A Model of Polarization on Social Media Caused by Empathy and Repulsion

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

SUMMARY In recent years, the ease with which social media can be accessed has led to the unexpected problem of a shrinkage in information sources. This phenomenon is caused by a system that facilitates the connection of people with similar ideas and recommendation systems. Bias in the selection of information sources promotes polarization that divides people into multiple groups with opposing views and creates conflicts between opposing groups. This paper elucidates the mechanism of polarization by proposing a model of opinion formation in social media that considers users' reactions of empathy and repulsion. Based on the idea that opinion neutrality is only relative, this model offers a novel technology for dealing with polarization.

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

1. Introduction

Today, social media services such as Twitter and Instagram have become widely adopted and are playing a significant role in information distribution. Unlike traditional mass media, social media allows users to efficiently focus on views that trigger their interest. Paradoxically, however, its very power raises the problem of excluding most information sources. This is because of the feature that allows users to choose the accounts they chose to follow and the existence of push-style recommendation systems. It is feared that social media will lead to polarization because it makes it easier to connect people with similar opinions which biases the information distribution [1]. Polarization is the division of opinion on a topic into basically two diametrically opposing positions.

Polarization tends to emerge when users express their opinions within a shuttered community; the opinions expressed are almost always seen as positive, with negative opinions being drowned or pushed out. As a result, a special understanding that is common only within the community becomes strengthened, and a gap forms between the thoughts and values of people outside the community. This reinforcement of belief within a community is called the echo chamber phenomenon. Due to such bias in information distribution, slander, and defamation between communities that have different opinions become more frequent. To prevent these social problems, we should elucidate the mechanism of polarization on social media.

The basic reactions of users to posts on social media are empathy and repulsion. The repulsive reaction is called the backfire effect in psychology [2], and one report has confirmed the backfire effect of social media use [3].

Reference [4] proposed a model of opinion formation that incorporates the backfire effect. In this model, polarization does not occur automatically in the distribution of biased opinions because the absolute origin is determined by the quantitative representation of opinions. This model corresponds to a situation where objective "neutral" positions are predetermined with regard to the distribution of opinions. Polarization is assumed to occur between users whose opinions are on the opposite sides of the neutral position.

In this paper, we propose a model that users are influenced by other users via social media such that their opinions change. A feature of this model is the origin of the quantitative representations of opinions is not necessarily objective neutrality. This feature reflects our view that opinion neutrality is relative. That is, polarization is generated from relative differences in opinion, but not from the predetermined standard of neutrality. Also, the proposed model is designed to yield different characteristics based on differences in opinion. There are two types of specific influence: one is a strong empathy for posts that are close to one's own opinion and dismissal of different opinions. The other is a strong repulsion for posts that differ from their own opinion and a ignoring of opinions that are close to their own.

The structure of this paper is as follows. In Section 2, we propose a model of opinion formation in which users show empathetic and repulsive reactions to posts on social media. In Section 3, we evaluate the model proposed in Section 2 and clarify the behaviors of the model for given parameters. Section 4 compares the proposed method with previous works and shows the advantages of the proposed method. We conclude in Section 5.

2. Modeling Polarization by Empathy and Repulsion

We model situations in which a user's posts on social media influence the opinions of other users; two kinds of reactions are considered, empathy and repulsion. When empathy occurs, 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 by posts that are more distant from their opinions, and less likely to be repulsed by posts that are closer to their own opinions. The following are the rules for changing opinions based on these reactions.

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

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 $o_i^p(t)$ of user i for a post at time t is $o_i^p(t) = o_i(t)$ reflects the user's opinion at that time. We assume that the opinion of user i is affected 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 due to $o_j^p(t^-)$. The opinion value of user i is given by $o_i(t) = o_i(t^-) + f(o_i(t^-), o_j^p(t^-))$. The second term on the right side will be explained later.

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

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 $f(o_i(t^-), o_j^p(t^-))$, which represents the change in the opinion value, as

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

Note that $\Delta = o_j^p(t^-) - o_i(t^-)$ is the difference between the opinion value of the post seen by user i in latest and the user's opinion value. Parameter $k\,(> 0)$ is a constant that determines the decay 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 be $o_i(t) \in [-1, 1]$, which satisfies $ck \leq 1$.

Next, consider the case of repulsion. In this case, it is necessary to develop a stronger repulsive reaction to the opinion values that are very different from their own opinion values, and the repulsion should be weaker when the opinion values are close. To express the repulsive reaction in the same form as in the case of empathy, we impose a periodic boundary condition on the two ends of $+1$ and -1 to make the difference between the opinion value of a post and the user's opinion $|\Delta'| = 2 - |\Delta|$; the smaller the difference $|\Delta'|$, the larger the repulsive reaction. Based on this, we design function $f(o_i(t^-), o_j^p(t^-))$, 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 1 illuminates the concrete actions of expressions (1), (2), and (3).

The expression (1) shows the change in the opinion value in the case of an empathy 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. 1(a) and the positive direction of Fig. 1(b).

Figure 1(a) is 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 a post by the proportion of $c\, k\, \exp(-k\,|\Delta|)$ of the difference $|\Delta|$. On the other hand, Fig. 1(b) is the case of $\Delta > 0$, where the sign of Δ denotes a positive change in opinion value in the opposite direction to that of figure 1(a).

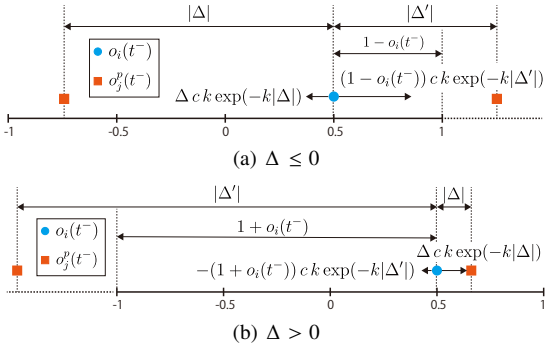


Fig. 1: Intuitive images of opinion value change.

The 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. 1(a), the opinion value of user i increases by the proportion $c\, k\, \exp(-k\,|\Delta'|)$ of the 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. 1(b), the opinion of user i decreases by the proportion of $-(1 + o_i(t^-))$ of the 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.

The parameter k (> 0) represents the decay 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.

3. Experimental evaluations of the proposed model

3.1 Preparation for evaluation experiments

We explain how to generate a sequence of posts and an index of polarization for the experiments conducted on the proposed model.

3.1.1 Generating a sequence of posts

Here, we generate artificial data using the multivariate Hawkes process [5]. 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 decay rate for user i due to user j's posts, respectively. Thus, the posting rate of user i is determined by the influence of the user-specific base rate and the previous posts of neighboring users.

We generate a sequence of posts using the multivariate Hawkes process. First, a social network with 10 nodes is created by the Barabási-Albert model (hereinafter referred to as the BA model) [6]. 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 $[0, 1]$. We evaluate the proposed model by generating a sequence of posts in the observation period $[t_1, t_1 + 2]$ with the first post occurrence time t_1.

3.1.2 Index of polarization

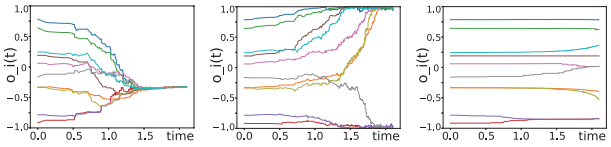
We introduce the index [7] for frequency distribution $(\boldsymbol{\pi}, \boldsymbol{y}) = (\pi_1, \ldots, \pi_n\,; y_1, \ldots, y_n)$ as the index of polarization:

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

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 [7], the stronger the sense a user has to 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 taken if half of the users belong to the lowest and half to the highest class, respectively, and the minimum value is taken if all users belong to the same class.

3.2 Parameter characteristics of the proposed model

We conduct evaluation experiments to elucidate the characteristics of the parameters of the proposed model. The experimental conditions are as follows. The initial opinion is given as a uniform random number of $[-1, 1]$. We apply the opinion value change rule to the sequence of contributions generated in Sect. 3.1.1, with the probability of an empathetic response and the probability of a repulsive reaction being $p_i = p$ and $1 - p_i = 1 - p$, respectively, for all i.



(a) $p = 0.9$, $k = 2$ (b) $p = 0.1$, $k = 2$ (c) $p = 0.9$, $k = 10$

Fig. 2: Examples of opinion value change.

Three typical simulation results with different parameters for the opinion change using the proposed model are shown in Figs. 2(a), 2(b) and 2(c). Parameter $k = 2$ is fixed for the cases of Figs. 2(a) and 2(b), and the probability of empathy p is set to $p = 0.9$ and $p = 0.1$, respectively. The difference in the probability of empathy reaction p causes a consensus of opinions in Fig. 2(a) and a polarization of opinions in Fig. 2(b). Figure 2(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.

We then evaluate the polarization of the final state at the final simulation time 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 $K = 0.5$, and the number of classes is set to $n = 10$. Thus the possible values of the polarization index are approximately $[0, 0.62]$. Assuming $c = 0.05$, Fig. 3 shows the average of 10 simulations for each combination of $p = 0, 0.1, \ldots, 1$ and $k = 1, \ldots, 10$ with different initial opinion values. According to Fig. 3, the polarization at the end of the simulation can be divided into three main types: concentrated into one opinion, polarized into two opinions, and little change in the opinions of each user. Opinions can be concentrated into one opinion when roughly $p \geq 0.4$ and $1 \leq k \leq 5$. This tendency occurred because they were more likely to show an empathic reaction even when the difference in opinion values was only slight. The polarization into two opinions occurs when $p < 0.4$ and $1 \leq k \leq 4$. This tendency occurred because they were more likely to show a repulsive reaction even when the difference in opinion values was only slight. The opinion of each user hardly changes when $k \geq 5$, regardless of the value of p. As k increases, it is due to the fact that we empathize only with those opinions that are very close and reject only those that are far enough away.

4. 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 [4]. The BEBA model is an expansion of classical opinion formation model, DeGroot model [8]; it contains bi-

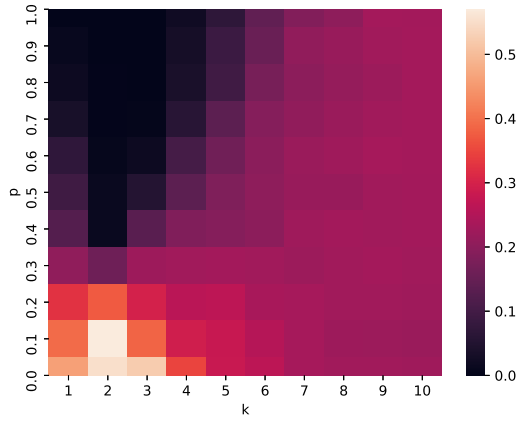


Fig. 3: Parameter characteristics of the proposed model.

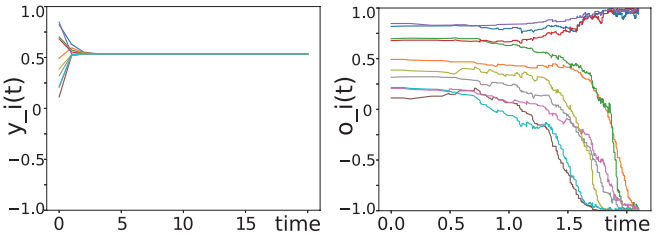
ased assimilation and the backfire effect. Biased assimilation means that we highly evaluate information that is consistent with our original opinion, and the backfire effect means that when we are exposed to an idea that is different from our own, we come to strongly believe our original opinion.

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 the opinion values $y_i(t)$ and $y_j(t)$ are the same sign, biased assimilation controls, and if they are different signs, backfire effect dominates. This means that the opinion value 0 is neutral and users have different opinions depending on 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. To clarify the difference between the two models, we give all the initial opinion values with the same sign in a comparative evaluation.

First, we describe the results of the experiments on the BEBA model. Figure 4(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 4(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 a repulsion occurred and polarization commenced. Depending on the choice of parameters, the proposed model was found to be able to handle cases of polarization in which repulsive reactions were active even with a biased opinion value distribution.

5. Conclusion

In this paper, we proposed a model of social media users' opinion formation to elucidate the mechanism of opinion polarization. The model is characterized by two types of reactions: empathy and repulsion, with different strengths of influence being created by differences in user's 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 with the concept of objective opinion neutrality, it has the advantage of being able to respond flexibly to a biased distribution 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.



(a) BEBA model (b) Proposed model

Fig. 4: Change in opinion values when initial opinion values given to all users have the same sign.

Acknowledgement

This research was supported by Grant-in-Aid for Scientific Research 19H04096 and 20H04179 from the Japan Society for the Promotion of Science (JSPS).

References

[1] C.R. Sunstein, *#Republic: Divided democracy in the age of social media*, Princeton University Press, 2018.

[2] B. Nyhan and J. Reifler, "When corrections fail: The persistence of political misperceptions," *Political Behavior*, vol. 32, no. 2, pp. 303–330, 2010.

[3] 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," *Proceedings of the National Academy of Sciences*, vol. 115, no. 37, pp. 9216–9221, 2018.

[4] X. Chen, P. Tsaparas, J. Lijffijt, and T. De Bie, "Opinion dynamics with backfire effect and biased assimilation," *arXiv preprint* arXiv:1903.11535, 2019.

[5] A.G. Hawkes, "Spectra of some self-exciting and mutually exciting point processes," *Biometrika*, vol. 58, no. 1, pp. 83–90, 1971.

[6] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

[7] J.-M. Esteban and D. Ray, "On the measurement of polarization," *Econometrica: Journal of the Econometric Society*, pp. 819–851, 1994.

[8] M.H. DeGroot, "Reaching a consensus," *Journal of the American Statistical Association*, vol. 69, no. 345, pp. 118–121, 1974.