Journal of Mathematical Sociology, 32:142–163, 2008

ISSN: 0022-250X print/1545-5874 online
DOI: 10.1080/00222500801932014

Routledge
Taylor & Francis Group

Dynamics and Stability of Collective Action Norms

James A. Kitts
Columbia University, New York, New York, USA

A set of computational experiments investigated a model of formal and informal control, showing how selective incentives to work for the collective good may paradoxically lead to enforcement of antisocial norms that oppose the collective good. In these conditions, the widely cited effects of selective incentives, group cohesiveness, and second-order free riding on collective action may be inverted. Mathematical analysis provides some certain bounds on the model's behavior and relaxes several restrictive assumptions used in the simulation research. This complementary view deepens our understanding of second order social control as a solution to problems of collective action.

Keywords: collective action, computational experiments, incentives, norms

Any group whose collective welfare depends on efforts of individual members has a "regulatory interest" (Heckathorn, 1988) in encouraging contributions (or discouraging free riding) among members. Rational choice scholars have described two approaches that may be used to ensure collective action among members: In one version, *formal control*, a central regime assigns "selective incentives" (Olson, 1965) to encourage individuals to work for the collective good. In the other version, *informal norms* are enforced naturally through everyday social interaction (Homans, 1961), as peers pressure each other to serve the collective good.

Scholars customarily assume that norms emerging within a group of rational actors advocate prosocial behavior, which is costly to the individual and beneficial for others. From this assumption, scholars have predicted that groups in which members are better able to regulate

I acknowledge the support of a National Science Foundation grant, SES0433086. For valuable feedback and suggestions, I thank William Litsch.

Address correspondence to James A. Kitts, Graduate School of Business, 704 Uris Hall, 3022 Broadway, Columbia University, New York, NY 10027. E-mail: jak2190@columbia.edu

each other's actions ("cohesive" groups) should be more successful at producing collective goods (Heckathorn, 1988). Researchers have also used the assumption of prosocial norms to posit the obverse argument, that "second-order free riding" (Oliver, 1980) – in which actors fail to pressure or sanction peers – makes groups less successful at producing collective goods. Studies of informal control, including cohesiveness and second-order free riding, have helped us understand collective action given prosocial norms, but take those norms as exogenous.

A recent study (Kitts, 2006) investigated the interplay of formal and informal control, showing how selective incentives to work for the collective good may paradoxically lead group members to enforce *antisocial* norms that discourage peers' work. When such norms emerge, conventional strategies for promoting collective action, increasing selective incentives, strengthening peer influence, and preventing second-order free riding, may all backfire. The study began with the observation that the incentives used to promote collective action are often *rival*; that is, each worker who receives a selective incentive thereby diminishes the incentive available to other workers. As a result, providing selective incentives may invert the regulatory interests of rational actors, making them prefer collectively irrational norms. Kitts (2006) specified a simple mathematical model and proved a set of general propositions about actors' inclinations to work for the collective good (without social influence) and their regulatory interests (given the choice to work or shirk) under this stylized model.

To derive hypotheses about outcomes when collective action and social influence dynamics operate simultaneously, he made further specifying assumptions about the operation of social influence and ran computer simulations based on this elaborated model. These simulations were organized as computational experiments; outcomes were mapped as parameters of the model were manipulated. Those experiments in the model showed that normative pressure can turn against the collective good in the condition where incentives are both rival and valuable. When such deleterious informal norms affect group members' behavior, then collective action suffers.

The computational experiments drew an intuitive map of equilibrium outcomes, but executing the simulations required a number of assumptions to flesh out the influence process. Further, the laudable flexibility of computer simulation came at the cost of generality of conclusions. In this article, I describe some generic constraints on inference from simulation and analytically develop new propositions about the dynamic properties of the model. These propositions are more modest in that they serve only as bounds on the observable behavior within the simulations, but they are more general in that

they do not depend on auxiliary assumptions about the form or strength of social influence and they do not depend on initial conditions. For portions of the parameter space, this simple analysis shows the outcome with certainty and without running a simulation.

MODEL

Assume a set of individuals who value a jointly produced good that is not "excludable" from those who fail to contribute. Specifically, each member faces a first-order choice to contribute to the collective good (*Work* or *Shirk*) and a second-order choice to influence others' first-order choices (*Promote* or *Oppose* peers' work). The production function, G, represents the net benefit received by each member i from all N members' work choices:

$$\textit{Production Function}: \qquad G(w_i, n) = n\, g + (g - c)w_i \tag{1}$$

where w_i denotes actor i's work choice ($w_i = 1$ for "Work"; $w_i = 0$ for "Shirk"), n is the total number of i's peers who are working ($0 \leq n \leq N - 1$), g is a parameter representing i's benefit created by each member's work, and c is a parameter representing the cost of working. Actor i receives a benefit g for each of n peers who works, and one more unit g if she chooses to work herself. The cost of working for the collective good, c, is also a constant decrement for each worker, but this cost is borne privately by i while the benefit g is enjoyed by all.

An incentive (μ) is awarded only to workers ($w_i = 1$), who work for the collective good. If the incentive is rival, each actor's choice to work diminishes the share of selective incentives remaining for other workers. The reward function R assumes that if and only if actor i works, she receives the same reward as her n working peers:

$$\textit{Reward Function}: \qquad R(w_i, n) = \mu\left(1 - \lambda \frac{n}{n+1}\right) w_i \tag{2}$$

where λ is a parameter representing the rivalness of the incentive ($0 \leq \lambda \leq 1$), μ is the total value of the incentive ($\mu \geq 0$), and n and w_i are as defined in Eq. (1). In the purely nonrival scenario ($\lambda = 0$), all workers receive the full value of the incentive (μ) regardless of others' choices. In the perfectly rival condition ($\lambda = 1$), worker i must share the selective incentive equally with n working peers, yielding a share equal to $\mu/(n + 1)$. Intermediate values of λ allow for a range of partially rival incentives.

An actor i's utility is the sum of the production function G and reward function R, given her own work choice (w_i) and the number

of her peers who work (n):

$$\textit{Utility Function}: \qquad U(w_i, n) = gn + w_i\left(g - c + \mu\left(1 - \lambda\frac{n}{n+1}\right)\right) \tag{3}$$

where the g and c parameters are as defined in (1) and μ and λ are as defined in (2). We may interpret actor i's utility as the total production by peers (gn), plus the personal cost and benefit of working $(g - c)w_i$ and i's share of the selective incentive, $\mu(1 - \lambda\frac{n}{n+1})w_i$.

An actor's Inclination to Work (IW) is the change in utility associated with the work choice:

$$\textit{Inclination to work}: \qquad IW(n) = \frac{\partial U}{\partial w_i} = (g - c) + \mu\left(1 - \lambda\frac{n}{n+1}\right) \tag{4}$$

A positive inclination to work ($IW > 0$) means that the actor will profit from choosing to work, while $IW < 0$ implies a net loss for choosing to work.[1] A larger number of peers working always implies a smaller share of a rival incentive for any worker and thus a lower inclination to work. If $\lambda = 0$ or no peers are working ($n = 0$), then there is no expected loss to peers and Eq. (4) reduces to $IW = g - c + \mu$.

Regulatory interests also may depend on the number of peers working and on an actor's own work choice. All members receive a personal benefit g from each peer's work for the collective good, so the baseline regulatory interest is positive. However, when a rival ($\lambda > 0$) incentive is valuable, it may also create a perverse regulatory interest in opposing peers' participation, because increasing the number of peers who work (n) will also increase the loss of the incentive to peers. The partial derivative of U with respect to n represents the change in utility due to a marginal change in the number of peers working. This yields the *regulatory interest function*:

$$\textit{Regulatory Interest Function}: \qquad RI(w_i, n) = \frac{\partial U}{\partial n} = g - w_i\frac{\lambda\mu}{(n+1)^2} \tag{5}$$

Kitts (2006) analyzed (4) and (5) to demonstrate five general propositions about actors' inclinations to work and their regulatory interests operating in isolation. Notably, he proved that increasing the value of

[1]The inclination to work remains the same whether the work choice is treated as binary $w_i \in \{0, 1\}$ or as a proportion of effort in the interval [0, 1]. Because an actor's utility depends linearly on w_i, she will never prefer an interior value of w_i (partial effort) to either extreme value.

a rival selective incentive (or increasing the rivalness of a valuable incentive) diminishes regulatory interests, regardless of the parameters of the production function, g or c, for any finite group size (N) or current level of collective action (n) in the model. However, increasing the value of the incentive obviously increases the inclination to work. To consider outcomes where both processes operate simultaneously, he used simulation.

Computational Experiments

Actors in the simulations make a first-order choice (*Work* or *Shirk*) followed by a second-order choice (*Promote* work, *Oppose* work, or *Abstain* from enforcement); in each case they are myopic, selecting the locally preferred option. Actors who Promote or Oppose must pay a cost of enforcement, e.

The choice to pressure peers requires each actor to predict the outcome of her enforcement efforts, but calibrating this prediction is complicated for actors: Changes in peers' work choices affect their own regulatory interests, possibly leading to changes in norms before an actor has observed the direct effect of her own pressure on peers' work choices. Assuming that actors have no means to accurately calculate the scope of their own influence, Kitts (2006) assigned a uniform random variable to represent actors' subjective *scope* of influence. For each actor i, θ_i is the maximum number of peers i expects to be able to influence, varying from the extreme belief that i's social pressure will have no effect ($\theta_i = 0$) to the extreme belief that i can convince all peers to change ($\theta_i = N - 1$).

The total force acting upon each actor to work or shirk in the simulation is the simple sum, V, of all actors' choices to *Promote* ($v_i = 1$), *Oppose* ($v_i = -1$), and *Abstain* ($v_i = 0$). This sum of normative pressure must be weighed against individuals' own inclinations (IW). The strength of social influence is represented as a parameter α, "group cohesiveness." This parameter ($0 \leq \alpha < 1$) weighs the extent that each actor's work choice (w_i) is influenced by social pressure (V) versus the member's own inclination to work as in (4). At $\alpha = 0$, social norms have no effect on behavior, but as α approaches 1.0, behavior is increasingly driven by peer pressure.

Two computational experiments let the first-order and second-order processes operate in tandem. Actors' choices to promote or oppose peers' work may depend on their own choices to work, and their choices to work may depend on their peers' choices to promote or oppose. The simulation used a conventional sequential decision model (cf. Heckathorn, 1990) to derive predictions from the model: The

work choice was updated in random sequence over actors, then the enforcement choice was updated in the same fashion, with these two steps repeated for a number of rounds sufficient to ensure a stable result.

Both experiments predicted the stable proportion of actors choosing to work (Participation) in the same baseline collective action problem ($g = 1, c = 5, e = 2, N = 10$) beginning from the same initial conditions (universal free riding at both the first and second levels). The first experiment investigated participation over the space of parameters *value* (μ) of incentives and *cohesiveness* (α), at both the minimum ($\lambda = 0$) and maximum ($\lambda = 1$) of rivalness of incentives. This investigation showed that in the model collective action depends on a three-way interaction of the *rivalness* of incentives (ranging from independence to zero-sum competition), *value* of selective incentives (ranging from worthless to very valuable), and group *cohesiveness* (ranging from ineffective peer influence to extremely powerful peer influence). Under the condition of high rivalness, selective incentives and cohesiveness each have positive main effects, but collective action collapses when both incentives and cohesiveness are strong. Under the condition of low rivalness, this interaction of cohesiveness and the value of incentives does not obtain.

The second experiment considered the middle range of cohesiveness (also allowing for actor-level heterogeneity in susceptibility to influence) so peer influence and individuals' inclinations were both important. Those experiments mapped the response surface over a broad range of incentive value and over the entire range of rivalness from $\lambda = 0$ to $\lambda = 1$. The resulting map revealed an intriguing nonmonotonic relationship of incentive value to collective action, and an interaction with rivalness. At low rivalness, collective action increases with incentive value, then crashes precipitously, and then rises again. Above a critical value of rivalness, this curious relationship of incentive value to collective action turns upside-down. Mathematical analysis in this article will provide a decisive explanation of this pattern and several other findings from the computational experiment.

Limitations of Simulation

After first becoming prominent in the physical and natural sciences, simulation is now proliferating in the social sciences as a tool for developing and refining theory (Macy and Willer, 2002). Indeed, freedom from concerns about analytical tractability has allowed scholars great flexibility to consider complex social dynamics in computational experiments. As a tradeoff, simulation has only a narrow capacity to

prove general statements about a model. Simulation based on a model can prove minimally that a result is a possible consequence of a set of assumptions (Axelrod, 1997). Simulation can also show how the model typically behaves under precisely specified conditions, but cannot show that an observed pattern will obtain in any conditions not explicitly examined. For example, sensitivity analysis may show us that a particular result obtains from a variety of different starting points; however, we cannot conclude on the basis of simulation that the result will obtain from *any* initial condition (unless of course the state space is narrow enough to explore all conditions). When the model is stochastic, we also cannot conclude that a particular outcome will *always* (or never) result from the same inputs even after observing many simulation results. Lastly, performing simulations over a very long time span may show us that results are remarkably consistent over time, but we cannot conclude on the basis of simulation that a pattern is dynamically stable—only that it appears stable over the time span observed. Mathematical analysis may sometimes be used as a complement to simulation to make such general claims about a model's generality, robustness, insensitivity to initial conditions, and dynamic stability.[2]

Kitts (2006) performed fine-grained manipulations of parameters to describe the shape of the model's response surface, and used a large number of replications to guard against sampling error on the model's stochastic response. He also provided sensitivity analyses in arrays of simulations that manipulated otherwise fixed (non-experimental) parameters, such as the distribution of influence scope (θ), the size of the group (N), and the cost of enforcement (e), or altered auxiliary model assumptions, such as protocols for updating actor choices. Such procedures increase our confidence in the robustness of qualitative conclusions, but simulation research cannot demonstrate anything strictly about the model's behavior in conditions not examined.

Having described some limitations of the simulation method, I will detail some of the specific assumptions that Kitts (2006) used for the

[2]For example, Kitts (2006) asserted that some simulation outcomes were "converged," or dynamically stable. He pointed out that in the sequential decision model if all actors pass up a turn without changing their work or enforcement choices, then the model has reached equilibrium. In this model, there is no force to change behavior if all actors are presently occupying their locally-preferred choice. Thus, if no neighbors have changed their behavior then the available choices remain the same to each and no actors will ever change. This conclusion is easy to show mathematically, but simulation alone would not allow this insight.

purpose of making the sequential decision model computationally tractable:

- Simulations began with standard initial conditions, with all actors shirking at the first order (not working for the collective good) and at the second order (not enforcing norms on their peers).
- The subjective scope of influence (θ) among actors was random, drawn from a uniform distribution in $[0, N-1]$ and unrelated to any other actor characteristics or behaviors.
- Susceptibility to peer influence (α) was either homogeneous across actors (Experiment 1) or randomly distributed across actors and unrelated to other actor characteristics (Experiment 2).
- Social relations were undifferentiated. All actors were influenced by the total pressure of group members (a simple sum of all members' enforcement choices), not differently by particular peers.

This article jettisons these assumptions and provides new general propositions about the model's behavior that will hold true regardless of the specification of the influence process, either as it is perceived by actors or as it plays out in behavior.

Analysis of Model

To give a more general characterization, I will now show that there are boundaries on the model's behavior that do not depend on several of the specifying assumptions used in the computational experiments. These boundaries can be used to predict outcomes analytically, without running simulations, in specified regions of the parameter space. They also improve our understanding of the model's behavior overall, accounting for the shape of the response surface and telling us where initial conditions or other details of the computational model may or may not influence outcomes.

The first boundary, Voluntary Participation, obtains where the actor's cost of working (c) equals the benefit of working when all $N-1$ peers are working, or $\mu(1-\lambda\frac{N-1}{N})+g$. It is convenient to express this boundary as an inequality on the range of the value of the incentive:

$$\textit{Voluntary Participation Boundary}:^{3} \qquad \mu > \frac{(c-g)N}{N-(N-1)\lambda} \qquad (6)$$

[3]By "voluntary," I mean only that behavior is not coerced by *peers*. It is obviously motivated by the selective incentive, so it is in no way analogous to "volunteerism."

When the incentive is on or below this curve, it is insufficient to compensate all members for working, so the temptation to free ride appears. When the incentive is above this curve, free riding will never be profitable in any state of the model, regardless of the other parameter values.

Proposition 1. *If the incentive value is above the Voluntary Participation boundary, no actor will be inclined to free ride.*

This is easily proven. Recall that an actor's inclination to work (*IW*) will be a decreasing function of the number of peers working (n) whenever $\lambda > 0$ and $\mu > 0$, because of the falling share of the selective incentive. Further, an actor's inclination to work will be independent of the number of peers working (n) whenever $\lambda = 0$ or $\mu = 0$, so *IW* must be a nonincreasing function of n in the model. If an actor is inclined to work ($IW > 0$) when all peers are working ($n = N - 1$), she will thus be inclined to work at all lesser levels of n. Therefore, all actors are personally inclined to work above the Voluntary Participation boundary (and will work unless antisocial norms prevent them from doing so).

If $N - 1$ is the number of i's peers and n is the number of peers who are working, then $N - n - 1$ is the number of peers who are shirking and $(N - n - 1)g$ is the quantity of collective good that is lost to i by their shirking. The next boundary, Futility of Promotion, exists where the quantity of collective good that i would receive by converting all of the shirking peers to working, or $(N - n - 1)g$, does not exceed i's cost of enforcing norms, e. This condition, $e \geq (N - n - 1)g$, guarantees that promoting work will never be profitable for any actor. It will be useful to place all fixed parameters on the right side to give a range of the variable n where promoting work can never be profitable:

$$\textit{Futility of Promotion Boundary}: \qquad n \geq N - 1 - \frac{e}{g} \tag{7}$$

This boundary is obviously defined with respect to the number of peers working (n), but does not depend on actors' influence scope (θ) or susceptibility to influence (α).

Proposition 2. *When collective action reaches or exceeds the Futility of Promotion boundary, no actor will choose to enforce prosocial norms.*

This proposition is also easy to prove. If (7) is satisfied, promoting work is not profitable for an actor who may convince all shirking peers to work. We may assume without loss of generality that i has

unlimited scope of influence because if i is unable to influence all shirking peers then the benefit of converting shirkers will be diminished and P2 will still hold. Similarly, competition over a rival incentive may make promoting work less attractive (only for workers who expect to receive the incentive), but it will not affect the absolute boundary where prosocial norms cannot appear. If the collective good is a strictly increasing function of the number of actors who participate, then if (7) is satisfied promoting working will never be profitable regardless of an actor's work choice (w_i), how many peers an actor expects to influence (θ), or how effective influence is in the model (α). This proposition is not as general as the others because it depends on the variable state of the model (n), but it will be useful for proving propositions about system-level behavior.

The other boundaries concern the conditions under which antisocial norms, opposing peers' work for the collective good, will never appear under the model. An actor i's willingness to oppose peers' work at a particular time depends dynamically on i's own work choice (w_i), on the current level of work among i's peers (n), and on i's expected scope of influence (θ_i). It is helpful then to find a boundary for the computational experiment where antisocial norms will *never* appear, regardless of w_i or θ_i. I can restrict the discussion here without loss of generality to the case where the scope of influence (θ_i) is unlimited, because constraining i's subjective scope of influence can only make opposing work less beneficial for i (e.g. if it makes a worker i think she is unable to hoard the entire incentive for herself by opposing others' work). If I identify a condition where opposing peers' work will not be profitable for an actor with unlimited scope of influence then it will also apply for actors of limited scope, including (trivially) actors who have no power to influence anyone.

An important boundary exists where the incentive that could be hoarded by forcing all n working peers to shirk, or $\frac{n}{n+1}\mu\lambda$, falls below the enforcement cost e plus the loss of n peers' work for the collective good, or $e + ng$. It is convenient to rearrange this inequality as a range of incentive value–a value of μ below which enforcing antisocial norms will never be profitable at a given level of collective action among peers (n):

$$\textit{Immunity From Opposition Boundary}: \qquad \mu \leq \frac{(n+1)(e+ng)}{\lambda n} \tag{8}$$

This inequality provides a boundary for μ (based on the parameters e, g, and λ) where antisocial norms cannot appear for each current level of collective action (n). To identify the lowermost boundary, where no

antisocial norms can ever occur, I find the value of n in $\{1, 2, 3, \ldots, N-1\}$ that minimizes the right side of the inequality and substitute this value for n. If the resulting inequality is satisfied for a value μ, then μ is below the Immunity from Opposition boundaries for all possible levels of collective action. The resulting inequality will define the global boundary over all states of the model, below which no antisocial norms can ever emerge. For the values of fixed parameters ($e = 2; g = 1$) used in the simulation study, $n = 1$ is such a critical value. I will refer to this lowermost curve as the Strong Immunity boundary:

$$\textit{Strong Immunity Boundary}: \qquad \mu \leq \frac{2(e+g)}{\lambda} \tag{9}$$

The Strong Immunity boundary here exists where the extra incentive that could be hoarded by forcing a single working peer to shirk ($\lambda\mu/2$) would fail to exceed the benefit lost (g) by doing so plus the cost (e) of enforcement. Proposition 3 follows:

Proposition 3. *If the incentive value is within the Strong Immunity boundary, no actor will enforce antisocial norms.*

Antisocial norms can never appear (or be stable if they are introduced exogenously) in the model when μ is on or below the Strong Immunity boundary. If (8) identifies a set of boundaries below which opposing work will never be profitable (at given levels of n) then the lowermost of these curves is the boundary below which opposing work will never be profitable, for any n. In this study, if antisocial norms are not profitable in the extreme case of $n = 1$—as specified by the Strong Immunity boundary (9)—they will never be profitable in any other state of the model. And this does not depend on heterogeneous parameters θ or α so if it is true for any actor, it is true for all others.

I define a more liberal boundary, where existing universal collective action ($w_i = 1$ and $n = N - 1$) is guaranteed protection from antisocial norms. Substituting $N - 1$ for n in (8) yields the Weak Immunity boundary:

$$\textit{Weak Immunity Boundary}: \qquad \mu \leq \frac{N(e + (N-1)g)}{\lambda(N-1)} \tag{10}$$

When all peers are already working for the collective good, no actor will profit from opposing work among peers if μ is below this boundary. In this condition the system can also maintain full productivity without social pressure. Thus, the All-Work/None-Enforce equilibrium in

this region is stable regardless of subjective effectiveness (θ) or objective effectiveness (α) of influence, but is not guaranteed to obtain from any initial level of collective action, n.

The Immunity from Opposition boundary (8) would generate comparable curves for every other level of n between 1 and $N-1$, but the given boundaries will be sufficient to reproduce and explain the overall shape of the response surface.

Previous propositions have defined specific constraints on individual actors' behavior with respect to regions of the parameter space. But knowing the actor-level constraints does not allow us to directly predict the system-level behavior. For example, knowing that actors will not be inclined to free ride above the Voluntary Participation boundary (as specified in P1) does not guarantee anything about the outcome of the dynamic system, as antisocial norms can force actors to shirk above the Voluntary Participation boundary and prosocial norms may prevent free-riding below the boundary. By combining the various constraints on actor choices, however, I can derive the ultimate outcome with certainty in some regions and also enrich our understanding of the model's behavior outside those regions.

To begin this broader analytic investigation of the model, I will demonstrate a region where All-Work/None-Enforce is a globally stable equilibrium. An equilibrium is globally stable when all trajectories of the model approach it and do not escape it, regardless of the initial conditions and following any exogenous perturbation. In this region, I show that universal voluntary participation will appear and remain under the model regardless of initial conditions, and regardless of the subjective effectiveness (θ) or objective effectiveness (α) of influence.

Proposition 4 describes the privileged condition where collective action flourishes without risk of either free riding or antisocial norms:

Proposition 4. *In the range of incentive value above the Voluntary Participation boundary and within the Strong Immunity boundary, All-Work/None-Enforce is a globally stable equilibrium.*

I have already shown that all actors will be inclined to work above the Voluntary Participation boundary (and will work unless prevented from doing so by antisocial norms). I have also already shown that no actors will enforce antisocial norms below the Strong Immunity boundary. If all actors are inclined to work (P1) and none will oppose work (P3)—where both inequalities are satisfied—then all actors will work. And once all actors are already working, then enforcement of prosocial norms on peers could never be profitable and thus no one will promote (P2). Although individuals may initially promote work by

peers (so prosocial norms may be found in this region, outside of equilibrium), this is transitory and enforcement will cease once the trajectory has passed the Futility of Promotion boundary on the path toward the cooperative equilibrium. If no actors will enforce any norm at equilibrium, then the final result is independent of susceptibility to influence (α). I have thus shown that this outcome obtains regardless of values for actors' scope of influence (θ), susceptibility to influence (α), or the initial or present state of working among actors (n or w).

We may appreciate the certainty that antisocial norms will never be stable in the Strong Immunity condition, but this does not guarantee that antisocial norms will be stable above that conservative boundary. At higher levels of incentive value, it will always be profitable for a worker to coerce a single working peer to shirk when $n = 1$, but it may not be profitable for her to enforce antisocial norms once a large number of peers are already working. If the current position of the model in state space may affect its qualitative trajectory, then its final resting place may depend on initial conditions or exogenous shocks. The outcome also depends on details of the influence process. For example, actors with low subjective scope of influence will be particularly sensitive to the current level of work among peers in their decision to enforce norms. Low scope workers will be more likely to free ride on enforcement of antisocial norms when collective action is widespread.

Between the Voluntary Participation and the Weak Immunity from Opposition boundaries, we can make a weaker (but no less certain) claim about the All-Work/None-Enforce equilibrium:

Proposition 5. *In the range of incentive value above the Voluntary Participation and Strong Immunity boundaries but within the Weak Immunity boundary, the All-Work/None-Enforce equilibrium is locally (but not globally) stable.*

That is, in this region of the parameter space all trajectories sufficiently near the All-Work/None-Enforce equilibrium will converge toward that equilibrium, but this result cannot be guaranteed for any arbitrary initial condition. Nor can I describe the size of the basin of attraction around that equilibrium or the out-of-equilibrium behavior outside that basin without making assumptions about the influence process. Proposition 5 is proven in the same way as P4, so I abbreviate the explanation. In this region, the incentive is valuable enough (given other parameters) to induce all members to work, but weak enough that none will profit from opposing peers' work when all peers are working.

Having defined these three general boundaries (Voluntary Participation, Strong Immunity, and Weak Immunity) and described the

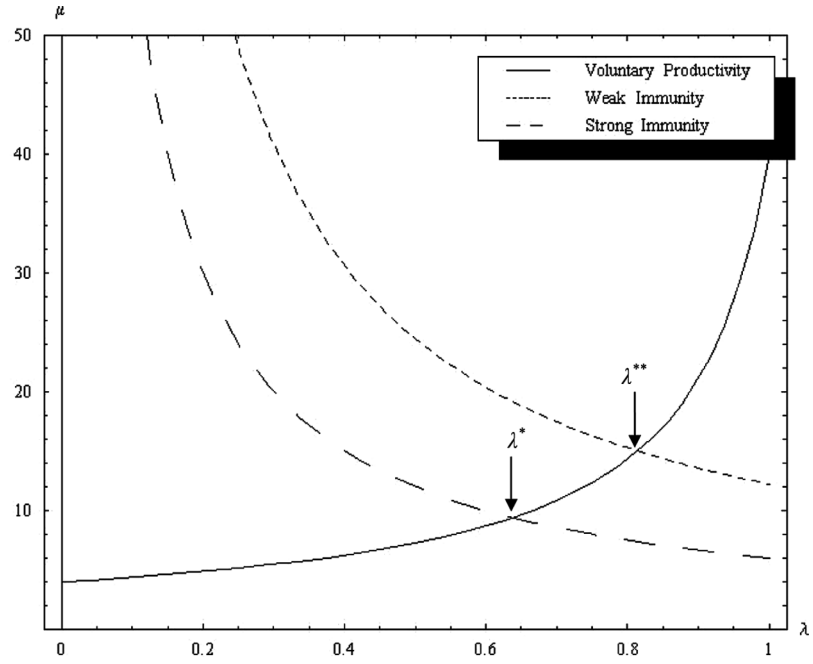

FIGURE 1 Qualitative model boundaries ($g = 1$, $c = 5$, $e = 2$, $N = 10$).

model's disposition with regard to these boundaries, I will plot them in a way that facilitates comparison to the simulation results.[4] Figure 1 depicts the three boundaries, given the numerical values of fixed parameters ($g = 1$, $c = 5$, $e = 2$, $N = 10$) used in the simulations:

The intersections of the Voluntary Participation boundary with the Strong Immunity and Weak Immunity boundaries occur at particular values of rivalness (λ) and these values may represent theoretically important bifurcation points in the model. On the left side of these critical values, the space between the boundaries represents effective selective incentives and immunity (weak or strong) from antisocial norms, whereas the space between the boundaries on the right side of the bifurcation points represents dual threats of free riding and antisocial norms. To find the intersection of Voluntary Participation and Strong Immunity, I set $\frac{(c-g)N}{N-(N-1)\lambda} = \frac{2(e+g)}{\lambda}$ and solve for λ, yielding $\lambda^* = \frac{2(e+g)N}{(c+2e+g)N-2(e+g)}$. Above this value of rivalness, the All Work/None

[4]The Immunity from Opposition and Futility of Promotion boundaries depend on the state of the model (the current number of peers working, n), so they cannot be included in this figure.

Enforce equilibrium cannot be guaranteed.[5] For the numerical case used in the experiments, this value is approximately $\lambda^* = .638$. To find the intersection of Voluntary Participation and Weak Immunity, I set $\frac{(c-g)N}{N-(N-1)\lambda} = \frac{(e+(N-1)g)N}{(N-1)\lambda}$ and solve for λ, yielding $\lambda^{**} = \frac{(e+(N-1)g)N}{(c+e+(N-2)g)(N-1)}$. When rivalness exceeds this critical value, antisocial norms can be individually profitable regardless of the current level of collective action (although they will still be enforced only by actors with sufficient scope).[6] For the numerical case used in the experiments, this value is approximately $\lambda^{**} = .815$.

In the simulation, the numerical range of incentive value for the region described in P4 is $40/(10 - 9\lambda) < \mu < 6/\lambda$ when $0 < \lambda < \lambda^*$; the range is simply $\mu > 4$ when rivalness is zero, because antisocial norms will never emerge in that condition, and the region does not exist when $\lambda \geq \lambda^*$. The range of incentive value for the region described in P5 is $6/\lambda < \mu < 12.22/\lambda$ when $0 < \lambda < \lambda^{**}$; the range is undefined when rivalness is zero, because in that condition there is no value of μ where the All-Work/None-Enforce equilibrium exists without being globally stable, and it does not exist when $\lambda \geq \lambda^{**}$.

Application to Simulation Results

I also performed computational experiments replicating the protocol specified in Kitts (2006), yielding response surfaces for working, promoting work, and opposing work over the same ranges of incentive value (μ) and rivalness (λ). Figure 2 shows the proportion working for the collective good (Participation), Figure 3 shows the proportion promoting work, and Figure 4 shows the proportion opposing work over this space of incentive value (μ) and rivalness (λ).

First, see that productivity is low on the right edge of Figure 2, below the Voluntary Participation boundary. Figures 3 and 4 show that peer pressure to work explodes in this region, and is responsible for much of the cooperation that emerges, but this area is free of antisocial norms.

Now compare the fin-shaped region in all three simulation figures, where all actors work and none enforce, to the analytical boundaries in Figure 1. The globally stable equilibrium described in P4 lies between the Voluntary Participation and Strong Immunity boundaries, to the

[5]This value approaches $\lambda = 2(e + g)/(c + 2e + g)$ as the group size (N) grows large. For the values of per capita productiveness, cost of working, and cost of enforcement here, this limiting value is 0.6.

[6]This value rapidly approaches $\lambda = 1.0$ as the group size (N) grows large. That is, for arbitrarily large groups, we can expect the locally stable equilibrium to exist unless the incentive is near perfect rivalness.

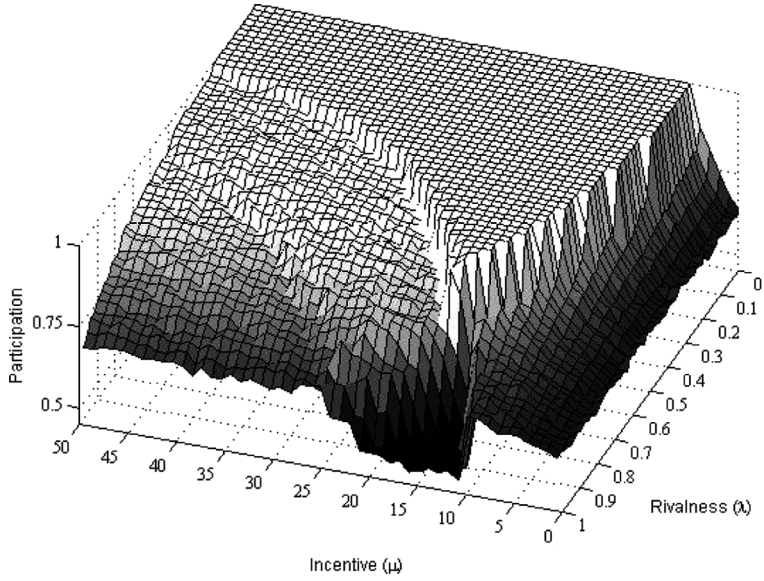

FIGURE 2 Collective action over range of rivalness and value of incentive ($g = 1$, $c = 5$, $e = 2$, $N = 10$).

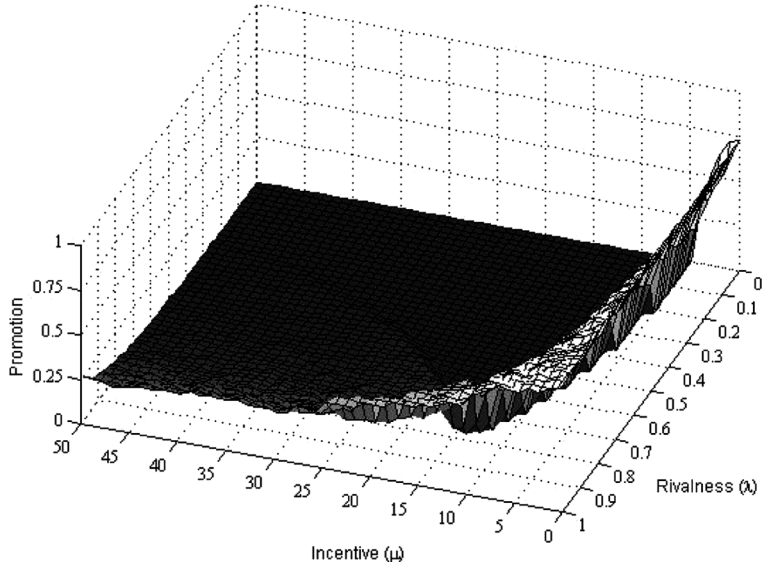

FIGURE 3 Prosocial norms over range of rivalness and value of incentive ($g = 1$, $c = 5$, $e = 2$, $N = 10$).

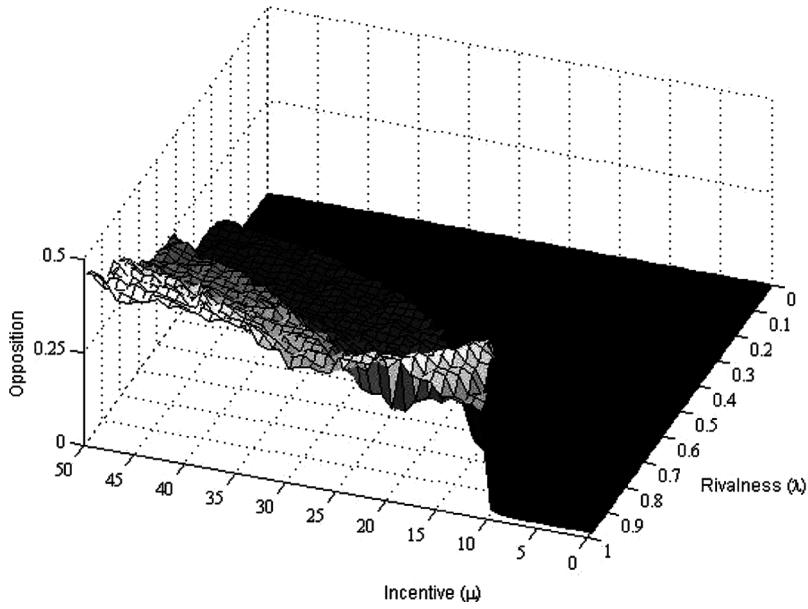

FIGURE 4 Antisocial norms over range of rivalness and value of incentive ($g = 1$, $c = 5$, $e = 2$, $N = 10$).

left of their intersection. This matches the fin-shaped region exactly on the right side (at the Voluntary Participation boundary) of Figure 2, but the plateau of cooperation in the simulations actually extends all the way to the Weak Immunity boundary. This shows that the simulations overwhelmingly reached the locally stable cooperative equilibrium under the simple influence model. In the strip between the Strong Immunity and Weak Immunity boundaries, this All-Work/None-Enforce equilibrium is in no sense determined. Of course, if the model begins with universal cooperation as an initial condition, this equilibrium would be reached by assumption, but the simulations began with universal free riding, the least favorable condition for collective action in the model.

At the intersection of Voluntary Participation and Weak Immunity boundaries (λ^{**}) the intriguing bifurcation in model behavior appears: The concavity of the relationship between productivity and the incentive value reverses when rivalness exceeds this bifurcation point, as the All-Work/None-Enforce equilibrium yields to the deleterious region where the liabilities of first-order free riding and of second-order antisocial norms both impede collective action. In this region, the individual share of the incentive is too small to support full productivity,

but yet is large enough to invite antisocial norms. This deleterious region has the lowest productivity in Figure 2.

DISCUSSION

It will be instructive to consider the dynamics of cooperation in the region between the Strong Immunity and Weak Immunity boundaries, where the cooperative equilibrium exists but the system may be trapped in a suboptimal equilibrium enforced by antisocial norms. Examination of model trajectories from a start of universal free-riding reveals that the early actor(s) to work in the deleterious region may often oppose work among peers to hoard the incentive. We can regard this as a novel and distinct "start-up problem" for collective action, where the early-movers have a perverse incentive to oppose the late-comers. Although the region described in P5 also faces this problem, the system in that region only needs a critical mass of members to begin working to sponsor a cascade of collective action that dissolves antisocial norms, yielding the All-Work/None-Enforce equilibrium.

If informal control can turn against the collective good, this suggests an intriguing twist on the second-order free-rider problem. Free-riding at the second order can provide a net benefit for the collective in some conditions as well as a net loss in others. The deleterious region plagued by both free riding and antisocial norms highlights a particularly interesting result: The role of second-order control depends not only on the parameters of the collective action problem, but on the current level of participation. Second-order control is a threat in lower levels of productivity, because a small number of workers are more likely to oppose competition from other workers, but it is necessary in higher levels of productivity, where the number of workers exceeds the level that can be motivated by the incentive. Thus, the effectiveness (and costs) of social influence may have opposite effects at low and high levels of collective action, holding constant all parameters of the collective action problem.

The original computational experiments were accompanied by sensitivity analyses that animated the model response surface as auxiliary parameters were varied. Those simulations included a broad range of enforcement costs, varying the incidence of second-order free riding.[7] Some of the results were intuitive, such as the negative effect of enforcement cost on participation in the lowest range of incentive value, where

[7]Kitts (2006) provided online appendices to animate the surface plots of Work, Promote, and Oppose choices as enforcement costs rise from $e = 0$ to $e = 5$: ***http://www.columbia.edu/~jak2190/ASR2006/***

collective action depends decisively on prosocial norms and antisocial norms cannot be found. The observed result for the high-value/high rivalness-corner was more sophisticated. Collective action was moderately widespread above the Voluntary Participation boundary (and outside the deleterious region) when enforcement entailed no cost. Then, as enforcement costs rose, the cooperative equilibrium spread out over a broader space, and thus a 'second-order free riding benefit' led some regions to overcome shirking. However, in other regions the level of participation fell markedly as enforcement costs rose. Thus, the sensitivity analyses suggested that the relationship between second order free riding and collective action may be more complex than previously understood. I pursue this question here.

To complement the previously published sensitivity analyses for the simulations, I will discuss some ways that the certain bounds on model behavior described here also vary with non-experimental parameters N and e. This web page provides animated versions of Figure 1, illustrating how the depicted boundaries depend on N and e: http://www.columbia.edu/~jak2190/JMS2008/

First consider the boundaries as enforcement costs (e) rise from 0 to 5, corresponding to the sensitivity analyses reported above. See that both Strong and Weak Immunity boundaries march outward into higher ranges of μ as e increases. The expanding All-Work/None-Enforce region accounts for the areas where enforcement costs increased cooperation in the computational experiment.[8] This happens because increasing the cost of enforcement broadens the conditions under which enforcement of antisocial norms will not be profitable. On the other hand, collective action outside the boundaries of the All-Work/None-Enforce region depends on prosocial norms, and thus rising enforcement costs have a negative impact on collective action virtually everywhere else. This is the familiar second-order free rider problem.

[8]Remember that the Strong Immunity boundary is simply the Immunity from Opposition boundary evaluated at the level of n that minimizes the right side of (8). It is necessary to return to this definition and slightly revise the Strong Immunity boundary for higher e, because when $e > 2$ then $n = 2$ is a strictly more favorable condition for antisocial norms than $n = 1$. In order for the Strong Immunity boundary to indicate the global boundary where no antisocial norms will appear in any state of the model, the Strong Immunity boundary is then piecewise-defined, $\mu < 2(e+g)/\lambda$ where $e \leq 2g$ and $\mu < 3(e+2g)/2\lambda$ where $e > 2g$. The first version applies to all of the simulations in Kitts (2006) and all other results here, but the piecewise version is needed for the online exploration of e over the range from 0 to 5. (This makes no substantive difference for any of my conclusions.)

Next note that the Weak Immunity boundary depends on group size, but the Strong Immunity boundary does not. As a group grows larger, the model predicts that universal productivity is locally stable at much higher ranges of incentive value. However, this does not imply that large groups are more likely to reach the high levels of cooperation required to attain this equilibrium. Notably, the Voluntary Participation boundary increases more steeply with rivalness at higher group sizes. Thus, as the group grows larger, a higher incentive value is necessary to support voluntary participation at high levels of rivalness.[9]

The assumptions that actors have equal power to influence and influence relations are undifferentiated likely play an important role in the model's tendency to reach the locally stable All-Work/None-Enforce equilibrium between the Strong Immunity and Weak Immunity boundaries. If power were unequally distributed, a small minority of early workers (only if they could be somehow impervious to antisocial norms exerted within their own coalition) could coerce peers to shirk while they hoard the incentive themselves. In those conditions, the system would be more easily trapped in a suboptimal equilibrium.

CONCLUSION

Previous analytical results proved that rivalness in selective incentives leads to perverse regulatory interests and computational experiments showed how this may lead to antisocial norms under well defined conditions: Where incentives are too weak to justify compliance, all actors have a regulatory interest in forcing peers to work. In this standard scenario, collective action depends on peer influence and will be undermined by second-order free riding or low cohesiveness. When presented with valuable rival incentives, however, members who receive the incentives will have a perverse interest in opposing work among peers. In this scenario, effective social influence can undermine collective action and second-order free riding or low cohesiveness can save it.

Whereas earlier computational experiments had mapped out the conditions where these two opposite scenarios obtain numerically

[9]The Voluntary Participation boundary always intersects the μ axis ($\lambda = 0$) at a positive value ($c - g$) and increases at an accelerating rate with rivalness. The limiting value of this boundary as N grows arbitrarily large is $\mu < (c - g)/(1 - \lambda)$, or $4/(1 - \lambda)$ in the numerical illustration here. I discuss group size here only to explore local robustness, not to make general arguments about the effect of group size.

under the model, this article described a set of constraints on actor behavior in the model that may be stated with certainty. It then derived system-level consequences of these actor-level constraints, yielding a set of bounds on micro-level and macro-level model dynamics. This analysis was able to describe much of the model's qualitative behavior while relaxing assumptions about social influence processes that had been required to make the simulations tractable. Doing so lends confidence to previous inferences from simulation research, although the results here are less comprehensive and intuitive.

More important, the analyses here explained patterns in model behavior that were not comprehensible without the mathematical investigation. For example, the simulation study had shown a more sophisticated relationship of second-order free riding to collective action than had been previously appreciated; increasing the enforcement costs amplified collective action in some conditions and inhibited it in others. The mathematical investigation showed how increasing the costs of social influence may both broaden the conditions where universal voluntary cooperation occurs by keeping antisocial norms at bay in some regions where the selective incentive can motivate collective action without prosocial norms, and yet diminish cooperation outside those conditions by dampening prosocial norms where they are needed. In regions where there is no globally stable equilibrium, the mathematical analysis helps us understand the dynamics of competition over rival incentives. For example, we may expect a distinct start-up problem in collective action, where the first-movers oppose later joiners in order to hoard rival incentives, such as prestige, associated with participation. Lastly, the boundaries specified here provide rigorous explanations for the shape of the response surface, including the curious nonmonotonic effect of incentive value on collective action, and its even more curious reversal of concavity at a critical value of rivalness.

Computer simulation is often regarded as a rescue from the unrealistic and restrictive assumptions that may be required to make mathematical analysis tractable. In investigating a previously published computational experiment, this article employs analytics as a complement to simulation. Rather than exhaustively specifying how the model will behave, it merely identifies boundaries on the universe of observable behavior in the simulation. Those boundaries have provided a lucid description of the qualitative shape of the response surface, making both the dynamics and the equilibrium outcomes in the simulation more intelligible. The analysis here also allowed for some more general proofs, which are difficult or impossible to obtain in simulation research.

REFERENCES

Heckathorn, D. D. (1988). Collective sanctions and the creation of prisoner's dilemma norms. *American Journal of Sociology*, *94*, 535–562.

Heckathorn, D. D. (1990). Collective sanctions and compliance norms: A formal theory of group-mediated social control. *American Sociological Review*, *55*, 366–384.

Homans, G. C. (1961). *Social Behavior: Its Elementary Forms*, New York: Harcourt, Brace, & World.

Kitts, J. A. (2006). Collective action, rival incentives, and the emergence of antisocial norms. *American Sociological Review*, *71*, 235–259.

Macy, M. W. & Willer, R. (2002). From factors to actors: Computational sociology and agent-based modeling. *Annual Review of Sociology*, *28*, 143–166.

Oliver, P. E. (1980). Rewards and punishments as selective incentives for collective action: Theoretical investigations. *American Journal of Sociology*, *85*, 1356–1375.

Olson, M. (1965). *The Logic of Collective Action: Public Goods and the Theory of Groups*, Cambridge, MA: Harvard University Press.