Confident moderates can make a population vulnerable to extremism

Anna Papush,[1] Seth A. Marvel,[2] Hyunsuk Hong,[3] and Steven H. Strogatz[1]

[1] *Department of Mathematics, Cornell University, New York 14853, USA*
[2] *Center for Applied Mathematics, Cornell University, New York 14853, USA*
[3] *Department of Physics and Research Institute of Physics and Chemistry, Chonbuk National University, Jeonju 561-756, Korea*

(Dated: September 2, 2011)

Common sense suggests that the more resistant that moderate-minded individuals are to the persuasion of those with more extreme views, the larger will be the initial 'seed' of evangelical extremists needed to convert the population to the extreme stance (assuming this conversion occurs). We find however that this is not the case in a minimalistic mean-field model of social opinion. Rather, the more steadfast that moderates are in their moderate position, the *fewer* proselytizing zealots are needed to win over the population. This conclusion holds up well under a variety of violations of the simplifying assumptions of our model, with one notable exception—if the moderates themselves begin to persuasively promote the moderate view, then the population as a whole becomes largely incapable of ideological takeover by the extremists.

I. INTRODUCTION

Transnational terrorism is currently a persistent reality in many parts of the world [1, 2]. As a result it has become critical to understand how extremist groups grow within local regions and how a population's receptiveness toward a committed extremist faction can aid or deter the spread of extremism through the population [3–5].

In this paper, we present a simple dynamical system that suggests a surprising conclusion in this emerging field of research: A population whose citizens have a stronger tendency to take a moderate rather than extreme stance can actually be *more* susceptible to evangelism by extreme factions. We consider several elaborations of the model that all point to this same conclusion, but we also find that when moderates promote moderation with enough success, extremism can no longer overwhelm the population. In other words, to remain an appreciable fraction of the population, moderates need champions too.

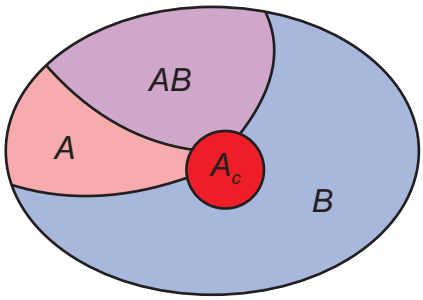

FIG. 1: The four subpopulations considered by the model. The A and B subpopulations are composed of uncommitted individuals who believe A and B, respectively. The AB subpopulation denotes the moderates (those that endorse neither A nor B). The A_c subpopulation is composed of individuals that believe A and are committed to it—they cannot be convinced by others in the population to change their mind. The relative sizes of the four regions in the schematic indicate the approximate relative sizes of the subpopulations at an intermediate point in a representative simulation. The fact that each pair of regions shares a boundary reflects that the four types of individuals can each interact with each other in the discrete-time dynamics described in the main text.

II. THE MODEL

We start by dividing the population into four different camps: There are those that hold extreme opinion A, those that hold the conceptually opposing opinion B (in the simplest case, just the negation of A), those that do not hold either A or B (we call these moderates), and those that hold A and are immune to the influence of others. We partially overload notation, using A, B, AB, and A_c, respectively, to denote the individuals in these four subpopulations and even sometimes the subpopulations themselves (Fig. 1).

The dynamics of the basic model we want to consider are deterministic, continuous and mean-field. One way to derive them is as a limit of the following random process: Suppose time is discrete and for each time step we select two nodes uniformly at random from the population (without replacement). We randomly choose one of them to be the speaker and the other the listener. If the speaker is an A or B and the listener is a B or A, respectively, then the listener is dissuaded from its extremist position and becomes an AB with probability 1. However if the listener is an AB, then with probability $\sigma(< 1)$ it remains a moderate when spoken to by an extremist from either camp, while with probability $1-\sigma$ it becomes an A if the speaker is an A and a B if the speaker is a B. Hence σ reflects a sort of stubbornness on the part of the AB listeners to migrate from their moderate position to an extreme view. In all other cases, there is no change in the state of the speaker and listener (Table I).

We now let n_A, n_B, and n_{AB} denote the fractions of the total population corresponding to the uncommitted A, B, and AB subpopulations, respectively. The fraction

Speaker	Listener	Outcome	Probability
A, A_c	B	A-AB	1
	AB	A-AB	σ
		A-A	$1-\sigma$
B	A	B-AB	1
	AB	B-AB	σ
		B-B	$1-\sigma$

TABLE I: A list of all possible interactions in our discrete-time finite-population model that can lead to a change of opinion. The outcome column gives the outcome of the interaction in the format Speaker-Listener. The value of σ lies on the unit interval and reflects the stubbornness of the moderates to stay moderate.

of the total population corresponding to the committed A subpopulation is represented by the parameter p to indicate that in any run of the dynamics, we will treat it as a constant, but we will consider how varying it leads to changes in the eventual state of the system. Using this new notation, the limits of a large population (growing like the total number of individuals N) and vanishing time step (shrinking like $1/N$) reduce our discrete dynamics to the pair of density rate equations:

$$\begin{aligned} \frac{dn_A}{dt} &= (1-\sigma)(p+n_A)n_{AB} - n_A n_B, \\ \frac{dn_B}{dt} &= (1-\sigma)n_B n_{AB} - (p+n_A)n_B. \end{aligned} \tag{1}$$

This system is more general than it may seem given the above derivation. For example, we can arrive at the same dynamics without assuming that every time an A talks to a B, the B converts to the AB position—it's enough to assume the conversion from B to AB happens with some probability $p' \leq 1$ (if we assume the same for conversions of A listeners to AB by B speakers). In fact, however, we don't even need to assume that conversion between the different states happens via one-on-one interactions. An equally plausible story could be proposed for individuals who spontaneously convert to a different state in response to the fractions of different subpopulations in their mean-field environment. In the strictest sense, the only assumption that we make for the behavior of the real-world system is the dynamics (1) themselves.

III. BEHAVIOR OF THE MODEL

If we run the system (1) to equilibrium for different values of p from the initial condition that all individuals are of opinion B except for a small fraction committed to opinion A, we find that the final state changes dramatically as we increase p through a critical value p_c (roughly 0.1 for values of σ around 0.5). For $p < p_c$ the system ends similar to how it started—almost all the individuals maintain B. However as p passes through p_c, the system undergoes a first-order (discontinuous) phase transition

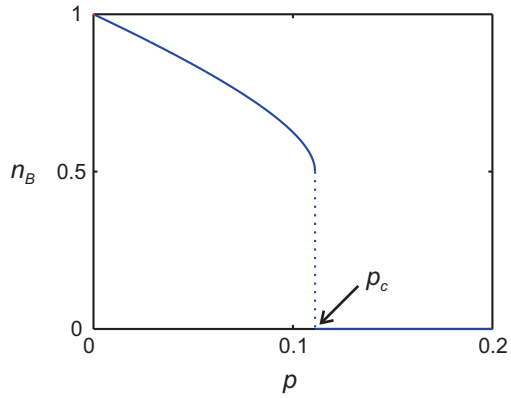

FIG. 2: The phase transition observed in the model system for $\sigma = 0.5$, as viewed in terms of the long-time fraction of B believers in the population (n_B). In partial analogy to physical systems exhibiting phase transitions, we might describe n_B as an order parameter for the system. A discontinuity in n_B occurs at $p_c = 1/9$ (the vertical dashed line is only meant to guide the eye; it is not part of the function). At higher values of p, n_B is zero. The curve for n_B can be computed exactly from (1).

and for $p > p_c$ the entire population quickly reaches a consensus on A (Fig. 2).

The critical value p_c is not a constant but a function of σ. Remarkably, we can solve (1) exactly to obtain this relationship in closed form (which has not been possible in previous studies on related models [3]). The phase portrait of (1) has two nullclines that osculate in a saddlenode bifurcation in the positive quadrant when $p = p_c$. From these two nullclines, we can derive four solutions of n_A and n_B in terms of p and σ. Two of these correspond to the two intersections resulting from the saddlenode bifurcation, so if we set the two corresponding expressions for n_A equal to each other, we obtain an expression for

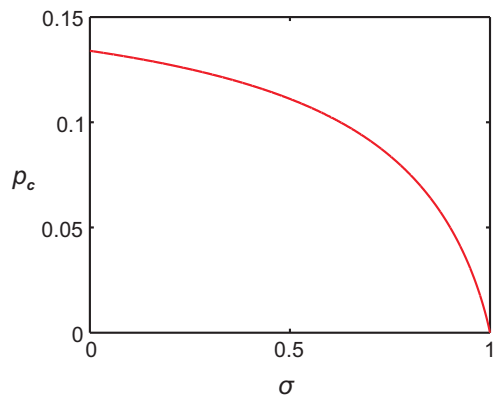

FIG. 3: A plot of the p_c-σ curve for the model of stubborn moderates. The mathematical form of this curve is given by Eq. (2).

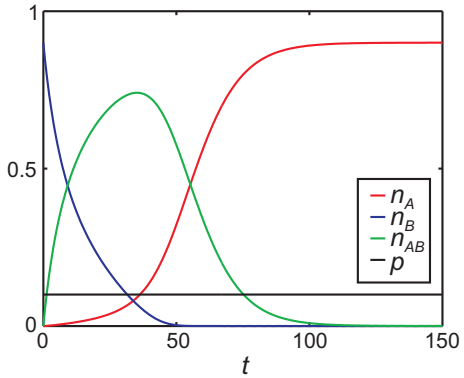

FIG. 4: Representative evolution of system (1) for $p = 0.1$ and $\sigma = 0.9$.

p_c in terms of σ. This simplifies to

$$\rho_c = \frac{4 - 7\sigma + 3\sigma^2 - 2\sqrt{3 - 11\sigma + 15\sigma^2 - 9\sigma^3 + 2\sigma^4}}{4 - 4\sigma + \sigma^2}. \tag{2}$$

This starts from $p_c = 1 - 3/\sqrt{2} \approx 0.134$ at $\sigma = 0$ and drops to $p_c = 0$ as σ approaches 1 (Fig. 3). Hence, not only does increased stubbornness of moderates make the population more vulnerable to ideological takeover by extremists, but it does so *arbitrarily*—for any $p > 0$, there is a σ large enough that this p is enough to sway the entire population to the extreme viewpoint.

It is interesting to ask why such an unintuitive relationship holds; is there a different way of viewing the problem that makes it more intuitive? First observe that increasing σ both reduces the flow of individuals to opinion B as well as the flow of individuals to opinion A. The former seems to suggest that an increase in σ should lower the p needed to convert the network to A (as we find) while the latter suggests that it should *raise* the p necessary to convert everyone to A. This quantitative tug-of-war is ultimately decided by the dynamics, but a more intuitive way of understanding the outcome is to regard σ as amplifying the extent to which the B's are drained away, allowing the committed A's to convert the AB population without significant competition from the B's. This explanation is consistent with the time evolution of the model system (Fig. 4).

We conclude this section by returning to the finite-population, discrete-time-step model that we introduced at the beginning. It is natural to consider the time till consensus on A when $p > p_c$ for the finite population model—the infinite population model suggests that it should be short. Extensive numerical simulations show that this consensus time, which we'll denote as T_c, scales like $\log N$ (Fig. 5). This agrees with a very recent study on a related model [3] and in fact has a simple informal explanation from dynamical systems theory: The finite dynamics are confined to the intersection of the half-spaces $n_A = 0$, $n_B = 0$ and $n_A + n_B < 1 - p$, and on that region they effectively jump between adjacent square regions with side lengths of order $1/N$. As N

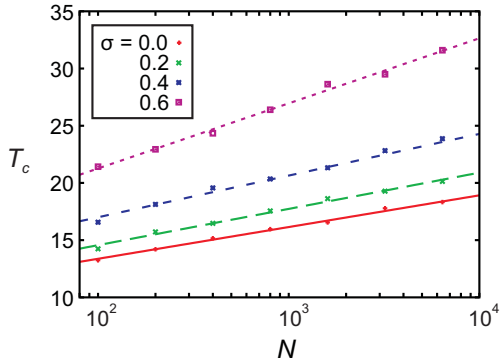

FIG. 5: The consensus time T_c as a function of N for the finite-population, discrete-time-step model with $p = 0.3$. The points indicate numerical data averaged over 100 trials, and the lines represent linear best fits in semilog space. The best fits corresponding to $\sigma = 0.0$, 0.2, 0.4, and 0.6 are $T_c = 1.20 \log N + 7.83$, $1.38 \log N + 8.23$, $1.58 \log N + 9.72$, and $2.47 \log N + 9.89$, respectively.

is increased, the finite dynamics approach (1), in which trajectories approach fixed points exponentially fast. We can examine when any one of these trajectories enters the square of size length $1/N$ centered on the fixed point (along the slower dimension) by setting $ae^{-bT_c} = c/N$ for constants a, b and c and solving for T_c as a function of N. As expected, we find $T_c \sim \log N$.

IV. RECONSIDERING THE ARBITRARY ASSUMPTIONS OF THE MODEL

In keeping the original model as simple as possible, we have neglected the possibility that a real social system could have additional dynamics at play that would weaken or eliminate the unintuitive relationship that we saw in the previous section. So we now turn to consider several extensions of the basic model that probe this idea directly, generalizing the model in various directions and then assessing whether the interesting features of the original model are robust in the generalized system.

First we examine the effect of including a spontaneous return to the AB state from the A and B states at a rate r. With this addition, the dynamics (1) become

$$\begin{aligned} \frac{dn_A}{dt} &= (1-\sigma)(p + n_A)n_{AB} - n_A n_B - r n_A, \\ \frac{dn_B}{dt} &= (1-\sigma)n_B n_{AB} - (p + n_A)n_B - r n_B. \end{aligned} \tag{3}$$

In this system, the new terms corresponding to a spontaneous deradicalization may reflect that extremist individuals could occasionally conclude by introspection or from observations in the nonsocial world around them that a moderate view makes more sense. Paradoxically however, we find from studying the phase portrait of (3) that increasing r increases the *per capita* impact of extreme factions—when we examine the phase portrait for

system (3), we find that increasing r reduces the number of extremists needed to convert the entire population.

Alternatively, we might include a committed fraction for B, B_c, with a fraction size of q. This changes the dynamics to

$$\begin{aligned} \frac{dn_A}{dt} &= (1-\sigma)(p+n_A)n_{AB} - n_A(q+n_B), \\ \frac{dn_B}{dt} &= (1-\sigma)(q+n_B)n_{AB} - (p+n_A)n_B. \end{aligned} \tag{4}$$

We again turn to phase portraits, now finding that if the difference between p and q is large enough, then two of the three intersection points of the two nullclines coalesce in a saddle node bifurcation and the population becomes either almost all convinced of A (if $p > q$) or almost all convinced of B (if $q > p$). The critical difference between p and q does depend on the actual values of p and q, and the exact mathematical relationship is complicated, but if q is fixed and small and we keep p greater than q, then the qualitative structure of this model is essentially the same as that of the original model (1).

We also might want to allow the moderates to influence extremists into giving up their extremism for a more balanced view. Assuming the moderates are effective at doing so with some (positive) rate constant r, the model equations become

$$\begin{aligned} \frac{dn_A}{dt} &= (1-\sigma)(p+n_A)n_{AB} - n_A n_B - r n_A n_{AB}, \\ \frac{dn_B}{dt} &= (1-\sigma)n_B n_{AB} - (p+n_A)n_B - r n_B n_{AB}. \end{aligned} \tag{5}$$

For small enough r, increases in r actually make the population more vulnerable to being overrun by extremism. However for larger r, increases in r make the population *less* vulnerable to extremist evangelicals and almost entirely moderate. This non-monotone relationship is due to a second bifurcation that produces a second phase transition. Although we leave a detailed study of this second bifurcation for future work, it will be critical to understand this structure to fully characterize the tradeoff between encouraging neutrality and incurring vulnerability toward system-wide extremism.

The way that we have used σ in the above systems keeps the rate constants of the reactions involving listening AB individuals below all other reactions. However we might want to consider the opposite case, where reactions involving AB listeners proceed *faster* than other reactions. This might make sense for a population that is highly balkanized into social groups where holding the opinion A or B is more comparable to group membership than intellectual conviction. In this social setting, individuals might migrate from one group to another for social rather than philosophical reasons—e.g. the new group better meets their relational needs or provides them with a stronger sense of social purpose. This complementary case in which AB's hastily abandon their unaffiliated status to join an extreme faction can be expressed formally by rescaling time such that the factor of $(1-\sigma)$ is switched to the reactions not involving AB individuals. From (1) this gives the system

$$\begin{aligned} \frac{dn_A}{dt} &= (p+n_A)n_{AB} - (1-\sigma)n_A n_B, \\ \frac{dn_B}{dt} &= n_B n_{AB} - (1-\sigma)(p+n_A)n_B. \end{aligned} \tag{6}$$

We again let σ range over the unit interval, and the resulting curve for p_c in terms of σ turns out to be

$$p_c = \frac{4-\sigma-2\sqrt{3-\sigma}}{4-4\sigma+\sigma^2}. \tag{7}$$

Oddly the final value of this curve, $3-2\sqrt{2}$, recently appeared in another study on a related but different model of opinion dynamics [4] and appeared in our own asymptotic analysis of a more complicated model of opinion dynamics than the one reported here. It remains to be seen whether this value might represent a universal limit for a certain class of opinion dynamics models.

V. FINAL REMARKS

Although we considered several important generalizations of our basic model, there are many more that could be investigated: We could look at the effect of including more intermediate states besides AB (e.g. spectra of opinion states like $A/AAB/ABB/B$). Or we could try further variations on the dynamics—for instance, allowing infrequent conversions directly from A to B and vice versa. We could also experiment with finite time dynamics that are more restrictive than allowing any pair of individuals to interact: Another study on a related model includes evidence that its results are robust to changes in connection topology [3] and our preliminary numerical work with random graphs suggests the same for our conclusions.

[1] J. L. Helgerson, in *Tenth Cosmos Club Spring Symposium* (National Intelligence Council, 2002).

[2] T. Sandler, D. G. Arce, and W. Enders, in *Copenhagen Consensus 2008* (Copenhagen Consensus Center, 2008).

[3] J. Xie, S. Sreenivasan, G. Korniss, W. Zhang, C. Lim, and B. K. Szymanski, Phys. Rev. E **84**, 011130 (2011).

[4] S. Galam and F. Jacobs, Physica A **381**, 366 (2007).

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