J Stat Phys
DOI 10.1007/s10955-012-0679-3

A Simple Model of Stability in Critical Mass Dynamics

Damon Centola

Received: 24 August 2012 / Accepted: 14 December 2012

Abstract Collective behaviors often spread via the self-reinforcing dynamics of critical mass. In collective behaviors with strongly self-reinforcing dynamics, incentives to participate increase with the number of participants, such that incentives are highest when the full population has adopted the behavior. By contrast, when collective behaviors have weakly self-reinforcing dynamics, incentives to participate "peak out" early, leaving a residual fraction of non-participants. In systems of collective action, this residual fraction constitutes free riders, who enjoy the collective good without contributing anything themselves. This "free rider problem" has given rise to a research tradition in collective action that shows how free riding can be eliminated by increasing the incentives for participation, and thereby making cooperation strongly self-reinforcing. However, we show that when the incentives to participate have weakly self-reinforcing dynamics, which allow free riders, collective behaviors will have significantly greater long term stability than when the incentives have strongly self-reinforcing dynamics leading to full participation.

Keywords Collective behavior · Threshold models · Stability dynamics · Enforcement · Norms · Collective action

1 Introduction

Collective behaviors ranging from social movements [16, 33] to consumer fads [3] all belong to a class of phenomena known as critical mass systems [43], in which participation by a small fraction of the population can trigger a snowball of activity. Whether one thinks of firms deciding whether to adopt a new business technology [41], of students deciding whether to leave a late-running lecture [43], or of outraged workers deciding whether to participate in a strike [28], the basic patterns of these collective behaviors can all be understood in terms of critical mass dynamics [16, 43]. Theoretical and empirical treatments of critical mass phenomena have traditionally emphasized the "start-up problem": all that is

D. Centola (✉)
Massachusetts Institute of Technology, E62-462 100 Main Street, Cambridge, MA 02142, USA
e-mail: dcentola@mit.edu

Published online: 05 January 2013

required to jump-start widespread activity is an initial critical mass, yet without the critical mass it is impossible to get anyone to participate.[1] This puzzle has led some researchers to develop arguments about the roles of heterogeneity [16, 33, 38], noise [31, 32], and social reinforcement [20, 21, 47] in initiating collective behavior.

Early research on critical mass dynamics suggested that once the start-up problem was solved, the collective behavior would naturally grow to a very high level, all but extinguishing non-participatory behavior. For instance, a cascade of students leaving a classroom, or the spread of a new business technology among firms, is likely to reach every actor once it passes the point of critical mass. This is because the incentives for participating in a mass exodus, or adopting a successful new technology, increase with the number of others who participate or adopt. "Incentives" here may be broadly thought of as the difference between the benefits for participating in an action and the costs associated with doing it. Thus, the risks of being singled out for punishment or retribution for leaving a lecture are reduced as the number of others who leave increases. Similarly, the rewards for adopting a new technology increase as others adopt it; this is especially true for complementary technologies such as telephones and fax machines, for which the incentives to adopt increase dramatically with the number of others who have adopted.

Yet, scholars have also shown that some systems that initially have snowball, or "bandwagon" growth patterns can nonetheless subsequently suffer "free rider" problems [17, 33, 39]. For instance, the mobilization of voluntary contributions to a public good (e.g., such as public radio, or a neighborhood clean-up effort) often relies on bandwagon dynamics, using each person's contribution to add to the sense of collective efficacy in providing the public good [14, 40]. However, after enough participation has been secured to ensure the maintenance of the public good, contributions will typically drop off even though additional participation would still benefit the collective effort [34, 36]. Past the point where the maintenance of the public good seems guaranteed, an individual's contribution appears to add little marginal benefit to the collective cause, and her calculations may well bring her to realize that she can gain the benefits of the collective good while forgoing the cost of paying for it herself.

In contrast to systems of collective behavior in which growth is strongly self-reinforcing, such that incentives to participate are the greatest when the greatest number are participating, systems that suffer free rider problems are weakly self-reinforcing.[2] Weakly self-reinforcing systems have the same initial growth dynamics as stronger systems, but have a non-monotonic relationship between the level of participation and the creation of incentives.[3] For weakly self-reinforcing collective behaviors, incentives to participate "peak out" before the behavior has spread to the full population. For example, the incentives to make

[1] We follow Schelling [43, 89] in defining "critical mass" as the level of activity above which a behavior becomes self-sustaining. Marwell and Oliver [33] introduce a more complex conception of critical mass into the study of collective action by showing that the role of the critical mass depends upon the shape of the production function. The present discussion focuses on systems of collective behavior that have a "critical mass point," below which cooperation fails, and above which participation increases to the cooperative equilibrium. "Critical mass" is thus defined as the level of activity that is required to push the population over the "critical mass point," initiating self-sustaining growth dynamics for the collective behavior.

[2] Olson [39] also identifies "free rider" problems in systems of collective action in which no one participates (i.e., N-Person prisoner's dilemmas [18]). Since these are not systems with critical mass dynamics, we do not address this kind of free rider problem in the present article.

[3] Leibenstein [29] refers to strongly self-reinforcing systems as having "bandwagon" effects and weakly self-reinforcing systems as having "snob" effects, which Granovetter and Soong [17] refer to as "reverse-bandwagon" effects.

voluntary contributions to a public good are highest just before the point where participation will be large enough to secure the maintenance of the public good. After this point, the individual benefits of joining the action begin to decrease. Thus, further increases in membership reduce the incentives for others to join [34].

For any given collective behavior, whether the system is strongly or weakly self-reinforcing depends upon the particular mechanisms governing individual incentives.

For instance, while the growth dynamics of a strike may sometimes be weakly self-reinforcing, as suggested by the existence of "scabs" and free-riders who cross the picket lines, there are frequently examples of strikes where strong normative pressure and feelings of group solidarity are the primary incentives for participating. In which case, incentives grow stronger as more people participate, making the collective behavior strongly self-reinforcing, and effectively eliminating the problem of free riding.

A large theoretical literature has been occupied with attempting to solve the problem of free riding by showing how systems of collective behavior can be changed from weakly to strongly self-reinforcing.[4] Axelrod [1, 2] and Heckathorn [20] show how norms can eliminate free riding, and Oliver [35] and Heckathorn [21] illustrate the role of enforcement systems (second order cooperation) in creating universal cooperation. Marwell and Oliver [33] focus on voluntary collective action, showing how different production functions determine whether systems are strongly or weakly self-reinforcing; and Kim and Bearman [26] use a network cascade model to show how local incentives can produce strongly self-reinforcing global systems of cooperation.

Yet, while strong incentives can increase participation and eliminate free riding, we show that there are important long term advantages to systems of collective action that are weakly self-reinforcing. We present a simple threshold model of collective behavior that demonstrates the effects of stochastic interactions on the long term stability of systems with critical mass dynamics. Our critical mass model of collective behavior encompasses the familiar dynamics of systems of critical mass collective action [21, 33], as well as the dynamics of other forms of collective behavior, such as crowd behavior [4, 23, 43], norm adoption [5, 16], and consumer behavior [3, 46]. We find that, on average, systems of collective behavior with weakly self-reinforcing dynamics are more stable than those with strongly self-reinforcing dynamics, and therefore can typically be expected to sustain greater long-term levels of participation. Our analysis indicates that the presence of nonparticipants in the collective behavior (i.e., "free riders" in systems of collective action) is a feature of systems with greater long-term stability.

2 Model

Individuals' willingness to participate in a collective behavior is represented through a distribution of individual thresholds for participation.[5] Each threshold, T_i, corresponds to the fraction of the population, p, that would need to participate in order to convince individual i to join the collective behavior. The cumulative distribution of thresholds for the population,

[4] Our approach assumes that strong incentives would substitute for weak ones if the motivations for behavior were altered [11, 15]. An interesting extension of this work would explore the complex effects on stability dynamics if strong incentives were "combined" with weak ones, creating hybrid systems of incentives.

[5] Following earlier scholars of collective behavior [5, 7, 8, 16, 23] and collective action [21, 26, 33], we assume a Gaussian distribution of thresholds. However, our results generalize to smooth, unimodal distributions that produce a "critical mass point" (see Young [47]).

$F(p)$, is given by Eq. (1), where erf is the error function. The function $F(p)$ reports the number of people whose thresholds are triggered when p people are participating, given the mean (μ), and standard deviation (σ) of the distribution of thresholds.

$$F(p) = \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{p-\mu}{\sigma\sqrt{2}}\right)\right] \tag{1}$$

The incentive functions (Eqs. (2) and (3)) map levels of participation onto the changing incentives for other actors to participate. They take as an argument $p = [0, 1]$, which is the fraction of the population already participating in a collective behavior, and return as a value $B(p) = [0, 1]$, the incentives for additional participation.[6] The dynamics of weak incentives are represented by the inverted U-shaped incentive function (i.e., the logistic equation) given in Eq. (2), which has often been used to model the growth dynamics of voluntary collective action. Equation (2) is the first derivative of the S-shaped production function used by collective action scholars (see Oliver [37]). This is the standard model used in the collective action literature action [6, 21, 22, 26, 31, 32, 37], and captures the basic dynamics of increasing and decreasing returns described by empirical researchers [34, 41]. It represents the change in the incentives to participate in terms of the marginal returns for participation. The parameter β is a shape parameter that controls the steepness of the logistic curve. When the parameter $\beta = 10$, this gives the standard logistic curve used by Macy [31, 32], Heckathorn [21, 22], Kim and Bearman [26], Kitts [27] and others to model voluntary collective action. In both Eqs. (2) and (3), the function $B(p)$ ranges over the interval $[0, 1]$, which ensures that both functions have the same minimum ($B(p) = 0$) and maximum ($B(p) = 1$) values. The solid line in Fig. 1 shows a representative growth trajectory (i.e., time series) of participation in a collective behavior with the incentive function in Eq. (2).

$$B(p) = \beta \frac{e^{\beta(0.5-p)}}{(1 + e^{\beta(0.5-p)})^2} \tag{2}$$

By contrast, while Eq. (2) models the change in incentives in weakly self-reinforcing systems, Eq. (3) models the change in incentives for strongly reinforcing systems. For strongly self-reinforcing systems, an often-used simplifying assumption is that incentives increase linearly with participation [16, 42]. We begin our analysis with the linear incentives model, and subsequently relax this assumption by exploring a family of non-linear models of increasing incentives. Equation (3) generates a range of monotonically increasing functional forms, ranging from convex to concave, that all represent possible incentive functions for strongly self-reinforcing systems [21]. The parameter $\alpha > 0$ controls the shape of the incentives and determines whether they are linear ($\alpha = 1$), convex ($\alpha < 1$) or concave ($\alpha > 1$). The dashed line in Fig. 1 shows a representative growth trajectory of the strongly self-reinforcing dynamics given in Eq. (3) (assuming linear incentives, $\alpha = 1$), which has a similar shape as the growth curve for the weakly self-reinforcing dynamics, but increases all the way to full participation—eliminating free riding.

$$B(p) = 1 - (1 - p)^{\alpha} \tag{3}$$

The general dynamics of collective behavior with varying incentive functions are given by combining Eqs. (1), (2), and (3). Equation (4) presents a simple threshold model of critical mass collective behavior [16, 43], in which individuals decide to participate in, or

[6] For mathematical simplicity, $B(p)$ has been normalized to a $[0, 1]$ scale such that $\max(B(p)) = 1$.

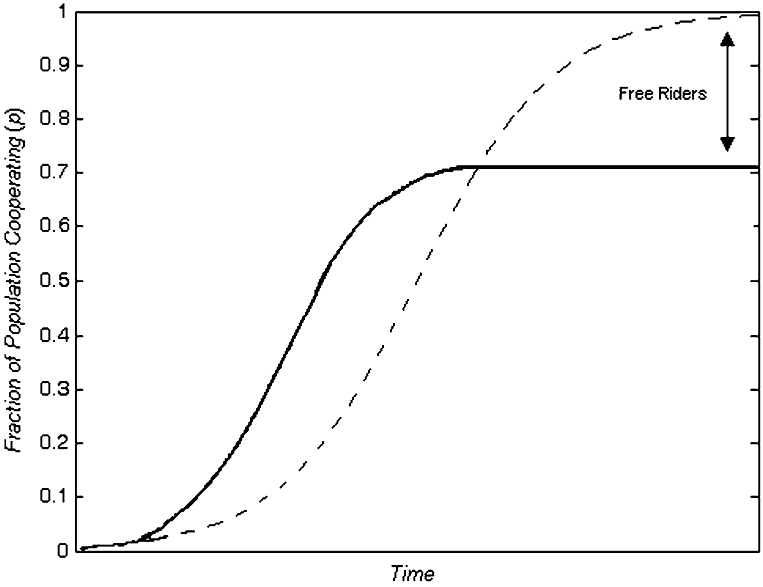

Fig. 1 Characteristic growth curves. For strongly self-reinforcing systems (*dashed line*), the growth of participation follows a characteristic S-shaped curve, which completes its growth trajectory when everyone in the population is participating. Weakly self-reinforcing dynamics (*solid line*), exhibit a similar S-shaped growth curve, but the growth trajectory finishes with only 70 % of the population cooperating. In systems of collective action, the remaining fraction of the population constitutes free riders

abstain from, the behavior based on the level of incentives created by the current number of participants, p.[7]

$$\dot{p} = \gamma\big[\big(F\big(B(p)\big) - p\big)\big] + \varepsilon \tag{4}$$

The basic dynamics of growth and decay are determined by the sign of $\dot{p}$, which represents the net change in participation over time. When the fraction of people willing to join the action, $F(B(p))$, is greater than the fraction who have already joined, p, the number of participants increases ($\dot{p} > 0$). Conversely, if the action becomes oversaturated, the number of current participants will be greater than the number of people who actually want to continue participating, and people will begin to drop out ($\dot{p} < 0$). The error term, ε, introduces continuous stochastic shocks to the level of participation every iteration of the numerical model. The magnitude of ε corresponds to the percent of the population that chooses a random action every iteration of the model.

3 Results

Our results show that weakly self-reinforcing systems of collective behavior are far more robust to perturbations than strongly self-reinforcing systems. The mechanisms governing these dynamics can be seen by comparing the phase portraits from systems of collective behavior with weak (Fig. 2) and strong (Fig. 3) self-reinforcing dynamics. While

[7] $\gamma > 0$ is a temporal scaling parameter.

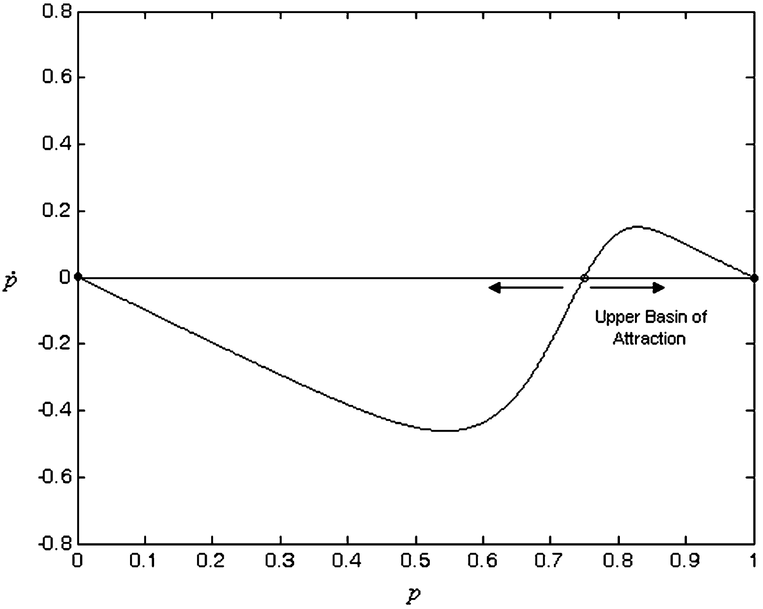

Fig. 2 Phase portrait for strongly self-reinforcing dynamics ($\mu = 0.3$, $\sigma = 0.1$, $\alpha = 1$). The *x-axis* represents the level of participation p, and the *y-axis* is the rate of change in participation, $\dot{p}$. The system is in equilibrium when $\dot{p} = 0$. The upper and lower equilibrium points are stable, while the middle point is a "critical mass point," above which participation grows, and below which it decays. The distance between the critical mass point and the upper attractor (0.25) is the size of the upper basin of attraction

the populations shown in Figs. 2 and 3 both have the same distribution of thresholds, their phase portraits reveal systematic underlying differences. As would be expected from the growth curves shown in Fig. 1, the upper attractor for the strongly self-reinforcing system is all the way at 1, indicating that once critical mass is achieved the self-reinforcing dynamics will drive the population to full cooperation. By contrast, the location of the corresponding attractor in the weakly self-self-reinforcing system is at 0.725, indicating that the maximum level of participation is significantly less than the full population. For systems of collective action, the remaining 27.5 % of the population constitutes free riders.

Further comparison shows that in the "weaker" system the size of the upper basin of attraction—the distance between the attractor (at 0.725) and the critical mass point (at 0.225)—is greater than it is in the strongly self-reinforcing system (in which the upper attractor and critical mass point are at 1 and 0.75, respectively). The size of the upper basin of attraction in the strongly self-reinforcing system is thus 0.25 (i.e., 1 – 0.75), while it is 0.5 in the weakly self-reinforcing system (i.e., 0.725 – 0.225). A larger basin of attraction implies that the system needs to be pushed farther from equilibrium in order to collapse. In other words, the system with weaker growth dynamics has a greater number of states of the system (i.e., values of p) for which participation increases back to the cooperative equilibrium. Consequently, while the strongly self-reinforcing dynamics can achieve greater levels of participation, it only takes a 25 % loss of the mobilized participants in order for participation to drop below the tipping point, causing a mass exodus from the collective behavior. Conversely, because the weakly self-reinforcing dynamics have a larger basin of attraction,

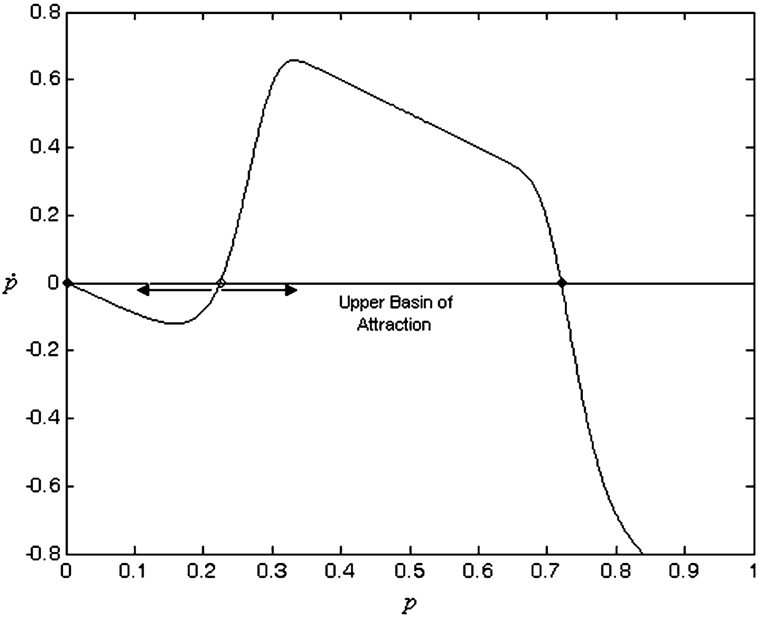

Fig. 3 Phase portrait for weakly self-reinforcing dynamics ($\mu = 0.3$, $\sigma = 0.1$, $\beta = 10$). Participation, p, is shown along the *x-axis*, and rate of change in participation, $\dot{p}$, is indicated by the *y-axis*. As in Fig. 4, this system has two stable equilibrium points, and one critical mass point. While the upper equilibrium is much lower (at 72 %) than it is in Fig. 4, this system has a larger basin of attraction (0.5), and a steeper slope approaching the attractor

a larger disturbance (in both relative and absolute sizes) is required in order to destabilize the collective behavior.[8]

While this analysis suggests that there may be an important comparative advantage for the system of collective behavior with weakly self-reinforcing dynamics, the size of the basin of attraction does not tell the entire story of stability. In addition to the size of the basin of attraction, we must also be concerned with the strength of the attractor.

The relative strength of the cooperative attractors in Figs. 2 and 3 can be seen by examining the slope of $\dot{p}$ as it approaches each of the upper equilibrium points. The steeper the (negative) slope, the stronger the attractor. In the system with strongly self-reinforcing dynamics (Fig. 2), the slope is approximately -1. This means that as errors push the system farther from equilibrium (p moves to the left) there is a gradual acceleration of the growth rate pushing the system back toward equilibrium, which increases up to a maximum rate of 0.15 as participation decreases. In the system with weakly self-reinforcing dynamics (Fig. 3), the slope is much steeper, quickly rising to 0.3. This difference in the slopes is crucial because it means that the speed at which the dynamics push participation back to equilibrium is much greater in the weakly self-reinforcing system. This implies that the strongly self-reinforcing system is more likely to permit additional errors and accidents to compound before the growth dynamics can restore equilibrium. In the weakly self-reinforcing system,

[8] Numerical analysis shows that this result generalizes across the entire range of threshold distributions that generate critical mass dynamics: For systems with identical distributions, weakly self-reinforcing systems have a larger upper basin of attraction than strongly self-reinforcing systems with linear incentives.

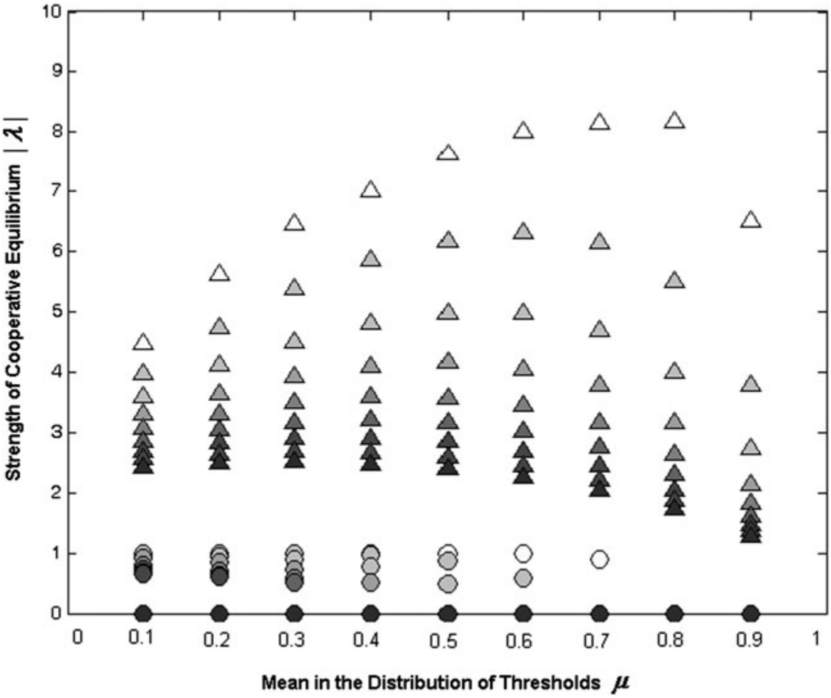

Fig. 4 Equilibrium strength for weak and strong self-reinforcing dynamics ($\alpha = 1$, $\beta = 10$). The mean of the threshold distribution is shown along the *x-axis*, ranging from $0.1 \geq \mu \geq 0.9$. The standard deviation of the threshold distribution is indicated by *shading*, ranging from *light* ($\sigma = 0.1$) to *dark* ($\sigma = 0.9$). Along the *y-axis*, the strength of the cooperative equilibrium ($|\lambda|$) is shown for systems with weak (*triangles*) and strong (*circles*) self-reinforcing dynamics. Results are shown across the full range of threshold distributions that generate critical mass dynamics. All of the systems with weak incentives have greater equilibrium strength than the systems with strong incentives. For all systems, equilibrium strength decreases with increasing standard deviation

small reductions in the number of contributors cause a steep acceleration of the growth rate, making it more likely that the system will return to equilibrium before further errors can accrue.

More generally, by analyzing the properties of the attractors in Eq. (4) we can determine the strength of the cooperative equilibrium for all critical mass systems with weak and strong self-reinforcing dynamics. We do this by taking the eigenvalue (λ) of the linearized system at the equilibrium points [44]. The larger the absolute value of λ, the stronger the force is that pulls the system back to equilibrium. Figure 4 presents the results of this analysis over the full range of threshold distributions that generate critical mass dynamics.

The x-axis in Fig. 4 shows the mean of the threshold distribution, μ, and the y-axis indicates the strength of the cooperative equilibrium, $|\lambda|$. Weakly self-reinforcing dynamics are represented by triangles, while the circles correspond to strongly self-reinforcing dynamics. The shading from light to dark indicates the standard deviation in the distribution of thresholds, σ, which ranges from 0.1 (light) to 0.9 (dark). For each value of the mean and standard deviation in the threshold distribution, there is a corresponding triangle and circle. The position of each triangle and circle along the y-axis indicates the strength of the cooperative equilibrium for that threshold distribution.

Comparing two equivalent systems (each similarly shaded triangle and circle in the same column), the triangle is always above the circle, indicating a stronger attractor, and greater resistance to de-stabilization. For many strongly self-reinforcing systems, the circle has a value of zero, indicating that cooperation is not a stable equilibrium; yet, the corresponding triangle always has $|\lambda| > 1$, indicating a strong attractor. Moreover, across all combinations of the mean and standard deviation of thresholds (i.e., across all x-values and shades) all of the triangles are above all of the circles. This means that for all systems that generate critical mass dynamics, collective behaviors with weakly self-reinforcing dynamics have stronger attractors than systems with strongly self-reinforcing dynamics. This is true regardless of whether the systems being compared have the same threshold distribution.

To show this more formally, the following proof demonstrates that for the model in Eq. (4), for systems that meet the following definitions of weakly and strongly self-reinforcing behaviors, the system with weak self-reinforcing dynamics will always have a larger value of $|\lambda|$, and therefore will have a stronger attractor at the cooperative equilibrium.

Proof. 1. In a strongly self-reinforcing system, for any fraction of the population p, ranging from 0 to 1, the incentive function $B(p)$ is increasing, such that the slope is always positive, i.e., $B'(p) > 0$ for $0 < p < 1$.

2. In a weakly self-reinforcing system, the incentive function $B(p)$ is non-monotonic in p, such that for p ranging from 0 to 1, $B(p)$ is a smooth single peaked function; i.e., it has an inverted U-shape with a single maximum p°, such that $B'(p^\circ) = 0$, $B'(p^\circ - \varepsilon) > 0$, and $B'(p^\circ) < 0$.

3. The slope at the fixed point p^* is obtained by taking the first derivative of Eq. (4) with respect to p and substituting p^* for p, which produces Eq. (5).

$$\frac{d\dot{p}}{dp} = \gamma\left[F'\left(B\left(p^*\right)\right)B'\left(p^*\right) - 1\right] \tag{5}$$

Equation (5) is equivalent to the slope of the function in Eq. (4). If the slope is positive at p^*, it indicates that the flow goes away from the fixed point in the vicinity of p^*. In which case, the fixed point is unstable. If the slope is negative, p decreases above the fixed point and increases below it, making the fixed point an attractor.

The greater the absolute value of the negative slope in the vicinity of an attractor, the faster the system goes to the fixed point, and the stronger the attractor is. For a strongly self-reinforcing system, the slope of $F(B(p^*))$ is approximately zero (but slightly positive), which makes $\frac{d\dot{p}}{dp} \approx -1$ (disregarding any scaling parameter).

For a weakly self-reinforcing system, the slope of $F(B(p^*))$ is steeply negative, making $\frac{d\dot{p}}{dp} \ll -1$. Thus, the upper fixed point in the weakly self-reinforcing system is a much stronger attractor than the upper fixed point in the strongly self-reinforcing system.

More generally, since the upper fixed point in weakly self-reinforcing systems always occurs while the slope of $B(p)$ is negative, and since $F(B(p))$ is a cumulative distribution function, which preserves monotonicity with respect to $B(p)$, $F(B(p))$ must be decreasing at the upper fixed point, making $\frac{d\dot{p}}{dp} < -1$. By similar reasoning, $F(B(p))$ must be increasing at the upper fixed point of a strongly self-reinforcing system, making $\frac{d\dot{p}}{dp} > -1$. Thus, slope of $\dot{p}$ at p^*, where p^* is the upper fixed point, will always be more steeply negative in weakly self-reinforcing systems than in strongly self-reinforcing systems. Therefore, the upper attractor in weakly self-reinforcing systems will always be stronger than the upper attractor in strongly self-reinforcing systems.

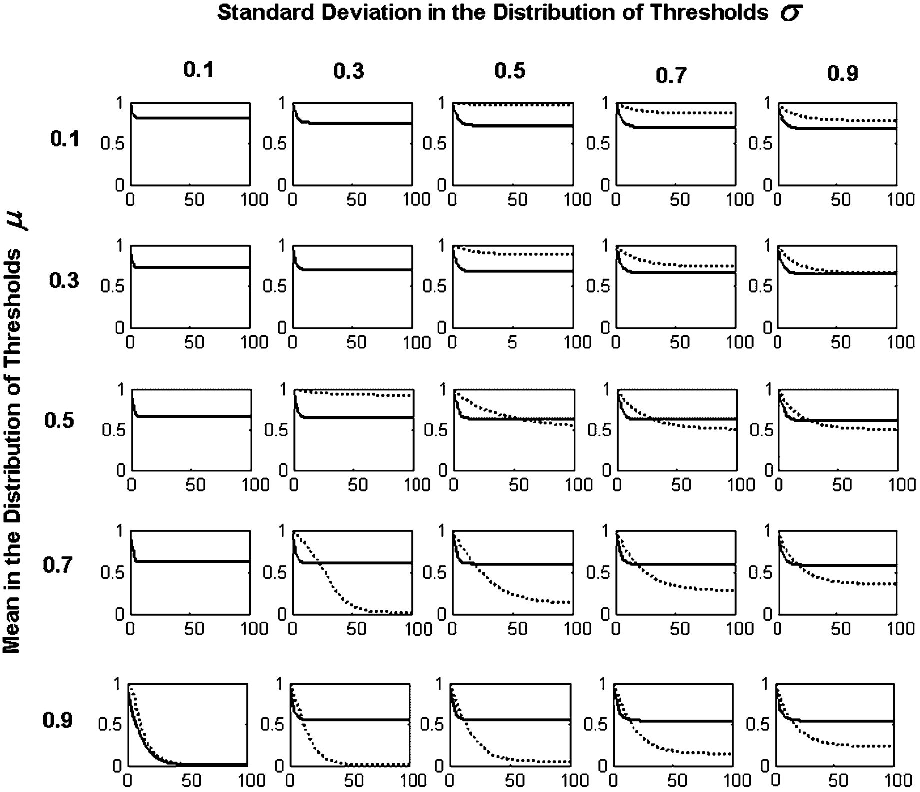

Fig. 5 Stability in weakly and strongly (linear) self-reinforcing dynamics ($\alpha = 1$, $\beta = 10$, $\varepsilon = 0.01$). In each window, the *y-axis* indicates the level of participation (p), and the *x-axis* indicates the time from 0 to 100 for weak (*solid line*) and strong (*dashed line*) self-reinforcing dynamics. Results are shown across the entire range of threshold distributions that generate critical mass dynamics. Both systems can be stable, and both can collapse. The differences in behavior reveal much greater stability in systems with weakly self-reinforcing dynamics

3.1 Dynamics of Stability

To illustrate the implications of these formal differences between weakly and strongly self-reinforcing dynamics for the stability of collective behaviors, we numerically integrate the model in Eq. (4) across the full range of threshold distributions that generate critical mass dynamics ($0.1 \geq \mu \geq 0.9$, $0.1 \geq \sigma \geq 0.9$). Figures 5, 6 and 7 show the results from these simulations. Each of the 64 panels in Figs. 5 through 7 has an x-axis that indicates time, and a y-axis that shows the level of participation, p. In each panel, the solid line represents weakly self-reinforcing dynamics, while the dashed line represents the corresponding system with strongly self-reinforcing dynamics. In all cases, time series were initialized with full participation ($p = 1$), and then iterated for 100 time periods with continuous perturbations ($\varepsilon = \pm 0.01$). For robustness, we also ran simulations over the range $0.01 \leq \varepsilon \leq 0.25$. The results were qualitatively similar to those reported below. Additionally, iterations for longer durations (1000 and 10,000 time periods) also produced qualitatively similar results.

Figure 5 compares the stability of weakly self-reinforcing dynamics with corresponding strongly self-reinforcing dynamics that have linearly increasing incentives ($\alpha = 1$). The re-

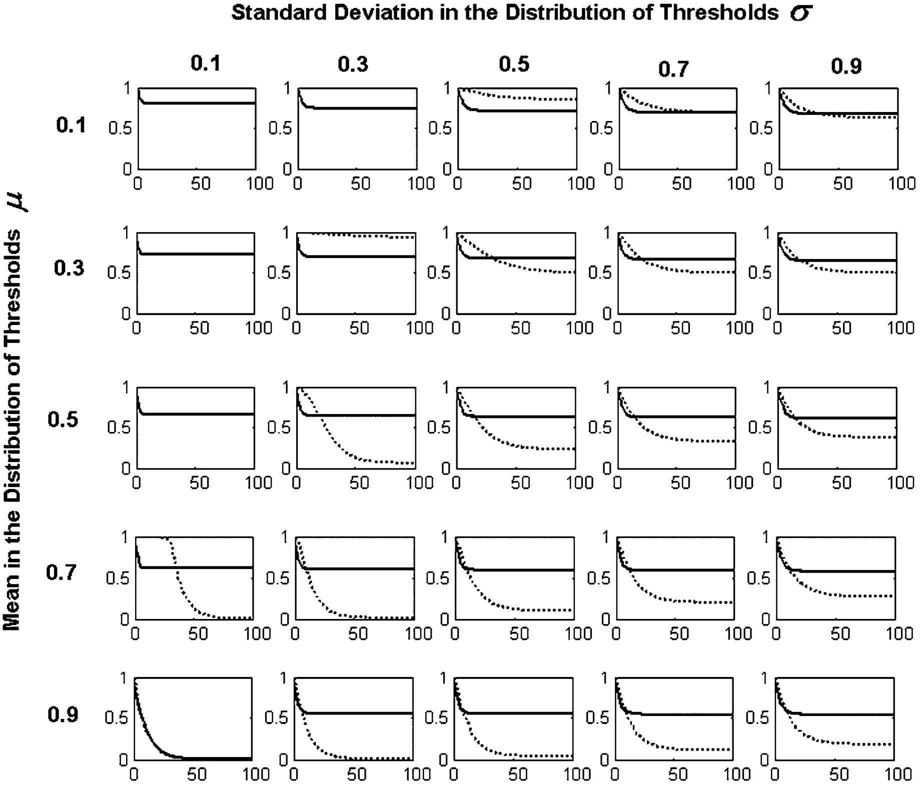

Fig. 6 Stability in weakly and strongly (convex) self-reinforcing dynamics ($\alpha = 0.5$, $\beta = 10$, $\varepsilon = 0.01$). In each window, the *y-axis* indicates the level of participation (p), and the *x-axis* indicates the time from 0 to 100 for weak (*solid line*) and strong (*dashed line*) self-reinforcing dynamics. Results are shown across the entire range of threshold distributions that generate critical mass dynamics. Both systems can be stable, and both can collapse. The differences in behavior show that greater convexity further weakens the relative stability of systems with strongly self-reinforcing dynamics

sults show that both systems can be stable (e.g., $\mu = 0.1$, $\sigma = 0.1$), and that both can be unstable (e.g., $\mu = 0.1$, $\sigma = 0.9$). However, where there are differences in stability, the results show greater stability in the systems with weakly self-reinforcing dynamics. Across a large range of the parameter space, strongly self-reinforcing dynamics collapse, while weakly self-reinforcing systems remains stable.

We relax the linearity assumption in the strongly self-reinforcing dynamics by exploring the shape parameter α in Eq. (3), over the range $0.1 \le \alpha \le 10$. When $\alpha < 1$, the incentives "accelerate" as more people participate [33]. This "convex" incentive function corresponds to strongly self-reinforcing dynamics in which a large majority of people need to participate before significant incentives are created for others. For example, when the behavior is risky—such as investing in a new stock, or buying a new technology—a large number of peers may need to adopt the behavior before it is highly credible. Similarly, in systems of normative collective action, when participation is low, the ability of cooperators to exert social influence on non-adopters is relatively weak and increases slowly. However, once membership becomes large, not only do additional recruits increase the social pressure on non-cooperators, but they also increase the visibility and legitimacy of the behavior—both

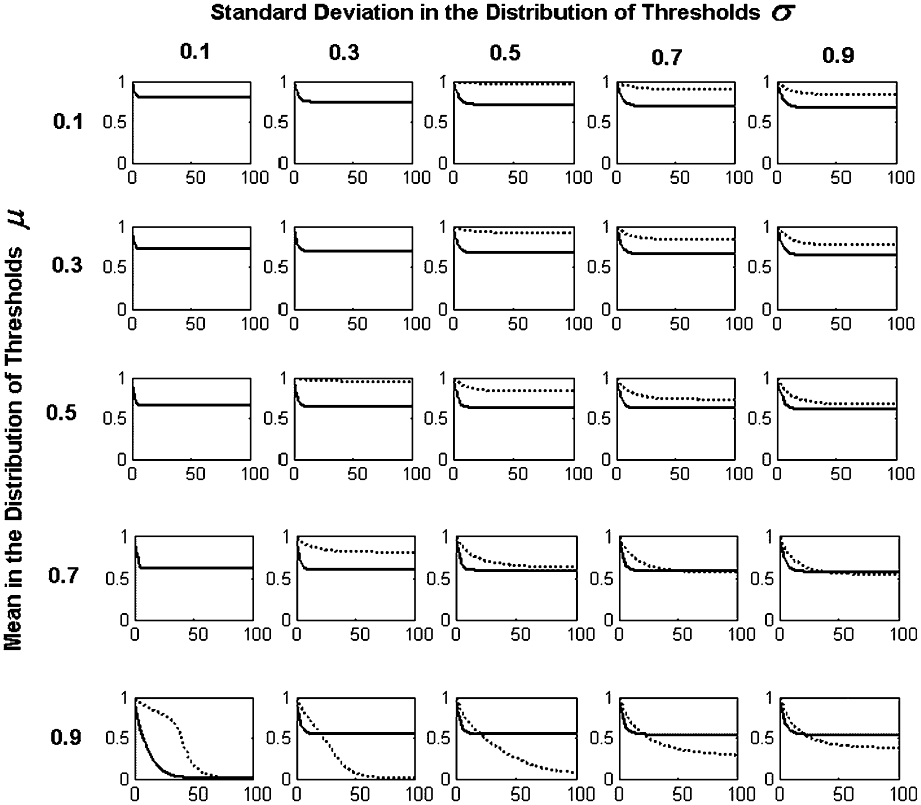

Fig. 7 Stability in weakly and strongly (concave) self-reinforcing dynamics ($\alpha = 5$, $\beta = 10$, $\varepsilon = 0.01$). In each window, the *y-axis* indicates the level of participation (p), and the *x-axis* indicates the time from 0 to 100 for weak (*solid line*) and strong (*dashed line*) self-reinforcing dynamics. Results are shown across the entire range of threshold distributions that generate critical mass dynamics. Both systems can be stable, and both can collapse. The differences in behavior show that greater concavity adds stability to the strongly self-reinforcing dynamics, but it is still relatively less stable than the system with weakly self-reinforcing dynamics

of which can create exponentially increasing levels of social pressure as the behavior reaches the full population [5, 19, 25, 35, 45].

The time series in Fig. 6 ($\alpha = 0.5$) show an even more pronounced difference in the stability of systems with weakly and strongly self-reinforcing dynamics than those observed for the linear system. Over almost the entire range of the parameter space, strongly self-reinforcing dynamics exhibit far less stability. This surprising lack of stability is due to the fact that very high levels of participation are needed in order to maintain strong incentives for participation. Stochastic drops in the level of activity can thus have a substantial impact on the incentives. For $\alpha = 0.5$, even a 10 % drop in participation from 100 % to 90 % results in a 30 % reduction in the incentives for other participants, making it likely that "high threshold" individuals will begin to drop out, initiating a cascade of defection.

By contrast, if $\alpha > 1$, the incentive function is "concave." While increasing participation still increases incentives, the marginal impact of each additional person decreases as participation increases. For $\alpha = 5$, the increase from 10 % to 20 % participation produces a dramatic gain in incentives, while an increase in the fraction of adopters from 80 % to

90 % has only a very small impact on the incentives for others. These dynamics correspond to collective behaviors with strong social interdependence—such as the spread of technologies like the Internet [12, 13], which only require a small critical mass to demonstrate the technology's usefulness, at which point widespread adoption quickly takes off. By the time 90 % of people have adopted, the marginal impact of new adopters is so small as to be almost irrelevant.

Figure 7 ($\alpha = 5$) shows that concave incentives are stable over a greater area of the parameter space than both the linear and convex functions. Yet, despite this relative robustness over other strongly self-reinforcing dynamics, these systems are still less stable than systems with weakly self-reinforcing dynamics. While there are differences in the stability dynamics of strong self-reinforcing systems with concave, linear, and convex incentives, in each case the powerful growth dynamics of strong self-reinforcement present a characteristic vulnerability to destabilization when compared with weakly self-reinforcing dynamics.

4 Discussion

The greater stability of systems with weakly self-reinforcing dynamics comes from the fact that slight perturbations away from equilibrium (i.e., "accidental" deviations from the cooperative behavior) create new incentives for people to start contributing. This sudden increase in incentives for participation "engages" the reserve of non-participants, which quickly increases the number of active members, and pulls the system back to the equilibrium level of participation. By contrast, in strongly self-reinforcing systems accidental defections lower the incentives for participation, which can cause further defections, which further reduce the incentives to participate. This feedback between reduced participation and lower incentives can slow down the system's recovery time—the time it takes for participants to re-join the collective behavior and restore the equilibrium level of participation—allowing more errors to accrue, which push the system farther from equilibrium. This increases the likelihood that a chain of stochastic events will eventually drive the system over the tipping point, resulting in a collapse of the collective behavior.

To illustrate how these dynamics can unfold, consider Coleman's example of a micro-credit lending "circle" [9, 10]. For illustrative purposes, let us assume that there are ten members of the lending group, who have a threshold distribution given by $\{0.05, 0.15, 0.25, 0.35, 0.45, 0.55, 0.65, 0.75, 0.8, 0.9\}$, and that the system has a linear incentive function with slope $= 1$. Assuming that the cooperative norm has been established ($p = 1$), the threat of high sanctions for defection provides assurances of cooperation, which allow even the highest threshold individuals (i.e., those with the greatest aversion to suffering loss), to confidently participate in the norm of exchanging loans for mutual benefit.

Given the stated distribution of thresholds, if the most risk averse member ($T = 0.9$) appears to be withdrawing from the cooperative norm, it will not cause anyone else to abandon the organization. Instead, this actor will receive strong pressure to rejoin the group, and the system will resume equilibrium. However, if the actor with $T = 0.8$ appears to be withdrawing, the perceived drop in participation will cause the actor with $T = 0.9$ to doubt the strength of the organization and decide to also withdraw. This will not necessarily destabilize the cooperative norm since even if the $T = 0.9$ member withdraws, the member with $T = 0.8$ still has sufficient confidence in the group to keep participating. Once the $T = 0.8$ actor corrects the misperception of having abandoned the group, the perturbation to the system is corrected, and there is once again a strong normative signal for the $T = 0.9$ actor. This reinforces the incentives for cooperation, and restores the equilibrium.

However, the timing is crucial. Once the $T = 0.8$ actor's lapse causes the $T = 0.9$ actor to withdraw, this will also cause the $T = 0.75$ actor to lose confidence in the organization and withdraw. If both of these withdraws occur before the $T = 0.8$ actor can correct the error, the reduced incentives for cooperation feedback to affect the $T = 0.8$ actor. Having accidentally triggered withdraws by the others, the $T = 0.8$ actor will now doubt the viability of the cooperative organization, and no longer feel compelled to participate. Once all three actors ($T = 0.75$, $T = 0.8$, and $T = 0.9$) have dropped out, confidence in the organization wanes and the $T = 0.65$ actor will also drop out, which, in turn, will cause the $T = 0.55$ actor to drop out, and so on. The dominos ineluctably fall, resulting in the "death" of the micro-credit organization.

Strongly self-reinforcing systems are inherently susceptible to this kind of unraveling process because errors (or "random" deviations from the behavior) and decreasing incentives (resulting in "intentional" deviations from the behavior) reinforce one another. This, in turn, increases the time it takes for the system to recover, allowing further defections to occur, and increasing the likelihood of participation being pushed below the tipping point.

Conversely, weakly self-reinforcing dynamics are just the opposite. As soon as participation drops below the equilibrium level, the number of people willing to cooperate increases. To see this, consider the example of public broadcasting, which is provided through charitable donations, and frequently suffers free riders who enjoy the public good without contributing. If illness or hardship prevents a subset of the usual donors from contributing, the drop in support may threaten to reduce the number and quality of programs that are available. However, unlike a system with strongly self-reinforcing dynamics, the drop in public support does not cause a downward spiral of incentives for participation. Instead, decreasing participation increases the incentives to contribute. Each individual's contribution now has a much greater marginal impact on the provision of the public good, which dramatically increases the value of their contributions. The immediate need for more participants creates a wave of new actors (i.e., the free riders), who are faced with the choice of either contributing or losing the valued public good. These actors thus become newly motivated members of the collective action, whose contributions restore participation to its equilibrium level.

The seeming weakness of weakly self-reinforcing dynamics is that incentives decrease as participation levels get too high, inevitably creating free riders. However, their hidden strength is that these free riders become a vital resource for stabilizing collective action when participation drops. This stabilizing effect does not come from a change in the moral sensibility of the free riders, nor from external sanctions for defection, but from the fact that the same dynamics of decreasing incentives that create free riders when provisions are high, in turn create strong incentives for the free riders to start participating when errors or accidents place a valued public good in jeopardy.

5 Conclusion

A general intuition in studies of collective behavior—ranging from the "viral" spread of new technologies [30] to the growth of social movements [20]—is that strongly self-reinforcing dynamics are desirable for their ability to marshal high levels of participation. The problem of social cooperation, in particular, has typically been framed as a problem of creating strong incentives (Olson [39], Hobbes [24])—the stronger the incentives are, the more people who will cooperate, and the greater the provision of collective goods. However, we have shown that there is an important difference between strong growth dynamics and strength in stability. Stronger growth dynamics produce greater levels of participation, but they can also create less stable systems of collective behavior. Conversely, weaker growth dynamics engender non-participatory behavior, but exhibit significantly greater long-term stability. This suggests an important trade-off implicit in the dynamics of critical mass: attempts to alter the growth dynamics of a collective behavior—in an effort to increase membership or reduce free riding—may also significantly compromise its long term stability.

Acknowledgements I would like to thank Michael Macy, Steve Strogatz, and Robert Gibbons for helpful comments and suggestions.

References

1. Axelrod, R.: The Evolution of Cooperation. Basic Books, New York (1984)
2. Axelrod, R.: An evolutionary approach to norms. Am. Polit. Sci. Rev. **80**(4), 1095–1111 (1986)
3. Bass, F.M.: A new product growth model for consumer durables. Manag. Sci. **13**, 215–227 (1969)
4. Berk, R.: A gaming approach to crowd behavior. Am. Sociol. Rev. **39**, 355–373 (1974)
5. Centola, D., Willer, R., Macy, M.: The Emperor's dilemma: a computational model of self-enforcing norms. Am. J. Sociol. **110**, 1009–1040 (2005)
6. Centola, D., Macy, M.: Complex contagions and the weakness of long ties. Am. J. Sociol. **113**, 702–734 (2007)
7. Centola, D.: Failure in complex networks. J. Math. Sociol. **33**(1), 64–68 (2009)
8. Centola, D., Eguiluz, V., Macy, M.: Cascade dynamics of complex propagation. Physica A **374**, 449–456 (2007)
9. Coleman, J.: Systems of trust: a rough theoretical framework. Angew. Soz.forsch. **10**, 277–287 (1982)
10. Coleman, J.: Foundations of Social Theory. Harvard University Press, Cambridge (1990)
11. Deci, E.L., Koestner, R., Ryan, R.M.: A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. Psychol. Bull. **125**, 627–668 (1999)
12. DiMaggio, P.: Social stratification, life-style, social cognition, and social participation. In: Grusky, D. (ed.) Social Stratification in Sociological Perspective, 2nd edn., pp. 542–552. Westview Press, Boulder (2001)
13. Economides, N., Himmelberg, C.: Critical mass and network size with application to the US fax market. Discussion Paper No. EC-95-11, Stern School of Business, N.Y.U. (1995)
14. Finkel, S.E., Muller, E.N., Opp, K.-D.: Personal influence, collective rationality, and mass political action. Am. Polit. Sci. Rev. **83**, 885–903 (1989)
15. Gneezy, U., Rustichini, A.: A fine is a price. J. Leg. Stud. **29**, 1–17 (2000)
16. Granovetter, M.: Threshold models of collective behavior. Am. J. Sociol. **83**, 1420–1443 (1978)
17. Granovetter, M., Soong, R.: Threshold models of interpersonal effects in consumer demand. J. Econ. Behav. Organ. **7**, 83–89 (1986)
18. Hardin, R.: Collective Action. Johns Hopkins University Press, Baltimore (1982)
19. Heckathorn, D.: Collective sanctions and the creation of Prisoner's dilemma norms. Am. J. Sociol. **94**(3), 535–562 (1988)
20. Heckathorn, D.: Collective sanctions and compliance norms: a formal theory of group-mediated social control. Am. Sociol. Rev. **55**, 366–384 (1990)
21. Heckathorn, D.: Collective action and group heterogeneity: voluntary provision versus selective incentives. Am. Sociol. Rev. **58**, 329–350 (1993)
22. Heckathorn, D.: Dynamics and dilemmas of collective action. Am. Sociol. Rev. **61**, 250–277 (1996)
23. Helbing, D., Farkas, I., Vicsek, T.: Simulating dynamical features of escape panic. Nature **407**, 487–490 (2000)
24. Hobbes, T.: In: Gaskin, J.C.A. (ed.): Leviathan [1651]. Oxford University Press, London (1998)
25. Horne, C.: Explaining norm enforcement. Ration. Soc. **19**(2), 139–170 (2007)
26. Kim, H., Bearman, P.: The structure and dynamics of movement participation. Am. Sociol. Rev. **62**, 70–93 (1997)
27. Kitts, J.: Collective action, rival incentives, and the emergence of antisocial norms. Am. Sociol. Rev. **71**, 235–259 (2006)
28. Klandermans, B.: Union action and the free-rider dilemma. In: Kriesberg, L., Misztal, B. (eds.) Social Movements as a Factor of Change in the Contemporary World. Research in Social Movements, Conflict and Change, vol. 10, pp. 77–92. JAI Press, Greenich (1988)
29. Leibenstein, H.: Beyond Economic Man: A New Foundation for Microeconomics. Harvard University Press, Cambridge (1976)

30. Leskovich, J., Adamic, L., Huberman, B.: The dynamics of viral marketing. ACM Trans. Web. **1**, 1 (2007)
31. Macy, M.W.: Learning theory and the logic of critical mass. Am. Sociol. Rev. **55**, 809–826 (1990)
32. Macy, M.W.: Chains of cooperation: threshold effects in collective action. Am. Sociol. Rev. **56**, 730–747 (1991)
33. Marwell, G., Oliver, P.: The Critical Mass in Collective Action: A Micro-Social Theory. Cambridge University Press, Cambridge, England (1993)
34. Oberschall, A.: Loosely structured collective conflicts: a theory and an application. In: Zald, M., McCarthy, J.D. (eds.) Research in Social Movements, pp. 45–70. Winthrop, Cambridge (1980)
35. Oliver, P.: Rewards and punishments as selective incentives for collective action: theoretical investigations. Am. J. Sociol. **85**, 1356–1375 (1980)
36. Oliver, P.: 'If you don't do it, nobody else will': active and token contributors to local collective action. Am. Sociol. Rev. **49**, 601–610 (1984)
37. Oliver, P.: Formal models of collective action. Annu. Rev. Sociol. **19**, 271–300 (1993)
38. Oliver, P., Marwell, G.: The paradox of group size in collective action: a theory of the critical mass II. Am. Sociol. Rev. **53**, 1–8 (1988)
39. Olson, M.: The Logic of Collective Action. Harvard University Press, Cambridge (1965)
40. Opp, K.-D., Gern, C.: Dissident groups, personal networks, and spontaneous cooperation: the east German revolution of 1989. Am. Sociol. Rev. **58**, 659–680 (1993)
41. Rogers, E.M.: Diffusion of Innovations. Free Press, New York (1995)
42. Samuelson, P.: The pure theory of public expenditure. Rev. Econ. Stat. **36**, 387–389 (1954)
43. Schelling, T.: Micromotives and Macrobehavior. Norton, New York (1978)
44. Strogatz, S.: Nonlinear Dynamics and Chaos. Perseus Books, Reading (1994)
45. Ullman-Margalit, E.: The Emergence of Norms. Clarendon Press, Oxford (1977)
46. Watts, D.J.: A simple model of global cascades on random networks. Proc. Natl. Acad. Sci. **99**, 5766–5771 (2002)
47. Young, P.: Innovation diffusion in heterogeneous populations: contagion, social influence, and social learning. Am. Econ. Rev. **99**, 1899–1924 (2009)