MATHEMATICAL ENGINEERING
TECHNICAL REPORTS

A Collective Opinion Formation Model under Bayesian Updating and Confirmation Bias

Ryosuke NISHI and Naoki MASUDA

METR 2013–22 September 2013

DEPARTMENT OF MATHEMATICAL INFORMATICS
GRADUATE SCHOOL OF INFORMATION SCIENCE AND TECHNOLOGY
THE UNIVERSITY OF TOKYO
BUNKYO-KU, TOKYO 113-8656, JAPAN

WWW page: http://www.keisu.t.u-tokyo.ac.jp/research/techrep/index.html

A Collective Opinion Formation Model under Bayesian Updating and Confirmation Bias

Ryosuke NISHI[1,2,3] and Naoki MASUDA[3,*]

[1]National Institute of Informatics,
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan

[2]JST, ERATO, Kawarabayashi Large Graph Project,
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan

[3]Department of Mathematical Informatics,
The University of Tokyo,
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan

*Corresponding author (masuda@mist.i.u-tokyo.ac.jp)

September 10th, 2013

Abstract

We propose a collective opinion formation model with a so-called confirmation bias. The confirmation bias is a psychological effect with which, in the context of opinion formation, an individual in favor of an opinion is prone to misperceive new incoming information as supporting the current belief of the individual. Our model modifies a Bayesian decision-making model for single individuals [M. Rabin and J. L. Schrag, Q. J. Econ. **114**, 37 (1999)] for the case of a well-mixed population of interacting individuals in the absence of the external input. We numerically simulate the model to show that all the agents eventually agree on one of the two opinions only when the confirmation bias is weak. Otherwise, the stochastic population dynamics ends up creating a disagreement configuration (also called polarization), particularly for large system sizes. A strong confirmation bias allows various final disagreement configurations with different fractions of the individuals in favor of the opposite opinions.

1 Introduction

There are various models of collective opinion formation in which agents modify their opinions according to interaction with other agents [1,2]. Opin-

ion formation is a dynamic process: for example, interaction between agents makes their opinions approach each other. An important problem in opinion dynamics is to examine when an agreement (i.e., consensus) among all the agents occurs. Complete agreement is rarely observed in the real world [3, 4]. However, it is an established fact that opinion dynamics under the voter model, a classical opinion model in statistical physics and probability theory, inevitably reaches agreement in finite populations [1,5–9]. The majority rule model has a similar feature [1,10,11]. Partly motivated by this discrepancy, various extensions of voter and majority rule models and different models of collective opinion formation have been proposed to account for the disagreement in finite populations. Examples include the Deffuant model [12], language competition models [13–20], voter-like models on adaptive networks [21–24], voter model under partisan bias (the assumption that agents naturally prefer one opinion) [25,26], and variations of Axelrod's cultural dynamics (see [1] for references). Theoretical models have also been proposed in social sciences to explain disagreement in the context of polarization. For example, prior beliefs or initially received signals can cause disagreement between agents, even if they receive the same public signals from then on [27–31].

Although there is a plethora of studies addressing the problem of agreement and disagreement in opinion dynamics, we propose a model incorporating two factors that are relevant to human behavior: Bayesian belief updating and confirmation bias. Bayesian belief updating is commonly used in studies of the decision making of agents receiving uncertain information [30,32,33]. The confirmation bias is a psychological bias inherent in humans, in which an agent inclined towards an opinion tends to misperceive incoming signals as supporting the agent's belief [34, 35]. A non-Bayesian model with the confirmation bias was previously proposed for explaining the influences of media and interactions between agents [36].

We are not the first to study opinion formation under the Bayesian updating and confirmation bias. In the framework of single agent opinion formation, Rabin and Schrag showed that the confirmation bias triggers overconfidence and can cause the individual to hold incorrect beliefs, even if it receives a series of external signals suggesting the true state of the world [37]. Orléan studied the Bayesian dynamics of agents subjected to the confirmation bias, interacting through the mean field [38]. The model yields agreement or disagreement depending on the parameter values.

In this study, motivated by the Rabin-Schrag model [37], we propose a model of collective opinion formation with a confirmation bias. We model direct peer-to-peer interactions between agents (not through the mean field) and their effects on the Bayesian updating of each agent. To study the pure effects of interactions among agents, we do not assume that agents receive signals from the environment as in previous studies [31]. We numerically simulate the model to reveal the conditions under which the populations of

agents agree and disagree, depending on the values of parameters such as the strength of the confirmation bias, fidelity of the signal, and the system size.

2 Model

Our model modifies the Bayesian decision-making model proposed by Rabin and Schrag [37] in two main ways. First, we consider a well-mixed population of Bayesian agents that interact with each other; Rabin and Schrag focused on the case of the single agent. Second, agents do not receive external signals from the environment in our model. In the Rabin-Schrag model, such an external signal, which represents the "correct" answer in the binary choice situation (i.e., the true state of nature), is assumed. By making the two changes, we concentrate on collective opinion formation by Bayesian agents, whereby there are two possible alternative opinions of equal attractiveness.

We label the N agents $1, \ldots, N$ and denote the opinion of agent i ($i = 1, \ldots, N$) by $x_i \in \{\mathrm{A}, \mathrm{B}\}$, where A and B are the alternative opinions. We assume that agents are not perfectly confident in their opinions. To model this factor, we adopt the Bayesian formalism used by Rabin and Schrag [37]. We denote by $\Pr(x_i = \mathrm{A})$ the strength of the belief (hereafter, simply the belief) with which agent i believes in opinion A. A parallel definition is applied to $\Pr(x_i = \mathrm{B})$. It should be noted that $\Pr(x_i = \mathrm{A}) \geq 0$, $\Pr(x_i = \mathrm{B}) \geq 0$, and $\Pr(x_i = \mathrm{A}) + \Pr(x_i = \mathrm{B}) = 1$. If $\Pr(x_i = \mathrm{A}) = 1/2$, agent i is indifferent to either opinion.

We update the agent's belief as follows. The time t starts from $t = 0$. Upon every updating of an agent's belief, we add $1/N$ to t such that the belief of each agent is updated once per time unit on average. In an updating event, we select an agent i to be updated with equal probability $1/N$. Agent i refers to agent j's opinion for updating i's belief $\Pr(x_i = \mathrm{A})$, where j ($\neq i$) is selected with equal probability $1/(N-1)$ from the population. Agent j imparts a signal $s \in \{a, b\}$, where a and b correspond to j's opinions A and B, respectively. We assume that the probabilities that agent j imparts $s = a$ and $s = b$ are given by

$$\begin{aligned}
\Pr(s = a) &= \Pr(s = a | x_j = \mathrm{A}) \Pr(x_j = \mathrm{A}) \\
&\quad + \Pr(s = a | x_j = \mathrm{B}) \Pr(x_j = \mathrm{B}) \\
&= \theta \Pr(x_j = \mathrm{A}) + (1 - \theta) \Pr(x_j = \mathrm{B}) \\
&= 1 - \theta + (2\theta - 1) \Pr(x_j = \mathrm{A}) \qquad (1)
\end{aligned}$$

and

$$\begin{aligned}\Pr(s=b) &= \Pr(s=b|x_j=\mathrm{A})\Pr(x_j=\mathrm{A}) \\ &\quad + \Pr(s=b|x_j=\mathrm{B})\Pr(x_j=\mathrm{B}) \\ &= (1-\theta)\Pr(x_j=\mathrm{A}) + \theta\Pr(x_j=\mathrm{B}) \\ &= \theta - (2\theta-1)\Pr(x_j=\mathrm{A}), \end{aligned} \tag{2}$$

respectively, where

$$\theta = \Pr(s=a|x_j=\mathrm{A}) = \Pr(s=b|x_j=\mathrm{B}) \tag{3}$$

represents the reliability of the signal, and $1/2 \leq \theta < 1$. If j is confident in its own opinion and the transformation from j's belief [i.e., $\Pr(x_j=\mathrm{A})$] to j's output signal (i.e., a or b) is reliable, signals a and b are likely to indicate opinions A and B, respectively. In the limit $\theta \to 1$, $\Pr(s=a) \approx \Pr(x_j=\mathrm{A})$ and $\Pr(s=b) \approx \Pr(x_j=\mathrm{B})$. If $\theta = 1/2$, $\Pr(s=a) = \Pr(s=b) = 1/2$ such that s does not convey any information about j's belief. We implicitly assume that all the agents share the same value of θ and that they know this fact when performing the Bayesian update, as described below.

When agent j imparts signal $s \in \{a, b\}$, agent i is assumed to perceive a subject signal $\sigma \in \{\alpha, \beta\}$, where α and β correspond to A and B, respectively. The flow of the signal conversion is depicted in Fig. 1. If agent i is not subject to the confirmation bias, α and β are equal to a and b, respectively. Otherwise, agent i may misinterpret the signal imparted by agent j, depending on the prior exposure of agent i to other signals. Following Rabin and Schrag [37], we define

$$\begin{aligned}&\Pr[\sigma=\alpha|s=a, \Pr(x_i=\mathrm{A}) \geq 1/2] \\ &= \Pr[\sigma=\beta|s=b, \Pr(x_i=\mathrm{A}) \leq 1/2] \\ &= 1 \end{aligned} \tag{4}$$

and

$$\begin{aligned}&\Pr[\sigma=\alpha|s=b, \Pr(x_i=\mathrm{A}) > 1/2] \\ &= \Pr[\sigma=\beta|s=a, \Pr(x_i=\mathrm{A}) < 1/2] \\ &= q, \end{aligned} \tag{5}$$

where q ($0 \leq q \leq 1$) parameterizes the strength of the confirmation bias. Equation (5) states that an agent preferring opinion A misinterprets an arriving b signal as A (i.e., $\sigma = \alpha$) with probability q. If $q = 0$, the confirmation bias is absent, and $s = a$ and $s = b$ are always converted to $\sigma = \alpha$ and $\sigma = \beta$, respectively. If $q = 1$, the agent perceives the signal that is consistent with its current preference [i.e., α if $\Pr(x_i=\mathrm{A}) > 1/2$ and β if $\Pr(x_i=\mathrm{A}) < 1/2$], irrespective of the signal imparted by agent j (i.e., a or

b). The other conditional probabilities can be readily derived from Eqs. (4) and (5). For example, Eq. (4) implies

$$\begin{aligned}
&\Pr[\sigma = \beta | s = a, \Pr(x_i = \mathrm{A}) > 1/2] \\
&= 1 - \Pr[\sigma = \alpha | s = a, \Pr(x_i = \mathrm{A}) > 1/2] \\
&= 0,
\end{aligned} \tag{6}$$

and Eq. (5) implies

$$\begin{aligned}
&\Pr[\sigma = \alpha | s = a, \Pr(x_i = \mathrm{A}) < 1/2] \\
&= 1 - \Pr[\sigma = \beta | s = a, \Pr(x_i = \mathrm{A}) < 1/2] \\
&= 1 - q.
\end{aligned} \tag{7}$$

Then, by using the Bayes' theorem, we update agent i's belief $\Pr(x_i = \mathrm{A}|\sigma)$ on the basis of the old belief $\Pr(x_i = \mathrm{A})$ $[= 1 - \Pr(x_i = \mathrm{B})]$ and the perceived signal (i.e., α or β). The perceived signal may be different from the received signal (i.e., a or b) because of their confirmation bias [Eqs. (4) and (5)]. We assume that agents are not aware that they may be subject to the confirmation bias. Agents use the subjective conditional probabilities given by

$$\overline{\Pr}(\sigma = \alpha | s = a) = \overline{\Pr}(\sigma = \beta | s = b) = 1 \tag{8}$$

and

$$\overline{\Pr}(\sigma = \alpha | s = b) = \overline{\Pr}(\sigma = \beta | s = a) = 0 \tag{9}$$

to perform the Bayesian update. The posterior belief $\Pr(x_i = \mathrm{A}|\sigma)$ is given by

$$\begin{aligned}
&\Pr(x_i = \mathrm{A}|\sigma) = \frac{\Pr(\sigma | x_i = \mathrm{A}) \Pr(x_i = \mathrm{A})}{\Pr(\sigma | x_i = \mathrm{A}) \Pr(x_i = \mathrm{A}) + \Pr(\sigma | x_i = \mathrm{B}) \Pr(x_i = \mathrm{B})} \\
&= \sum_{s=a,b} \overline{\Pr}(\sigma | s) \Pr(s | x_i = \mathrm{A}) \Pr(x_i = \mathrm{A}) / \\
&\left\{ \sum_{s=a,b} \overline{\Pr}(\sigma | s) \Pr(s | x_i = \mathrm{A}) \Pr(x_i = \mathrm{A}) \right. \\
&\left. + \sum_{s=a,b} \overline{\Pr}(\sigma | s) \Pr(s | x_i = \mathrm{B}) \Pr(x_i = \mathrm{B}) \right\} \\
&= \begin{cases}
\dfrac{\theta \Pr(x_i = \mathrm{A})}{\theta \Pr(x_i = \mathrm{A}) + (1 - \theta) \Pr(x_i = \mathrm{B})} & (\sigma = \alpha), \\
\dfrac{(1 - \theta) \Pr(x_i = \mathrm{A})}{(1 - \theta) \Pr(x_i = \mathrm{A}) + \theta \Pr(x_i = \mathrm{B})} & (\sigma = \beta).
\end{cases}
\end{aligned} \tag{10}$$

It should be noted that $\Pr(x_i = \mathrm{B}|\sigma) = 1 - \Pr(x_i = \mathrm{A}|\sigma)$. Then, we increment the time by $1/N$ such that each agent is updated once per unit time on average. Iterative application of Eq. (10) leads to

$$\begin{aligned}\Pr(x_i = \mathrm{A}) &= \frac{\theta^{n_{\alpha i}}(1-\theta)^{n_{\beta i}}}{\theta^{n_{\alpha i}}(1-\theta)^{n_{\beta i}} + (1-\theta)^{n_{\alpha i}}\theta^{n_{\beta i}}} \\ &= \left\{1 + \left(\frac{1-\theta}{\theta}\right)^{n_{\alpha i} - n_{\beta i}}\right\}^{-1} \end{aligned} \tag{11}$$

and

$$\begin{aligned}\Pr(x_i = \mathrm{B}) &= \frac{(1-\theta)^{n_{\alpha i}}\theta^{n_{\beta i}}}{(1-\theta)^{n_{\alpha i}}\theta^{n_{\beta i}} + \theta^{n_{\alpha i}}(1-\theta)^{n_{\beta i}}} \\ &= \left\{1 + \left(\frac{1-\theta}{\theta}\right)^{n_{\beta i} - n_{\alpha i}}\right\}^{-1}, \end{aligned} \tag{12}$$

where $n_{\alpha i}$ ($n_{\beta i}$) is the accumulated number of signals $\sigma = \alpha$ ($\sigma = \beta$) that agent i has perceived. The state of each agent i is uniquely determined by $n_{\alpha i} - n_{\beta i}$, which is consistent with basic Bayesian theory [37, 39].

3 Results

3.1 Setup for numerical simulations

Unless otherwise stated, we set $1/2 < \theta < 1$ and assume a neutral initial condition $\Pr(x_i = \mathrm{A}) = 1/2$ ($1 \le i \le N$), or, equivalently, $n_{\alpha i} = n_{\beta i}$ ($1 \le i \le N$). The agents exchange signals and update their beliefs, possibly under a confirmation bias. After a transient, the agents believe in either opinion with a strong confidence, i.e., $\Pr(x_i = \mathrm{A}) \approx 0$ or 1. We halt a run when $|n_{\alpha i} - n_{\beta i}| \ge \Delta n_{\mathrm{c}}$ is satisfied for all i for the first time, where Δn_{c} is the threshold. In other words, a run continues if at least one agent i has the $|n_{\alpha i} - n_{\beta i}|$ value smaller than Δn_{c}.

3.2 Case without the confirmation bias

We first consider the case without a confirmation bias (i.e., $q = 0$). We investigate the dynamics of the mean belief $\overline{P}_{\mathrm{A}}(t) \equiv \sum_{i=1}^{N} \Pr(x_i = \mathrm{A})/N$ at time t by drawing a return map, i.e., $\overline{P}_{\mathrm{A}}(t)$ as a function of $\overline{P}_{\mathrm{A}}(t-1)$ [10, 40, 41]. The return map for $N = 100$, $\theta = 0.99$, and $\Delta n_{\mathrm{c}} = 500$ based on 1000 runs is shown in Fig. 2. Because $\overline{P}_{\mathrm{A}}(t) > \overline{P}_{\mathrm{A}}(t-1)$ when $0.5 < \overline{P}_{\mathrm{A}}(t-1) < 1$ and $\overline{P}_{\mathrm{A}}(t) < \overline{P}_{\mathrm{A}}(t-1)$ when $0 < \overline{P}_{\mathrm{A}}(t-1) < 0.5$, the dynamics is in accordance with majority rule behavior. All 1000 runs finished with an agreement of opinion A [i.e., $\Pr(x_i = \mathrm{A}) \approx 1$ for all i] or opinion B [i.e., $\Pr(x_i = \mathrm{A}) \approx 0$ for all i]. Each case occurred approximately half the time.

3.3 Case with the confirmation bias

We turn on the confirmation bias to examine the possibility that it induces disagreement among agents. At least for large q (i.e., $q \approx 1$), disagreement is expected to be reached because the first perceived signal would determine the final belief of each agent and is equally likely to be α and β for many agents.

In the following numerical simulations, we measured the degree of disagreement, which we defined as follows. We determined that agreement was reached in a run if the final signs of $n_{\alpha i} - n_{\beta i}$ were the same for $i = 1, \ldots, N$. Otherwise, we said that disagreement was reached. We denoted the fraction of runs that finished with disagreement by $F_{\rm d}$.

We set $\Delta n_{\rm c} = 500$ and the number of runs to 1000. In Figs. 3(a) and 3(b), $F_{\rm d}$ is shown as a function of q and θ for $N = 2$ and 100, respectively. First, $F_{\rm d}$ monotonically increases with q and decreases with θ for both $N = 2$ and $N = 100$. It should be noted that disagreement occurred in at least one run in the regions right to the solid fractured lines in Fig. 3. Second, $F_{\rm d}$ for $N = 2$ [Fig. 3(a)] is smaller than $F_{\rm d}$ for $N = 100$ [Fig. 3(b)] for all the q and θ values. Therefore, disagreement seems to be a likely outcome of the model for large N, particularly for large q and small θ. When $N = 100$, perfect agreement, i.e., $F_{\rm d} = 0$, is realized only for q close to zero. In other words, even a small degree of confirmation bias elicits disagreement among the agents.

3.4 Probability flow analysis for $N = 2$

To obtain analytical insights into the model, we performed an annealed approximation for $N = 2$ by averaging out fluctuations of the dynamics for different times and runs. The configuration of the population is specified by $(m_1, m_2) \equiv (n_{\alpha 1} - n_{\beta 1}, n_{\alpha 2} - n_{\beta 2})$. The stochastic dynamics of the model can be mapped to a random walk on the two-dimensional lattice; a walker is initially located at $(m_1, m_2) = (0, 0)$ and randomly hops to one of the four neighboring lattice points in each time step. We defined $f_R(m_1, m_2)$, $f_L(m_1, m_2)$, $f_U(m_1, m_2)$, and $f_D(m_1, m_2)$ as the probabilities that the walker located at (m_1, m_2) moves to $(m_1 + 1, m_2)$, $(m_1 - 1, m_2)$, $(m_1, m_2 + 1)$, and $(m_1, m_2 - 1)$, respectively. The four probabilities are given by

$$f_R(m_1, m_2) = \begin{cases} \dfrac{(1-q)\Pr(s = a|m_2) + q}{2} & (m_1 \geq 1), \\ \dfrac{\Pr(s = a|m_2)}{2} & (m_1 = 0), \\ \dfrac{(1-q)\Pr(s = a|m_2)}{2} & (m_1 \leq -1), \end{cases} \tag{13}$$

$$f_L(m_1, m_2) = \frac{1}{2} - f_R(m_1, m_2), \tag{14}$$

$$f_U(m_1, m_2) = \begin{cases} \dfrac{(1-q)\Pr(s=a|m_1)+q}{2} & (m_2 \geq 1), \\ \dfrac{\Pr(s=a|m_1)}{2} & (m_2 = 0), \\ \dfrac{(1-q)\Pr(s=a|m_1)}{2} & (m_2 \leq -1), \end{cases} \tag{15}$$

and

$$f_D(m_1, m_2) = \frac{1}{2} - f_U(m_1, m_2), \tag{16}$$

where

$$\begin{aligned} \Pr(s=a|m_j) &= 1-\theta+(2\theta-1)\Pr(x_j=\mathrm{A}|m_j) \\ &= 1-\theta+(2\theta-1)\left\{1+\left(\frac{1-\theta}{\theta}\right)^{m_j}\right\}^{-1} \end{aligned} \tag{17}$$

is the probability that agent j with $n_{\alpha j} - n_{\beta j} = m_j$ imparts signal $s = a$. $f_R(m_1, m_2)$ and $f_U(m_1, m_2)$ increase with m_1 and m_2, and $f_L(m_1, m_2)$ and $f_D(m_1, m_2)$ decrease with m_1 and m_2.

In the following, we study the mean dynamics of the random walk driven by the drift terms. Because the transition probability of the random walk is symmetric with respect to the lines $m_1 = m_2$ and $m_1 = -m_2$, we focus on the region given by $-m_2 \leq m_1 \leq m_2, m_2 > 0$. We define $m_{1\mathrm{c}}$ and $m_{2\mathrm{c}}$, which are not integers in general, as the values satisfying $f_U(m_{1\mathrm{c}}, m_2) = f_D(m_{1\mathrm{c}}, m_2)\ {}^\forall m_2 > 0$ and $f_R(m_1, m_{2\mathrm{c}}) = f_L(m_1, m_{2\mathrm{c}})\ {}^\forall m_1 < 0$, respectively. They are given by

$$\begin{aligned} m_{2\mathrm{c}} &= -m_{1\mathrm{c}} \\ &= \frac{\ln\left[(2\theta-1)\left\{\dfrac{1}{2(1-q)}-(1-\theta)\right\}^{-1}-1\right]^{-1}}{\ln\dfrac{\theta}{1-\theta}}. \end{aligned} \tag{18}$$

Note that $m_{1\mathrm{c}}$ and $m_{2\mathrm{c}}$ exist if and only if $(2\theta-1)\left\{1/2(1-q)-(1-\theta)\right\}^{-1}-1 > 0$, i.e.,

$$q < 1 - \frac{1}{2\theta}. \tag{19}$$

First, we consider the case $q < 1-1/2\theta$. We partition the upper quadrant of the lattice (given by $-m_2 \leq m_1 \leq m_2, m_2 > 0$) into five regions: region 1 ($0 < m_1 \leq m_2$), region 2 ($m_1 = 0, m_2 > 0$), region 3 ($-m_2 \leq m_1 < 0, 0 < m_2 < m_{2\mathrm{c}}$), region 4 ($m_{1\mathrm{c}} < m_1 < 0,\ m_2 > m_{2\mathrm{c}}$), and region 5

$(-m_2 \leq m_1 < m_{1c})$, as shown in Fig. 4(a). We obtain from the condition $1/2 < \theta < 1$

$$f_R(m_1, m_2) - f_L(m_1, m_2) \geq f_U(m_1, m_2) - f_D(m_1, m_2) > 0 \qquad (20)$$

in region 1,

$$f_R(m_1, m_2) - f_L(m_1, m_2) > 0, \quad f_U(m_1, m_2) - f_D(m_1, m_2) = \frac{q}{2} \qquad (21)$$

in region 2,

$$f_U(m_1, m_2) - f_D(m_1, m_2) \geq f_L(m_1, m_2) - f_R(m_1, m_2) > 0 \qquad (22)$$

in region 3,

$$f_R(m_1, m_2) - f_L(m_1, m_2) > 0, \quad f_U(m_1, m_2) - f_D(m_1, m_2) > 0 \qquad (23)$$

in region 4, and

$$f_R(m_1, m_2) - f_L(m_1, m_2) \geq f_D(m_1, m_2) - f_U(m_1, m_2) > 0 \qquad (24)$$

in region 5. The probability flow of the walker after the annealed approximation, i.e., $(f_R(m_1, m_2) - f_L(m_1, m_2), f_U(m_1, m_2) - f_D(m_1, m_2))$ inferred from Eqs. (20)-(24) is shown schematically in Fig. 4(a). If the walker is in the second quadrant (i.e., regions 3, 4, and 5) where the two agents disagree with each other, the random walker is likely to eventually escape and enter the first quadrant (i.e., region 1) where the two agents agree with each other. In fact, Fig. 5(a), which shows the actual probability flow, indicates that the agreement necessarily occurs. Therefore, agreement is the expected outcome when $q < 1 - 1/2\theta$.

Second, if $q > 1 - 1/2\theta$, regions 4 and 5 are absent because m_{1c} and m_{2c} diverge. Regions 1 and 2, in which inequalities (20) and (21) are satisfied, respectively, are the same as those in the case $q < 1 - 1/2\theta$. Region 3, in which inequality (22) is satisfied, is modified to $-m_2 \leq m_1 < 0$. The probability flows are schematically shown in Fig. 4(b). $f_L(m_1, m_2) > f_R(m_1, m_2)$ and $f_U(m_1, m_2) > f_D(m_1, m_2)$ are satisfied in region 3. Therefore, once the walker is deep in the second quadrant, it is likely to move toward $m_1 \to -\infty$ and $m_2 \to \infty$, which implies that two agents finally disagree. The actual probability flow shown in Fig. 5(b) is consistent with this prediction.

The transition line $q = 1 - 1/2\theta$ is shown by the dashed line in Fig. 3(a). It accurately predicts the parameter region in which disagreement can occur, i.e., the region right to the solid line.

The same transition line is also derived for the Rabin-Schrag model, which is concerned with a single agent subjected to a confirmation bias [37]. In their model, the agent forms a belief by repetitively receiving a stochastic signal $s \in \{a, b\}$ from nature, according to $\Pr(s = a|x = \mathrm{A}) =$

$(s = b|x = \mathrm{B}) = \theta$. Rabin and Schrag calculated the probability that the agent eventually misunderstands the state of the nature (i.e., A or B), starting from neutral belief. This probability is equal to zero when $q \leq 1 - 1/2\theta$ and positive when $q > 1 - 1/2\theta$ (see proposition 4 in [37]). Our results obtained in this section are consistent with theirs because disagreement in our model roughly corresponds to misunderstanding in the Rabin-Schrag model.

3.5 Different disagreement configurations for large N

In general, there are $N-1$ disagreement configurations, as distinguished by the number of agents that finally believe in opinion A, which ranges from 1 to $N-1$. To distinguish different disagreement configurations, we examined the fraction of agents that believed in the minority opinion at the end of a run. We averaged this fraction over the runs ending with disagreement. We called this quantity the average size of the minority.

Figures 6(a) and 6(b) show the average size of the minority for $N = 10$ and $N = 100$, respectively. The black regions indicate the parameter values for which the average size of the minority is undefined because all 1000 runs end with agreement. When q is small, the average size of the minority monotonically decreases with q and monotonically increases with θ for both $N = 10$ and 100. Therefore, small q and large θ values allow only balanced disagreement configurations, in which the numbers of the agents believing in the opposite opinions are close to $N/2$.

However, the average size of the minority increases when q is large. This is particularly the case for $N = 100$ [Fig. 6(b)]. This increase occurs for the following reason. With a strong confirmation bias, agents end up with an opinion consistent with a small number of signals perceived in the early stages, and both signals are equally likely to be observed in the early stages under neutral initial conditions. In the extreme case in which $q = 1$, agents reinforce the opinion that is consistent with their first perceived signal. Therefore, unbalanced disagreement configurations are rarely realized when q is large.

3.6 Effects of the system size and initial condition

Figures 3 and 6 suggest that the agreement is unlikely to be reached in a large population. To examine the effect of the population size, we defined q_{c} as the value of q such that a threshold number of runs among 10^4 runs end with agreement. For a given θ value, we determined q_{c} by the bisection method. The number of agreement runs may not monotonically change in q because the number of runs is finite. Therefore, the bisection method does not perfectly work in general. However, we corroborated that the following results were negligibly affected by the lack of monotonicity.

The dependence of q_c on N is shown in Fig. 7(a) for three threshold values. For example, the results for the threshold value 100 (shown by circles) indicate that at least 100 runs among the 10^4 runs end up with disagreement when $q > q_c$. We set $\theta = 0.99$ and $\Delta n_c = 500$. To explore the possibility of disagreement in large populations, we set θ close to 1. It should be noted that Fig. 3 indicates that the probability of disagreement is small for a large θ value. In Fig. 7(a), q_c quickly decreases for $N \leq 10$ and gradually decreases for $N \geq 100$. Disagreement often occurs for large N unless q is small. Nevertheless, Fig. 7(a) suggests that the range of q for which agreement always occurs survives for diverging N.

In generating Fig. 7(a), we used an initial condition in which all the agents had a neutral belief [i.e., $\Pr(x_i = \mathrm{A}) = 0.5, 1 \leq i \leq N$]. To check the effect of the initial condition, we investigated the dependence of q_c on N under two other initial conditions. In the bimodal initial condition, we initially set $n_{\alpha i} - n_{\beta i} = 100$ $(1 \leq i \leq N/2)$ and $n_{\alpha i} - n_{\beta i} = -100$ $(N/2+1 \leq i \leq N)$. We assumed that N was even for this initial condition. In the so-called most unbalanced initial condition, we set $n_{\alpha 1} - n_{\beta 1} = 100$ and $n_{\alpha i} - n_{\beta i} = -100$ $(2 \leq i \leq N)$.

The numerical results for the two initial conditions are shown in Figs. 7(b) and 7(c). The parameter values $\theta = 0.99$, $\Delta n_c = 500$ are the same as those used in Fig. 7(a). The transition point q_c decreases with N more rapidly with the bimodal initial condition [Fig. 7(b)] than with the neutral initial condition [Fig. 7(a)]. This result is intuitive: the bimodal initial condition paves the way to disagreement. In contrast, q_c under the most unbalanced initial condition is almost constant near 0.5 irrespective of N. Therefore, disagreement is highly unlikely unless the confirmation bias is strong (i.e., q is greater than 0.5). The results shown in Fig. 7 suggest that the eventual behavior of the model strongly depends on the initial condition even after the results are averaged over runs.

4 Discussion

Our numerical results are summarized as follows. When the confirmation bias is absent (i.e., $q = 0$), the opinion dynamics under the Bayesian update rule leads to the complete agreement among agents. The behavior of the model is similar to majority rule dynamics (Fig. 2). When the confirmation bias is present, disagreement is a likely outcome, particularly for a strong confirmation bias (i.e., large q). Disagreement is also more likely for a lower fidelity of the signal (i.e., $\theta \approx 1/2$) and a larger system size. The transition line separating the parameter region in which both agreement and disagreement can occur and that in which only agreement occurs is approximately given by $q = 1 - 1/2\theta$ when $N = 2$. This line is identical to the one determined by Rabin and Schrag for their model for a single agent's decision

making [37]. Finally, the behavior of the model strongly depends on the initial condition.

Our model and results are different from Orléan's [38], although Orléan's model employs multiple agents that perform the Bayesian updates under a confirmation bias. First, the belief of each agent is binary in Orléan's model, whereas our model introduces an infinite range of discrete beliefs, as in [37]. Second, interaction between agents is introduced differently in the two models. In Orléan's model, each agent refers to the global fraction of agents believing in one of the two opinions. In our model, agents refer to other opinions by peer-to-peer interaction, i.e., by receiving a binary signal that is correlated with the belief of the sender. Third, the stochastic dynamics of Orléan's model is ergotic when the collective opinion does not reach agreement. The collective opinion obeys a stationary distribution, irrespective of the initial condition. In contrast, in all our simulations, the stochastic dynamics of our model was nonergotic, such that the final configuration depended on the initial condition in a wide parameter region.

In social science studies of polarization, several authors analyzed Bayesian models in which different agents receiving a series of common signals end up in disagreement. The proposed mechanisms governing disagreement include different initial beliefs or factors that affect perception of later incoming signals [27–31], different update rules [28], and ambiguity aversion [28, 42]. These models and ours are different in three major ways. First, a ground truth opinion corresponding to the state of nature is assumed in these models but not in ours. Second, public signals commonly received by different agents are assumed in these models but not in ours. Third, the agents do not have direct peer-to-peer interaction in these models, but they do in ours. Models with interacting Bayesian agents, which show disagreement (reviewed in Ref. [30]), are also different from our model in the first respect. It should be noted that Zimper and Ludwig discussed confirmation bias with their Bayesian model [28]. However, they derived a confirmation bias from their model, rather than assuming one, such that their results pertaining to confirmation bias were also distinct from ours.

Extending our model to the case of networks is straightforward. For example, we can select a recipient of the signal with probability $1/N$ and then select the sender with equal probability among the neighbors of the recipient on the network. Another possible update rule is to select the sender first and then the recipient among the sender's neighbors. Yet another possibility is to select a link with equal probability and designate one of the two agents as sender and the other as recipient. On heterogeneous networks, the results may depend on the update rule because it is the case in the voter model [7–9, 43]. Extension of the model to the case of confirmation bias heterogeneity may also be interesting. Neurological evidence shows that different individuals have different confirmation bias strengths [44]. The strength of the confirmation bias and the position of the node in a social

network may be correlated and affect the dynamics. It is also straightforward to extend the model to the case of multiple opinion cases. These and other extensions, along with the study of analytically tractable models that capture the essence of the present study, warrant future work.

5 Acknowledgments

We thank Mitsuhiro Nakamura, Taro Takaguchi, and Shoma Tanabe for critical reading of the manuscript. This work is supported by Grants-in-Aid for Scientific Research [Grant No. 23681033 and Innovative Areas "Systems Molecular Ethology" (Grant No. 20115009)] from MEXT, Japan.

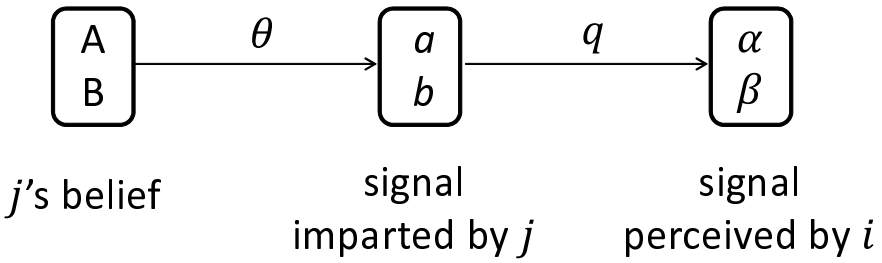

Figure 1: Schematic of signal conversion. Signals a and α correspond to A. Signals b and β correspond to B. θ is the reliability of the signal, and q is the strength of the confirmation bias.

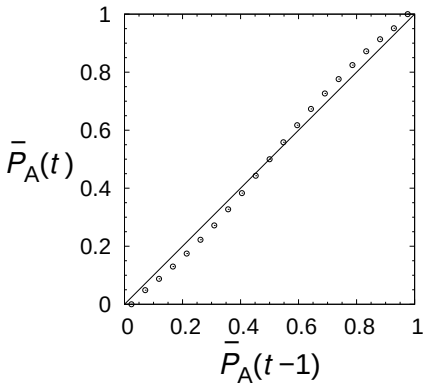

Figure 2: Return map of the mean belief. We set $N = 100$, $q = 0$, and $\theta = 0.99$. We recorded the values of $(\overline{P}_{\mathrm{A}}(t-1), \overline{P}_{\mathrm{A}}(t))$ for $t = 1, 1+1/N, 1+2/N, \ldots$ for 1000 runs and divided the recorded pairs into 21 classes. The kth class ($k = 1, \ldots, 21$) was composed of the pairs satisfying $(k-1)/21 \leq \overline{P}_{\mathrm{A}}(t-1) < k/21$. We obtained the mean value $\left\langle \overline{P}_{\mathrm{A}}(t) \right\rangle_k$ for the kth class by averaging $\overline{P}_{\mathrm{A}}(t)$ over all the pairs contained in the kth class. Finally, we plotted $\left\langle \overline{P}_{\mathrm{A}}(t) \right\rangle_k$ against $(k-0.5)/21$ for $k = 1, \ldots, 21$. The diagonal $\overline{P}_{\mathrm{A}}(t) = \overline{P}_{\mathrm{A}}(t-1)$ is also shown as a guide.

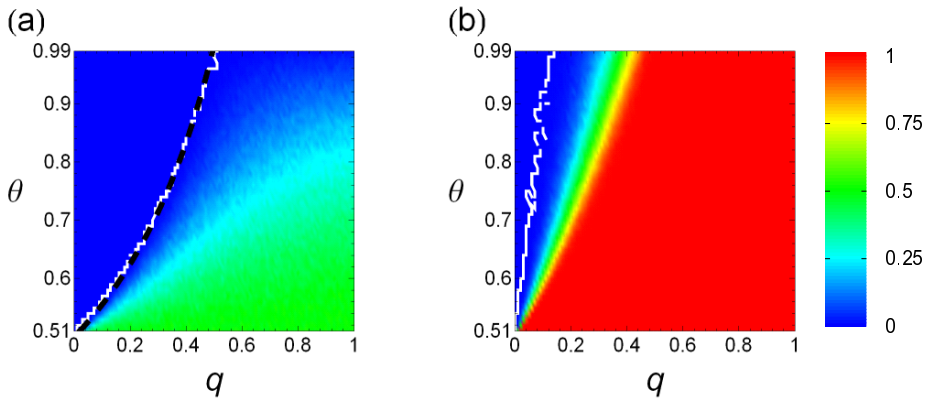

Figure 3: Fraction of disagreement F_{d}. (a) $N = 2$. (b) $N = 100$. Solid lines represent the boundary between $F_{\mathrm{d}} = 0$ and $F_{\mathrm{d}} > 0$. The dashed line in (a) represents $q = 1 - 1/2\theta$. The dashed line is not drawn in (b) because this theoretical estimate is valid only for $N = 2$. In (a), the two lines almost overlap each other. The initial belief of each agent was assumed to be neutral [i.e., $\Pr(x_i = \mathrm{A}) = 0.5$, $1 \leq i \leq N$].

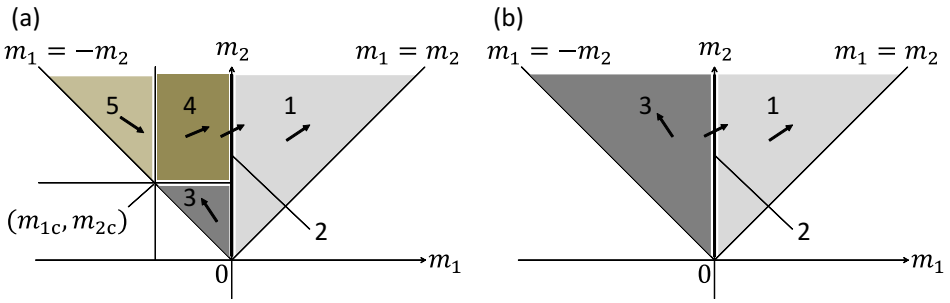

Figure 4: Schematic of the probability flow given by $f_R(m_1, m_2) - f_L(m_1, m_2)$ and $f_U(m_1, m_2) - f_D(m_1, m_2)$. (a) $q < 1 - 1/2\theta$. (b) $q > 1 - 1/2\theta$. The labels from 1 to 5 correspond to the five regions.

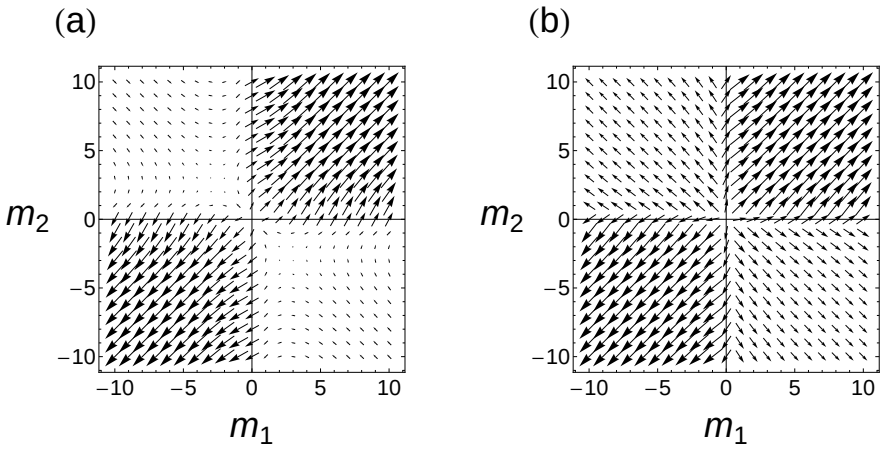

Figure 5: Probability flow of the opinion dynamics when $N = 2$. Vector $(f_R(m_1, m_2) - f_L(m_1, m_2), f_U(m_1, m_2) - f_D(m_1, m_2))$ is shown by an arrow of proportional size at each position of the random walker (m_1, m_2). (a) $q = 0.15$ and $\theta = 0.64$, which satisfies $q < 1 - 1/2\theta$. (b) $q = 0.4$ and $\theta = 0.64$, which satisfies $q > 1 - 1/2\theta$. The size of the vectors is manually normalized for clarity, independently for the two panels.

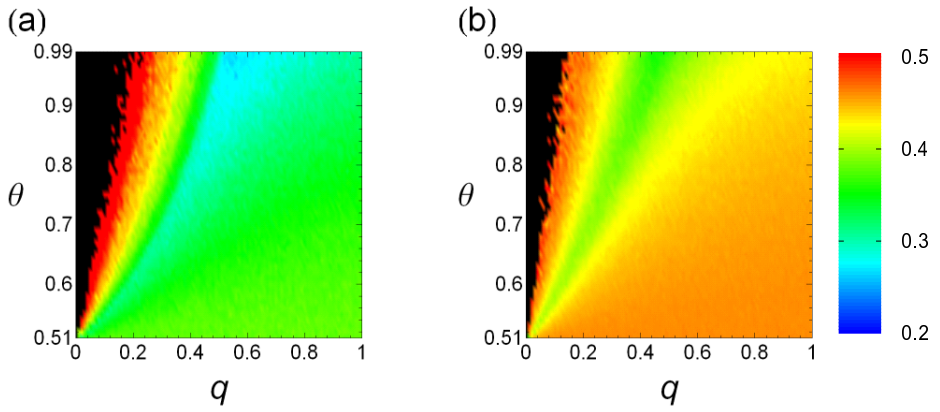

Figure 6: Average size of the minority. (a) $N = 10$ and (b) $N = 100$. The initial belief of each agent is assumed to be neutral [i.e., $\Pr(x_i = \mathrm{A}) = 0.5$, $1 \leq i \leq N$]. The black region represents the case where all the 1000 runs end with agreement such that the average size of the minority is undefined.

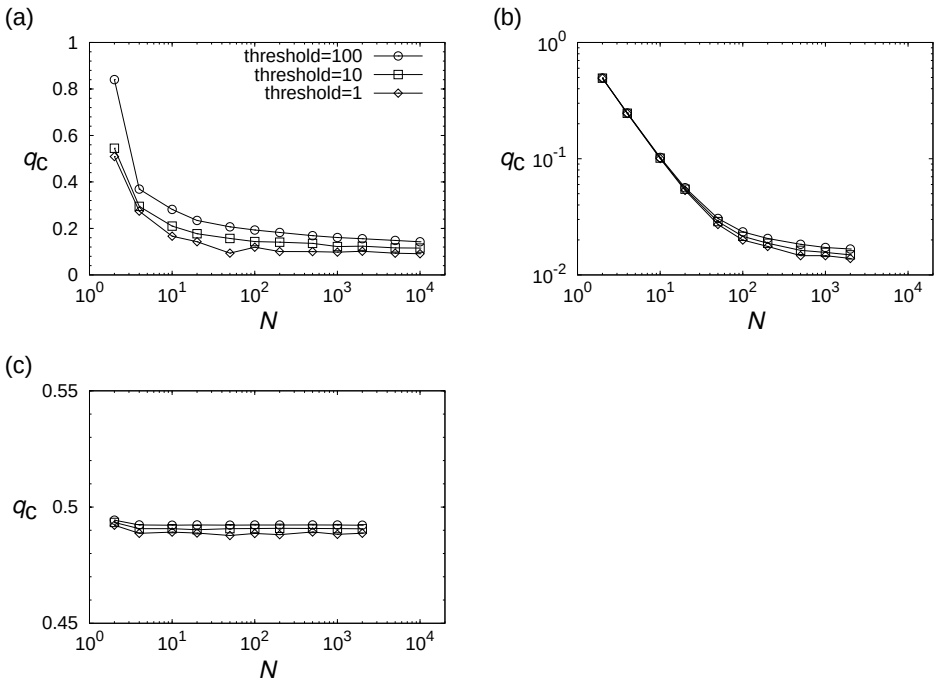

Figure 7: Threshold for the agreement-disagreement transition (q_c) as a function of N under the (a) neutral, (b) bimodal, and (c) most unbalanced initial conditions. We set $\theta = 0.99$.

References

[1] C. Castellano, S. Fortunato, and V. Loreto, Rev. Mod. Phys. **81**, 591 (2009).

[2] P. L. Krapivsky, S. Redner, and E. Ben-Naim, *A Kinetic View of Statistical Physics* (Cambridge University Press, Cambridge, 2010).

[3] T. Kuran, *Private Truths, Public Lies: The Social Consequences of Preference Falsification* (Harvard University Press, Cambridge, MA, 1995).

[4] R. Huckfeldt, P. E. Johnson, and J. Sprague, *Political Disagreement: The Survival of Diverse Opinions within Communication Networks* (Cambridge University Press, Cambridge, 2004).

[5] P. Donnelly and D. Welsh, Math. Proc. Cambridge Philos. Soc. **94**, 167 (1983).

[6] V. Sood and S. Redner, Phys. Rev. Lett. **94**, 178701 (2005).

[7] T. Antal, S. Redner, and V. Sood, Phys. Rev. Lett. **96**, 188104 (2006).

[8] V. Sood, T. Antal, and S. Redner, Phys. Rev. E **77**, 041121 (2008).

[9] N. Masuda and H. Ohtsuki, New J. Phys. **11**, 033012 (2009).

[10] S. Galam, Eur. Phys. J. B **25**, 403 (2002).

[11] P. Chen and S. Redner, Phys. Rev. E **71**, 036101 (2005).

[12] G. Deffuant, D. Neau, F. Amblard, and G. Weisbuch, Adv. Complex Syst. **3**, 87 (2000).

[13] M. Patriarca and T. Leppänen, Physica A **338**, 296 (2004).

[14] M. Patriarca and E. Heinsalu, Physica A **388**, 174 (2009).

[15] J. Mira and Á. Paredes, Europhys. Lett. **69**, 1031 (2005).

[16] J. Mira, L. F. Seoane, and J. J. Nieto, New J. Phys. **13**, 033007 (2011).

[17] X. Castelló, V. M. Eguíluz, and M. San Miguel, New J. Phys. **8**, 308 (2006).

[18] F. Vazquez, X. Castelló, and M. San Miguel, J. Stat. Mech. P04007 (2010).

[19] L. Chapel, X. Castelló, C. Bernard, G. Deffuant, V. M. Eguíluz, S. Martin, and M. San Miguel, PLOS ONE **5**, e8681 (2010).

[20] R. Fujie, K. Aihara, and N. Masuda, J. Stat. Phys. **151**, 289 (2013).

[21] F. Vazquez, V. M. Eguíluz, and M. San Miguel, Phys. Rev. Lett. **100**, 108702 (2008).

[22] B. Kozma and A. Barrat, Phys. Rev. E **77**, 016102 (2008).

[23] P. Holme and M. E. J. Newman, Phys. Rev. E **74**, 056108 (2006).

[24] C. Nardini, B. Kozma, and A. Barrat, Phys. Rev. Lett. **100**, 158701 (2008).

[25] N. Masuda, N. Gibert, and S. Redner, Phys. Rev. E **82**, 010103(R) (2010).

[26] N. Masuda and S. Redner, J. Stat. Mech. L02002 (2011).

[27] A. K. Dixit and J. W. Weibull, Proc. Natl. Acad. Sci. U.S.A. **104**, 7351 (2007).

[28] A. Zimper and A. Ludwig, J. Risk Uncertainty **39**, 181 (2009).

[29] D. Acemoglu, C. Victor, and Y. Muhamet, "Fragility of asymptotic agreement under bayesian learning", http://econwww.mit.edu/files/3795 (2009, Accessed: Febrary 1, 2013).

[30] D. Acemoglu and A. Ozdaglar, Dyn. Games Appl. **1**, 3 (2011).

[31] J. Andreoni and T. Mylovanov, Am. Econ. J. Microecon. **4**, 209 (2012).

[32] K. Binmore, *Rational Decisions* (Princeton University Press, Princeton, NJ, 2008).

[33] A. C. R. Martins, J. Stat. Mech. P02017 (2009).

[34] S. Plous, *The Psychology of Judgment and Decision Making* (McGraw-Hill, New York, 1993).

[35] R. S. Nickerson, Rev. Gen. Psychol. **2**, 175 (1998).

[36] G. Deffuant and S. Huet, Complexity **15**(5), 25 (2010).

[37] M. Rabin and J. L. Schrag, Q. J. Econ. **114**, 37 (1999).

[38] A. Orléan, J. Econ. Behav. Organ. **28**, 257 (1995).

[39] A. Pérez-Escudero and G. G. de Polavieja, PLOS Comput. Biol. **7**, e1002282 (2011).

[40] J. M. Pacheco, F. C. Santos, M. O. Souza, and B. Skyrms, Proc. R. Soc. B **276**, 315 (2009).

[41] J. M. Pacheco, F. L. Pinheiro, and F. C. Santos, PLOS. Comput. Biol. **5**, e1000596 (2009).

[42] S. Baliga, E. Hanany, and P. Klibanoff, "Polarization and ambiguity", Am. Econ. Rev. (to be published).

[43] N. Masuda, J. Theor. Biol. **258**, 323 (2009).

[44] B. B. Doll, K. E. Hutchison, and M. J. Frank, J. Neurosci. **31**, 6188 (2011).