Can Partisan Voting Lead to Truth?

Naoki Masuda[1,2] and S. Redner[3]

[1] Graduate School of Information Science and Technology, The University of Tokyo, 7-3-1 Hongo, Bunkyo, Tokyo 113-8656, Japan

[2] PRESTO, Japan Science and Technology Agency, 4-1-8 Honcho, Kawaguchi, Saitama 332-0012, Japan

[3] Center for Polymer Studies and Department of Physics, Boston University, Boston, MA USA 02215

E-mail: masuda@mist.i.u-tokyo.ac.jp

Abstract. We study an extension of the voter model in which each agent is endowed with an innate preference for one of two states that we term as "truth" or "falsehood". Due to interactions with neighbors, an agent that innately prefers truth can be persuaded to adopt a false opinion (and thus be discordant with its innate preference) or the agent can possess an internally concordant "true" opinion. Parallel states exist for agents that inherently prefer falsehood. We determine the conditions under which a population of such agents can ultimately reach a consensus for the truth, a consensus for falsehood, or reach an impasse where an agent tends to adopt the opinion that is in internal concordance with its innate preference so that consensus is never achieved.

PACS numbers: 02.50.-r, 05.40.-a, 89.75.Fb

There has been considerable recent public attention in the US on the following question: "Is President Obama a Muslim?" A Pew Research Center public opinion poll in the US found that 19% of respondents answered this question affirmatively in August 2010 [1]. Independent of one's personal views, there are two incontrovertible facts on this matter: (i) the question "Is President Obama a Muslim?" has a definite answer, and (ii) there is a lack of consensus on the answer to what seems to be a clear-cut question.

What is the source of this lack of consensus? Does conflicting evidence exist? Do people interpret the same evidence differently? What are the roles of the mass media and fellow peers in influencing the beliefs of individuals? In this work, we focus on peer influence by formulating a voter-like model in which we account for the possibility that each agent's opinion may be in concordance or discordance with an internal and fixed set of beliefs. Our model incorporates two distinct mechanisms for an agent to change its opinion: one is the desire for an individual to be in concordance with this inviolate belief set, while the second is the tendency to agree with neighbors to minimize conflict. Our goal is to shed light on whether global truth can ultimately emerge in a socially-interacting population where the desire for internal concordance may conflict with global truth. Our approach is in the spirit of work by physicists in applying minimalist modeling to real social phenomena (see, e.g., [2, 3, 4, 5]).

The dichotomy between one's private beliefs and publicly-expressed opinions can take many forms [6]. One such example is H. C. Andersen's fairy tale "The Emperor's New Clothes" [7], in which the Emperor's absence of clothes is finally exposed by the innocent remark of young child. The phenomenon of the spread of false norms has recently been elucidated by a simple computational model [8]. In a related vein, Asch [9] found that respondents to an uncomfortable question would conform to a clearly false consensus judgment rather than risk the stigma of being viewed as deviant. Similarly, white Americans overestimated the degree of support for forced racial segregation during the 1960's and 70's [10]. While 18% of whites indicated that they favored segregation, 47% of these respondents believed that most whites favored segregation. Another relevant example is the maddening feature that the general public has strong divergences of opinion on issues for which there is agreement within the scientific community [11].

Motivated in part by these examples, we construct a simple generalization of the voter model — the "partisan" voter model — to account for the competing influences of social consensus and personal concordance. The classic voter model accounts for the evolution to consensus in a population of binary and spineless agents that repeatedly evolve by adopting the opinion state of a randomly-selected neighbor [12]. In our model, we posit that each individual has an internal ethos — namely, a fixed beliefs on a set of fundamental issues, and that each agent is more likely to alter its opinion to be in concordance with this internal ethos [13]. We also ascribe a value to the opinion states that we label as "truth" (T) and "falsehood" (F). Individuals evolve by voter model dynamics but the rate of an update step depends on the direction of the opinion change. An opinion change that makes an agent concordant with its internal ethos occurs preferentially to a change that causes discordance. As a result of interactions among neighboring agents by voter model dynamics, an agent that intrinsically prefers the truth can thus actually have a false opinion. The resulting model bears some resemblance to a multi-state voter model of Page et al. [14], in which an agent can self-adjust its vector set of opinions to be internally consistent in addition to regular voting dynamics.

In general, there are four possibilities for the opinion state of each agent (figure 1):

- An agent intrinsically prefers truth and is in the true state. Such an agent is *concordant*.
- An agent intrinsically prefers falsehood and is in the false state. Such an agent is also concordant.
- An agent intrinsically prefers truth but is in the false state. Such an agent is *discordant*.
- An agent intrinsically prefers falsehood but is in the true state. Such an agent is also discordant.

We define the densities of agents in these four states as T_+, F_+, T_-, and F_-, respectively, each of which evolves with time t. The subscripts $+$ and $-$ indicate concordance and discordance, respectively. The total density of agents that intrinsically believe the truth (independent of their instantaneous state) is the time-independent quantity $T = T_+ + T_-$ (figure 1), and similarly the density of agents that intrinsically believe falsehood is $F = F_+ + F_-$. However, the density of agents that happen to be in the true state at a particular time is time-dependent quantity $T_+ + F_-$; this includes concordant agents that prefer truth and discordant agents that prefer falsehood but happen to be in the true state. Similarly, the density of agents that happen to be in the false state is $F_+ + T_-$. The total density of agents of any type, *i.e.*, $T + F$, is fixed and may be set to 1 without loss of generality. Our goal is to understand if global truth can emerge by this voting dynamics.

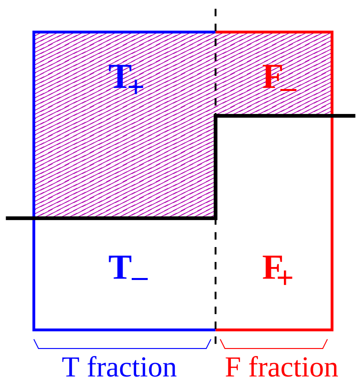

Figure 1. Illustration of the densities of the four types of agents, T_+, F_+, T_-, and F_-. The shaded region shows the fraction of the population in the true state; this includes a fraction T_+ of concordant agents that intrinsically prefer the truth and a fraction F_- of discordant agents that intrinsically prefer falsehood but happen to agree with the truth.

We define the evolution of the agents by the following extension of the voter model: first, an agent is picked at random and then one of its neighboring agents is picked. When these two agents are in opposite opinion states (truth or falsehood), the initial agent is updated with the following rates:

- A T agent that happens to have a false opinion changes to the true (concordant) state with rate $1+\epsilon_t$; a T agent that happens to have a true opinion changes to the false (discordant) state with rate $1-\epsilon_t$.
- An F agent that happens to have a true opinion changes to the false (concordant) state with rate $1+\epsilon_f$; an F agent that happens to have a false opinion changes to the true (discordant) state with rate $1-\epsilon_f$.

That is, an agent changes to a concordant state with enthusiasm (as quantified by the enhanced rates $1+\epsilon_t$ or $1+\epsilon_f$ for changing to the T or F states, respectively, where $\epsilon_t, \epsilon_f > 0$), while an agent changes to a discordant state with reluctance (reduced rates $1-\epsilon_t$ or $1-\epsilon_f$).

To understand the evolution of the population by this partisan voter dynamics we first study the mean-field approximation, where all agents can be viewed as neighbors of each other. This approximation provides a useful starting point for investigating the dynamics of the model on more realistic social networks. In this mean-field description, the densities of agents in each of the four possible states evolve by the rate equations:

$$\begin{aligned}
\dot{T}_+ &= (1+\epsilon_t)T_-\left[T_+ + F_-\right] - (1-\epsilon_t)T_+\left[T_- + F_+\right] ,\\
\dot{T}_- &= (1-\epsilon_t)T_+\left[T_- + F_+\right] - (1+\epsilon_t)T_-\left[T_+ + F_-\right] ,\\
\dot{F}_+ &= (1+\epsilon_f)F_-\left[F_+ + T_-\right] - (1-\epsilon_f)F_+\left[F_- + T_+\right] ,\\
\dot{F}_- &= (1-\epsilon_f)F_+\left[F_- + T_+\right] - (1+\epsilon_f)F_-\left[F_+ + T_-\right] ,
\end{aligned} \tag{1}$$

where the overdot denotes time derivative. The terms in (1) have simple meanings. For example, in the equation for $\dot{T}_+$, the first term on the right side accounts for the increase (with rate $1+\epsilon_t$) in the density of concordant agents that are in the true state. Such a change arises when a discordant agent that prefers truth interacts with an agent that currently is in the true state — either an agent that prefers truth and is concordant with this internal preference, or an agent that prefers falsehood but happens to be discordant with its internal preference. The equation for $\dot{T}_-$ can be understood similarly. The equations for $\dot{F}_\pm$ are the same as those for $\dot{T}_\pm$ by the transformation $T_\pm \leftrightarrow F_\pm$ and $t \leftrightarrow f$. Notice, as expected, that the densities of agents that intrinsically prefer truth or falsehood are constant, that is, $\dot{T} = \dot{T}_+ + \dot{T}_- = 0$ and similarly $\dot{F} = \dot{F}_+ + \dot{F}_- = 0$.

Without loss of generality and as one might reasonably expect, we assume that a majority of the agents intrinsically prefer the truth; that is, $T \geq \frac{1}{2}$. To analyze the rate equations (1), we introduce the linear combinations $S = T_+ + F_+$ (that ranges between 0 and 1), and $\Delta = T_+ - F_+$ (that ranges between $T-1=-F$ and T). Using $T_+ = \frac{1}{2}(S+\Delta)$ and $F_+ = \frac{1}{2}(S-\Delta)$, as well as $T_- = T - T_+$ and $F_- = F - F_+$, we rewrite the rate equations (1) in terms of Δ, S, and T. Straightforward algebra leads to

$$\dot{S} = (2+\epsilon_t+\epsilon_f)T(1-T) - (\epsilon_t+\epsilon_f)\Delta^2 + \Delta\left[2(1+\epsilon_t+\epsilon_f)T - (1+\tfrac{1}{2}\epsilon_t+\tfrac{3}{2}\epsilon_f)\right]$$
$$-S + (\epsilon_t-\epsilon_f)\left[S(T-\Delta) - \tfrac{1}{2}S\right] \qquad (2a)$$
$$\dot{\Delta} = (\epsilon_t+\epsilon_f)(T-\Delta)S + \epsilon_f(\Delta - S) +$$
$$+(\epsilon_t-\epsilon_f)\left[T(1-T) + 2T\Delta - \Delta^2 - \tfrac{1}{2}(S+\Delta)\right]. \qquad (2b)$$

To understand the dynamical behavior of these equations, first consider the symmetric (and simpler) case of $\epsilon_t = \epsilon_f \equiv \epsilon$, in which each type of agent has the same degree of preference toward internal concordance and the same reluctance for discordance. For this special case, the nullclines of equations (1), $\dot{S} = 0$ and $\dot{\Delta} = 0$, respectively, are:

$$\begin{aligned} &S = 2(1+\epsilon)T(1-T) - 2\epsilon\Delta^2 + (1+2\epsilon)(2T-1)\Delta \equiv f(\Delta)\,, \\ &\left(S-\tfrac{1}{2}\right)\left[\Delta - \tfrac{1}{2}(2T-1)\right] = \tfrac{1}{4}(2T-1)\,. \end{aligned} \qquad (3)$$

The locus $S = S(\Delta)$ is monotonically increasing when $\epsilon < 2T-1$ but has a single maximum when $\epsilon > 2T-1$, while the locus $\dot{\Delta} = 0$ consists of two hyperbolae in the Δ-S plane (figure 2). The intersections of these loci give the fixed points of the dynamics. As illustrated in figure 2, there are two generic situations. For weak intrinsic bias, $\epsilon < 2T-1$, there is a stable fixed point at $(\Delta^*, S^*) = (T, T)$ that corresponds to truth consensus. In terms of the original densities $T_\pm$ and $F_\pm$, this fixed point is located at $T_+ = T$ and $F_+ = 0$. Thus, truth consensus is reached in which the population consists of: (i) agents that intrinsically prefer the truth and are concordant, and/or (ii) agents that intrinsically prefer falsehood but are discordant (see figure 1). There is also an unstable fixed point at $(\Delta^*, S^*) = (T-1, 1-T)$ that corresponds to falsehood consensus. For $\epsilon > 2T-1$, there exists an additional stable fixed point located at $(\Delta^*, S^*) = \left(\frac{1}{2\epsilon}(1+\epsilon)(2T-1), \frac{1}{2}(1+\epsilon)\right)$ in the interior of the Δ-S plane. This fixed point corresponds to a mixed state of *impasse*, where each agent tends to be aligned with its own internal preference so that global consensus is not possible.

This same qualitative picture about the nature of the fixed points continues to hold for general bias parameters $\epsilon_t \neq \epsilon_f$ — consensus can be reached only when agents are not too strongly biased toward internal concordance, while an impasse is reached otherwise. A striking consequence of this model is that there is a substantial range for the bias parameters ϵ_t and ϵ_f — and for values that seem socially realistic — for which a consensus of the truth is *not* reached. To quantify this statement, we fix the density of agents that intrinsically prefer the truth and study the final outcome of the dynamics as a function of ϵ_t and ϵ_f. We make the natural assumption that a majority of agents intrinsically prefer truth, $T \geq \frac{1}{2}$. The transition between impasse and falsehood consensus occurs when the fixed points corresponding to these two collective states coincide. This gives the criterion

$$\Delta^* = T - 1 = \frac{T(2\epsilon_t\,\epsilon_f + \epsilon_t + \epsilon_f) - \epsilon_t\,\epsilon_f - \epsilon_f}{2\epsilon_t\epsilon_f}\,. \qquad (4)$$

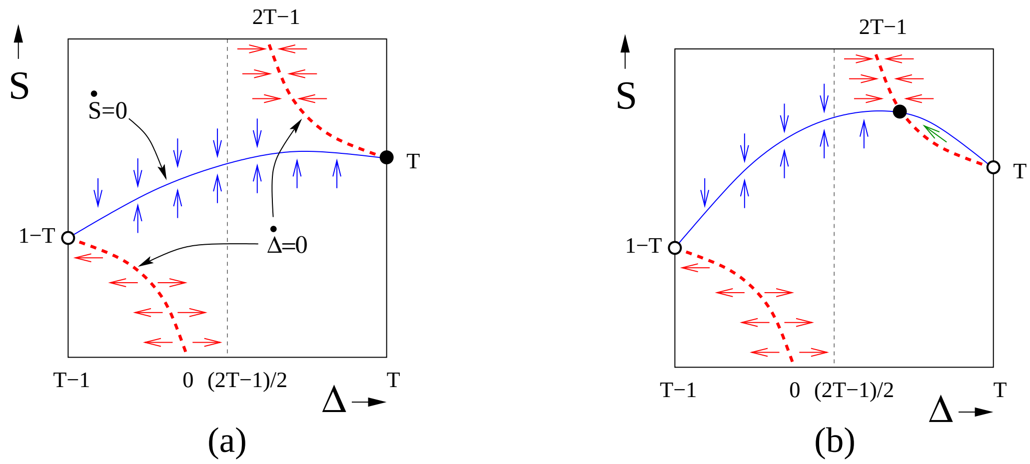

Figure 2. Flow diagrams corresponding to the rate equations (1) when $\epsilon_t = \epsilon_f \equiv \epsilon$ in the cases of: (a) weak bias $\epsilon < 2T-1$ and (b) strong bias $\epsilon > 2T-1$. The nullclines defined by $\dot{S}=0$ (solid) and $\dot{\Delta}=0$ (dashed) are shown, as well as the local flows near each nullcline. The stable and unstable fixed points are denoted by filled and open circles, respectively.

Similarly, the transition between impasse and truth consensus is defined by

$$\Delta^* = T = \frac{T(2\epsilon_t\,\epsilon_f + \epsilon_t + \epsilon_f) - \epsilon_t\,\epsilon_f - \epsilon_f}{2\epsilon_t\epsilon_f}\,. \tag{5}$$

These two equalities are satisfied when

$$\epsilon_f = T\epsilon_t/(1-T-\epsilon_t)\,, \tag{6a}$$

$$\epsilon_f = T\epsilon_t/(1-T+\epsilon_t)\,, \tag{6b}$$

respectively. These two loci are plotted in figure 3(a) for the representative case of $T = 0.8$. The region between these two curves defines the part of the parameter space where the final state is that of impasse. Surprisingly (or perhaps not), even when 80% of the population intrinsically prefers truth, the impasse state can be reached even when the bias ϵ_t toward the truth is stronger than the bias ϵ_f toward falsehood. Moreover, this impasse state is reached for a wide range of bias parameters ϵ_t and ϵ_f.

To test these predictions and obtain additional insights into the partisan voter model, we simulated the partisan voter model on the ring and on the periodic square lattice. We studied two natural initial conditions: (i) a random system and (ii) a droplet state. In the former, we fix the fraction of voters that intrinsically believe the truth to be $T = 0.8$. For this population, one-half of the agents are assigned the true opinion, independent of their intrinsic beliefs. In the latter case, the system contains a small cluster of agents that believe falsehood and also possess the false opinion, while all other agents believe the truth and possess the true opinion. We study the evolution from these two initial states.

Figure 3(b) gives a temperature plot of the fractions of realizations that reach consensus (either truth or falsehood) within a cutoff time 5000 as a function of ϵ_t and

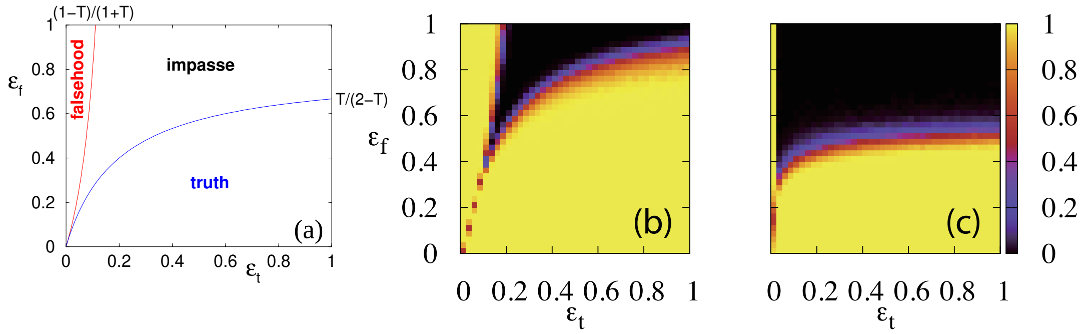

Figure 3. Phase diagrams of the partisan voter model for the representative case of $T = 0.8$. (a) The attractor state of the rate equations (1) as a function of the biases ϵ_t and ϵ_f. (b) Simulation results on a 20×20 square lattice for the randomly distributed initial state and preference. We set $T = 0.8$. (c) Simulation results on a 20×20 square lattice for the clustered initial state and preference of 4×4 agents.

ϵ_f for the random initial condition on the square lattice; qualitatively similar results occur on the ring. The bright region for small ϵ_t and large ϵ_f corresponds to the system reaching falsehood consensus within the cutoff time. Conversely, the bright region for large ϵ_t and small ϵ_f corresponds to reaching truth consensus within the cutoff. In the dark region, consensus has not yet been reached by the cutoff; we interpret this behavior as corresponding to impasse. The simulation result qualitatively mirrors that predicted by the rate equations in figure 3(a). Since consensus must eventually occur in simulations of a finite system, the lack of consensus by the cutoff indicates that the consensus time grows exponentially with the system size in the dark region.

Figure 3(c) illustrates the evolution for an initial 4×4 droplet of falsehood (both in belief and in current opinion) in a background of agents that believe the truth and possess the true opinion; again, we obtain qualitatively similar results for the corresponding one-dimensional system. Although a fraction $T = 1 - 16/400 = 0.96$ of the agents prefer the truth and are initially in the true state, impasse (in the sense outlined above) is the long-time outcome for a substantial range of bias parameters. That is, a small fraction of provocateurs — in the form of agents that steadfastly believe falsehood — can forestall the attainment of consensus. Moreover, the range of bias parameters for which impasse is reached is substantial and impasse can arise even when partisans for truth are more strident than partisans for falsehood.

To summarize, we extended the voter model to include a competition between conformity — the tendency to agree with one's neighbors — and partisanship — the desire to be in concordance with one's fixed personal ethos. Analysis of the rate equations shows that consensus can be prevented over a substantial range of model parameters. A sobering implication of our model is that a small minority of sufficiently partisan agents that are inherently predisposed to falsehood can prevent the attainment of consensus to truth.

Acknowledgments

We thank James Fowler for correspondence that helped motivate this work, and Damon Centola for discussions and literature advice. NM acknowledges financial support by the Grants-in-Aid for Scientific Research (No. 20760258) from MEXT, Japan. SR acknowledges support from the US National Science Foundation grant DMR0906504.

References

[1] http://pewresearch.org/pubs/1701/
[2] S. Galam and S. Moscovici, Eur. J. Soc. Psych. **21**, 49 (1991).
[3] S. Galam and F. Jacobs, Physica A **381**, 366 (2007).
[4] C. Castellano, S. Fortunato, and V. Loreto, Rev. Mod. Phys. **81**, 591 (2009).
[5] N. Masuda, N. Gibert, S. Redner, Phys. Rev. E **82**, 010103(R) (2010).
[6] T. Kuran, *Private Truths, Public Lies* (Harvard University Press, Cambridge, 1995).
[7] H. C. Andersen, The Emperors New Clothes, (Harcourt Brace, Orlando, FL, 1837).
[8] D. Centola, R. Willer, and M. Macy, "The Emperors Dilemma: A Computational Model of Self-Enforcing Norms", Am. J. Sociol. **110**, 1009–40 (2005).
[9] S. E. Asch, in *Groups, Leadership, and Men*, pp. 177-90, ed. H. Guetzkow (Carnegie Mellon University Press, Pittsburgh, 1951).
[10] H. J. O'Gorman, "Pluralistic Ignorance and White Estimates of White Support for Racial Segregation", Public Opinion Quarterly **39** 313–330 (1975).
[11] D. Kahan, D. Braman, and H. Jenkins-Smith, J. Risk Res. (in press 2010).
[12] T. M. Liggett, *Interacting Particle Systems*, (Springer-Verlag, New York, 1985).
[13] L. Festinger, *A Theory of Cognitive Dissonance*, (Row, Peterson & Company, 1957).
[14] S. Page, L. M. Sander, and C. M. Schneider-Mizell, J. Stat. Phys. **128**, 1279 (2007).