. On the other hand, Fig. 2(b) shows the case of $\Delta > 0$, where the sign of Δ

denotes a positive change in opinion value in the opposite direction to that of Fig. 2(a).

Expressions (2) and (3) show the change in opinion value under a repulsion reaction. First, we will discuss the case of $\Delta \leq 0$. As shown in Fig. 2(a), the opinion value of user i increases by the proportion $c\,k\,\exp(-k\,|\Delta'|)$ of difference $1 - o_i(t^-)$ between the opinion value of user i and the most extreme opinion value. In this case, the amount of change in the opinion value becomes the expression (2) and the user's opinion value moves away from the posted opinion value. We now discuss the case of $\Delta > 0$. As shown in Fig. 2(b), the opinion of user i decreases by the proportion of $c\,k\,\exp(-k\,|\Delta'|)$ of difference $-(1 + o_i(t^-))$ between the opinion value $o_i(t^-)$ of user i and the most extreme opinion -1. In this case, the amount of change in the opinion value is given by the expression (3), where the user's opinion value moves away from the posted opinion value.

Parameter k (> 0) represents the attenuation rate of the strength of the effect, and the value of k characterizes the rate of change in the strength of the effect based on the difference in opinion values. Parameter k allows us to design opinion value change rules where users show a stronger empathetic reaction to posts with close opinion values and a stronger repulsive reaction to posts with distant opinion values.

The limitation of the proposed model is that the range of the opinion values can only be a finite interval. This is because it is not possible to define periodic boundary conditions. However, this limitation can be relaxed by introducing an appropriate scaling. The typical scaling is stereographic projection, which gives a bijection between points in $(-\infty, \infty)$ and points on an open section on a circle with a finite radius.

4. Evaluations of the Proposed Model

4.1 Preparation for Evaluations

We explain how we generated sequences of posts and the index of polarization used in the evaluations of the proposed model.

4.1.1 Generating a Sequence of Posts

Here, we employ the multivariate Hawkes process to generate artificial data [23]. The multivariate Hawkes process is a point process in which multiple processes mutually excite the occurrence of events. To use this for the generation of user post sequences on social media, we assume that each process is considered to be a user and each event is a post to social media, and that posts are promoted between connected users.

The posting rate, λ_i, of user i ($i = 1, \ldots, N$) is expressed as follows.

$$\lambda_i(t) = \mu_i + \sum_{j \in \partial i} \sum_{t_h \in H_j} \alpha_{ji}\, \mathrm{e}^{-\beta_{ji}(t - t_h)}, \tag{4}$$

where μ_i is the base rate, ∂i is the set of neighbors of user i, H_j is the set of past posting times of user j, and t_h is its element. α_{ji} and β_{ji} ($j = 1, \ldots, N$, $i = 1, \ldots, N$) represent the rate jump and attenuation rate for user i due to user j's posts, respectively. Thus, the posting rate of user i is determined by the user-specific base rate and the previous posts of neighboring users.

We generate a sequence of posts using the multivariate Hawkes process. First, social networks with 100 nodes are created by the Barabási-Albert model (hereinafter referred to as the BA model) [27] and complete graph. In Eq. (4), α_{ji} and β_{ji} are given as uniform random numbers in the range of $[0, 1]$ for each combination of j and i ($j, i = 1, \ldots, N$). The base rate, μ_i, is given as a uniform random number in the range of $[0, 1]$. The following experiments use the 10000 postings generated by the multivariate Hawkes process.

Simulations are conducted according to thinning method [29]. This method is briefly explained in Appendix.

4.1.2 Index of Polarization

We introduce the index [28] for frequency distribution $(\boldsymbol{\pi}, \boldsymbol{y}) = (\pi_1, \ldots, \pi_n\,;\, y_1, \ldots, y_n)$ as the index of polarization:

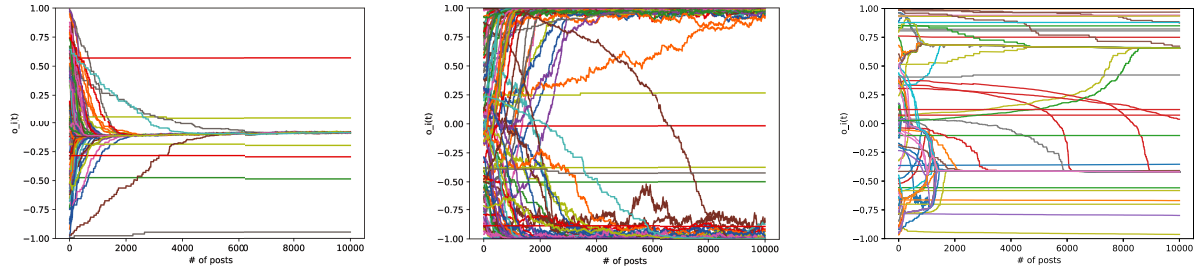
$$P = K \sum_{i=1}^{n} \sum_{j=1}^{n} \pi_i^{1+\theta}\, \pi_j\, \left|y_i - y_j\right|, \tag{5}$$

where n is the number of classes that equally divide the opinion value space $[-1, 1]$, y_i is the i-th class value from the bottom, π_i is the number of users belonging to the i-th class, K is a parameter for normalization, and θ is a parameter called the polarization sensitivity, which takes a value in the range of $(0, \theta^* \simeq 1.6]$. According to [28], the stronger the sense a user has of belonging to a class and the greater the degree of hostility toward another class, the greater is the degree of polarization. The index of polarization (5) is designed to satisfy the appropriate axioms for a valid index of polarization. For example, the maximum value is achieved if half of the users belong to the lowest and half to the highest class, and the minimum value is taken if all users belong to the same class.

It is often observed that users are divided into groups with conflicting opinions and slander each other on OSNs. The polarization index (5), which has a higher value when the group size is larger and the opinions between groups are farther apart, is considered to be effective in evaluating the polarization of people's opinions.

In the following evaluation experiments, we set the parameter $\theta = 0.5$ and the number of classes to 10 for the follow reasons. This index can quantitatively measure the degree of polarization in the distribution of opinion values if the parameters used to identify the degree of polarization of each opinion value distribution are properly set.

If the value of parameter θ, called the polarization sensitivity, is small, the value of the polarization index P will not be large unless the bias of the distribution is large. However, if the sensitivity to polarization takes a value of $(0, \theta^* \simeq 1.6]$,

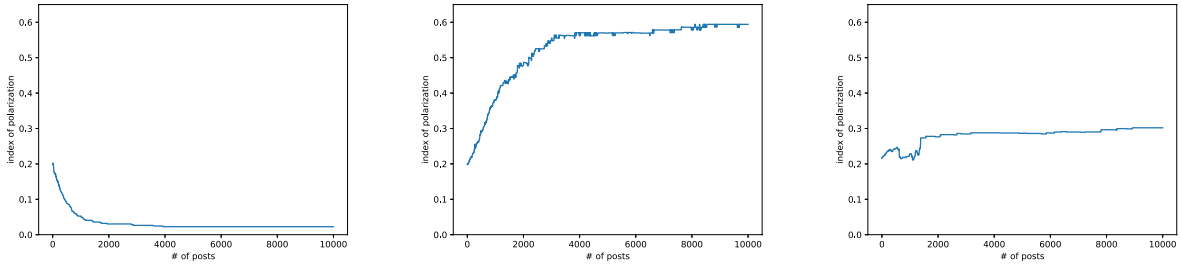


(a) Parameters were $p = 0.9$, $k = 2$. Almost all users were aggregated.

(b) Parameters were $p = 0.1$, $k = 2$. Two large groups were formed, each containing approximately half of all users.

(c) Parameters: $p = 0.9$, $k = 10$. Opinions were less likely to change, and several relatively small groups were formed.

Fig. 3 Examples of opinion value change conducted in BA graph. These are typical patterns of opinion formation and these indices of polarization are as follows: (a) low, (b) high, and (c) middle.



(a) Parameters were $p = 0.9$, $k = 2$. Index of polarization decreased toward 0.

(b) Parameters were $p = 0.1$, $k = 2$. Index of polarization increased to about 0.6.

(c) Parameters were $p = 0.9$, $k = 10$. Index of polarization did not change significantly.

Fig. 4 Changing of polarization index corresponding to Fig. 3.

it can work as an indicator of polarization. Also, if the number of classes is too small, such as 2 or 3, there will be no difference among the users belonging to each class in the distribution of different opinion values, and the polarization index is likely to have a high value even if the opinion values are randomly distributed. The polarization index works well as long as the number of classes is not set to such extreme settings.

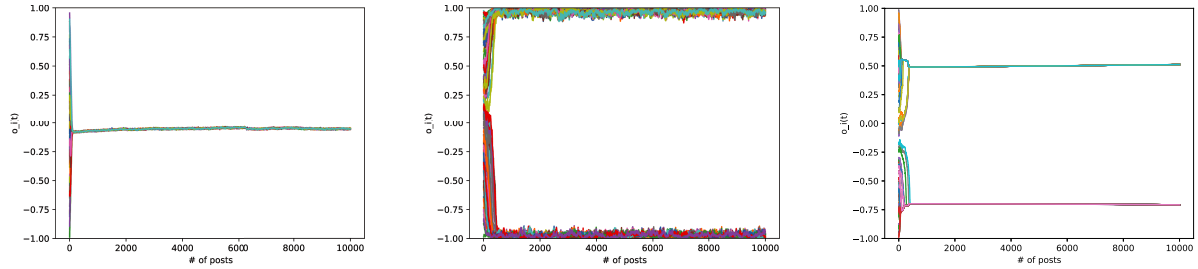
4.2 Parameter Characteristics of the Proposed Model

We conduct evaluation experiments to elucidate the characteristics of the parameters of the proposed model. The evaluation conditions are as follows. The initial opinion is given as a uniform random number in $[-1, 1]$. We apply the opinion value change rule to the sequence of posts generated in Sect. 4.1.1, with the probability of an empathetic reaction and the probability of a repulsive reaction being $p_i = p$ and $1 - p_i = 1 - p$, respectively, for all i.

Three typical simulation results with different parameters for the opinion change using the proposed model in the case of BA graph are shown in Fig. 3 and the case of complete graph are shown in Fig. 5. Besides, the changes of polarization index corresponding to opinion changing are shown in Fig. 4 and Fig. 6, respectively.

We discuss the simulation results for BA graph first. The probability of empathy p is set to $p = 0.9$ and $p = 0.1$ for the cases of Figs. 3(a) and 3(c), respectively; parameter k is fixed at 2. The difference in the probability of empathy reaction p yields the consensus of opinions in Fig. 3(a) and the polarization of opinions in Fig. 3(b). Figure 3(c) shows the case of $k = 10$. Since the influence of other users' posts is smaller than when $k = 2$, the opinion values of each user do not change much from their initial values. Some users did not join large groups in 3(a), 3(b), and 3(c). This is because the number of interactions is different for each user by using a sparse network model, the BA graph. Although Fig. 3(c) does not show complete polarization, the simulation is considered to be converged, as can be seen from the fact that the polarization index in Fig. 4(c) does not change much. Even if the polarization progresses further from this state, there are many users whose opinions do not change due to the influence of the graph structure, so the polarization index does not fluctuate significantly to reach the maximum value.

Next, we discuss simulation results for a complete graph. Experimental conditions are same as simulations of BA graph; the parameter k is fixed at 2 and the probability of empathy p is set to $p = 0.9$ and $p = 0.1$ for the cases of Figs. 5(a) and 5(b), respectively. The results are similar to the case of the BA graph, where consensus and polarization occur due to differences in p. Figure 5(c) shows the case where the parameters are $p = 0.9$ and $k = 10$. Two groups are formed, but each group does not have the opinion value ± 1. When k takes a relatively large value, it strongly empathizes only with users whose opinions are very close and strongly repels only with users whose opinions are far apart and hardly influence each other in the middle. Therefore, the two groups have little influence and converge to an opinion

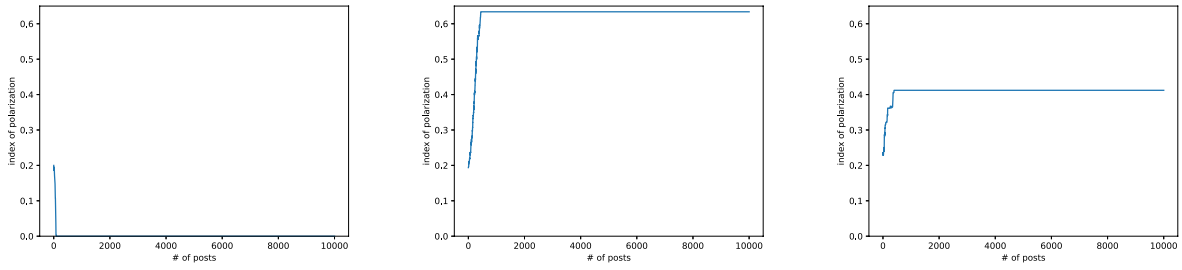


(a) Parameters were $p = 0.9$, $k = 2$. All users were aggregated.

(b) Parameters were $p = 0.1$, $k = 2$. All users were divided into two groups, each containing approximately half of all users.

(c) Parameters were $p = 0.9$, $k = 10$. All users were divided into two groups, each containing approximately half of all users. Two group's opinion values dose not go to extremes.

Fig. 5 Examples of opinion value change conducted in complete graph. These are typical patterns of opinion formation and these indices of polarization are as follows: (a) low, (b) high, and (c) middle.



(a) Parameters were $p = 0.9$, $k = 2$. Index of polarization rapidly decreased toward 0.

(b) Parameters were $p = 0.1$, $k = 2$. Index of polarization rapidly increased toward about 0.63.

(c) Parameters: $p = 0.9$, $k = 10$. Index of polarization converged to approximately 0.41.

Fig. 6 Changing of polarization index corresponding to Fig. 5.

value other than ± 1.

Comparing the case of BA graphs (Fig. 3) with the case of complete graphs (Fig. 5), the difference is that the convergence is faster in the case of complete graphs, and all users belong to the group. This is because in the case of BA graph, the number of interactions is biased for each user due to the non-uniformity of node degree, and in the case of complete graphs, anyone's posts affect all other users.

We then evaluate the index of polarization of the final state at the end of simulation using the index of polarization (5), for each combination of parameters p and k. The normalization parameters are set to $K = \left(\sum_{i=1}^{n} \pi_i\right)^{-(2+\theta)}$, the polarization sensitivity is set to $\theta = 0.5$, and the number of classes is set to $n = 10$. Thus the possible values of the polarization index lie approximately in the range $[0, 0.64)$.

We conducted 30 experiments with different initial opinions and evaluated the probability of the consensus reached and the probability of the polarization attained. Consensus reached is defined as satisfying $P \leq 0.064$ at the end of the simulation, and polarization attainment is defined as satisfying $P \geq 0.576$ (top and bottom ten percent of the possible values of the polarization index $[0, 0.64)$).

Figure 7 shows the probability of consensus reached for each combination of p and k. When k was relatively large, consensus was not reached. This is because empathy decays more strongly when k is high. When k is relatively small,

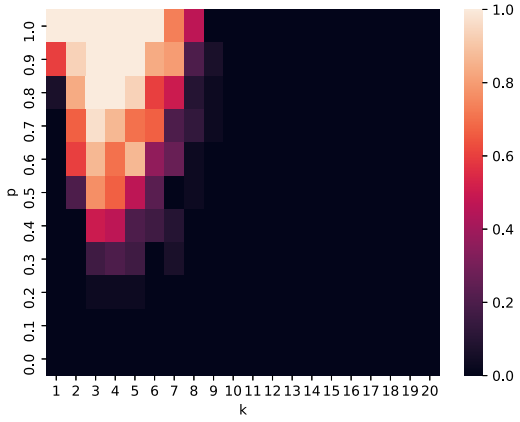


Fig. 7 Consensus probability of each combination of parameters. We conducted 30 simulations for each combination of p and k in BA graph, and evaluated the percentage of users reaching consensus ($P \leq 0.064$).

the probability of consensus depends on probability p. The closer p is to 1, the higher in the probability of consensus.

Figure 8 shows the probability of polarization for each combination of p and k. When k is relatively large, no polarization occurs. This is the same as in the probability of consensus since users are less influenced by differing opinions on posts given relatively large k values. When k is relatively small, the probability of polarization depends on

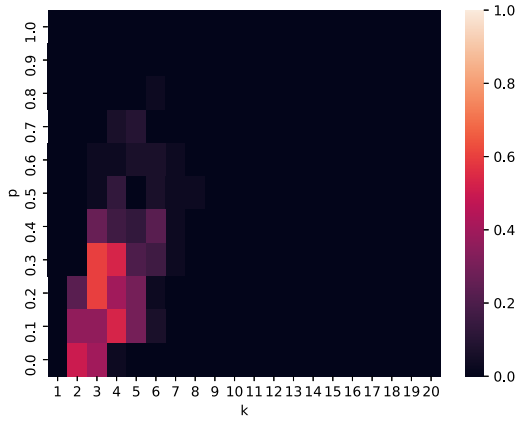


Fig. 8 Polarization probability of each combination of parameters. We conducted 30 simulations for each combination of p and k in BA graph, and evaluated the percentage of polarization ($P \geq 0.576$).

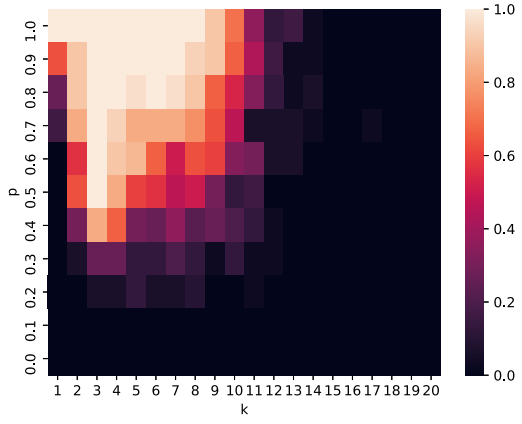


Fig. 9 Consensus probability of each combination of parameters. We conducted 30 simulations for each combination of p and k in complete graph, and evaluated the percentage of users reaching consensus ($P \leq 0.064$).

the probability, p, of an empathetic reaction. In contrast to the probability of consensus, polarization probability rises as p approaches 0.

We also investigated parameter characteristics in a complete graph. Figure 9 shows the consensus rate for the case of complete graphs. The characteristic that the consensus rate is high when the parameter k is relatively small and the empathy probability p is relatively large is the same as in the case of the BA graph. However, there is a difference that the consensus rate is higher even in regions where k is larger. Figure 10 shows the polarization rate for the complete graph. As in the case of the BA graph, the polarization rate is high when the parameter k is relatively small and the empathy probability p is relatively small. However, there is a difference that the polarization rate is high even in regions where k is larger. The reason for this difference is due to the structure of BA graphs and complete graphs, as described in the comparison between Fig. 3 and Fig. 5.

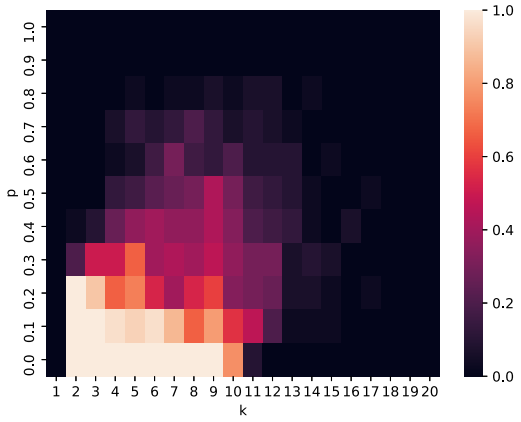
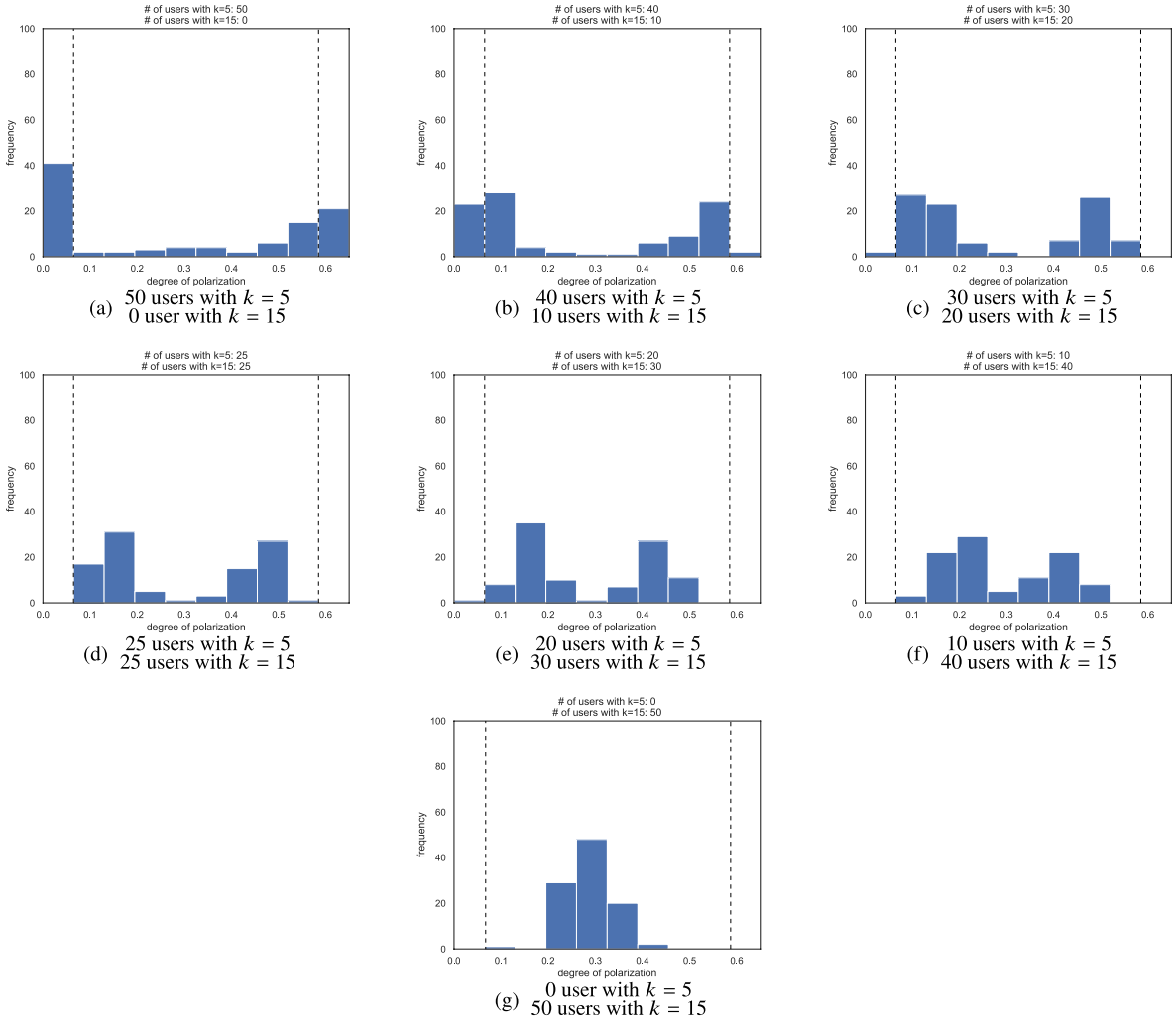


Fig. 10 Polarization probability of each combination of parameters. We conducted 30 simulations for each combination of p and k in complete graph, and evaluated the percentage of users reaching consensus ($P \geq 0.576$).

4.3 Model Behavior under Heterogeneous Parameter Combinations

In this section, we conduct simulations under the condition that the parameters are different for each user. The specific conditions of the experiment are as follows. We prepare a network with 50 users and each user's parameter k is set to $k = 5$ or $k = 15$. We simulate using different settings of mixing ratio of the population of users with $k = 5$ and $k = 15$. Specifically, we conduct experiments with each combination of the following: (a) 50 users with $k = 5$, (b) 40 users with $k = 5$ and 10 users with $k = 15$, (c) 30 users with $k = 5$ and 20 users with $k = 15$, (d) 25 users with $k = 5$ and 25 users with $k = 15$, (e) 20 users with $k = 5$ and 30 users with $k = 15$, (f) 10 users with $k = 5$ and 40 users with $k = 15$, and (g) 50 users with $k = 15$. We conduct 100 simulations and compare the distribution of the index of polarization at the end of each simulation for the different mixing ratios. The probability of an empathetic reaction is $p = 0.5$ for all users. Parameters and the number of classes of polarization index (5) are the same settings used for the experiments in Sect. 4.2.

Figure 11 shows the distribution of polarization at the end of the simulation for each mixture ratio with different values of parameter k. In Fig. 11(a), the parameter $k = 5$ for all users. Then, the lowest class of polarization index is the most frequent and the highest class is the second most frequent. The reason why the consensus is often achieved is that the attenuation in the strength of empathy k is relatively small. However, if the opinions are too far apart, the strength of empathy decreases, and the effect of repulsion increases. Once users start to get close to each other, they often reach consensus, whereas interactions between users with quite different opinions lead to polarization. As the mixing ratio of the population of $k = 5$ and $k = 15$ changes, the number of cases taking lower and higher polarization decreases, and the number of cases taking middle polarization index increases.



(a) 50 users with $k = 5$, 0 user with $k = 15$
(b) 40 users with $k = 5$, 10 users with $k = 15$
(c) 30 users with $k = 5$, 20 users with $k = 15$
(d) 25 users with $k = 5$, 25 users with $k = 15$
(e) 20 users with $k = 5$, 30 users with $k = 15$
(f) 10 users with $k = 5$, 40 users with $k = 15$
(g) 0 user with $k = 5$, 50 users with $k = 15$

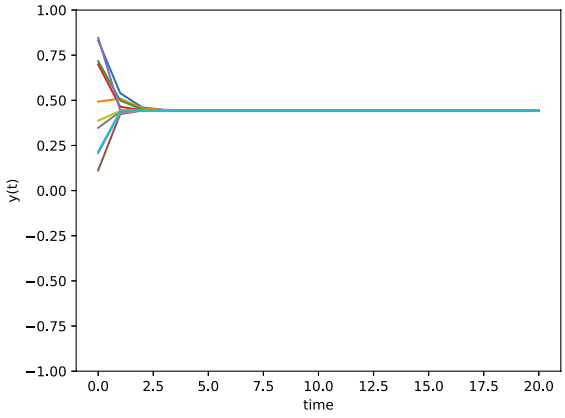
Fig. 11 Frequency distribution of the polarization index at the end of the simulation under mixed conditions of users with different parameters. In Fig. 11(a), all users' parameters are $k = 5$. In the order of Fig. 11(b), 11(c), . . . , 11(g), the ratio of users with $k = 5$ decreases and the ratio of users with $k = 15$ increases. When the ratio of $k = 5$ is higher, it is easy to form a large group so that the polarization index takes lower and higher values. As the ratio of $k = 15$ increases, users tend to form multiple small groups so that the polarization index takes middle values. Besides, dashed lines are threshold to determine consensus and polarization.

Users with a relatively large attenuation in empathy strength, k, are more empathetic to posts that are very close to their own but are less influenced by opinions slightly further apart. As the number of users with relatively large k increases, the variability of the opinion values in the whole network decreases, and middle polarization index cases were more frequent at the end of the simulation. From this result, it can be seen when the proportion of users with $k = 5$ is large, it is easy to form a large population while an increasing proportion of users with $k = 15$ tends to form multiple small groups.

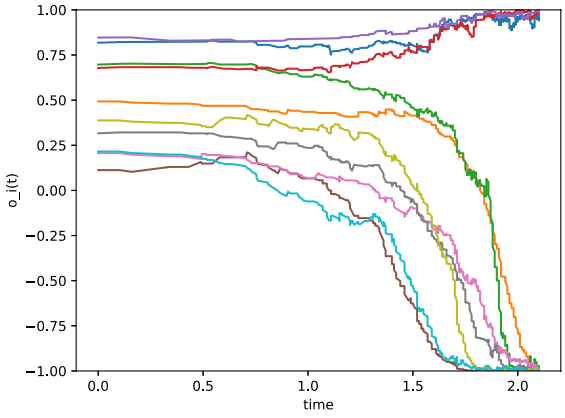
5. Comparison with Previous Research

We compare our model with the opinion formation model, which takes into account the reactions corresponding to empathy and repulsion. As an example, we take the BEBA model [12]. The BEBA model is based on the classical opinion formation model of DeGroot model [14]; BEBA contains biased assimilation and the backfire effect. Biased assimilation means that we highly evaluate information that is consistent with our original opinion, and the backfire effect means that when we are exposed to an idea that is different from our own, we come to believe more strongly in our original opinion.

The BEBA model is briefly described as follows. First, consider a social network with N users. User i ($i = 1, \ldots, N$) has opinion value $y_i(t) \in [-1, 1]$ at time t. For adjacent users i and j, if opinion values $y_i(t)$ and $y_j(t)$ are the same sign, biased assimilation dominates, but if they are different signs, backfire effect dominates. This means that opinion value 0 is neutral and users have different opinions depending on



(a) Evaluation results from BEBA model under biased initial opinion value distribution, all users' initial opinion values have the same sign. Opinions were aggregated.



(b) Evaluation results from the proposed model under biased initial opinion value distribution, all users' initial opinion values have the same sign. Polarization was confirmed.

Fig. 12 Change in opinion values under the condition such that all users' opinion values have the same sign.

whether they have positive or negative opinion values. On the other hand, in the proposed model, the amount of change in the opinion value is determined based on the difference between the post's and user's opinion values. In order to clarify the difference between the two models, we set all initial opinion values to have the same sign in a comparative evaluation.

Figure 12 shows comparison of experimental results of BEBA model and the proposed model. First, we describe the results of the experiments on the BEBA model. Figure 12(a) shows the change in opinion values yielded by the BEBA model. Since the initial opinion values with the same sign were given to all nodes, the backfire effect was quiescent, and the opinions became concentrated at the end of the simulation. Next, we describe the results of the experiments on the proposed model. Figure 12(b) shows the change in opinion values for $p = 0.1$ and $k = 2$ yielded by the proposed model. Although the initial opinion values with the same sign were given to all nodes, the amount of change in the opinion values was determined based on the difference in the opinion values, so that repulsion occurred and polarization was created. Depending on the parameter values used, the proposed model was found to yield cases of polarization in which repulsive reactions were active even with a biased opinion value distribution.

In the case of a biased initial distribution, such that all users' opinion values are the same sign, the BEBA model always reaches consensus. Besides, [13] also predetermines the opinion neutrality and the sign of opinion value never changes. On the other hand, the proposed model can predict cases of polarization depending on the parameter settings. Also, the proposed model can handle different cases of opinion change, including cases where a consensus is reached if we set different parameters. The difference in the results between the two models is due to the difference in the idea of opinion neutrality. Note that this comparative experiment is only conducted with BEBA model but other models that need to artificially introduce opinion neutrality in advance basically behave in a similar way to BEBA model. As we discussed in Sect. 2, other models that artificially introduce opinion neutrality in advance include [14]–[18]. The proposed model reflects the idea that opinion neutrality is not artificially introduced but by the relative difference in users' opinions; it can handle a wider range of opinion change cases.

6. Conclusion

This paper proposed a model to simulate polarization of social media users. The model is characterized by two types of reactions: empathy and repulsion, with different strengths of influence being created by differences in user opinion values. It also reflects the idea that relative differences in opinion determine the neutrality of opinion. In particular, since our idea of opinion neutrality dispenses that the artificially introduced concept of opinion neutrality, it has the advantage of being able to respond flexibly to biased distributions of opinion values. These features allow a wide range of user characteristics to be appropriately modeled so as to well cover the phenomenon of polarization. The dependence of the change in the opinion value on the initial values of the opinions will be examined in the future.

Acknowledgments

This research was supported by Grant-in-Aid for Scientific Research (B) No. 19H04096 (2019–2021), No. 20H04179 (2020–2022), and No. 21H03432 (2021–2023), and Grant-in-Aid for Challenging Research (Exploratory) No. 21K19775 (2021–2023) from the Japan Society for the Promotion of Science (JSPS), and TMU local 5G research support.

References

[1] C.R. Sunstein, "#Republic: Divided Democracy in the Age of Social

Media," Princeton University Press, 2018.
[2] W.J. Brady, J.A. Wills, J.T. Jost, J.A. Tucker, and J.J. Van Bavel, "Emotion shapes the diffusion of moralized content in social networks," National Academy of Sciences, vol.114, no.28, pp.7313–7318, 2017.
[3] I. Weber, V.R.K Garimella, A. Batayneh, "Secular vs. Islamist polarization in Egypt on Twitter," IEEE/ACM International Conference on Social Networks Analysis and Mining (ASONAM), pp.290–297, 2013.
[4] A.V. Proskurnikov and R. Tempo, "A tutorial on modeling and analysis of dynamic social networks. Part II," Annual Reviews in Control, vol.45, pp.166–190, 2018.
[5] M. Jalili and M. Perc, "Information cascades in complex networks," Journal of Complex Networks, vol.5, no.5, pp.665–693, 2017.
[6] R.S. Nickerson, "Confirmation bias: A ubiquitous phenomenon in many guises," Rev. Gen. Psychol., vol.2, no.2, pp.175–220, 1998.
[7] L.A. Adamic and N. Glance, "The political blogosphere and the 2004 US election: Divided they blog," Proc. 3rd international workshop on Link discovery, pp.36–43, 2005.
[8] M.D. Conover, B. Gonçalves, A. Flammini, and F. Menczer, "Partisan asymmetries in online political activity," EPJ Data science, vol.1, p.6, 2012.
[9] W. Cota, S.C. Ferreira, R. Pastor-Satorras, and M. Starnini, "Quantifying echo chamber effects in information spreading over political communication networks," EPJ Data Sci., vol.8, p.35, 2019.
[10] B. Nyhan and J. Reifler, "When corrections fail: The persistence of political misperceptions," Polit. Behav., vol.32, no.2, pp.303–330, 2010.
[11] C.A. Bail, L.P. Argyle, T.W. Brown, J.P. Bumpus, H. Chen, M.F. Hunzaker, J. Lee, M. Mann, F. Merhout, and A. Volfovsky, "Exposure to opposing views on social media can increase political polarization," Proc. National Academy of Sciences, vol.115, no.37, pp.9216–9221, 2018.
[12] X. Chen, P. Tsaparas, J. Lijffijt, and T. De Bie, "Opinion dynamics with backfire effect and biased assimilation," arXiv preprint arXiv:1903.11535, 2019.
[13] C. Altafini and F. Ceragioli, "Signed bounded confidence models for opinion dynamics," Automatica, vol.93, pp.114–125, 2018.
[14] M.H. DeGroot, "Reaching a consensus," J. Am. Stat. Assoc., vol.69, no.345, pp.118–121, 1974.
[15] N.E. Friedkin and E.C. Johnsen, "Social influence and opinions," Journal of Mathematical Sociology, vol.15, no.3-4, pp.193–206, 1990.
[16] P. Dandekar, A. Goel, and D.T Lee, "Biased assimilation, homophily, and the dynamics of polarization," Proc. National Academy of Sciences, vol.110, no.15, pp.5791–5796, 2013.
[17] U. Chitra and C. Musco, "Analyzing the impact of filter bubbles on social network polarization," Proc. 13th International Conference on Web Search and Data Mining, pp.115–123, 2020.
[18] F. Baumann, P. Lorenz-Spreen, I. Sokolov, and M. Starnini, "Modeling echo chambers and polarization dynamics in social networks," Phys. Rev. Lett., vol.124, no.4, p.048301, 2020.
[19] G. Deffuant, D. Neau, F. Amblard, and G. Weisbuch, "Mixing beliefs among interacting agents," Adv. Complex Syst., vol.3, no.01n04, pp.87–98, 2000.
[20] R. Hegselmann and U. Krause, "Opinion dynamics and bounded confidence models, analysis, and simulation," Journal of Artificial Societies and Social Simulation, vol.5, no.3, pp.1–33, 2002.
[21] H.D. Aghbolagh, M. Zamani, S. Paolini, and Z. Chen, "Balance seeking opinion dynamics model based on social judgment theory," Physica D: Nonlinear Phenomena, vol.403, Art. no.132306, 2020.
[22] G. Deffuant, F. Amblard, G. Weisbuch, and T. Faure, "How can extremism prevail? A study based on the relative agreement interaction model," Journal of Artificial Societies and Social Simulation, vol.5, no.4, 2002.
[23] A.G. Hawkes, "Spectra of some self-exciting and mutually exciting point processes," Biometrika, vol.58, no.1, pp.83–90, 1971.
[24] M.A. Rizoiu, Y. Lee, S. Mishra, and L. Xie, "Hawkes processes for events in social media," Frontiers of Multimedia Research, C. Shih-Fu, ed., Chapter 8, pp.191–218, Association for Computing Machinery and Morgan & Claypool, Cham, 2017.
[25] A.H. Zadeh and R. Sharda, "Hawkes point processes for social media analytics," Reshaping Society through Analytics, Collaboration, and Decision Support, L.S. Iyer, D.J. Power, eds., Chapter 5, pp.51–66, 2015.
[26] N. Hirakura, M. Aida, and K. Kawashima, "A model of polarization on social media caused by empathy and repulsion," arXiv preprint arXiv:2011.08141, 2020.
[27] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," Science, vol.286, no.5439, pp.509–512, 1999.
[28] J.-M. Esteban and D. Ray, "On the measurement of polarization," Econometrica, vol.62, no.4, pp.819–851, 1994.
[29] Y. Ogata, "On Lewis' simulation method for point processes," IEEE Trans. Inf. Theory, vol.27, no.1, pp.23–31, 1981.
[30] D. Sabin-Miller and D.M. Abrams, "When pull turns to shove: A continuous-time model for opinion dynamics," Phys. Rev. Research, vol.2, no.4, p.043001, 2020.

Appendix: Simulation Method of Multivariate Hawkes Process for Generating a Sequence of Posting

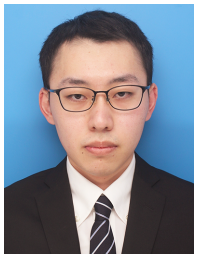
We used the thinning method [29] for simulating the multivariate Hawkes process and each user have posting rate:

$$\lambda_i(t) = \mu_i + \sum_{j \in \partial i} \sum_{t_h \in H_j} \alpha_{ji}\, \mathrm{e}^{-\beta_{ji}(t-t_h)}. \tag{A·1}$$

Using summation of all users' posting rates, $\lambda = \sum_{i=1}^{N} \lambda_i$, we determine the time of next posting. Next, we determine who post by ratio of posting rate λ_i/λ. Simulation detail is described in Algorithm 1.

Algorithm 1 Simulation Method of Multivariate Hawkes Process for Generating a Sequence of Posting

$\lambda^* \leftarrow \sum_{i=1}^{N} \mu_i$
$t^* \leftarrow 0$
$l \leftarrow 0$
let T be the end time of simulation
while $t^* < T$ **do**
  let E be a random number generated from exponential distribution $\lambda^* \exp(-\lambda^* x)$
  let $t^* \leftarrow t^* + E$ be a candidate of next posting time
  if $t^* > T$, terminate the simulation
  let $r \leftarrow \lambda(t^*)/\lambda^*$ be an adoption rate
  let u be a random number generated from uniform distribution from 0 to 1
  if $u \leq r$ **then**
    adopt the posting time t^* as next posting time
    $l \leftarrow l + 1$
    $t_l \leftarrow t^*$
    let v be a random number generated from probability distribution $P^* \leftarrow \lambda_i(t^*)/\sum_{i=1}^{N} \lambda_i(t^*)$, thus v indicates posting user
    $i \leftarrow v$
    $\lambda^* \leftarrow \lambda(t^*) + \sum_{j \in \partial i} \alpha_{ji}$
  else
    reject the posting time t^* as next posting time
    $\lambda^* \leftarrow \lambda(t^*)$
  end if
end while



Naoki Hirakura received his B.E. and M.E. degrees in Engineering from Tokyo Metropolitan University, Japan, in 2018 and 2020, respectively. Currently, he is a Ph.D. student of the Graduate School of Systems Design, Tokyo Metropolitan University.

Masaki Aida received his B.S. degree in Physics and M.S. degree in Atomic Physics from St. Paul's University, Tokyo, Japan, in 1987 and 1989, respectively, and his Ph.D. in Telecommunications Engineering from the University of Tokyo, Japan, in 1999. In April 1989, he joined NTT Laboratories. From April 2005 to March 2007, he was an Associate Professor at the Faculty of Systems Design, Tokyo Metropolitan University. He has been a Professor of the Graduate School of Systems Design, Tokyo Metropolitan University since April 2007. His current interests include analysis of social network dynamics and distributed control of computer communication networks. He received the Best Tutorial Paper Award and the Best Paper Award of IEICE Communications Society in 2013 and 2016, respectively, and IEICE 100-Year Memorial Paper Award in 2017. He is a fellow of IEICE, a senior member of IEEE, and a member of ACM and ORSJ.

Konosuke Kawashima received his B.E. degree from the University of Tokyo, Japan, in 1969, and Dr. Eng. degree from the same university in 1993. He is currently a Visiting Professor at Tokyo Metropolitan University, and a Professor Emeritus at Tokyo University of Agriculture and Technology (TUAT). He joined NTT Laboratories in 1969, where he engaged in the research and development of teletraffic engineering for various networks and systems. From 1997, he was Director of Teletraffic Research Center at NTT Advanced Technologies Corp. From 2002, he was a Professor at Department of Computer and Information Sciences of TUAT until March 2012. He received the Young Engineer Award and Paper Award from the IEICE in 1978, 1982 respectively. He also received the Best Paper Award, Distinguished Achievement and Contributions Award from the Operations Research Society of Japan (ORSJ) in 1986, 2007 respectively, and the Technical Award from the Telecommunications Advancement Foundation in 1996. He is a member of IEEE, and a fellow of ORSJ and IEICE.