CODEE Journal

Volume 12 *Linking Differential Equations to Social Justice and Environmental Concerns*

Article 2

2-13-2019

Consensus Building by Committed Agents

William W. Hackborn
University of Alberta

Tetiana Reznychenko
Vasyl` Stus Donetsk National University

Yihang Zhang
East China Normal University

Consensus Building by Committed Agents

William W. Hackborn
University of Alberta, Augustana Campus

Tetiana Reznychenko
Vasyl` Stus Donetsk National University

Yihang Zhang
East China Normal University

Keywords: Social dynamics, committed agents, bifurcations, center manifold
Manuscript received on October 3, 2018; published on February 13, 2019.

Abstract: One of the most striking features of our time is the polarization, nationally and globally, in politics and religion. How can a society achieve anything, let alone justice, when there are fundamental disagreements about what problems a society needs to address, about priorities among those problems, and no consensus on what constitutes justice itself? This paper explores a model for building social consensus in an ideologically divided community. Our model has three states: two of these represent ideological extremes while the third state designates a moderate position that blends aspects of the two extremes. Each individual in the community is in one of these three states. A constant fraction of individuals are *committed agents* dedicated to the third, moderate state, while all other moderates and those from either extreme are uncommitted. The states of the uncommitted may change as they interact, according to prescribed rules, at each time step with their neighbors; the *committed agents*, however, cannot be moved from their moderate position, although they can influence neighbors. Our main objective is to investigate how the proportion of committed agents affects the large-scale dynamics of the population: in other words, we examine the special role played by those committed to embracing both sides of an ideological divide. A secondary but equally important goal is to gently introduce important dynamical systems concepts in a natural setting. Finally, we briefly outline a model with different interaction rules, a fourth state representing those who loathe the other three states, and agents who may be committed to any one of the four states.

1 Introduction

Never doubt that a small group of thoughtful, committed citizens can change the world. Indeed, it is the only thing that ever has.

— Attributed to Margaret Mead in [11, p. 158]

How can a society build tolerance, respect, and consensus when it is deeply divided along political or religious lines? It can't, we believe, because deep divisions entail, by their nature, intolerance and disrespect between those on opposite sides, and consensus cannot occur in that environment. This paper examines how social consensus might be built in a community sharply divided by two ideological extremes, denoted A and B, representing opinions or beliefs that are oppositional or opposite. In the American political context, some examples might be Democrat and Republican or pro-choice and pro-life. Some examples in the global religious context might be Christian and non-Christian, Muslim and non-Muslim, or even Christian and Muslim. This latter binary ignores other faiths and the fact that Christians and Muslims live amicably in many parts of the world, but it is a semblance of the kind of reality that exists in some communities and to which our model might be applied. Similar statements can be made about the Democrat-Republican binary and other conceivable binaries. In this paper, we assume further that the adherents of A and B rarely examine their own viewpoint and seldom engage in productive dialogue with those who hold the opposing viewpoint. Of course, many Democrats and Republicans think critically about their own opinions and engage in polite and reasoned dialogue with those who hold differing ones. Nevertheless, hard-line Democrat and Republican voters are prevalent in many electoral districts, just as fundamentalist Christians and Muslims are common in various countries, and such extremists often attempt to impose their collective wills on those who hold different views. We make the plausible assumption in this paper (and make no attempt to justify it from the sociology or political science literature) that the polarization in public discourse resulting from ideological extremism obstructs efforts to identify and collectively address inequities in a society and, in some cases, may contribute substantially to those inequities.

Our model is a variation of the *binary agreement model* of Xie et al. [13], which is in turn a two-word version of the naming game [2]. Players in this game may use word A to name an object, word B, or both words, {A,B} (abbreviated as AB), and interactions between a speaker and a listener can cause their name(s) for the object to change according to a prescribed set of rules. Similarly in [13], agents may hold opinion A, opinion B, or opinion AB, and their opinions may change via speaker-listener interactions exactly as in the two-word naming game. However, [13] adds a fourth kind of agent committed to opinion A: *committed agents* cannot be swayed from their opinion but may influence other agents to change their opinions according to the given rules.

The work in [13] is a kind of *mathematical sociology*, also called *social physics*, a subject (born in the mid-twentieth century) that employs mathematical tools, such as graph theory and differential equations, and models often drawn from physics (see Galam [5] for examples) to understand social behavior. Xie et al. [13] provide an informative discussion of the background literature relevant for their work. Like others, they regard the agents in their model as the nodes of a *graph* (also known as a *network*) whose edges indicate relationships conducive to speaker-listener interactions. They simulate social interactions on a finite graph with N nodes; in each unit time step of the simulation, each of the N agents represented in the graph interacts with one of its neighbors, chosen at random (the graph is assumed to be connected, so each node must have at least one neighbor). The main analytical tool used in [13], however, is a limiting case in which $N \to \infty$, called

the *mean-field approximation*[1]. In this limiting case, changes in opinion states can be modeled by a system of ordinary differential equations (ODEs), the *mean-field equations*. The mean-field approximation is often used (usually without much explanation) in a first course on ODEs to model, for example, the interaction of predator-prey populations and chemical reactions that obey the *law of mass action*: it justifies the assumption that the probability of, say, a predator meeting a prey or a molecule of one kind colliding with that of another is proportional to the size of the subpopulations to which they belong.

The most important finding of [13] is the tipping point associated with the proportion of the population committed to belief A: let p be this proportion, so $1-p$ is the proportion of the uncommitted population free to adopt beliefs A, B, or AB in a mutable way as they interact with other agents, some of whom are committed to A. Xie et al. [13] found that when p exceeds a critical value $p_c \approx 0.0979$ (a root of a cubic polynomial), the mean-field equations have only one fixed point, a stable (attracting) *consensus state* in which all uncommitted agents adopt belief A. When $p < p_c$, there are two additional fixed points, a stable non-consensus state and an unstable saddle point; as $p \to p_c^-$, these two fixed points approach and annihilate each other (in what is known as a *saddle-node bifurcation*), leaving only the attracting consensus state. Hence, a sufficient proportion of agents committed to belief A tips the entire population to eventually adopt that belief.

The work in [13] has inspired much subsequent research. Marvel et al. [8], for example, consider models with agents committed to opinion A and others free to adopt A, B, or AB, as in [13], but for which opinion dynamics are governed by simpler rules (e.g. speakers never change their opinions); they examine seven such models, attempting to find conditions that allow the mean-field equilibrium AB subpopulation (regarded as moderates) to thrive rather than be destroyed by those committed to A (seen as revolutionaries). Of these seven models, only one involving an external stimulus (such as a media campaign) that discourages the extremism of opinions A and B permits the equilibrium AB population to thrive. As another example, Verma et al. [12] extend the naming game dynamics used in [13] to the case where there are two subpopulations of committed agents, some committed to extreme A and others to B, in addition to agents free to believe A, B, or AB; they show that in some cases a minority of *zealots* committed to one extreme can win over a majority of the population to their side against a larger group of zealots committed to the other extreme, while in other cases neither extreme wins a majority.

In this paper, as in [8], we want to discourage extremism and encourage an increase in the size of the equilibrium AB subpopulation. This is what we mean by *consensus building*, as opinion AB is a moderate, "consensus" state even at the individual level; by contrast, we do not regard a situation in which the entire population adopts an extreme viewpoint, A or B, as the kind of harmonious consensus (conducive to social justice) that we'd like to achieve. Viewpoint AB represents for us the position of those who are able to understand and critically weigh the merits of both A and B. In the American political realm, AB

[1]The mean-field approximation used here and in [13] involves a *complete* graph, i.e. there is an edge connecting each node to every other node for a total of $\binom{N}{2} = N(N-1)/2$ edges, and each edge is equally likely to be chosen at random. Because there are N social interactions per unit time step, a discrete-time simulation becomes continuous as $N \to \infty$. See [7, pp. 127-130] for mean-field approximations in the context of social network analysis. An idea of how mean-field theory is used in physics can be found at Wikipedia permalink https://en.wikipedia.org/w/index.php?title=Mean_field_theory&oldid=855120670

might represent the viewpoint of an independent or, more generally, anyone who does not vote along party lines but instead casts each ballot only after a careful examination of candidates and ideas; AB might also represent a third, *centrist* political party, as in [9] where it is denoted by C. The approach used here to encourage agents to adopt opinion AB is to assume that a proportion of the total population is committed to AB (and that no agents are committed to A or B). After an extensive search of the literature, we found [9] to be the only other paper that has agents committed to the moderate state situated between extremes A and B. Mobilia [9] uses what we call a *binary persuasion* model (different from our *binary consensus* model, described below) to represent interactions between voters supporting three parties, A, B, and C (the centrist party to which some agents are committed); in his model, supporters of party A and those of party B never interact, but a supporter of A or B might persuade a supporter of C to join their side, and similarly a supporter of C might persuade a supporter of A or B to join side C, but supporters of C are assumed to have less persuasive power than those who support A or B. Therefore, Mobilia describes his model as a struggle between the commitment of those devoted to C and the greater persuasiveness of those who support A and B.

Section 2 presents and analyses the mean-field equations for a *binary consensus* (BC) model, ending with a brief comparison between our results and those of [9]. Section 3 introduces a *binary persuasion* (BP) model, with far more parameters than our BC model. After giving hints about the most significant results of the BP model, we leave the completion of its analysis to the reader (along with a promise to reveal the missing analysis on request). Finally, we make a few final observations and draw some conclusions in Section 4. As we put the finishing touches on this paper, the confirmation hearings for Brett Kavanaugh to become an Associate Justice of the Supreme Court of the United States are taking place. We offer this as an extreme example of the polarization that seems to have gripped American politics and the kind of dysfunction, posturing, and very partisan politics it has produced at the highest levels of that nation's government. In our opinion, social justice suffers in such a bitterly divided country, partly because people often fail to notice suffering when they're fighting a perceived enemy. It is similarly clear from history that religious minorities often suffer from injustice in regions sharply polarized along religious lines. Can those who resolutely straddle both sides of an ideological divide help to build the consensus that seems necessary for communities to move in positive directions? That is the question we explore in this paper.

2 A Binary Consensus Model

> *[T]he test of a first-rate intelligence is the ability to hold two opposed ideas in the mind at the same time, and still retain the ability to function.*
>
> — F. Scott Fitzgerald, from an *Esquire* magazine essay [4, p. 41]

The interaction dynamics for our binary consensus (BC) model are given in Table 1. These rules differ only slightly from those in [13]. The only differences lie in rows 2 and 4 of the table: the results after interaction for the corresponding rows in [13] are A–AB and B–AB, respectively, and 1 is the relative probability of these interactions in [13] compared

to λ in our model, where $0 < \lambda < 1$. In this way, we have made the probability of these two interactions relatively low, as it seems fairly unlikely that transformative conversations will occur between a speaker and listener on opposing ideological sides – imagine, for example, the likelihood of productive dialogue between hard-nosed Republicans and Democrats. We have also chosen the outcomes of these two interactions so that both speaker and listener revise their positions and agree afterwards. This seems more true for the kind of two-way conversation (rather than an alternating one-way argument) that might actually transform hard-line opponents. Notice that the speaker and listener agree after their interaction in all rows of Table 1 and that the opinion of the speaker changes in six of the twelve rows. This is why we call these rules a *binary consensus* model: the speaker and listener come to a consensus (on A, B, or AB) at the end of each transformative interaction.

In most previous studies (e.g. [13] and [8]), the only committed agents are those who hold opinion A. In our BC model, however, committed agents must hold opinion AB. Let

Dynamics of the Binary Consensus Model		
Before Interaction	After Interaction	Relative Probability
$A \xrightarrow{A} A$	A – A	not applicable
$A \xrightarrow{A} B$	AB – AB	λ
$A \xrightarrow{A} AB$	A – A	1
$B \xrightarrow{B} A$	AB – AB	λ
$B \xrightarrow{B} B$	B – B	not applicable
$B \xrightarrow{B} AB$	B – B	1
$AB \xrightarrow{A} A$	A – A	$\frac{1}{2}$
$AB \xrightarrow{B} A$	AB – AB	$\frac{1}{2}$
$AB \xrightarrow{A} B$	AB – AB	$\frac{1}{2}$
$AB \xrightarrow{B} B$	B – B	$\frac{1}{2}$
$AB \xrightarrow{A} AB$	A – A	$\frac{1}{2}$
$AB \xrightarrow{B} AB$	B – B	$\frac{1}{2}$

Table 1: Speaker-Listener interaction rules for our binary consensus model. The speaker may hold any one of three opinions (A, B, and AB), and there is an equal chance of an AB speaker voicing opinions A or B. The *Before Interaction* column indicates the speaker, listener, and the opinion voiced (above the arrow). The *After Interaction* column shows the opinions of speaker and listener after their conversation: note that the opinion of the speaker may change. The final column indicates the relative probability of a transformative interaction per unit time; conversations for which this probability is *not applicable* (as no transformation occurs) are listed for completeness. Note that committed adherents of AB *cannot* be moved from their opinion by their interactions with others, but these interactions *can* change the opinions of others.

q, with $0 \leq q < 1$, be a constant denoting the proportion of all agents committed to AB, and x, y, z be the proportions of agents who hold (but are not committed to) opinions A, B, AB, respectively, at a moment t in time. So, x, y, and z all lie in the interval $[0,\ 1]$, and

$$x + y + z + q = 1\,. \tag{2.1}$$

Consequently, the variables x and y must lie in the triangular region Λ defined by

$$\Lambda = \{\,(x, y) \mid x \geq 0\,,\ y \geq 0\,,\ x + y \leq 1 - q\,\}\,. \tag{2.2}$$

See Figure 1. The last inequality in (2.2) is implied by equation (2.1) with $z \geq 0$.

Now, using the interaction rules in Table 1 and the fact that the mean-field rate of a particular interaction is proportional to the fractions of agents involved in that interaction, we find the mean-field equations to be

$$\begin{aligned} \frac{dx}{dt} &= xz + \tfrac{1}{2}zx - \tfrac{1}{2}zx - \tfrac{1}{2}qx + 2\left(\tfrac{1}{2}z^2\right) + \tfrac{1}{2}qz + \tfrac{1}{2}zq - \lambda xy - \lambda yx\,, \\ \frac{dy}{dt} &= yz + \tfrac{1}{2}zy - \tfrac{1}{2}zy - \tfrac{1}{2}qy + 2\left(\tfrac{1}{2}z^2\right) + \tfrac{1}{2}qz + \tfrac{1}{2}zq - \lambda yx - \lambda xy\,. \end{aligned} \tag{2.3}$$

Each term in system (2.3) corresponds to one row in Table 1. The terms involving q correspond to interactions between agents committed to AB and other agents; committed agents can be involved in interactions that transform other agents but cannot be transformed themselves. Note also that there is no need to include an equation for dz/dt in (2.3), as $z = 1 - x - y - q$ from equation (2.1). Denoting the right-hand sides of equations (2.3)

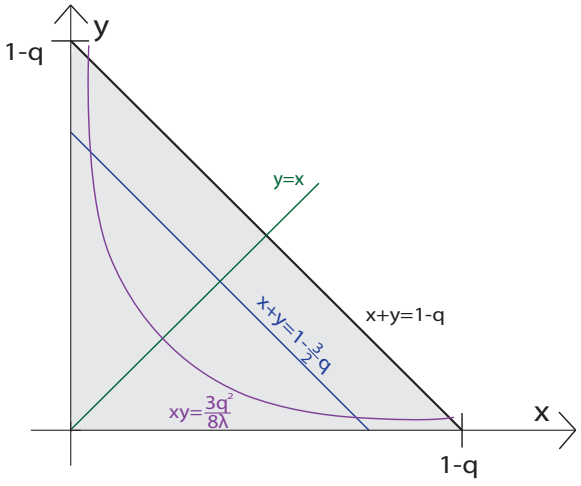

Figure 1: The flow domain Λ in the phase plane associated with our binary consensus model: a triangle bounded by $x = 0$, $y = 0$, and $x + y = 1 - q$. One fixed point in the flow domain lies on the line of symmetry, $y = x$, and two more are placed symmetrically about this line on the hyperbola $xy = 3q^2/(8\lambda)$ when $0 < q < 2/(3 + \sqrt{6/\lambda})$.

by $f(x,y)$ and $g(x,y)$ and eliminating z using (2.1), we find

$$\begin{aligned}\frac{dx}{dt} &= f(x,y) = xz - \tfrac{1}{2}qx + z(z+q) - 2\lambda xy \\ &= 1 - q - (1+\tfrac{1}{2}q)x - (2-q)y + y^2 + (1-2\lambda)xy\,, \qquad (2.4a)\\ \frac{dy}{dt} &= g(x,y) = yz - \tfrac{1}{2}qy + z(z+q) - 2\lambda xy \\ &= 1 - q - (1+\tfrac{1}{2}q)y - (2-q)x + x^2 + (1-2\lambda)xy\,. \qquad (2.4b)\end{aligned}$$

By examining the component of the vector field $\langle f(x,y),\ g(x,y)\rangle$ in the inward-normal direction on each of the boundary lines of the domain Λ and, separately, the direction of this field at each corner point of Λ, we have verified that none of the values of $\langle f(x,y),\ g(x,y)\rangle$ on the boundary of Λ point to the exterior of Λ. Therefore, the flow associated with system (2.4) is trapped within Λ. Now, subtracting (2.4b) from (2.4a) to exploit symmetry yields

$$\frac{d(x-y)}{dt} = -(x-y)(x+y+\tfrac{3}{2}q-1)\,, \qquad (2.5)$$

which reveals that the line of symmetry, $y=x$, is an invariant set of the flow; see [6, p. 33] for more about this concept. We observe from (2.5) that fixed points[2] of the flow (at which f and g vanish) can lie only on the lines $y=x$ or $x+y=1-\frac{3}{2}q$; see Figure 1. Consider the line $x+y=1-\frac{3}{2}q$: it lies in the interior of the flow domain Λ when $0<q<\frac{2}{3}$, and (2.1) implies $z=\frac{1}{2}q$ on this line. But, using (2.4a), $f(x,y)=0$ and $z=\frac{1}{2}q$ imply $xy=3q^2/(8\lambda)$, the equation of a hyperbola. Hence, two fixed points (x,y) lie at the intersections of this hyperbola and the line $x+y=1-\frac{3}{2}q$, where

$$x = \frac{2-3q\pm\sqrt{(2-3q)^2-6q^2/\lambda}}{4}\,,\quad y=\frac{3q^2}{8\lambda x}\,;\quad 0<q<\frac{2}{3+\sqrt{6/\lambda}}\,. \qquad (2.6)$$

These fixed points (x,y), contingent on the intersections that define them, exist when q satisfies the condition in (2.6). If they exist, they are mirror images of each other in the line of symmetry $y=x$, as seen in Figure 1. On the line of symmetry itself,

$$\frac{dx}{dt} = f(x,x) = 2(1-\lambda)x^2 - \tfrac{1}{2}(6-q)x + 1 - q, \qquad (2.7)$$

using equation (2.4a). It follows that the flow has two fixed points (x,x) corresponding to each root x of the quadratic $f(x,x)$ in (2.7). Regarding the invariant set $y=x$ as a phase line, the smaller (larger) root corresponds to a stable[3] (unstable, respectively) fixed point; in other words, the line $y=x$ is a stable (unstable) manifold of the fixed point corresponding to the smaller (larger, respectively) root; see [6, pp. 12-16]. Naming the smaller root x_0,

$$x_0 = \frac{1}{8(1-\lambda)}\left[6-q-\sqrt{(6-q)^2-32(1-\lambda)(1-q)}\right]. \qquad (2.8)$$

[2] In this paper, fixed points are positions at which the right-hand sides of each equation in an autonomous system of ODEs vanishes. They are also known as equilibria or steady-state solutions.

[3] Unless otherwise indicated, the term *stable* means *locally asymptotically stable*, and *unstable* means *not* stable. See [6, p. 3] for more information.

To examine the behavior of x_0 as $\lambda \longrightarrow 1^-$, we find

$$x_0 = \frac{2(1-q)}{6-q} + \frac{16(1-\lambda)(1-q)^2}{(6-q)^3} + \frac{256(1-\lambda)^2(1-q)^3}{(6-q)^5} + O(r^4), \tag{2.9}$$

where $r = 32(1-\lambda)(1-q)/(6-q)^2$, when we write the square root in equation (2.8) as $(6-q)(1-r)^{1/2}$ and use the binomial series. The infinite series indicated in (2.9) converges in all valid cases, since $0 < r < \frac{8}{9}(1-\lambda) < 1$ for $0 < q < 1$ and $0 < \lambda < 1$; its first term is independent of λ and approaches $\frac{1}{3}$ as $q \longrightarrow 0^+$.

Clearly $x_0 > 0$ from (2.8) or (2.9), and $f(x,x) = -\frac{1}{2}qx - 2\lambda x^2 < 0$ on the boundary line $x + y = 1 - q$ (on which $z = 0$) from (2.4a). Hence, the fixed points (x,x) associated with the two roots of $f(x,x)$ in equation (2.7) straddle the line $x + y = 1 - q$: the fixed point associated with the smaller root x_0 lies inside the flow domain Λ and the one associated with the larger root lies outside. We already know that the line $y = x$ is a stable manifold of the fixed point (x_0, x_0). To examine the two-dimensional stability of this point, first observe from (2.5) that trajectories close to the line $y = x$ converge towards it as time proceeds when $x + y > 1 - \frac{3}{2}q$ and diverge away from it when $x + y < 1 - \frac{3}{2}q$. To see this more clearly, we choose coordinates (ξ, η) for which ξ and η are constant along lines parallel to the lines $y = x$ and $x + y = 1 - \frac{3}{2}q$, respectively, and such that ξ and η vanish on these latter lines. Coordinates of this kind, together with their inverse, are given below:

$$\xi = x - y\,, \qquad x = \tfrac{1}{2}(\xi + \eta - \tfrac{3}{2}q + 1)\,, \tag{2.10a}$$

$$\eta = x + y + \tfrac{3}{2}q - 1\,, \qquad y = \tfrac{1}{2}(\eta - \xi - \tfrac{3}{2}q + 1)\,. \tag{2.10b}$$

Using (2.10), system (2.4) becomes

$$\frac{d\xi}{dt} = f(x,y) - g(x,y) = F(\xi,\eta) = -\xi\eta\,, \tag{2.11a}$$

$$\begin{aligned}\frac{d\eta}{dt} &= f(x,y) + g(x,y) = G(\xi,\eta) \\ &= \left[(3q-2)\lambda - \tfrac{5}{2}q - 1\right]\eta + (1-\lambda)\eta^2 + \lambda\xi^2 + K\,,\end{aligned} \tag{2.11b}$$

where $F(\xi,\eta)$ and $G(\xi,\eta)$ denote the right-hand sides of equations (2.11), and

$$K = \tfrac{1}{4}\left[6q^2 - \lambda(2-3q)^2\right]\,. \tag{2.12}$$

A Jacobian matrix represents the linearization of a vector field near a point. For the field $\langle F(\xi,\eta),\, G(\xi,\eta)\rangle$, the Jacobian is

$$J(\xi,\eta) = \begin{pmatrix} \frac{\partial F}{\partial \xi} & \frac{\partial F}{\partial \eta} \\ \frac{\partial G}{\partial \xi} & \frac{\partial G}{\partial \eta} \end{pmatrix} = \begin{pmatrix} -\eta & -\xi \\ 2\lambda\xi & (3q-2)\lambda - \frac{5}{2}q - 1 + 2(1-\lambda)\eta \end{pmatrix}. \tag{2.13}$$

Let η_0 be the η coordinate of the fixed point on the line of symmetry $\xi = 0$ in the flow region Λ defined by (2.2). The (x,y) coordinates of this fixed point are (x_0, x_0), and η_0 can therefore be determined from (2.8) and (2.10b). However, it is better for our purposes to use the fact that $G(0,\eta_0) = 0$: in this way, equation (2.11b) yields

$$\eta_0 = \frac{-C - \sqrt{C^2 - 4(1-\lambda)K}}{2(1-\lambda)}\,, \quad C = -(2-3q)\lambda - \tfrac{5}{2}q - 1\,, \tag{2.14}$$

where K is given in (2.12). On the line $\xi = 0$, the Jacobian matrix in (2.13) is diagonal; the eigenvalues of $J(0, \eta_0)$, namely $-\eta_0$ and $C + 2(1-\lambda)\eta_0$ with C as in (2.14), lie on its main diagonal and correspond to, respectively, eigenvectors parallel to the lines $\eta = 0$ (i.e. $x + y = 1 - \frac{3}{2}q$) and $\xi = 0$ (i.e. $y = x$, the line of symmetry). From an earlier result (i.e. that the line $\xi = 0$ is a stable manifold of the fixed point on this line in the region Λ), we know that the latter eigenvalue must be negative, but it is easy to verify this directly: from the first equation in (2.10b), we see that the boundary line $x + y = 1 - q$ coincides with $\eta = \frac{1}{2}q$, so $\eta \leq \frac{1}{2}q$ in the flow domain Λ, and thus

$$C + 2(1-\lambda)\eta_0 \leq C + 2(1-\lambda)\tfrac{1}{2}q = -2\lambda(1-q) - \tfrac{3}{2}q - 1 < -1\,, \tag{2.15}$$

using the expression for C in (2.14). The sign of the other eigenvalue, $-\eta_0$, is opposite to that of η_0 itself. Since $C < 0$, as seen in (2.15), it follows from (2.14) that $\eta_0 > 0$ if and only if $K > 0$, and the equivalencies below can then be inferred from (2.12):

$$\eta_0 > 0 \quad \Leftrightarrow \quad K > 0 \quad \Leftrightarrow \quad \lambda < \frac{6q^2}{(2-3q)^2} \quad \Leftrightarrow \quad q > q_c = \frac{2}{3 + \sqrt{6/\lambda}}\,. \tag{2.16}$$

For planar flows like those associated with systems (2.4) and (2.11) (see [6, pp. 42-60] for details on such flows), a fixed point is stable if the real parts of both eigenvalues of its Jacobian matrix at that point are negative. If its eigenvalues are real, then it is a (stable) *sink node* if both eigenvalues are negative, an (unstable) *source node* if both are positive, and an (unstable) *saddle* if they have different signs; if its eigenvalues are complex conjugates, it is a *spiral* sink or source, respectively, if the sign of their real parts is negative or positive. Hence, the fixed point $(\xi, \eta) = (0, \eta_0)$ is a sink node when $q_c < q < 1$ because, according to (2.16), $\eta_0 > 0$ in this case and so the eigenvalue $-\eta_0$ is negative; equivalently, it is a sink node when λ satisfies the condition in (2.16) (in addition to the earlier requirement that $0 < \lambda < 1$). When $0 < q < q_c$, however, the fixed point $(\xi, \eta) = (0, \eta_0)$ is a saddle, as the eigenvalue $-\eta_0$ is positive, and two other fixed points (x, y) given in (2.6) also exist in this case. Figure 2 depicts these interesting flow dynamics in the phase plane via two cases, one for $q < q_c$ and the other for $q > q_c$.

To investigate further the emergence of the two fixed points in (2.6) from the fixed point $(\xi, \eta) = (0, \eta_0)$ as q decreases through q_c (holding λ constant), we employ the *center manifold* techniques of Guckenheimer and Holmes [6], pp. 123-138. A center manifold is an invariant set tangent to the eigenspace corresponding to eigenvalues with zero real parts of the Jacobian matrix at a fixed point. Our flow has a center manifold tangent to the line $\eta = \eta_0$ at the fixed point $(\xi, \eta) = (0, \eta_0)$ when $q = q_c$, or equivalently, using (2.16), when $\eta_0 = 0$ or $K = 0$. This center manifold becomes an unstable manifold of fixed point $(0, \eta_0)$ if $0 < q < q_c$ (when $\eta_0 < 0$) and is an invariant subset of the stable manifold of point $(0, \eta_0)$ if $q_c < q < 1$ (when $\eta_0 > 0$), as we have seen. Following [6, p. 130 ff.], we approximate the invariant manifold tangent to the line $\eta = \eta_0$ at the fixed point $(0, \eta_0)$ by

$$\eta = h(\xi) = a + b\xi^2 + O(\xi^4)\,. \tag{2.17}$$

Note that odd powers of ξ need not be included in (2.17), owing to symmetry about the axis $\xi = 0$. The technique for finding the unknown coefficients a and b in (2.17) first

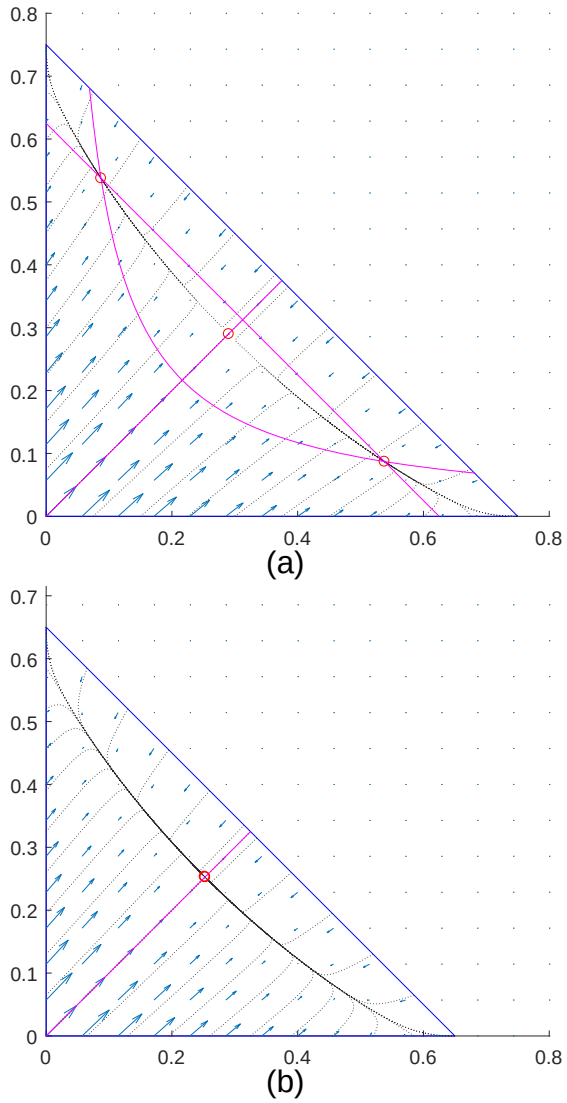

Figure 2: The phase plane flow for two cases of our binary consensus model showing fixed points (circles, red online), some trajectories (dotted), and the direction field (blue online). (a) $q = 0.25,\ \lambda = 0.5$: a fixed point (a saddle) exists on the line of symmetry, $y = x$, connected via its unstable manifold (clearly visible) to two fixed points (sink nodes) lying at the intersections of the line $x + y = 1 - \frac{3}{2}q$ and the hyperbola $xy = 3q^2/(8\lambda)$ (violet online). (b) $q = 0.35,\ \lambda = 0.5$: only one fixed point (a sink node) lies on the line $y = x$. A pitchfork bifurcation occurs at $q = q_c$, where $q_c = 2/(3 + \sqrt{6/\lambda})$; $q_c \approx 0.3094$ for $\lambda = 0.5$.

involves differentiating (2.17) with respect to t, substituting the expression for $d\xi/dt$ in equation (2.11a) into the result, and using (2.17) again. This produces

$$\frac{d\eta}{dt} = \frac{dh}{d\xi}\frac{d\xi}{dt} = -\frac{dh}{d\xi}\xi\eta = -\xi\frac{dh}{d\xi}h(\xi) = -\xi\left[2b\xi + O(\xi^2)\right]\left[a + b\xi^2 + O(\xi^4)\right]\ . \qquad (2.18)$$

However, we obtain another expression for $d\eta/dt$ by combining (2.11b) and (2.17):

$$\begin{aligned}\frac{d\eta}{dt} &= Ch(\xi) + (1-\lambda)\left[h(\xi)\right]^2 + \lambda\xi^2 + K \\ &= C\left[a + b\xi^2 + O(\xi^4)\right] + (1-\lambda)\left[a + b\xi^2 + O(\xi^4)\right]^2 + \lambda\xi^2 + K\,, \end{aligned} \tag{2.19}$$

where K and C are given in (2.12) and (2.14). Equating the coefficients of corresponding powers of ξ in (2.18) and (2.19) yields

$$a = \eta_0\,, \quad b = \frac{\lambda}{(2-3q)\lambda + \frac{5}{2}q + 1 + (2\lambda - 4)\eta_0}\,, \tag{2.20}$$

with η_0 as in (2.14). The value of a in (2.20) is exactly what we would expect from (2.17). Furthermore, in the same way as we proved the inequality in (2.15) using the fact that $\eta \le \frac{1}{2}q$ in the flow domain Λ, we can show that the denominator of b in (2.20) is bounded below by $2\lambda(1-q) + \frac{1}{2}q + 1 > 1$. So $0 < b < \lambda$, and the resulting concavity in the curve described by (2.17) is evident in Figure 2.

Although the manifold tangent to the line $\eta = \eta_0$ at the fixed point $(\xi, \eta) = (0, \eta_0)$, as given in (2.17), is a center manifold of system (2.11) only if $q = q_c$ (when $K = 0$ and $\eta_0 = 0$), Guckenheimer and Holmes [6], p. 134 point out that it can be regarded as a "family of center manifolds" parametrized by K within the center manifold of a larger system that includes the equation $\frac{dK}{dt} = 0$ (which has its own zero eigenvalue) in addition to system (2.11). Provided that K, ξ, and η are sufficiently close to zero, the power series in (2.17) approximates this family of center manifolds that vary with K via the coefficients a and b in (2.20), which depend on η_0 and thus K due to (2.14). Since this family is associated with the eigenvalue $-\eta_0$, which is close to zero and therefore associated with much *slower* flow dynamics than those associated with the strongly negative eigenvalue $C + 2(1-\lambda)\eta_0$, the local evolution of the flow described by system (2.11) can be reduced to what happens on this family; the slow dynamics implied by the fact that eigenvalue $-\eta_0 \approx 0$ also explains why these center manifolds are so visible in Figure 2. The following equation, obtained by substituting η in (2.17) into (2.11a) and using $a = \eta_0$ in (2.20), thus captures the local evolution (i.e. when ξ and η are sufficiently close to zero, and q is sufficiently close to q_c) of the flow projected onto the family of center manifolds:

$$\frac{d\xi}{dt} = -\xi\left[\eta_0 + b\xi^2 + O(\xi^4)\right]\,. \tag{2.21}$$

Equation (2.21) represents a *pitchfork bifurcation* [6, pp. 145-150] at $q = q_c$. If $q > q_c$ then $\eta_0 > 0$ from (2.16), and so the expression $\eta_0 + b\xi^2 + O(\xi^4)$ in (2.21) is positive for sufficiently small ξ values since $b > 0$; in this case, (2.21) has an evidently stable fixed point at $\xi = 0$; note that $\eta_0 = 0$ when $q = q_c$, and so this fixed point is stable, but only marginally so, in this case too. However, if $q < q_c$ then $\eta_0 < 0$, and the right-hand side of (2.21) vanishes when $\xi \approx \pm\sqrt{-\eta_0/b}$; these two ξ values are stable fixed points of (2.21) in this case and $\xi = 0$ is unstable. This shows that, as q decreases through q_c (holding λ constant), the sink node at $(\xi, \eta) = (0, \eta_0)$ in system (2.11) becomes a saddle and two new sink nodes emerge from it.

Our analysis of the flow associated with system (2.11) [or equivalently, system (2.4)] will be complete after we employ one more important tool: the divergence of the flow

field, $\operatorname{div}\langle F(\xi,\eta), G(\xi,\eta)\rangle = \frac{\partial F}{\partial \xi} + \frac{\partial G}{\partial \eta}$, which is identical to the trace of the Jacobian (2.13). Using the fact that $\eta \leq \frac{1}{2}q$ in the flow region Λ, as seen earlier, and that $\eta \geq \frac{3}{2}q - 1$, from the first equation in (2.10b) with $(x, y) = (0, 0)$, it is easy to show from (2.13) that $\operatorname{div}\langle F(\xi,\eta), G(\xi,\eta)\rangle \leq -2q - 1 < -1$ on Λ. Bendixson's Criterion [6, p. 44] can thus be used to infer that the flow has no closed orbits. This indicates that the flow is characterized by the pitchfork bifurcation and fixed points we have examined above. Note that, although our analysis of the bifurcation is necessarily localized to values of q near q_c, the absence of closed orbits and existence of the saddle at point $(0, \eta_0)$ for all $0 < q < q_c$ indicate (and would be key parts of a rigorous proof) that the two fixed points (x, y) given in (2.6) persist as sink nodes for all $0 < q < q_c$. This is supported by computations of the trace $T = \operatorname{trace}[J(\xi,\eta)]$, the determinant $D = \det[J(\xi,\eta)]$, and the discriminant $T^2 - 4D$ at numerous such points (x, y) [associated with (ξ,η) via (2.10)] using the fact that the eigenvalues of $J(\xi,\eta)$ are real and negative if $T < 0$, $D > 0$, and $T^2 - 4D > 0$.

We are now in a position to summarize the results of our binary consensus model. In the event that $q \geq q_c$, there is only one fixed point, $(x, y) = (x_0, x_0)$, a global attractor to which all flow trajectories are drawn. Moreover, it is clear from equation (2.9) that $x_0 \longrightarrow 0^+$ monotonically as $q \longrightarrow 1^-$ (while holding λ constant). But equation (2.1) implies that the total (stable for $q \geq q_c$) equilibrium proportion of AB supporters, both committed and uncommitted, is $1 - 2x_0$, and $1 - 2x_0 \longrightarrow 1^-$ as $q \longrightarrow 1^-$. Hence, in the mean-field limit of our binary consensus model with a given proportion q committed to opinion AB, a consensus on AB is continuously built towards total consensus as q grows (with λ held constant); this seems possibly more realistic than the model of [9] for which a consensus on C (the centrist position) occurred discontinuously at a tipping point.

When $q < q_c$, there are three fixed points and only the two placed symmetrically about the line $y = x$ are stable (see Figure 1). In this case, making q larger (while holding λ constant) serves to bring these stable fixed points closer together; this can be seen from the hyperbola, $xy = 3q^2/(8\lambda)$, on which these points lie. Since flow trajectories are drawn to one or the other of these fixed points (except trajectories that start on the line of symmetry $y = x$), this indicates that the equilibrium results (in an election, say) are less volatile when q is larger (which is probably a good thing for an election). Making λ smaller (while holding q constant, with $q < q_c$) also pushes the two stable fixed points closer. This seems counter-intuitive, as λ is the relative probability of producing an AB supporter in interactions between adherents of A and B (see rows 2 and 4 of Table 1): larger values of λ imply more AB supporters produced in this way. On the other hand, an interaction between uncommitted supporters of AB always converts them to A or B (see the final two rows of Table 1, which come directly from the two-word naming game). So, high numbers of uncommitted AB supporters are hard to sustain, and it is therefore tempting to change the relative probability of transformation in the last two rows of Table 1 from $\frac{1}{2}$ to $\frac{1}{2}\mu$, where $0 < \mu \leq 1$; we leave this possible modification as an exercise for the reader. The wonderful thing about our model (and models in general) is that they can be modified and explored. Breaking the symmetry with respect to A and B in our binary consensus model would be an interesting exploration. One of the commendable features of [9] is the symmetry-breaking included in its model: supporters of opinion A are more persuasive (at converting supporters of C) than are supporters of B.

3 A Binary Persuasion Model

Don't vote: It just encourages the bastards.

— P. J. O'Rourke, from the title of his book [10]

Our second model, which we only outline in this section, is a significant departure from our BC model in Section 2. One important difference is the addition of a new kind of opinion, ∅, that represents the absence of any opinion (i.e. the null opinion: neither A nor B nor any opinion in between). In the American context, for example, those who despise politicians of all stripes and wouldn't consider voting for any of them might be said to hold viewpoint ∅. We think there is some merit to a political model that includes non-voters, as otherwise they are ignored, despite the fact that they usually represent a significant proportion of citizens (more than 40% of eligible voters in recent Canadian federal elections). Counting those who choose not to vote because they are, say, appalled by political corruption or because no candidate truly represents their position might actually lead to better government, as election campaigns might begin to court these potential voters by, respectively, refusing to accept donations from powerful lobbies or revising their platforms to accord more with the views of the disenchanted. It is worth noting that four Canadian provinces [3] and some other jurisdictions[4] allow voters to publicly decline their vote or indicate "none of the above" on their ballots. In the realm of religious beliefs, the ∅ viewpoint might include atheists or at least those who are repulsed by mainstream religious systems represented by A and B. It is important to count such people, we think, because even some religious leaders[5] have argued that it's preferable to abandon religion when it harbors bigotry and fosters discord. Furthermore, as with non-voters, religious non-conformists form a large segment of many societies.

While the ∅ opinion is an important new feature of the *binary persuasion* (BP) model considered here, the essential difference between our BP model and the BC model of Section 2 lies in the disparity between the words *persuasion* and *consensus*. The interaction rules for our BC model involve speaker and listener reaching a consensus (on A, B, or AB) at the end of each transformative interaction: the opinion of the speaker and that of the listener may change to achieve this consensus. In contrast, for our BP model, only the opinion of the listener (never that of the speaker) changes at the end of a transformative interaction. Our BP model is therefore more like proselytization than dialogue. In any case, one of the reasons we have chosen to consider a persuasion model in addition to a consensus model is to illustrate the diverse possibilities for speaker-listener models. Another reason is that the interaction rules for a binary persuasion model can be represented visually, and more simply, using a *compartment diagram*. One further simplification we make in our BP model is to combine states A and B. How can we do this? We assume that the relative probabilities of transformative interactions between opinions A, ∅, and AB are identical to the relative probabilities of such interactions between opinions B, ∅, and AB; in other words, we assume that the compartments for

[4]See Wikipedia permalink https://en.wikipedia.org/w/index.php?title=None_of_the_above&oldid=847838101

[5]For example, 'Abdu'l-Bahá states, "If religion becomes a cause of dislike, hatred and division, it were better to be without it, ..." [1, p. 130]

A and B (in a compartment diagram) are linked to those for ∅ and AB in the same way, so we can combine the compartments for A and B into a single compartment, which we label A△B (i.e. the symmetric difference of A and B, denoting agents holding opinion A or opinion B but not both). What happens inside the A△B compartment (such as whether an agent can change her opinion from A to B or vice versa) need not concern us, as far as our binary persuasion model is concerned: we have combined the extremists into a single compartment and can think of the residents of that compartment as having the same opinion, namely one-sided extremism. Figure 3 thus represents our BP model.

Our BP model has three parameters, α, β, and γ, which are regarded as positive. These three parameters make our BP model more complicated than our BC model, but several simplifying assumptions were still required to reduce the number of parameters to three. The model represented in Figure 3 is the simplest, analytically tractable model that we were able to handle, within the context of persuasive interactions between agents of the three states ∅, A△B, and AB. For example, our decision to make *gamma* the relative probability for conversions from state ∅ to state AB and for the opposite conversions simplifies our BP model, and it seems realistic. Note also that, because $\alpha > 0$ and $\beta > 0$ (so $1 + \alpha > 1$ and $1 + \beta > 1$), agents in the extremist state A△B are more persuasive at converting those in states ∅ and AB than the other way around. This reflects our perception that opinions A and B offer a kind of social safety that ∅ and AB do not offer. In the American political realm, this might mean that voting Democrat or Republican is socially safer than not voting at all or voting for some fringe party of independent thinkers. Mobilia [9] does something similar: he makes supporters of parties A and B more persuasive at converting centrists (party C) to their side than vice versa (but the similarity of our BP model to that of [9] is coincidence – we built and completely analyzed the former before we discovered the latter).

Unlike the BC model of Section 2, our BP model allows for committed agents of all kinds: let p, c, and q be constants in the interval $[0, 1)$ denoting the proportions of all agents committed to opinions ∅, A△B, and AB, respectively. Also, let w, u, and z

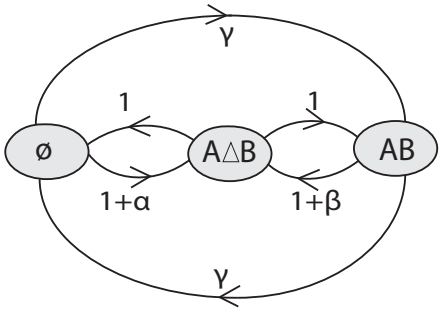

Figure 3: Compartment diagram for our binary persuasion model showing the relative probabilities of transitions between compartments ∅, A△B, and AB. The label on a directed edge from compartment X to compartment Y is the relative probability (per unit time) of a speaker in compartment Y persuading a listener in compartment X to move to compartment Y. Compartments represent sets of agents of the same type.

be the proportions of all agents who hold (but are not committed to) opinions $\emptyset$, A$\triangle$B, and AB, respectively, at a given time t; each of these variables lies in the interval $[0,\ 1]$. Furthermore, all of these proportions must add up to 1, so

$$w + z + u + p + q + c = 1\,. \tag{3.1}$$

The mean-field equations for our BP model can now be written and analyzed as we did for the BC model in Section 2. We have included with this paper (at https://scholarship.claremont.edu/codee/vol12/iss1/2/) MATLAB code that plots phase portraits for our BC and BP models; the mean-field equations for the latter can be seen in a MATLAB function. It is not difficult to show that the resulting flow has at most three fixed points. Furthermore, in the case when $c = 0$ (i.e. no agents are committed to state A$\triangle$B) but $p > 0$ and/or $q > 0$, one fixed point lies on the line $u = 0$ (a boundary of the flow domain where support for opinion A$\triangle$B is zero); this fixed point is stable in some cases, unstable in others, and there is a sharp boundary between these cases (a set of tipping points) that can be described in a simple, geometric way; the transition of this fixed point from stability to instability occurs as it collides with another fixed point in a *transcritical bifurcation* on a center manifold. As we do not want this paper to be overlong and want others to enjoy the thrill of discovery, we leave the (sometimes messy) analysis of our BP model as an exercise for the reader. We are happy to make our results available to anyone who requests them from the corresponding author.

4 Some Observations and Conclusions

The results of this paper give some insight into human social behavior. Although our binary consensus model is only an imitation of reality, it indicates that consensus in a divided community can be gradually achieved if enough people are committed to embracing both sides of the divide. It also suggests that the distance between two stable but opposing equilibria can be reduced by a group of committed moderates. These results make this paper relevant to creating an environment in which social justice might flourish.

Finally, we hope that our paper helps students and others understand some of the beautiful aspects of flows associated with systems of nonlinear ODEs. We have identified three different kinds of bifurcations that can occur as a parameter is varied. In Section 1, we mentioned the saddle-node bifurcation that occurs in the model of [13]; the center manifold on which the three fixed points exist when $p < p_c$ can be seen clearly in Fig. 2(a) of [13]. A pitchfork bifurcation arises in the BC model of Section 2; these bifurcations often occur naturally in flows that have a line of symmetry. We also mentioned the transcritical bifurcation that occurs in the BP model of Section 3. Bifurcations and center manifolds are two important keys for understanding the deeper aspects of flows associated with ODEs.

References

[1] ‘Abdu’l-Bahá. *Paris Talks: Addresses Given by ‘Abdu’l-Bahá in 1911*. UK Bahá’í Publishing Trust, London, 11th edition, 1979.

[2] Andrea Baronchelli, Maddalena Felici, Vittorio Loreto, Emanuele Caglioti, and Luc Steels. Sharp transition towards shared vocabularies in multi-agent systems. *Journal of Statistical Mechanics: Theory and Experiment*, 2006(06):P06014, 2006. http://stacks.iop.org/1742-5468/2006/i=06/a=P06014.

[3] Terra Ciolfe. Ontario election 2018: How to decline your vote. *Maclean’s Magazine*, June 2018. This did not appear in print. It is from the 2 June 2018 online edition of *Maclean’s* magazine available at https://www.macleans.ca/politics/ontario-election-2018-how-to-decline-your-vote/.

[4] F. Scott Fitzgerald. The Crack-Up. *Esquire*, page 41ff, Feb 1936. The first part of a three-part essay published in the February, March, and April 1936 issues of *Esquire* magazine.

[5] S. Galam. Sociophysics: A review of Galam models. *Int. J. Mod. Phys. C*, 19(3): 409–440, 2008. https://doi.org/10.1142/S0129183108012297.

[6] J Guckenheimer and P. Holmes. *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*. Springer, New York, 1st edition, 1983.

[7] Matthew O. Jackson. *Social and Economic Networks*. Princeton University Press, Princeton, 2008.

[8] Seth A. Marvel, Hyunsuk Hong, Anna Papush, and Steven H. Strogatz. Encouraging moderation: Clues from a simple model of ideological conflict. *Phys. Rev. Lett.*, 109: 118702, Sep 2012. https://link.aps.org/doi/10.1103/PhysRevLett.109.118702.

[9] M. Mobilia. Commitment versus persuasion in the three-party constrained voter model. *J. Stat. Phys.*, 151:69–91, 2013. https://doi.org/10.1007/s10955-012-0656-x.

[10] P.J. O’Rourke. *Don’t Vote: It Just Encourages the Bastards*. Grove/Atlantic, New York, 2010.

[11] F.G. Sommers and T. Dineen. *Curing Nuclear Madness: A New-age Prescription for Personal Action*. Methuen, London, 1984.

[12] Gunjan Verma, Ananthram Swami, and Kevin Chan. The impact of competing zealots on opinion dynamics. *Physica A: Statistical Mechanics and its Applications*, 395:310–331, 2014. ISSN 0378-4371. https://doi.org/10.1016/j.physa.2013.09.045.

[13] J. Xie, S. Sreenivasan, G. Korniss, W. Zhang, C. Lim, and B. K. Syzmanski. Social consensus through the influence of committed minorities. *Phys. Rev. E*, 84(1):011130, 2011. https://doi.org/10.1103/PhysRevE.84.011130.